TRISTAN II

TRIennal Symposium on Transportation ANalysis

PREPRINTS



C.N.R
Progetto Finalizzato
Trasporti Due



C.N.R.-I.A.S.I.
Istituto di Analisi
dei Sistemi ed Informatica

A.I.R.O. Associazione Italiana
Ricerca Operativa

Vol. II

CAPRI, ITALY
Congress Center
June, 23-28 1994

Vol. II

TRISTAN II

TRIennal Symposium on Transportation ANalysis

PREPRINTS

Vol. II

This compendium was prepared in the interest of timely information dissemination. It contains the preprints of the papers presented by the authors during the Conference.

Capri, June 23-28, 1994

TRISTAN II

TRIennal Symposium on
Transportation ANalysis

PROGRAM COMMITTEE

Lucio Bianco	*Chairman*	University of Rome Tor Vergata, Italy
Paolo Toth	*Chairman*	University of Bologna, Italy
Maurizio Bielli		I.A.S.I. - C.N.R., Italy
Ennio Cascetta		University of Naples, Italy
Teodor Gabriel Crainic		University of Montreal, Canada
Warren B. Powell		Princeton University, USA

ADVISORY COMMITTEE

Michael O. Ball	Toshihide Ibaraki
Jaime Barcelo	Martine Labbe'
Alex Belenky	Gilbert Laporte
Carlos Daganzo	Jan Karel Lenstra
Mark S. Daskin	Amedeo R. Odoni
Pierre Dejax	Jose' Paixao
Jacques Desrosiers	Markos Papageorgiou
Awi Federgruen	Harilaos N. Psaraftis
J. Enrique Fernandez	Donald H. Ratliff
Michael Florian	Alexander Rinnooy Kan
Terry L. Fries	Werner Rothengatter
Giorgio Gallo	Jean-Marc Rousseau
Patrick T. Harker	Francois Soumis

ORGANIZING COMMITTEE

M. Bielli G. E. Cantarella A. Scognamiglio

INVITED LECTURERS

J. Barcelo	University of Barcelona, Spain
M. Ben-Akiva	Massachussets Institute of Technology, Cambridge, USA
E. Cascetta	University of Naples, Italy
M. Florian	Centre for Research on Transportation, Montreal, Canada
W. Powell	Princeton University, USA
H.D. Ratliff	Georgia Tech., Atlanta, USA

SECRETARIAT of TRISTAN II

- **Maurizio Bielli** I.A.S.I. - C.N.R.
  Viale Manzoni, 30 - 00185 - Rome - Italy
  Phone + 39 - 6 - 77161 - Fax + 39 - 6 - 7716461
  E-MAIL: BIANCO@IASI.RM.CNR.IT

Volume stampato con il contributo del Consiglio Nazionale delle Ricerche, Comitato Nazionale Scienze di Ingegneria ed Architettura

INDEX

VOLUME II

MODELLING DYNAMICS IN TRAFFIC ASSIGNMENT (*CHAIRMAN*: M. LABBE')

Control Strategies Simulation on Urban Networks: Evaluation Tools and Techniques
G.E. Cantarella, E. Cascetta, M. Di Gangi 495

Dynamic Network Equilibrium Problem Formulated as an Infinite Dimensional Variational Inequality Problem
J.H. Wu 497

Flow Propagation Description in Dynamic Network Assignment Models
J.E. Fernández, J. de Cea 517

CREW SCHEDULING PROBLEMS (*CHAIRMAN*: J. DESROSIERS)

An Evolutionary Algorithm for the Crew Rostering Problem
F. Catanas, J. Paixão 533

Exact and Approximate Algorithms for the Generalized Crew Scheduling Problem
M. Fischetti, S. Martello, P. Toth 539

Hybrid Genetic Algorithms for Bus Driver Scheduling
R.S.K. Kwan, A. Wren 543

Polyhedral Methods for Set-covering/Partitioning Problems
S. Ceria, C. Mannino, P. Nobili, A. Sassano 547

ALGORITHMS FOR IVHS (*CHAIRMAN*: J. BARCELO)

Calibrating Incident Detection Algorithms by Simulated Annealing
S. Cohen 549

Neural Networks for the Feedback Optimal Control of Freeway Systems
A. Di Febbraro, D. Di Noto, T. Parisini, V. Recagno, S. Sacone, R. Zoppoli 561

An On-line Dynamic Traffic Prediction Model for an Inter-urban Motorway Network
M. Ben-Akiva, E. Cascetta, H. Gunn 577

The Impact of Predictive Information on Guidance Efficiency: An Analytical Approach
M. Ben-Akiva, A. de Palma, I. Kaysi 579

URBAN TRANSPORTATION PLANNING (*CHAIRMAN*: C. BARNHART)

Application of Contextual Technology for Supporting Decision Making in Urban Transportation Networks
L. Bianco, P. Dell'Olmo, I. Ezhkova 591

Intermodality in Urban Transportation: Modelling and Analysis
A. Di Febbraro, V. Recagno, S. Sacone 603

A Reliability Model for Urban Transportation Planning
B. Sanso', L. Milot 617

Relationship between Parking Localization and Traffic Flows in Urban Area
S. Carrese, S. Gori, T. Picano 623

METHODS FOR VEHICLE ROUTING PROBLEMS (*CHAIRMAN*:A. WREN)

Deterministic and Probabilistic Sales-Delivery Man Problems
I. Averbakh, O. Berman 639

An Exact Algorithm for the Vehicle Routing Problem Based on Set Partitioning
N. Christofides, E. Hadjiconstantinou, A. Mingozzi 643

Incorporating Facet-Inducing Inequalities into Graphical-Construct- Based Lagrangean Relaxation Methodologies
A.I. Ali, S. Shmerling 645

AIR TRANSPORTATION (*CHAIRMAN*: M. BIELLI)

Optimization in Air Traffic Flow Management
D. Bertsimas, A.R. Odoni 661

Experiments with Complex Integer Programming Models for the Airport Ground Delay Problem
T. Mifflin, A. Assad, M. Ball, L. Stone, T. Thompson 689

Some Models and Algorithms for En Route Air Traffic Flow Management
V. Tosic, O. Babic, M. Cangalovic, D. J. Hohlacov 693

EQUILIBRIUM AND ASSIGNMENT MODELS (*CHAIRMAN*: Y. ASAKURA)

A Dynamic Assignment Model for Congested Radial Transit Networks
S. Saidi, M. Gendreau, F. Soumis 711

An Equilibrium Assignment Model for Intercity Transit Networks
A. Nuzzolo, F. Russo 715

Spatial Externalities and Networks
J. van den Bergh, P. Nijkamp 731

RELAXATION METHODS FOR VEHICLE ROUTING PROBLEMS (*CHAIRMAN*: F. SOUMIS)

A New Exact Algorithm for the Vehicle Scheduling Problem with Time Window and other Constraints
L. Bianco, V. Maniezzo, A. Mingozzi, S. Ricciardelli 747

A Multiplier Adjustment Method for the Multi-Depot Vehicle Scheduling Problem
M. Mesquita, J. Paixão 749

An Optimization Algorithm for the Vehicle Routing Problem with Time Windows based on Lagrangean Relaxation
N. Kohl, O.B.G. Madsen 763

Lower Bounds for the VRP via TSSP + 1 Decomposition
C. Noon 779

SIGNAL CONTROL (*CHAIRMAN*: G. E. CANTARELLA)

Queuing Optimization of a Signalized Intersection
A. Ceder, I. Reshetnik 781

REALBAND: An Approach for Real Time Coordination of Traffic Flows
P. Dell'Olmo, P.B. Mirchandani 805

A Method for Integrated Demand Responsive Signal Control
B. Friedrich 819

A Decentralized Traffic Control System Based on Logic Programming
G. Rinaldi, K. Truemper 823

NETWORK DESIGN (*CHAIRMAN*:H.S. MAHMASSANI)

Simplex-based Tabu Search for the Multicommodity Capacitated Fixed Charge Network Design Problem
T.G. Crainic, M. Gendreau, J. M. Farvolden, P. Soriano 825

Parallel Implementations of Bounding Procedures for Multicommodity Capacitated Network Design
B. Gendron, T.G. Crainic 833

A Multicriteria Analysis for Urban Network Design and Parking Location
G.E. Cantarella, A. Vitetta 839

Network Design Problems with Applications to Transportation Industries
P. Jaillet, G.Yu, M. Song 853

SPECIAL CASES OF VEHICLE ROUTING PROBLEM (*CHAIRMAN*: *A. MINGOZZI*)

The Traveling Salesman and the Vehicle Routing Problems with Delivery and Backhauls
S. Anily, G. Mosheiov 859

A Computerized Approach to the New York City School Bus Routing Problems
J. Braca, J. Bramel, B. Postner, D. Simchi-Levi 877

Managing Urban Snow Removal and Disposal Operations
J.F. Campbell, A. Langevin 881

TRAFFIC PLANNING MODELS (*CHAIRMAN*: S. PALLOTTINO)

Study on O-D Demand Adjustment Problem: A Bilevel Programming Approach
M. Florian, Y. Chen 887

An Efficient Algorithm for a Bicriterion Traffic Assignment Problem
P. Marcotte, D. Zhu 891

Modelling and Algorithmic Improvements in DSS for Urban Traffic Planning
D. Inaudi, V. Velardi 899

PUBLIC TRANSPORT (*CHAIRMAN*: D. H. RATLIFF)

A Planning and Design Model for Transit Route Networks with Coordinated Operations
M.C. Shin, H.S. Mahmassani 903

Distributed Dispatching Control in Public Transport Network
A. Adamski 917

Rural Public Transport Optimization
J. Cerny 935

ALGORITHMS FOR VEHICLE ROUTING PROBLEMS (*CHAIRMAN*: N. CHRISTOFIDES)

Algorithms for Special Classes of Sparse Directed Rural Postman Problems
M. Ball, H. Gun 939

The Traveling Salesman Problem under Accessibility Constraints
F. Semet 941

TRANSPORT PLANNING MODELS (*CHAIRMAN*: J. BRAMEL)

Initial Results of Theoretical and Experimental Research Concerning Road Traffic Flow Analysis and Models
V. Torrieri, V. Astarita, D. Gattuso, A. Vitetta 945

A DSS for the Urban Path Planning Integrating Parking Management, Public and Private Transportation Systems
S. Bertocchi, A. Sciomachen 957

Italian Research on Modelling Transport/Environment Interactions
M. Bielli, P. Reverberi 963

INDEX OF AUTHORS 981

CONTROL STRATEGIES SIMULATION ON URBAN NETWORKS: EVALUATION TOOLS AND TECHNIQUES

G.E. Cantarella* - E. Cascetta** - M. Di Gangi**

*Istituto di Ingegneria Civile ed Energetica - Università degli Studi di Reggio Calabria

**Dipartimento di Ingegneria dei Trasporti - Università degli Studi "Federico II" - Napoli

Large use of traffic engineering measures and/or the application of informative systems have been done in the last period in order to enhance traffic conditions. Depending on their way of application and their purposes, control strategies can supply messages concerning with either information or suggestions or indications, to be given either pre-trip or en-route. Whatever is the nature of messages, the way and the time they are supplied, messages themselves can vary both from a day to another (day-to-day dynamics) and from a time period to another within the same day (within-day dynamics).

Applied control strategies can vary from simple (adaptive or not) traffic lights to Variable Message Signals (VMS), to Root Guidance Systems (RGS) and to Trip Planning Systems (TPS). The evaluation of the performances of such systems needs a modelling tool able to simulate their effects.

In this paper a dynamic assignment model for a general network is presented. It follows a nonequilibrium approach in which flow fluctuation are modeled as a deterministic or stochastic dynamic process. This model includes a recently proposed Dynamic Network Loading model which adopts continuous distribution of packets over time for computing within-day variable link flows (and other related parameters) from path flows.

The proposed model seems able to evaluate the effects of control measures and strategies and can be used in designing and valuating informative systems. Applications to real cases are also shown in order to test the sensitivity and the operational characteristics of the model.

Dynamic network equilibrium problem formulated as an infinite dimensional variational inequality problem

Jia Hao WU *

May 9, 1994

Abstract

The dynamic network equilibrium problem is to find time-varying network flows given an origin-destination (O-D) demand departure rate for a time period in a congested network. In this paper, we formulate the problem mathematically as an infinite dimensional variational inequality problem. Then for the computational purpose it is converted into a finite dimensional variational inequality problem through a discretization. A standard simplicial projection method is adapted so that the solution method is highly decomposable by each origin-destination (O-D) and time period (interval).

Keywords: Dynamic traffic assignment problem, Equilibrium problem, Mathematical programming, Variational inequalities, Projection method, Dynamic network equilibrium problem.

Running head: Dynamic network equilibrium problem

The paper (first version) was first presented at Second EURO working group meeting on the urban traffic and transportation, Paris, France, 1993.

*Center for Research on Transportation, Université de Montréal, Montréal, Québec, Canada H3C 3J7.
This research was supported in part by an individual grant OGP0157735 from Natural Sciences and Engineering Research Council of Canada

1 Introduction

The dynamic network equilibrium problem is defined to find time-varying network flows given an origin-destination (O-D) demand departure rate for a fixed time period in a congested network (Smith, 1993, Friesz et al, 1993). This problem is also a direct extension of the static network equilibrium problem or the traffic equilibrium assignment problem (Dafermos, 1980) in certain sense.

The static network equilibrium problem has attracted extensive studies from researchers in traffic engineering, operations research, mathematical programming and economics for more than thirty years. See, for example, Wardrop (1952), Smith (1979) and Dafermos (1980). This problem is now well understood and there are many efficient algorithms which have been incorporated or adapted into some transportation planning softwares, which facilitates the wide applications of these methods. In general, they are well accepted for the strategic transportation planning. In the static network equilibrium problem, there exist two well defined equilibrias in the network: system optimal and user optimal. By the former, we mean that the network flows are defined such that total travel time (cost) is minimal while, by the later, we mean that each user is not able to change its route so that its travel time is less than the current one. They are both Wardrop's principles (Wardrop, 1952). However, there are several limitations in applications of the static network equilibrium problem as more and more transportation planners use it for some operational studies such as an interaction between the traffic signal control and the network equilibrium problem, the traffic management system, the passenger information system, the origin-destination matrix estimation using traffic counts and the variable demand departure rate analysis. The most serious one that one faces is the following. When the static network equilibrium problem is used for a short time period, the flows in the static network equilibrium problem may not be consistent with the flows observed during that time period. This is due to the travel time delay in the congested network, more specifically, the flows assigned may not be the same flows observed for the same time period.

This issue has motivated the research in the dynamic network equilibrium problem for

more than ten years after the papers by Merchant and Nemhauser (1978a, 1978b) on the system optimal dynamic traffic assignment. This problem is very challenging for several reasons. First, the definitions of the system and the user optimal dynamic network equilibrium problems are not unique unlike the static network equilibrium problem. Second, the problems formulated so far for the dynamic network equilibrium problem are usually not convex (either in the objective function or the feasible region). This results in many heuristic algorithms specially for the problem. Throughout the paper, we make no difference between the travel time and the travel cost for simplicity. .

In the literature, there exist two basically different versions of the dynamic network equilibrium problem based on definitions of the path travel time. We define the path travel time of version-I (called the instantaneous path time) as the path time along the path from an origin (O) to a destination (D) at an instant time t, which is calculated as the summation of arc travel times at the instant time t along the path k, and that of version-II (called the actual path travel time) as the path travel time along the path k departuring from an origin (O) at time t to a destination (D), which is equal to the summation of arc travel times where the arc travel time is determined at the time when the flows of the path k arrive at the arc rather than at the instant time t. Their mathematical formulations are given later (in (1) and (2) respectively). The definition of version-I is based on several published papers (Merchant and Nemhauser, 1978a, 1978b, Carey, 1987, Friesz et al, 1989, Ran et al, 1993).

Definition 1 (System optimal dynamic network equilibrium-I) *The flows in the network over a fixed time period [0, T] is said to be in system optimal dynamic equilibrium if the total instantaneous travel time of the transportation over the period is minimized.*

Definition 2 (User optimal dynamic network equilibrium-I) *The flows in the network over a fixed time period [0, T] is said to be in user optimal dynamic equilibrium if the total instantaneous travel time of the transportation for each user is minimized.*

It is noted that these two equilibrias are the same if there is no congestion in the network. We have the following observations for the two dynamic network equilibrias of version-I.

1. For some O-D pairs, users departuring at time t near T (near end of the time period) may not finish their trips at T.

2. Each user has full information over the time period.

3. Adjustments from one state to another occur instantaneously as conditions on the network change.

4. Users of same O-D pairs, departuring at same time, may come up with different travel time.

Due to the last two observations, the definitions of version-I have been very controversial and challenged. Current trend tends to the following definitions about the dynamic network equilibrium. See Janson (1991a, 1991b) and Friesz et al (1993).

Definition 3 (System optimal dynamic network equilibrium-II) *The flows in the network over a fixed time period [0, T] is said to be in system optimal dynamic equilibrium if the total actual travel time of the transportation over the period is minimized.*

Definition 4 (User optimal dynamic network equilibrium-II) *The flows in the network over a fixed time period [0, T] is said to be in user optimal dynamic equilibrium if the total actual travel time of the transportation for each user is minimized.*

It is noted that the basic difference between the version-I and the version-II lies in the definition of the path travel times (actual or instantaneous). The basic assumptions for version-II are the following.

- A1) Each user has full information over the time period.
- A2) The fixed time period is only applied for the departure time.

The assumption A1) can be justified if the network is in a stability condition and users have many driving experiences and are fully informed.

While author is revising the paper, I was brought into attentions of two recent related works. Bernstein et al (1992) developed three diagonalization related heuristics for a discretized variational inequality problem. Bernstein et al (1993) developed a variational control formulation of the simultaneous route and departure-time choice equilibrium problem with a solution procedure.

The study to be introduced is mainly theoretical, although it is designed with empirical applications for large scale networks in mind. In this paper, based on the definitions of version-II, we develop a different method to deal with the dynamic network equilibrium problem. The study of this paper is different from existing literature in several respects.

- The paper by Friesz et al (1993) is closely related with this study. Their model is more general in the sense that it considers the variable departure time. We note by a (slightly) different mathematical formulation that the dynamic system there is in fact a kind of the continuous network loading system. This is in fact a fixed point problem. Depending on the solution procedure, we may obtain a discretized network loading system. It is noted that in static network equilibrium the network loading system is no longer a fixed point problem and easy to solve.

- In this paper, a possible way of solving the infinite dimensional variational inequality problem is proposed by a method of discretization so that a discretized variational inequality problem, i.e., finite dimensional variational inequality problem, is obtained. Then a solution method, i.e., a simplicial projection method is developed where there exist three important computational problems, i.e., quadratic programming problem, time-dependent shortest path and assignment problem and the network loading system problem. Since the travel time mapping is determined by the fixed point problem, this approach is remarkablely complicated than that of standard classic methods for traditional finit-deminsional variational inequality problem.

The paper is organized as follows. In Section 2, basic concepts of the dynamic network are introduced. In Section 3, the dynamic network equilibrium problem of version-II is formulated and its several equivalent mathematical formulations are discussed. Section 4 is

devoted to the solution algorithms. In Section 5, some conclusions are given and the future studies are indicated.

2 Concepts of the dynamic network

We are given a network $G = (N, A)$ where the N is the set of nodes and a subset of N is the zone set where trips are generated and attracted and the A is the set of arcs. The network is a graphic representation of the traffic network modeling system under the consideration. The network here is very general which may include the signal control policies at various road intersections. We let

k	=	a path of a set of consecutive arcs in A;
K_i	=	a set of paths from an origin p to a destination q ($i = (p, q)$);
K	=	a set of paths;
δ	=	the arc-path incidence matrix where δ_{ak} being 1 if the path k passes the arc a and 0 otherwise;
$[0, T]$	=	the fixed time period under the consideration for departure only;
T'	=	the finishing time ($T' \geq T$), i.e., the time all the flows reach their destinations;
I	=	the set of total O-D pairs;
a^+	=	the successor arc of the arc a along a path;
$g_i(t)$	=	the departure rate for the O-D pair i at time t;
$h_k(t)$	=	a path flow rate of path k departuring at time t;
$h(t)$	=	a set of path flow rates departuring at time t;
h	=	$(h(t))$;
$v_a(t)$	=	arc flows on arc a at time t;
$v(t)$	=	a set of arc flows at time t;
v	=	$(v(t))$;
$s_a(v(t))$	=	the arc travel time on arc a that user takes at entering time t;
$s(v(t))$	=	a set of arc travel times at time t;
s	=	$(s(v(t)))$;
$S_k(t, h)$	=	the actual travel time for path k given that the departure from the origin occurs at time t; and
S	=	$(S_k(t, h))$.

Here $s_a(v(t))$ is an estimated arc travel time rather than the actual travel time. We assume that the arc travel time is determined at time t of entering the arc. However, once a user moves onto the arc, the time t changes and thus the arc travel time may change. Based on these definitions, we have following relationships. According to the definition of the actual path travel time of the path k departuring at time t, i.e., $S_k(t, h)$, we assume the following

formula:

$$S_k(t,h) = \sum_{a \in A} \delta_{ak} s_a(v(t + S_{ak}(t,h))), \quad \forall\, k \in K_i,\ i \in I, t \in [0,T] \tag{1}$$

where $S_{ak}(t,h)$ is the actul travel time that the user takes along the path k departuring at time t arriving at the tail of arc a and $\delta_{ak} = 1$ if the path k passes the arc a and $\delta_{ak} = 0$ otherwise. The matrix $\delta = (\delta_{ak})$ is constant. It is clear that the actual path travel time is basically different from the instantaneous path travel time which is given as

$$S_k(t,h) = \sum_{a \in A} \delta_{ak} s_a(v(t)), \quad \forall\, k \in K_i,\ i \in I, t \in [0,T]. \tag{2}$$

In the paper, we use the actual path travel time instead of the instantaneous path travel time. For simplicity, in the following, we drop the word 'actual'.

We assume that $v(0) = 0$ and $v(T') = 0$. Suppose that one departures at time $\tilde{t}$ and arrives at the tail of arc a at t along the path k, i.e.,

$$t(\tilde{t}) = \tilde{t} + S_{ak}(\tilde{t}, h).$$

Then we define $\theta_{ak}(t)$ to be the inverse function of t, i.e.,

$$\tilde{t} = \theta_{ak}(t) = (t(\tilde{t}))^{-1} = (\tilde{t} + S_{ak}(\tilde{t}, h))^{-1}. \tag{3}$$

Thus $\theta_{ak}(t)$ implies the departure time of the user who arrives at the tail of arc a at t along the path k. Let $O(a)$ be the tail of the arc a and the path k starts at o. If $O(a) = o$, then $\theta_{ak}(t) = t$.

As can be seen, the regularity of the function t requires that if $\tilde{t}' < \tilde{t}''$ then $t(\tilde{t}') < t(\tilde{t}'')$. This implies that no vehicle can overtake another on a given path. Friesz et al (1993, section 3) showed that the path-specific travel time function $t(\tilde{t})$ is regular for certain classes of functions $s(v)$. Here throughout the paper, we assume that $t(\tilde{t})$ is regular.

Let $v_{ak}(t)$ be an accumulative arc flows of arc a induced by the path flow rate $h_k(y)$ with y from time $\theta_{a+k}(t)$ to $\theta_{ak}(t)$, i.e.,

$$v_{ak}(t) = \int_{\theta_{a+k}(t)}^{\theta_{ak}(t)} h_k(y)dy \forall a \in A_k, k \in K_i, i \in I, t \in [0,T] \tag{4}$$

where A_k is the set of arcs of the path k.

The following equation is a straight forward summation:

$$v_a(t) = \sum_{i \in I} \sum_{k \in K_i} v_{ak}(t), \forall\, i \in I,\ a \in A, t \in [0, T']. \tag{5}$$

Define $\Omega(t) = \{h(t) : \sum_{k \in K_i} h_k(t) = g_i(t), \forall i \in I, h_k(t) \geq 0, \quad \forall\, k \in K_i, i \in I\}$ for all $t \in [0, T]$, which is the region of feasible path flow rates. Let $\Omega = (\Omega(t))$.

In the following, we illustrate some notations by a specific example (Figure 1).

The path from the node origin 1 to the node destination 4 is called the path k. The departure time from the node origin 1 of the path k is t. The travel time to the node 3 is $t + S_{ak}(t, h)$. The arc (3,4) travel time is determined as $s_{(3,4)}(v(t + S_{ak}(t, h)))$. The departure time from the node origin 1 of the path k given that the user arrives at node 3 at time $\tilde{t}$ is $\theta_{ak}(\tilde{t})$ where $\tilde{t} = t + S_{ak}(t, h)$. The total travel time for the path k given that the departure from the node origin 1 occurs at time t is $S_k(t, h)$.

It is interesting to consider the concept of the network loading system (NLS) in the infinite dimensional space. We define NLS as finding v, s, S given feasible h. We denote it as $F(v, s, S|h) = (v, s, S)$ which is really a fixed point problem. Here this system is characterized by equations (1) and (4) where all the variables are interrelated each other as shown in Figure 2. Cascetta and Cantarella (1991) have a similar fixed point problem in their within-day model. It is this system that makes the mapping S more complicated to analyse. We note that in static network equilibrium problem NLS is much simpler and no longer a fixed point problem which is given as $v = \delta h$ and $S = \delta^t s$ where v is a vector of arc flows, h is a vector of path flows, s is a vector of arc travel time, S is a vector of path travel time and δ is arc-path incidence matrix. The relationships of the variables are shown in Figure 3.

3 Dynamic network equilibrium problem formulations

In this section, we first define mathematically the user optimal dynamic network equilibrium-II conditions and present several other equivalent formulations.

Consider the following problem (P0):

$$(P0) \quad S_k(t,h) \begin{cases} = U_i(t) & \text{if } h_k(t) > 0 \\ \geq U_i(t) & \text{otherwise} \end{cases} \quad \forall k \in K_i, i \in I, t \in [0,T] \tag{6}$$

where $h_k(t) \in \Omega(t)$ and $U_i(t)$ is the minimal travel time for O-D pair i departuring at time t. This is a mathematical statement of the user optimal dynamic network equilibrium condition (Definition 4). The problem (P0) can be rewritten as the following problem (P1):

$$(P1) \quad \begin{cases} S_k(t,h) - U_i(t) - \beta_k(t) = 0 \\ h_k(t) \in \Omega(t) \\ \beta_k(t) \geq 0 \\ h_k(t)\beta_k(t) = 0. \end{cases} \quad \forall k \in K_i, i \in I, t \in [0,T] \tag{7}$$

It is well known that the problem (P1) is a necessary condition of the following infinite dimensional variational inequality problem (P2): find $h^*(t) \in \Omega(t), \forall t \in [0,T]$ such that

$$(P2) \quad S(t,h^*)^t(h(t) - h^*(t)) \geq 0, \quad \forall h(t) \in \Omega(t), \forall\, t \in [0,T]. \tag{8}$$

The next Proposition establishs the equivalence relationships among the three problems.

Proposition 3.1 *The problems (P0), (P1) and (P2) are equivalent.*

Proof. It is easy to see that (P2) $\rightarrow$ (P1) $\rightarrow$ (P0). Now we show that the problem (P0) $\rightarrow$ (P2) also. Suppose $h^*(t)$ satisfies the user optimal dynamic network equilibrium condition and $h(t) \in \Omega(t)$. We have

$$\begin{aligned} 0 &= \sum_{i \in I} \sum_{k \in K_i} (h_k(t) - h_k^*(t)) \\ &\quad (\text{ since both } h_k(t) \text{ and } h_k^*(t) \text{ are both feasible path flow rates}) \\ &= \sum_{i \in I} \sum_{k \in K_i} (h_k(t) - h_k^*(t)) U_i(t) \\ &\leq \sum_{i \in I} \Big(\sum_{k \in K_i | h_k(t) > h_k^*(t)} (h_k(t) - h_k^*(t)) S_k(t,h^*) + \end{aligned}$$

$$
\begin{aligned}
& \sum_{k \in K_i | h_k(t) < h_k^*(t)} (h_k(t) - h_k^*(t)) U_i(t)) \\
= & \sum_{i \in I} \sum_{k \in K_i} (h_k(t) - h_k^*(t)) S_k(t, h^*) \\
= & S(t, h^*)^t (h(t) - h^*(t))
\end{aligned}
$$

where the last second equality comes from $U_i(t) = S_k(t, h^*)$ if $h_k^*(t) > 0$. Therefore the proof is complete. □

It is easy to see that (P2) can be shown to be equivalent to find $h^*(t) \in \Omega(t), \forall t \in [0, T]$ such that

$$
\int_0^T S(t, h^*)^t (h(t) - h^*(t)) dt \geq 0, \quad \forall h(t) \in \Omega(t), \forall\, t \in [0, T]. \tag{9}
$$

Now we consider a formulation of a system optimal dynamic network equilibrium problem of version II:

$$
\begin{aligned}
& \min \quad \textstyle\sum_{i \in I} \int_0^T \sum_{k \in K_i} S_k(t, h) h_k(t) dt \\
& s.t. \qquad h(t) \in \Omega(t), \forall t \in [0, T].
\end{aligned}
$$

The first order necessary condition is given as follows:

$$
\begin{cases}
S_k(t, h) + S_k'(t, h) h_k(t) - U_i(t) + \beta_k(t) = 0 & \\
h_k(t) \in \Omega(t) & \\
\beta_k(t) h_k(t) = 0 & \forall k \in k_i, i \in I, t \in [0, T] \\
\beta_k(t) \geq 0 &
\end{cases} \tag{10}
$$

which is equivalent to

$$
U_i(t) \begin{cases} = S_k(t, h) + S_k'(t, h) h_k(t) & \text{if } h_k(t) > 0 \\ \leq S_k(t, h) + S_k'(t, h) h_k(t) & \text{if } h_k(t) = 0. \end{cases} \tag{11}
$$

4 Solution algorithms

In the finite dimensional static network equilibrium problems defined in the path flow space, there exist some efficient algorithms. See Bertsekas and Gafni (1982) for the traffic equilibrium problem and Wu, Florian and Marcotte (1991) for the transit equilibrium problem, where some kinds of restriction strategies are used. In fact, the well known simplicial decomposition method (Lawphongpanich and Hearn, 1984, Wu and Florian, 1992) can be

applied. Our basic idea is to first convert the infinite dimensional variational inequality problem (P2) into a finite dimensional variational inequality problem approximately by a discretization and then to use a modified restriction method to solve it. The discretization was discussed in Rockafellar (1982).

4.1 An finite dimensional variational inequality problem

The departure time period $[0,T]$ is first subdivided into M small time intervals. The length of each interval is T/M denoted by Δt and each time interval is indexed by m. The number M should be carefully determined based on the real network on hand which is not discussed here. However, it is always true that the smaller the Δt the more accurate the discretization. This feature is not shared in some existing discretized models (Janson, 1991a, 1991b) where Δt has to be larger than a certain value.

Let $h_k(m)$ = the path flow departuring from the interval m, $v_a(m)$ = the arc flows on arc a at the interval m, $S_k(m,h)$ = the total travel time for path k given that the departure from the origin occurs at the interval m. Let A_k be the set of consecutive arcs of the path k, $v=(v(m)), s=(s(m)), S=(S(m,h))$, $h=(h(m))$ and $\Omega=(\Omega(m))$. A finite dimensional variational inequality problem (P3) is adapted from (P2) and formulated as follows: find $h^*(m)\in\Omega(m), m=1,2,...,M$, such that

$$\text{(P3)}\quad S(m,h^*)^t(h(m)-h^*(m))\geq 0, \forall\ h(m)\in\Omega(m), \tag{12}$$

where $\Omega(m)=\{h(m)|\sum_{k\in K_i}h_k(m)=g_i(m),\forall i\in I, h_k(m)\geq 0,\forall k\in K_i, i\in I,\}$ for all m. It is clearly seen that this problem is a finite dimensional variational inequality problem over a Cartesian product of spaces. Let $\Omega=\Pi_{m=1}^{M}\Omega(m)$ where $\Omega(m)$ is a compact and convex subset (Bertsekas and Tsitsiklis, 1989, Cohen and Chaplais, 1988). Let $h^*=(h^*(1),h^*(2),...,h^*(M))$ and $S(h^*)=(S(1,h^*),S(2,h^*),...,S(M,h^*))$. Thus (P3) is rewritten as a system of M coupled variational inequality problems: find $h^*\in\Omega$ such that

$$S(h^*)^t(h-h^*)\geq 0,\forall\ h\in\Omega. \tag{13}$$

Now let us consider the basic relationships of the dynamic network in the discretized model

corresponding to (1) and (4), i.e., the discretized NLS:

$$S_k(m,h) = \sum_{a \in A} \delta_{ak} s_a(v(m + S_{ak}(m,h)/\Delta t)), \ \forall k \in K_i, i \in I, m = 1, ..., M, \quad (14)$$

$$v_{ak}(m) = \sum_{r=\theta_{a+k}(m)}^{\theta_{ak}(m)} h_k(r)\Delta t, \quad (15)$$

$$\forall a \in A_k, k \in K_i, m = 1, ..., M', \tilde{m} \leq m,$$

$$v_a(m) = \sum_{i \in I} \sum_{k \in K_i} v_{ak}(m), \ \ \forall a \in A, m = 1, ..., M', \quad (16)$$

where $\theta_{ak}(m) = (\tilde{m} + S_{ak}(\tilde{m},h)/\Delta t)^{-1}, \forall k \in K_i, m = 1, ..., M', \tilde{m} \leq m$ and $\tilde{m}$ is the departure interval of the path k with $v_{ak}(0) = 0, M' \geq M$ and $v(M') = 0$.

This discretization is complicated by the fact that the time period is discretized while the path travel time is not a real number in general. Specifically, both $S_{ak}(m,h)$ and $\theta_{ak}(m)$ may not be integer. If the interval $m + S_{ak}(m,h)/\Delta t$ and $\theta_{ak}(m)$ are not integer, we may determine them by rounding them to smallest integers. In order to cope with this issue, we may use certain average techniques (Drissi-Kaitouni and Gendreau, 1992). It is noted that the variational inequality problem (P3) is very much different from the variational inequality problem of the static network equilibrium problemin that the discretized NLS (14-15) has to be taken into account in (P3).

4.2 A model algorithm

In fact, the solution methods for the standard variational inequality problem (P3) have been discussed (Pang, 1985, Cohen and Chaplais, 1988). We propose a similar projection method to solve it. Let the matrix B be symmetric and positive definite and the parameter α be positive. Starting with $h^0 \in \Omega$, for $n = 0, 1, ...$, solve the following problem (P4):

$$\text{(P4)} \quad \min S(h^n)^t(h - h^n) + \frac{1}{2\alpha}(h - h^n)^t B(h - h^n) \quad s.t. \ h \in \Omega \quad (17)$$

and let the solution be h^{n+1}. If the matrix B is diagonal and positive definite, then the problem (P4) can be decomposed by each time interval and O-D pair, i.e., (P4) is equivalent to solving the following problem (P5), i.e., for $i \in I, m = 1, ..., M$,

$$
\begin{aligned}
\min \quad & \textstyle\sum_{k\in K_i(m)} \; S_k(m,h^n)^t(h_k(m)-h_k^n(m)) + \frac{1}{2\alpha}(h_k(m)-h_k^n(m))^t B_k(m)(h_k(m)-h_k^n(m)) \\
(P5) \qquad s.t \quad & \sum_{k\in K_i(m)} h_k(m) = g_i(m), \\
& h_k(m) \geq 0, \forall k \in K_i(m),
\end{aligned}
$$

where the element $B_k(m)$ is a positive constant for all k and m and $K_i(m)$ is the set of paths for all the O-D pair i and the time interval m. This decomposition allows for the parallel computation at the same time. At this point, it is still not clear about the convergence of this algorithm since the property of the travel time mapping S is not fully understood due to the discretized NLS.

4.3 A restriction method

Now we consider the solvability of the problem (P3). In general, the set $K_i(m), i = 1, ..., |I|$, is very large, which prohibits the application of the model algorithm. In order to cope with this issue, a well known restriction (a kind of simplicial decomposition) strategy is adopted. This strategy has been successfully used (Bertsekas and Gafni, 1982, Wu, Florian and Marcotte, 1991). The basic idea of this strategy is the following. Starting with a subset of $K_i(m)$, i.e., $K_i^0(m)$ for all i and m, and a initial path flow h^0, we solve the quadratic program (P4), and find h, v, s, S as well as a new time-dependent shortest path and assignment problem based on the $s(v)$. Then we add the path into the subset $K_i^0(m)$ and solve (P4) again and repeat. The disadvantage of this strategy is that given the subset, say $K^l(m)$, the following restricted variational inequality problem(P6), i.e., find $h(m)^* \in \Omega^l(m), m = 1, ..., M$, such that

$$
(P6) \quad S(m,h^*)^t(h(m)-h^*(m)) \geq 0, \forall h(m) \in \Omega^l(m), \; m = 1, ..., M, \tag{18}
$$

where $\Omega(m)^l = \{h(m) : \sum_{k\in K_i^l(m)} h_k(m) = g_i(m), \forall\, i \in I, h_k(m) \geq 0, k \in K_i^l(m), \forall i \in I\}$ is only approximated by one quadratic program problem. This approximation can be improved by a sequence of the quadratic program problems.

We note that if $\Omega(m)^l = \Omega(m)$ then (P6) is the same as (P3). Our modified restriction strategy is that given $\Omega(m)^l$, we solve (P6) with a sequence of the following quadratic programming problems (P7) instead of just one: for $m = 1, ..., M, i \in I$,

$$(P7) \quad \begin{array}{lll} \min & \sum_{k \in K_i^l(m)} & S_k(m, h^n)^t (h_k(m) - h_k^n(m)) + \frac{1}{2\alpha}(h_k(m) - h_k^n(m))^t B_k (h_k(m) - h_k^n(m)) \\ s.t & \sum_{k \in K_i^l(m)} & h_k(m) = g_i(m), \\ & & h_k(m) \geq 0, \forall k \in K_i^l(m), \end{array}$$

where $K_i^l(m)$ is a set of time-dependent paths found for the O-D pair i and the time interval m.

Now given h we consider the computation of v, s and S, i.e., the discretized NLS (14)-(15). In the case of the static network equilibrium problem, they are straightforward to compute. However in the dynamic network equilibrium problem, the computation of these variables are complicated since we have to solve the discretized NLS. This system can not be solved analytically. Therefore we propose a method to solve it approximately. Given h, v^0, S^0, for $j = 0, 1, ...,$ compute $\theta_{ak}(m)$ for all a, k, m with S^0, then calculate $v_{ak}^{j+1}(m), \forall a \in A_k, k \in K_i^l(m), i \in I, m = 1, ..., M'$ and then compute

$$v_a^{j+1}(m) = \sum_{i \in I} \sum_{k \in K_i^l(m)} v_{ak}^{j+1}(m), \forall a \in A, m = 1, ..., M'$$

and

$$S_k^{j+1}(m) = \sum_{a \in A} \delta_{ak} s_a(v^{j+1}(m + S_{ak}^j(m)/\Delta t)), \forall k \in K_i^l(m), i \in I, m = 1, ..., M$$

and repeat.

Next we consider the time-dependent shortest path and assignment problem. The problem is stated as

$$\min \sum_{m=1}^{M} \sum_{i \in I} \sum_{k \in K_i(m)} S_k(m, h) \tilde{h}_k(m) \quad s.t. \quad \tilde{h}(m) \in \Omega(m) \tag{19}$$

which can be decomposed by each time interval m and O-D pair i, i.e.,

$$\min \sum_{k \in K_i(m)} S_k(m,h)\tilde{h}_k(m) \quad s.t. \sum_{k \in K_i(m)} \tilde{h}_k(m) = g_i(m), \tilde{h}_k(m) \geq 0, \forall k \in K_i(m). \quad (20)$$

For this problem, there are efficient labeling setting algorithms available such as a modified Dijkstra's algorithm (Cooke and Halsey, 1966, Yucaoglu, 1973, Orda and Rom, 1990). Now we are ready to present the restriction algorithm:

(step 1) Start with two paths $\Omega^2(m), l = 2,\ m = 1, ..., M$ with h^0.
(step 2) Solve the restricted variational inequality problem(P6) to find $h^*(m) \in \Omega^l(m), m = 1, 2, ..., M$, i.e., repeat
(step 2.1) Compute v, s, S by solving the discretized NLS (14)-(15).
(step 2.2) Solve the quadratic programming problem (P7).
(step 3) Computing the time-dependent shortest path and assignment problem (19) with new path $\bar{k}_i(m), i \in I, m = 1, ..., M$.
Let $K_i^{l+1}(m) = K_i^l(m) \bigcup \bar{k}_i(m), \forall i \in I, m = 1, ..., M$. Goto (step 2).

The step 2 is a sequence of step 2.1 and step 2.2. The convergence test is carried out by using the gap function, i.e.,

$$gap = \sum_{m=1}^{M} S(m,h)(\tilde{h}(m) - h(m)).$$

If the *gap* is close to zero, then we find an approximate solution and stop. It is noted that this test is done very easily.

In summary, the restriction method involves three subproblems: the quadratic programming problem, the discretized NLS and the time-dependent shortest path and assignment problem.

5 Conclusions

In this paper, we presented an infinite dimensional variational inequality problem for the dynamic network equilibrium problem and proved their equivalence. Then the problem is discretized into a finite dimensional variational inequality problem for the computational purpose which is solved by a restriction method. In general the solution method is heuristic

unless under some restrictive assumptions although the method is similar in the theory of the variational inequality problem. We hope that the work would provide some insights into the difficult dynamic network equilibrium problem, from which the future research may benefit.

Several possible extensions are the following. We may incorporate the variable demand departure rate over time in the model. This issue has been under investigations. And then it is possible to extend it to the dynamic transit equilibrium problem (Wu, Florian and Marcotte, 1991, Wu and Florian, 1992) with fixed or variable demand departure rates. Since the model provides very detailed information, the dynamic traffic signal control with the dynamic network equilibrium problem is also an interesting research.

Acknowledgement The author thanks Professor Micheal Florian, Professor Michel Gendreau, Professor Patrice Marcotte and Mr. Yang Chen for several discussions.

References

[1] Aubin, J.P. and Ekeland, I., Applied Nonlinear Analysis, (John Wiley and Sons, New York, 1984).

[2] Bernstein, D., Friesz, T., Tobin, R., Shenoi, R. and Wei, B-W, "Solving a variational inequality formulation of the simultaneous route and departure-time choice equilibrium problem", Presented at The 39th North American Meeting of the RSAI (1992).

[3] Bernstein, D., Friesz, T., Tobin, R., and Wei, B-W, "A variational control formulation of the simultaneous route and departure-time choice equilibrium problem", *Transportation and Traffic Theory /C.F. Daganzo (Editor)* (Elsevier Science Publishers B.V., 1993).

[4] Bertsekas, D.P. and Tsitsiklis, J., Parallel and distributed computation: Numerical methods (Prentice-Hall, 1989).

[5] Bertsekaş, D.P. and Gafni, E.M.," Projection methods for variational inequalities with application to the traffic assignment problem", *Mathematical Programming Study* 17 139-159 (1982).

[6] Cascetta, E. and Cantarella, G.E., "A day-to-day and within-day dynamic stochastic assignment model", *Transportation Research* **25A** 277-291 (1991).

[7] Cohen, G. and Chaplais, F.,"Nested monotony for variational inequalities over product of spaces and convergence of iterative algorithms", *Journal of Optimization Theory and Applications* **59** 369-390 (1988).

[8] Cooke, K.L. and Halsey, E.,"The shortest route through a network with time-dependent internodal transit times", *Journal of Optimization Theory and Applications* 14 493-498 (1966).

[9] Dafermos, S., "Traffic equilibria and variational inequalities," *Transportation Science* 14 42-54 (1980).

[10] Dafermos, S., "An iterative scheme for variational inequalities", *Mathematical Programming* **26** 40–47 (1983).

[11] Drissi-kaitouni, O. and Gendreau, M "A new dynamic traffic assignment model", Publication 854, Centre de recherche sur les transports, Université de Montréal (1992).

[12] Friesz, T.L., Luque, J., Tobin, R.L. and Wie, B.W., "Dynamic network traffic assignment considered as a continuous time optimal control problem", *Operations Research* **37** 893-901 (1989).

[13] Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L. and Wei, B.W., "A variational inequality formulation of the dynamic network user equilibrium problem", *Operations Research* **41** 179-191 (1993).

[14] Janson, B.N.,"Convergent algorithm for dynamic traffic assignment", *Transportation Research Record* **1328** 69-80 (1991a).

[15] Janson, B.N., "Dynamic traffic assignment for urban road network", *Transportation Research-B* **25B** 143-161 (1991b).

[16] Lawphongpanich, S. and Hearn, D.W.,"Simplicial decomposition of the asymmetric traffic assignment problem", *Transportation Research B* 123-133 (1984).

[17] Merchant, D.K. and Nemhauser, G.L.," A model and an algorithm for the dynamic traffic assignment problem", *Transportation Science* **12** 183-199 (1978a).

[18] Merchant, D.K. and Nemhauser, G.L.," Optimality conditions for a dynamic traffic assignment model", *Transportation Science* **12** 200-207 (1978b).

[19] Orda, A. and Rom, R.,"Shortest-path and minimum-delay algorithms in networks with time-dependent edge-length", *Journal of the Association for Computing Machinery* **37** 607-625 (1990).

[20] Pang, J.S. and Yu, C.S., "Linearized simplicial decomposition methods for computing traffic equilibria on networks", *Networks* 14 427-438 (1984).

[21] Pang, J.S., "Variational inequality problems over product sets: applications and iterative methods", *Mathematical Programming* **31** 206-219 (1985).

[22] Ran, B, Boyce, D.E. and LeBlance, L.J., "A new class of instantaneous dynamic user-optimal traffic assignment models", *Operations Research* **41** 192-202 (1993).

[23] Rockafellar, R.T., "Lagrange multipliers and variational inequalities," in : R.W. Cottle, F. Giannessi and J.L. Lions, eds., Variational inequalities and Complementarity problems: Theory and Applications (Wiley, New York, 1980) pp303-322.

[24] Smith, M.J., "The existence, uniqueness and stability of traffic equilibria", *Transportation Research* 13B 295-304 (1979).

[25] Smith, M.J., "A new dynamic traffic model and the existence and calculation of dynamic user equilibria on congested capacity constrained road network", *Transportation Research* 27B 49-63 (1993).

[26] Wardrop, J.G., "Some theoretical aspects of road traffic research", *Proc. Inst. Civil Engineers*, Part II, 325-378 (1952).

[27] Wu, J.H., Florian, M. and Marcotte, P., "Transit equilibrium assignment: a model and solution algorithm", *Publication* 721, CRT, Université de Montréal (1991) and Transportation Science (forthcoming).

[28] Wu, J.H. and Florian, M., "A simplicial decomposition method for the transit equilibrium assignment problem", Annals of Operations Research 44 245-260 (1993)

[29] Yucaoglu, E., Solution algorithms for resource and route constrained shortest path problems in time-dependent transportation networks, A Ph.D dissertation, the department of industrial engineering of Stanford University, 1973.



Figure 1: An example for notations

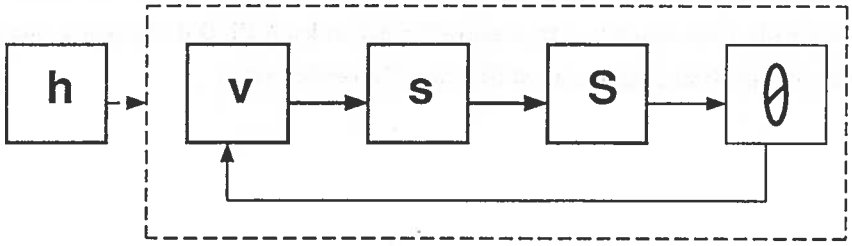


Figure 2: Dynamic network loading system

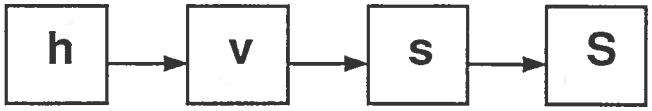


Figure 3: Static network loading system

FLOW PROPAGATION DESCRIPTION IN DYNAMIC NETWORK ASSIGNMENT MODELS

J. Enrique Fernández L. and Joaquín de Cea Ch.

Depto. Ingeniería de Transporte, Universidad. Católica de Chile

Casilla 306, FAX: (56-2) 552 4054 Santiago 22, CHILE.

ABSTRACT

In the present paper we analyze the causes and peculiarities of the flow propagation problems derived from the use of exit function formulations in dynamic traffic assignment models; conditions for a correct formulation are defined and consistent exit function formulations are developed; the corresponding flow propagation characteristics are studied through the use of some simple examples. Special consideration is given to the treatment of FIFO disciplines in the dynamic propagation of flows. Finally, a system optimum version of dynamic assignment is formulated using optimum control theory and the corresponding solution characteristics are analyzed. The paper described contributes to the understanding of the characteristics of dynamic propagation of flows in transportation networks and helps to find appropriate formulations in order to advance in the development of more realistic dynamic network assignment models.

1. Introduction

During the last years an increasing number of transportation science researchers have started to work in the development of the so called "dynamic traffic assignment models"; these, try to replace the static view of traffic prediction, used by the traditional network equilibrium models, based on Beckmann's mathematical programming formulations of Wardrop's static traffic equilibrium principles (Beckmann et al., 1956), with a dynamic description of network flows during a given period of analysis [0,T].

Some relevant contributions that have analyzed the dynamic assignment problem for transportation networks include the papers by Merchant and Nemhauser (1978a, 1978b), Carey (1986,1987,1992), Friesz et al. (1989), Wie, Friesz and Tobin (1990), Papageorgiou (1990), Janson (1991a, 1991b), Ran, Boyce and LeBlanc (1993), Addison and Heydecker (1993), (Hao, 1993), and others.

The main difference with static network equilibrium models is that flows change with time and therefore an evolution and propagation of them through the network is obtained; flows travel through network links and paths at variable speeds with time (flow propagation) and the magnitud of flows going through a given point in the network also changes with time (flow evolution). Following the formulation used by Merchant and Nemhauser (1978a), in most of the first models studied the description of flow dynamics (evolution and propagation) is obtained by including a dynamic equation in the problem constraint set, in which a key element is the so called "link exit function". This defines a relation between the flows that exit a link at a given time and the number of vehicles inside the link.

However, the formulations so far proposed on the basis of this link exit functions have presented serious problems producing non realistic descriptions of the flow propagation phenomena. For continuous time formulations, the main problem is the instantaneous flow propagation through the network, resulting in some proportion of the flows reaching their destination at the same time t at which they are leaving their origin (Fernández and De Cea, 1993 and 1994). Consequently with this, gross violations of the FIFO flow discipline (Carey, 1992) and solutions that satisfy non-realistic behavioral principles, as the so called "Boston equilibrium" (Friesz et al. (1989)) or similar principles (Ran et al. 1993), are obtained.

Some authors have tried to avoid the use of an exit function (Janson, 1990; Ran et al. 1993; Hao, 1993) but this approach has lead to considerable more complicated and less tractable formulations.

In this paper we take the approach of developing consistent and realistic formulations of flow dynamics, without renouncing to the use of exit functions, to describe flow evolution and propagation through the network. Therefore, conditions for a correct formulation are defined and consistent exit function formulations are developed and the corresponding flow propagation characteristics are studied through the use of some simple examples. Special

consideration is given to the treatment of FIFO disciplines in the dynamic propagation of flows.

All dynamic network assignment formulations considered assume that O-D demands are perfectly known and given for the whole analysis period [0,T].

2. Basic Definitions and Notation

We will assume a transportation network represented by a directed graph $G = (\mathcal{N}, \mathcal{A})$, with $\mathcal{N}$ the set of nodes and $\mathcal{A}$ the set of links; nodes represent street intersections, special points where link design characteristics suffer changes and centroids corresponding to spatial points where origin and destination flows are located for modeling purposes; links represent city streets. A general index a will be used to identify a link and i, j, k, n will identify nodes, with k, n reserved for centroids (k for Origins (O) and n for Destinations (D)). In addition, we will adopt the following basic notation:

W	: set of O–D pairs in the network.
w	: a general element of set W, with $w = (k, n)$ and k, n centroids.
P	: set of paths perceived over network G.
P_w	: subset of paths associated with O–D pair w.
p	: a general element of set P.

3. Dynamic network flows description

The analysis will be performed for a period of time $[0, T]$ corresponding to a typical operating period (morning peak); we will use variable t to represent a time instant within the period, $t \in [0, T]$.

Static transportation equilibrium models assume steady state operating conditions over the network (Wardrop, 1952, Beckmann McGuire and Winstein, 1956, LeBlanc, Morlok and Pierskalla, 1975, Florian and Nguyen, 1978, Abdulaal and Leblanc, 1979, Aashtiani and Magnanty 1981, Dafermos, 1982, Fernández and Friesz, 1983, etc.).Therefore the value of O–D demands, flows and level of service variables is stationary and independent of time. In such case, the flow u_a, entering a given link a, is equal to the flow g_a, exiting the link and equal to the constant equilibrium link flow for the whole analysis period; no need appears to use in the model formulation the link traffic density, $d_a = [f_a/l_a]$ where, f_a represents the number of vehicles within the link and l_a is the link length.

By definition, dynamic models relax this assumption and all the mentioned variables u_a, g_a and f_a, are considered functions of time $(u_a(t), g_a(t), f_a(t))$; this introduces the new concepts of "flows evolution through time" and "flows propagation over the network", that are irrelevant for static formulations.

3.1 The Merchant and Nemhauser exit funtions

In order to describe the evolution of flows through time in a link, Merchant and Nemhauser (1978a), using a discrete time formulation, introduce the concept of "exit function", $g_a[f_{a,t}]$; this gives the number of vehicles leaving link a, during

time interval t, as a function of the number of vehicles inside the link, $f_{a,t}$. Then, they state that the number of vehicles that are inside a link, $f_{a,t}$, at any time interval t, is equal to the number of vehicles that were inside the link at the beginning of the preceding interval, $f_{a,t-1}$, plus the number of vehicles entering the link during interval t, $u_{a,t}$, minus the vehicles leaving it during t, $g_{a,t}$.

Using a continuos time optimal control model, Friesz et al., (1989), reformulate the same dynamic equation as the following differential equation:

$$\frac{df_a(t)}{dt} = u_a(t) - g_a[f_a(t)], \qquad a \in \mathcal{A}, \quad t \in [0,T], \tag{1}$$

where the left hand side represents the instantaneous rate of change, in the number of vehicles that are inside link a at time t, with the other variables having the same meaning as in the discrete time formulation. The initial value of the number of vehicles in each link of the network, represented by $f_a(0), \forall a \in \mathcal{A}$, is assumed known. Therefore, the corresponding value for any time t is given by:

$$f_a(t) = \int_0^t \{u_a(t) - g_a[f_a(t)]\}dt + f_a(0), \qquad a \in \mathcal{A}, \quad t \in [0,T], \tag{2}$$

It is here convenient noticing that in the preceding equations, u and g represent flows expressed as number of vehicles that go through an specific point of the network (in our case, entrance and exit link nodes), per unit of time. However, f represents only the number of vehicles inside the link at a given time; then, the derivative of f with respect to time, that appears in the left hand side of equation (1), has also flow units.

An important characteristic of the dynamic models that use equations (1) is that they use the link as modeling unit, assuming that once a flow enters a link it is immediately and homogeneously distributed over its whole length; then, given that each link a has a length l_a, it is possible to define an average density for the link: $d_a(t) = [f_a(t)/l_a]$; notice that at any given time t, d_a is the same for any point along the link, nevertheless its value will continuously change with t.

The definition of exit function g_a, allows Merchant and Nemhauser (1978a), and Friesz et al., (1989) to describe flows evolution and propagation using dynamic equations (1), which have a simple formulation, linear in the controls u_a and in the exit function g_a. However, this dynamic representation of network flows is not as general as it could seem at first glance and implies a very special, and rather restrictive, dynamic description of the evolution and propagation of flows over the network.

In general, the problems mentioned above are related to the determination of flow speeds that is implicit in the definition of exit functions. For the continuous time formulation we have that:

i) from the application of equation (1), follows that flows $u_a(t)$, entering link a at t, are immediately added to vehicles already inside the link, $f_a(t)$.

ii) given that the number of vehicles leaving the link at t, $g_a[f_a(t)]$, is a function of the number of vehicles inside the link at that same moment, some vehicles

will experience an infinite propagation speed, reaching their destination n at the same moment that they enter the first link connected to the origin centroid k. This instantaneous propagation is produced because when a flow enters a link, it is immediately homogeneously distributed along its whole length.

iii) given the flow propagation characteristics described in the two preceding points, FIFO rule does not hold in the dynamic processing of flows through the network.

3.2 Conservation of flows in dynamic assignment

A main difference between the static and dynamic assignment problems is that, for the former, the number of trips reaching each destination during the analisis period is assumed known and given, nevertheles in the dynamic problem it is not possible to know 'a priori' the ammount of trips that must reach each destination at each time $t \in [0, T]$, neither during the whole period $[0, T]$; actually, it is not even possible to know (without solving the problem) what will be the total length of the period that must be considered in order to make sure that all the flows originated in $[0, T]$ reach their destinations.

In the dynamic case, the input demand data is given by a matrix of O-D travel demands $\{T_w(t)\}, w = (k, n), t \in [0, T]$, were each element, $T_w(t)$, gives the instantaneous rate of trips starting from origin k with destinaton n, at t. Then, although there is a precise time at which flows leave their origin (or in other words, the number of trips leaving each origin at each $t \in [0, T]$, is precisely known), the time at which they reach their destination will depend on travel times over the network, that are unknown variables of the problem, to be determined; also, some flows that leave their origin close to time T will not be able to reach their destination within the interval of analysis $[0, T]$.

Therefore, different to the case of static equilibrium assignment models, where conservation of flows at destinations is considered as part of the constraints, in the dynamic case, the magnitude of flows that reach each destination n, at each time $t \in [0, T]$, $\overline{D}^n(t)$, is part of the variables of the problem.

$$\overline{D}^n(t) = \sum_{a \in B(n)} g_a(t), \quad (n \in \mathcal{N}), \quad t \in [0, T] \tag{3}$$

Then, for a network with multiple destinations, in order to make sure that each flow $T_w, w = (k, n)$ ends up in its proper destination, a multicopy representation is necessary. In this case we must define variables f_a^n, u_a^n, and exit functions g_a^n, which allows us to have n different network representations, each of them with only one destination m defined over a common set $(\mathcal{N}, \mathcal{A})$. Then, if we identify the origin centroids by index k, the destination centroids by index n and the network nodes by index j, we can use the following multicopy constraints formulation for dynamic flows conservation:

$$T_w(t) = \sum_{a \in A(k)} u_a^n(t), \quad \forall k, n \in \mathcal{N}, \quad w = (k, n), \quad t \in [0, T], \tag{4}$$

$$\sum_{a \in A(j)} u_a^n(t) = \sum_{a \in B(j)} g_a^n(t), \quad j \in \mathcal{N}, \quad t \in [0,T], \tag{5}$$

$$u_a^n(t) = 0, \quad \forall a \in B(m), \quad n \neq m, \quad t \in [0,T], \tag{6}$$

where m is also an index for destinations. Conditions (6) are necessary in order to make sure that each flow reaches its corresponding destination and replace the tipycal flow conservation constraints at destination nodes, in static models. Notice that:

$$f_a(t) = \sum_n f_a^n(t), \quad u_a(t) = \sum_n u_a^n(t), \quad g_a(t) = \sum_n g_a^n(t),$$

3.3 Consistent dynamics and exit function formulation

Equations (1) have the attractive characteristic of maintaining the link as the basic operational element of the network, nevertheless, as we have seen in section 3.2, its present formulation implies some inconsistencies or limitations in the representation of space-time relations that define the progression of flows through the network. There is no explicit consideration of the flow propagation speed inside each link, which nevertheless is implicit in the specification of exit function $g_a[f_a(t)]$; as we saw in section 3.2, this produces unrealistic flow propagation characteristics. However, the elimination of the exit function does not seem to present many advantages (Ran et al, 1993; Janson, 1991a, 1991b), generating the need to indirectly introduce other, more complex, flow propagation constraints.

We think that the basic problem of equation (1) formulation is in the definition of the exit function g_a, as depending on the value of the present number of vehicles inside the link at time t; this is equivalent to assume that congestion affecting a given vehicle depends on the number of vehicles which are behind it, instead of being affected by those in its front.

We will assume that the link crossing time. and therefore the average operating speed of a vehicle inside an link, is determined by the number of vehicles which are inside the link, in other words, in front of him, at the moment t^0 that it enters the link[1]. Then, we will assign a link crossing time $\tau_a(t^0)$ to the flow $u_a^n(t^0)$ that enters the link at t^0; this is determined by the corresponding average flow speed, that will be a function of the total average link traffic density $d_a(t^0)$, with:

$$d_a(t) = \frac{\sum_n f_a^n(t)}{l_a} = \sum_n \frac{f_a^n(t)}{l_a} = \sum_n d_a^n(t)$$

Therefore, at the time that flow $u_a^n(t)$ enters the link we can determine the exit time $t^0 + \tau_a(t^0)$.

[1] A more complete formulation, that we will analyze in a future publication, considers that g_a^n is function of all the flows that are inside each of the links leading to link a end node, at t^0; thus, we can consider the congestion produced, in the link exit node, by the interaction of all link flows sharing the capacity of the common intersection

We will assume that the link travel time value $\tau_a(t)$, for some vehicle (whatever be the destination) entering link a at time t, can be obtained using a function of the form:

$$\tau_a(t) = \tau_a^0 \left[1 + \beta \left(\frac{d_a(t)}{\overline{d}_a}\right)^\eta\right] \tag{7}$$

where $d_a(t)$, represents the average total traffic density inside the link, at the moment t that the vehicle considered enters the link; τ_a^0 is the travel time corresponding to free flow travel speed, $\overline{d}_a$ is the traffic density corresponding to the link capacity and β, η are calibration parameters. If we know $\tau_a(t^0)$ we can also calculate the average traveling speed $v_a(t^0)$, that will be experienced by a flow entering link a at t^0. Notice that we are assuming that the travel time is the same for all flows, independent of their destination.

Now we can reformulate dynamic equation (1) using the new exit function definition as follows:

$$\frac{df_a^n(t)}{dt} = u_a^n(t) - g_a^n(t), \qquad a \in \mathcal{A}, \quad n \in \mathcal{N}, \quad t \in [0,T], \tag{8}$$

$$g_a^n(t) = g_a^n[f_a^n(t - \tau_a(t^0)), \tau_a(t^0)], \qquad t^0 = t - \tau_a(t^0) \tag{9}$$

where $\tau_a(t^0)$ is given by (7). Notice that if $f_a^n(t^0) = 0$, then $g_a^n(t) = 0, \forall t \leq t^0 + \tau_a(t^0)$; with the definition of g_a^n given in (9), $u_a^n(t^0)$ will not have any influence on the exit rate value, g_a, until time $t^0 + \tau_a(t^0)$. In this way, we introduce the explicit consideration of travel time τ_a in the definition of exit function g.

Given some initial values for the variables considered, equation (1) has only the role of keeping track of the number of vehicles inside the link at any time $t \in [0,T]$, $f_a^n(t)$, which will influence the link travel time through equation (7); in turn, travel time will influence the exit rate at any time $t + \tau_a$. This eliminates the problems produced by the implicit determination of link travel time (or speed) through the definition of g_a and its use in equation (1).

Making use of the fundamental relation of traffic engineering (Drew, 1965) we propose adopting the following explicit definition for the exit function:

$$g_a^n(t) = v_a[d_a(t - \tau_a(t^0))]d_a^n(t - \tau_a(t^0)), \qquad t^0 = t - \tau_a(t^0) \tag{10}$$

Then, if $u_a^n(t^0)$ is the flow entering link a at time t^0 with destination n, we can calculate, using (7), the link crossing time $\tau_a(t^0)$ and therefore, also the average speed $v_a(t^0)$ relevant for that flow, and then, using (10), we can determine the exit rate at time $t^0 + \tau_a(t^0)$. Notice that by replacing d_a^n and v_a in terms of f_a^n, τ_a and l_a the exit function can also be expresed as:

$$g_a^n(t) = \frac{f_a^n(t - \tau_a(t^0))}{\tau_a(t^0)}, \tag{11}$$

Notice that as total traffic density inside the link, d_a, increases, flow is slowed down (τ_a increases, according with (7)) and therefore because of (11), the exit rate decreases (the contrary happens if d_a decreases).

3.4Practical considerations

As was mentioned before, a limitation of an approach that uses the link as basic element of the analysis is that it is not possible to differentiate among operating characteristics for different positions inside the link; therefore, we have to work with average link values for speed and density. Nevertheless, those values are considered continuously variable with time. These formulations have received the name of "link flow models" and some authors have considered them unsuitable for dynamic assignment, "because (they) respond slowly to changes in the inflow" (Addison and Heydecker, 1993), in addition to the problems of instantaneous propagation analyzed in section 3.1.

In the preceding section, we have presented a new exit function formulation that overcomes the problems mentioned. The introduction of link travel time τ_a, in the formulation of the exit function, eliminates the most serious problem, of instantaneous propagation of flows; the other problem, of slow response to inflow changes, is of a more practical nature and can be minimized if the links are properly defined: they should be as short as possible (using intermediate dummy nodes when necessary), to represent homogeneous operating characteristics inside each of them.

This will produce an increase in the number of variables considered, however, the routing decisions and flow conservation constraints will be trivial for dummy nodes, reducing the computational impact of their introduction. Here it is obviously applicable the usual trade off between modeling accuracy and computational burden. In the experiments performed by the authors, with the exit function formulation proposed in (10) and (11), it has been found that the slow response problem is not serious in practical applications (see the results presented in Appendix A).

In the formulation proposed, the propagation of flows characteristics depend mainly on the link travel time values τ_a; therefore it is important in practical applications to calibrate properly the functions (3), for the different link categories defined according with their operational characteristics.

Notice also that though no explicit link capacities are defined these are indirectly considered through the introduction of an increasing travel time function that depends on the level of congestion in the link. Thus, when traffic density increases, link flow will slow down (travel times increases) reducing the link exit rates.

4. FIFO and flows propagation consistency conditions

It has been postulated by some authors (Carey, 1992) that the behavior of flows in their progression through the network links should satisfy FIFO conditions; in other words, the vehicles should exit a link in the same order that they entered it; no overpassing is allowed. This assumption seems reasonable when network congestion levels are high, given the difficulties to overpassing in those conditions; however, a FIFO constraint can be not so reasonable under low congestion operating conditions and it is obviously non necesary for free flow conditions.

FIFO progression of flows can be represented by the following analytical relations among incoming, exiting and link travel times:

$$\forall t_i^0 < t_i^f, \quad t_i^0, t_i^f \in [0,T], \quad t_i^0 + \tau_a(t_i^0) < t_f^0 + \tau_a(t_f^0), \tag{12}$$

that can also be written as:

$$(t_f^0 - t_i^0) > (\tau_a(t_i^0) - \tau_a(t_f^0)), \tag{13}$$

and given that we are using a continuos time formulation we can derive the same condition when t_f^0 goes to t_i^0. If we define $\Delta t = (t_f^0 - t_i^0)$, we then divide (13) by Δt and multiply it by -1, we obtain:

$$\frac{(\tau_a(t_i^0 + \Delta t) - \tau_a(t_i^0))}{\Delta t} > -1, \tag{14}$$

which when Δt goes to cero, gives the condition:

$$\dot{\tau} = (d\tau_a(t)/dt) > -1, \tag{15}$$

It is easy to see that condition (15) implies that, when the link traffic density is increasing and therefore (according with (7)), the congestion level is also increasing, vehicles entering later will experience higher travel times and therefore FIFO will be satisfied. However, if the level of traffic density is decreasing, travel times experienced by vehicles entering later will be lower and therefore overpassing could happen if (15) is not satisfied, violating FIFO.

To analyze the consequences of this condition in our case, we can calculate the expression of $\dot{\tau}$ from (7):

$$\dot{\tau} = \left(\frac{\tau_a^0 \beta \eta}{l_a \overline{d}_a^{\eta}}\right) d_a^{\eta-1}(u_a - g_a), \tag{16}$$

where we replaced $\dot{d}_a = (u_a - g_a)/l_a$ making use of (1). From (16) we obtain that if $u_a(t) = g_a(t)$, then the travel time is constant over time ($\dot{\tau} = 0$); if $u_a > g_a$, travel time is increasing ($\dot{\tau} > 0$), and if $u_a < g_a$, travel time is decreasing ($\dot{\tau} < 0$). Then, if we replace (16) in (14) we can obtain the following condition for FIFO flows:

$$(g_a - u_a) < \frac{v_a^0 \overline{d}_a}{\beta \eta} \left(\frac{\overline{d}_a}{d_a(t)}\right)^{\eta-1}, \tag{17}$$

where $v_a^0 = l_a/\tau^0$, is the free flow link average (commercial) speed.

Therefore, given that FIFO could be violated only if condition (15) is not satisfied, it could happen only if $u_a < g_a$, when, according to (16), link flow average speed is decreasing. Actually, from (17) we can see that the worst case corresponds to a situation in which $d_a(t) = \overline{d}_a$ and link entrant flow u_a comes

suddenly down close to zero; Then in order not to violate FIFO, the link exit flow must satisfy the constraint:

$$g_a < \frac{v_a^0 \overline{d}_a}{\beta\eta}, \qquad \Longrightarrow \qquad v_a(t-\tau_a)d_a(t-\tau_a) < \frac{v_a^0 \overline{d}_a}{\beta\eta}, \tag{18}$$

If, according with the assumption of the worst case made above, $d_a(t-\tau_a) = \overline{d}_a$, it is easy to check from (7), that the travel speed $v_a(t-\tau_a)$ would be $v_a^0/(1+\beta)$. Therefore, introducing in (18) we obtain that, for FIFO not to be violated, in the worst case analyzed, it would be necessary that $\beta \leq (\eta-1)^{-1}$. Taking a value of $\eta = 4$, normal for BPR functions used in static assignment models, the condition would be $\beta \leq 1/3$.

Then, we can not make sure that FIFO would not be violated in some cases, nevertheless, it is also easy to see that the violations will happen rarely and in those operating conditions for which FIFO is not essential, or could even be not reasonable.

Notice that while FIFO holds it is reasonable that:

$$\int_{t_i^0}^{t_f^0} u_a(t)dt = \int_{t_i^0+\tau_a(t_i^0)}^{t_f^0+\tau_a(t_f^0)} g_a(t)dt, \tag{19}$$

However, this consistency condition in the processing of link flows will not be satisfied if FIFO does not hold[2].

5. Network Operating Costs

We will assume that the only relevant cost for network users is travel time. Using link travel times τ_a, we can calculate the total travel cost experienced by a vehicle traveling through the whole path $p = (a_1, a_2, \ldots, a_m)$, $a_1 = (k,i)$, $a_m = (j,n)$, $p \in P_w$, $w = (k,m)$ starting at node k at time t^*, will be given by:

$$\Gamma_p(t^*) = (t^{**} - t^*) = \sum_{i=1}^{m} \tau_{a_i}(t_i^0), \tag{20}$$

were t^{**} represents the time at which the vehicle reaches destination n; with $t_1^0 = t^*$, and $t_j^0 = \sum_1^{j-1} \tau_{a_i}(t_i^0)$.

It is also possible to define the total operating cost over arc a at t, as the sum of the travel costs experienced by the vehicles that are leaving the arc at

[2] For example, given an interval $[t_i^0, t_f^0]$ if three vehicles enter arc a, the first at t_i^0, the second at $t_m^0 (> t_i^0)$, the third at $t_f^0 (> t_m^0)$, and the second overpasses the first in its travel inside arc a, then during interval $[t_i^0, t_f^0]$, three vehicles enter the arc, but during interval $[t_i^0 + \tau(t_i^0), t_f^0 + \tau(t_f^0)]$ only two of them exit it, which violates condition (19).

that time, $\hat{C}_a(t^0) = \tau_a[f_a(t^0)]g_a[t^0 + \tau_a(t^0)]$; then, the total cost over the network for period $[0, T]$ can be written as:

$$Z = \sum_a \int_0^T \tau_a(t^0) g_a[t^0 + \tau_a(t^0)] dt^0, \tag{21}$$

Remember that, according to the definition of function $g_a(t)$ given in (11):

$$g_a[t^0 + \tau_a(t^0)] = \frac{f_a(t^0)}{\tau_a(t^0)}, \tag{22}$$

Therefore, (21) can be written as:

$$Z = \sum_a \int_0^T f_a(t^0) dt^0, \tag{23}$$

Notice that although some flows starting from their origins at a time t close to final time T are still inside the network this is taken into account and penalized by expression (23).

6. Model Formulation

If we wish to minimize the total cost of transportation over the analysis period $[0, T]$, we should solve the following control problem:

$$\text{(P1):} \quad \min_{\{f,u\}} \quad Z = \sum_{a \in \mathcal{A}} \int_0^T f_a(t) dt, \tag{24}$$

$$\text{s.t. :} \quad \frac{df_a^n(t)}{dt} = u_a^n(t) - g_a^n(t), \quad \forall (a \in \mathcal{A}), (n \in \mathcal{N}), \quad t \in [0, T], \tag{25}$$

$$g_a^n(t) = \frac{f_a^n(t - \tau_a))}{\tau_a(t - \tau_a)},$$

$$T_w(t) = \sum_{a \in A(k)} u_a^n(t), \quad \forall (k, n \in \mathcal{N}), w = (k, n), \tag{26}$$

$$\sum_{a \in B(j)} g_a^n(t) = \sum_{a \in A(j)} u_a^n(t), \quad \forall (j, n \in \mathcal{N}), \quad t \in [0, T], \tag{27}$$

$$f_a^n(0) = f_a^n, \qquad \forall (a \in \mathcal{A}), (n \in \mathcal{N}), \tag{28}$$

$$u_a^n(t) = 0, \quad \forall a \in B(m), \quad n \neq m, \quad t \in [0, T], \tag{29}$$

$$u_a^n(t) \geq 0, \quad f_a^n(t) \geq 0, \quad \forall (a \in \mathcal{A}), (n \in \mathcal{N}), \quad t \in [0, T]. \tag{30}$$

The necessary conditions for an optimal solution of (P1) are given by Pontryagin's maximum principle (Pontryagin et al, 1962). The Hamiltonian expression for (P1) will be:

$$H(f, u, \lambda, t) = \sum_{a \in \mathcal{A}} f_a(t) + \sum_{a \in \mathcal{A}} \sum_{n \in \mathcal{N}} \lambda_a^n [u_a^n(t) - g_a^n(t)], \tag{31}$$

where, $\lambda_a^n(t)$ are adjoint variables that correspond to a generalization, for the dynamic case, of lagrange multipliers and are associated with dynamic equations (25). We next construct the Lagrangean:

$$\mathcal{L} = H(f,u,\lambda,t) + \sum_{k\in\mathcal{N}}\sum_{n\in\mathcal{N}} \mu_k^n \left[T_w(t) - \sum_{a\in A(k)} u_a^n(t) \right] + \\ + \sum_{j\in\mathcal{N}}\sum_{n\in\mathcal{N}} \mu_j^n \left[\sum_{a\in B(j)} g_a^n(t) - \sum_{a\in A(j)} u_a^n(t) \right] \tag{32}$$

where, $\mu_i^n(t)$, are the dual variables corresponding to node flow conservation equations (26) (if $i = k$) and (27) (if $i = j$).

Then, we can obtain the following first order necessary conditions for problem (P1):

$$\lambda_a^n(t) - \mu_i^n(t) \geq 0, \quad \forall (i,n\in\mathcal{N}),(a\in A(i)),(t\in[0,T]),(i\neq n) \tag{33}$$

$$u_a^n(t)[\lambda_a^n(t) - \mu_i^n(t)] = 0, \quad \forall (i,n\in\mathcal{N}),(a\in A(i)),(t\in[0,T]),(i\neq n) \tag{34}$$

$$\lambda_a^n(T) = 0, \quad \forall a\in\mathcal{A}, \tag{35}$$

$$-\dot{\lambda}_a^n(t) = 1 - \lambda_a^n(t)/g_a'(t) + \mu_j^n(t)/g_a'(t), \quad (j\neq k,n), \\ \forall (j,n\in\mathcal{N}), a\in B(j), \quad (t\in[0,T]), \tag{36}$$

where in (36), $g_a'(t)$ is equal to the derivative of the exit function with respect to $f(t)$, and it is easy to show that is equal to the derivative with respect to $f(t^0), t = t^0 + \tau_a(t^0)$[3].

Aknowledgments

This research has been financed by **FONDECYT–Chile** and the **Pontificia Universidad Católica de Chile.**

APPENDIX A

In order to see the flow progression and evolution characteristics produced by the exit function formulation proposed in (10) and (11) some experiments were performed with the following simple network, constituted by four links that form a path from an origin k to a destination n:



[3] This can be seen using expression (2) to obtain the derivative $\frac{\partial f_a(t)}{\partial f_a(t^0)} = 1$; this follows from: $\frac{\partial g_a(t'-\tau_a)}{\partial f_a(t^0)} = 0, t'\in[t^0,t], t = t^0 + \tau_a(t^0)$.

The following parameter values were assumed for the travel time and exit functions: $l_a = 0.2[km], v_a^0 = 48[km/hr], \tau_a^0 = 15sec, \beta = 3, \eta = 4, \overline{d}_a = 65[veh/km]$. The experiment was performed in discrete time, assuming time periods of 5 secs.

Fig 1

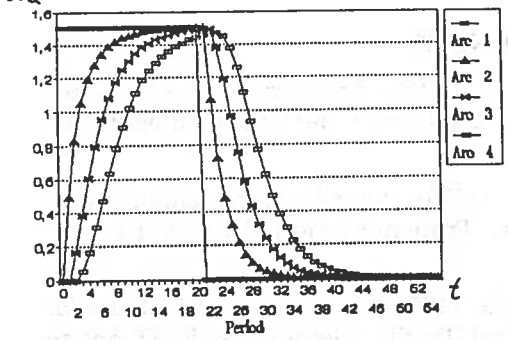


Fig 4

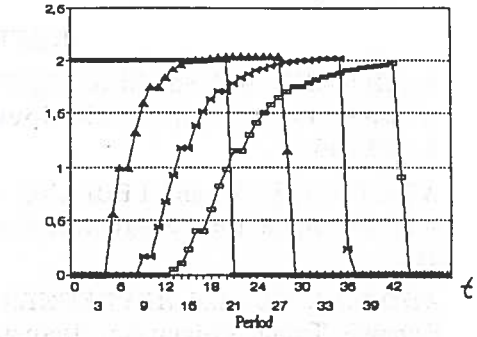


Fig 2

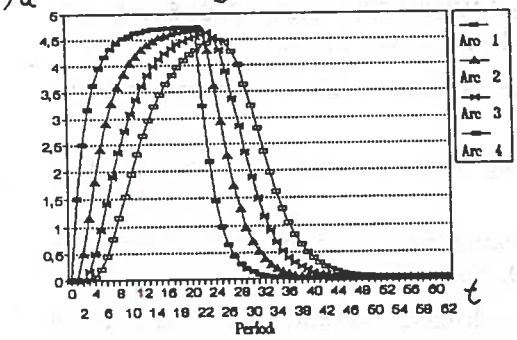


Fig 5

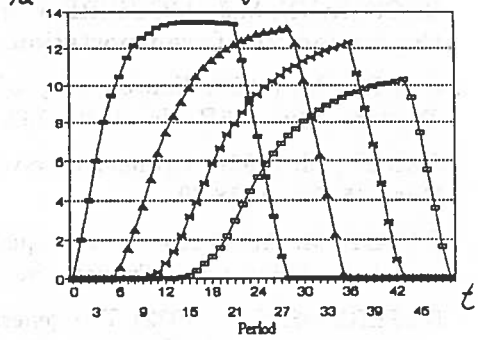


Fig 3

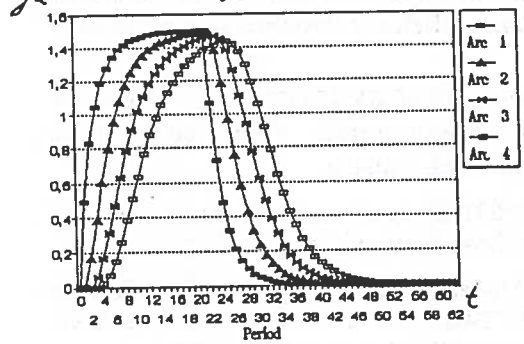
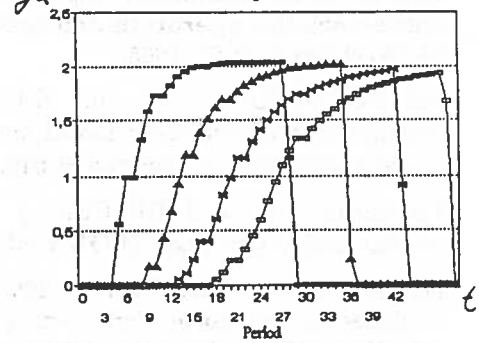


Fig 6



Figures 1,2 and 3 present the time evolution of variables u_a, f_a, and g_a, when an exit function of the form $g_a(f_a(t))$ is used; a constant input flow of 1.5 is assumed to enter the first link (at node k) during the period of time [0,20]. We can see that instantaneous propagation of flows is obtained. On the other side, Figures 4,5,6 present the results when the exit function formulation (11) is used; a constant input flow of 2 is now assumed at k, for the same time period [0,20]. We can see that instantaneous propagation has now disapeared and a reasonable propagation and evolution of flows is obtained.

REFERENCES

AASHTIANY, H.Z. and MAGNANTY, T.L. (1981) Equilibrium on a congested transportation network. **SIAM Journal, Algebraic and Discrete Methods 2**, 213–216.

ABADULAAL, M. and LEBLANC, L.J. (1979) Methods for combining modal split and equilibrium assignment models. **Transportation Science 13**, 292–214.

ADDISON, J.D. and HEYDECKER, B.G. (1993) A Mathematical Model for Dynamic Traffic Assignment. **Transp. and Traffic Theory / C.F. Daganzo (Editor** 171–184, Elsevier Sc. Pub.B.V.

BECKMANN, M.J., MCGUIRE, C.B. and WINSTEIN, C.B. (1956) **Studies in the economics of transportation**. Yale University Press, New Haven.

CAREY, M. (1992) Nonconvexity of the dynamic traffic assignment problem. **Transp. Res., 26B**, No. 2, 127-133.

CAREY, M. (1987) Optimal time-varying flows on congested networks. **Op. Res. 35**, No. 1, 58-69.

CAREY, M. (1986) A constraint qualification for a dynamic traffic assignment model. (technical note) **Transp. Sc. 20**, No.1, 55-58.

DAFERMOS, S.C. (1982) The general multimodal network equilibrium model with elastic demand. **Networks 12**, 57–72.

DREW, D.R., Deterministic aspects of freeway operation and control. **Freeway characteristics operation and accidents.** Highway Research Record, 99, 48–58, Washington D.C., 1965.

DRISSI-KAÏTOUNI, O. and HAMEDA-BENCHEKROUN, A. (1989) A dynamic traffic assignment model and solution algorithm. Centre de recherche sur les transports, Publicación # 649, Montreal, Canada.

Fernandez, j.e. and DE CEA, J. (1993) Dynamic assignment models for transportation networks. **IFORS 93**, Lisboa, Portugal, 12-16 of July.

Fernandez, j.e. and DE CEA, J. (1994) Modelos Dinámicos para la Planificación de Redes de Transporte, (forthcoming) ACTAS DEL VI CONGRESO CHILENO DE INGENIERIA DE TRANSPORTE.SOCHITRAN, Santiago, Chile.

FERNANDEZ, J.E. and FRIESZ, T.L. (1983) Equilibrium predictions in transportation markets: the state of the art. **Transp. Res. 17B**, No. 2, 155-172.

FLORIAN, M. and NGUYEN, S. (1978) A combined trip distribution modal split and assignment model. **Transp. Res. 12,** 241–246.

FRIESZ, T.L., LUQUE, J., TOBIN, R.L. and WIE, B-W.. (1989) Dynamic network traffic assignment considered as a continuous time optimal control problem. **Op. Res. 37**, No. 6, 893-901. HAO WU, J. (1993) Dynamic network equilibrium problem formulated as an infinite dimensional variational inequality problem. Publication # 952, CRT, Université de Montreal, Montreal, Canada.

JANSON, B.N.(1991a) Dynamic traffic assignment for urban road networks. *Transp. Res. 25B*, Nos. 2/3, 143-161.

JANSON, B.N.(1991b) Convergent algorithm for dynamic traffic assignment. *Transp. Res. Rec. 1328*, 69-80.

LBBLANC, L.J., MORLOK, E. and PIERSKALLA, W. (1975) An efficient approach to solving the road network equilibrium traffic assignment problem. **Transp. Sc. 9,** 309–318.

MAHMASSANI, H.S. and JAYAKRISHNAN, R. (1991) System performance and user response under real–time information in a congested traffic corridor. *Transp. Res. 25A*, No. 5, 293–307.

MERCHANT, D.K. and NEMHAUSER, G.L. (1978-a) A model and an algorithm for the dynamic traffic assignment problem. **Transp. Sc. 2**, No.3, 183-199.

MERCHANT, D.K. and NEMHAUSER, G.L. (1978-b) Optimality Conditions for a Dynamic Traffic Assignment Model. **Transp. Sc. 2**, No.3, 200-207.

PAPAGEORGIOU, M. (1990) Dynamic Modeling, Assignment and Route Guidance in Traffic Networks. **Transp. Res. 24B**(6), 471-495.

PONTRYAGIN, L.S., BOLTYANSKI, V.A., GRAMkRELIDZE, R.V. and MISCHENKO, E.F. (1962) **The Mathematical Theory of Optimal Processes.** John Wiley, New York.

RAN, B., BOYCE, D.E., and LEBLANC, L.J. (1993) A new class of instantaneous dynamic user-optimal traffic assignement models. Op. Res., Vol. 41, No. 1, pp. 192–201.

WARDROP, J.G. (1952) Some Theoretical aspects of road traffic research. *Proc. Inst. Civil Engrg.* (Part II) **1**, 325–378.

WIE, B-W, FRIESZ, T.L. and TOBIN, R.G. (1990) Dynamic user optimal traffic assignment on congested multidestination networks. **Transp. Res. 24B,** No. 6, 431-442.

An Evolutionary Algorithm for the Crew Rostering Problem

Fernando Catanas
Departamento de Estatística e Investigação Operacional
Faculdade de Ciências de Lisboa
Lisboa, Portugal

José Paixão
Departamento de Estatística e Investigação Operacional
Faculdade de Ciências de Lisboa
Lisboa, Portugal

1. Introduction

In this paper, we present an Evolutionary Algorithm for determining feasible solutions for the Crew Rostering Problem (CRP). This is a problem that usually occurs in situations where it stands the need of rotating personnel in a certain period, say m days, and over a set on nt tasks to be accomplished in that period. Along with each task it is defined a starting time and an ending time, which we will denote, respectively, by st_k and et_k (k=1,...,nt).

Typically, this situation arises in public transport companies where one aims to find the best assignment of rosters (sequences of tasks) to crews. In general, one has nr types of roster, each of them characterized by a triplet (ctr_t, str_t, etr_t), stating the cost, starting time and ending time for t=1,...,nr.

The criteria for selecting the rosters depends very much on in-company restrictions. For instance, in North America, the selection of rosters is usually done on a seniority basis, while in the context of most of the European companies, the planners try to find the set of rosters that maximize the highest workload. In our case, we consider a weighted objective function combining costs and duration for the rosters.

So far, this kind of problem has not received as much attention as other scheduling problems. A number of works have focused on a simplified rostering problem known as 'cyclic scheduling' [*Baker*, 1974 and 1976], [*Baker & al.*, 1977], [*Bartholdi & al.*, 1980], [*Bartholdi*, 1981], [*Koop*, 1986]. [*Carraresi & al.*, 1984a and 1984b] instead, have dealt with the general rostering problem by considering minimization of the maximum workload. They stated this problem as a Multilevel Bottleneck Assignment (MBA) problem and proved that it is NP-hard. A heuristic algorithm has been proposed which provides 'asymptotically optimal' solutions. Other works have been produced by other authors which mainly refer to practical experience in solving problems (e.g. [*Jachnik*, 1981]) [*Bianco & al.*, 1992] proposed an MBA based heuristic that consists in finding a minimum cost combination of different types of rosters with the constraints of controlling starting and ending times of each roster and of assuring a minimum rest time between two consecutive working days.

In a previous paper, [*Catanas & al.*, 1993] have formulated the CRP as a Set Covering Problem with additional constarints. There, Column Generation is embedded in a tree search procedure in order to solve the

CRP. Now, we describe a meta-heuristic for the CRP based on the class of Evolutionary Algorithms (see [*Michalewicz*, 1992]). With such an approach one aims to find good quality upper bounds for the optimum that may lead to significant improvements in the tree-search scheme proposed in [*Catanas & al.*, 1993] and in the MBA approach proposed in [*Bianco & al.*, 1992].

In the next section, we recall the Set Covering formulation for the CRP. Section 3 is devoted to the description of the new meta-heristic which is computationnaly assessed in section 4.

2. Mathematical Formulation

As presented in [*Catanas & al.*, 1992], the CRP can be formulated as Set Covering Problem with additional constraints:

$$
\begin{aligned}
(F1) \quad \min Z_{F1} = & \sum_{j \in \Re} cr_j\, y_j + \Theta \cdot \Delta \\
& \sum_{j \in \Re} a_{kj}\, y_j \geq 1 \qquad k = 1, \cdots, nt. \\
& dr_j\, y_j \leq \Delta \qquad j \in \Re. \\
& y_j \in \{0,1\} \qquad j \in \Re.
\end{aligned}
$$

In the above mathematical formulation, one has the following:

- $\Re$: set of feasible rosters.
- cr_j: cost of roster j, ($j \in \Re$).
- dr_j: duration of roster j, ($j \in \Re$).
- a_{kj}: $\begin{cases} 1; & \text{if task } k \text{ is covered by roster } j. \\ 0; & \text{otherwise.} \end{cases}$, ($j \in \Re$; $k = 1, \cdots, nt$).
- y_j: $\begin{cases} 1; & \text{if roster } j \text{ is in the solution.} \\ 0; & \text{otherwise.} \end{cases}$, ($j \in \Re$).
- Δ: $\max_{j \in \Re} \{dr_j \cdot y_j\}$

Roster feasibility is defined by the following conditions:

C1: The roster must correspond to one of the types of roster in terms of working days, starting and ending times.

C2: There must exist a minimum time length H between each two consecutive tasks in the roster.

C3: The total duration of the roster (i. e., the sum of the durations of the composing tasks), cannot exceed a certain time limit D.

The objective function in model F1 combines the cost of the rosters with the maximum corresponding workload, denoted by Δ. That is, if S is a set of rosters corresponding to a feasible solution for F1, the corresponding cost is given by: $f(S)=\sum_{j\in S} cr_j y_j+\Theta\cdot\Delta$, with $\Delta=\max_{j\in S}\{dr_j\cdot y_j\}$

3. Solving the problem

It can be shown that F is a NP-hard problem, [*Catanas & al.*, 1993], displaying two immediate difficulties: the first one is that F is a 0/1 ILP, the second one is the potencially largeness of #[$\Re$], i. e., the dimension of the set $\Re$. Given real life problems, usually very large, it is not possible to derive explicitly every column of the set $\Re$. In [*Bielli & al.*, 1992] it is proposed an heuristic method to solve the problem (under a different theoretical approach, though) which unfortunatelly doesn't deal with condition C3. Implicit Column Generation Methods [*Catanas & al.*, 1993] has proven to be very time and increasingly memory consuming from an iteration to another; moreover, this latter approach deals with very large LPs and the available packages didn't work out but just a few thousand of columns and rows which is a very low dimension compared to the $O(10^6)$ dimension common to medium to large scale problems.

The basic idea of the new algorithm is to construct an initial set of random feasible solutions which is called the ***initial population***, of size p. Each ***member***, m=1,...,p, of the population is a feasible solution, S_m, of F. The value of each feasible solution is given by (1a-1b), which will be denoted by f_m. The value of the feasible solution corresponds to the ***fitness*** of the associated member. Since the CRP is a minimization problem, the lowest f_m is, the higher is the fitness. The algorithm is very simple and consists of the following: first of all, we create an initial random population of size p. Now, instead of trying to improve the population by means of "natural selection", closely following Darwin's theory, we use a "breeding" scheme as [*Muehlenbein*, 1993] depicted. This technique consists in producing offspring with the $100\times P_b$% best members of the population, disregarding the remaining members. So, the worst members always have to be discarded. Notice that in the usually GAs, all members have a probability of mating (e.g. proportional to its fitness), and that doesn't happen with our approach. It is expected that, generation after generation, the total fitness of the population increases as well as the best fitness.

We have detected that, even for medium size problems, premature convergence, happens very quickly (few tens of generations); this happens when all the members of the population are the same. This situation

might lead us to a solution corresponding to a local optimum of the search space, which can be very hilly. So, what we have to do is to somehow remake the population. As [*Clement & al.*, 1993] proposed, mutation or migration are the operators that might avoid (or at least postpone) the problem. We have adopted the method proposed in [*Clement & al.*, 1993].

Each member of the population is represented by a quasi-matrix, where each line corresponds to a roster (allele). This representation has proven to be both quite natural and very efficient. The generation of feasible solutions is based on a multi-level graph representation of the feasible rosters (disregarding condition C3). Applying a Shortest Path algorithm, modified to deal with constraint C3, to this graph, it is possible to efficiently generate feasible solutions for the CRP. For solving these shortest paths problems we have used algorithm PDM (see [*Syslo & al.*, 1983]), adapted to support the additional constraint C3.

Mating (or crossover) is the operator that enables us to produce offspring. We have developed a method to do it in a very starightforward way. After performing the evaluation of the population, the fitnesses are sorted out in ascending order (we shall denote the sorted array of fitnesses by $f_{(m)}$); therefore, fitness $f_{(1)}$ corresponds to the fittest member of the population and fitness $f_{(p)}$ to the less fitted one. We choose the best (regarding fitness) $100 \times P_b$% members of the population (this will correspond to the integer part of the value $p \times P_b$). Over this subset of members two are randomly selected to mate; thus, we select a "male" and a "female". With a certain probability P_m this members will be able to mate. If so the crossover operator works the following way: we choose alternatively and randomly alleles from the male and the female until all tasks are covered.

4. Computational Results

We have considered a set of 20 random generated problems with the following characteristics: m=5; H=480; T=2100 min. (35 hours). We have considered one type of roster only [1,2,3,4,5] with it's cost following a discrete uniform distribution U[10,20]. The number of tasks per day follows a discrete uniform distribution U[3,5]. The beginning of each task follows a discrete uniform distribution U[300,900] (in minutes), i.e., each task starts somewhere between 5h00m a.m. and 3h00m p.m. The duration of each task follows a discrete uniform distribution U[405,435], i.e., the duration of each task is a value between 6h45m hours and 7h15m. The following table shows the values of Θ, the cost of the type of roster and the number of tasks for each of the 20 random generated tests used.

prob.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Θ	2,16	6,34	9,89	2,9	7,19	6,34	2,05	2,15	6,39	9,02	0,48	6,8	2,67	4,13	8,28	1,72	6,44	2,75	4,81	8,03
cr_1	12	17	16	13	13	11	18	17	15	20	13	20	12	19	10	20	16	18	20	11
nt	19	17	20	22	23	20	23	19	21	21	20	22	21	18	19	21	23	22	21	20

studying the impact of "relaxing" these procedures to a less elitist approach where chromossome diversity could be exploited out. Our research will also concentrate in other shemes rather than breeding schemes as considered here. The possibility of controling premature convergence by means of introducing a tabu operator might be useful too and some research is also being done in this subject. However, the main conclusion of the work done until now is that the use of weaker algorithms than the classical problem-specific ones might be very successful for the CRP.

References

- *Baker, K. R.* (1974), **Scheduling a full-time workforce to meet cyclic staffing requirements**, Management Science 22/12, 1561-1568.
- *Baker, K. R.* (1976), **Workforce allocation in cyclic scheduling problems**, Operational Research Quarterly 27/1, 155-167.
- *Baker, K. R. & Magazine, M. J.* (1977), **Workforce scheduling with cyclic demands and day-off constraints**, Management Science 24, 161-167.
- *Bartholdi, J. J.* (1981), **A guaranteed accuracy round-off algorithm for cyclic scheduling and set covering**, Operations Research 29, 501-510.
- *Bartholdi, J. J.; Orlin, J. & Ratliff, H. D.* (1980), **Cyclic scheduling via integer programs with circular ones**, Operations Research 28, 1074-1085.
- *Bielli, M.; Bianco, L.; Mingozzi, A. & Spadoni, M.* (1992), **A Heuristic Procedure for the Crew Rostering Problem**, European Journal of Operational Research 58 272-283, North-Holland.
- *Carraresi, P. & Gallo, G.* (1984a), **Network models for vehicle and crew scheduling**, European Journal of Operational Research 16, 139.151.
- *Carraresi, P. & Gallo, G.* (1984b), **A multilevel bottleneck assignment approach to the bus drivers' rostering problem**, European Journal of Operational Research 16, 163-173.
- *Catanas, F. & Paixão, J. M.* (1993), **A New Approach for the Crew Rostering Problem**, Proceedings of the Sixth International Workshop on Computer-Aided Scheduling of Public Transport, Lisbon.
- *Clement, R. & Wren, A.* (1993), **Greedy Genetic Algorithms, Optimizing Mutations, and Bus Driver Scheduling**, Proceedings of the Sixth International Workshop on Computer-Aided Scheduling of Public Transport, Lisbon.
- *Darwin C* (1950), **On the Origin of Species by means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life.** (Reprint) Watts, London.
- *Jachnik, J.* (1981), **Attendance and rostering systems**, in: A. Wren (ed.), Computer Scheduling of Public Transport, North-Holland, Amsterdam, 337-344.
- *Koop, G. J.* (1986), **Cyclic Scheduling of offweekends**, Operations Research Letters 4, 259-263.
- *Michalewicz, Z.* (1992), **Genetic Algorithms+Data Structures=Evolution Programs**, Springer Verlag
- *Muehlenbein, H.* (1993), **New Research Directions for Genetic Algorithms**, Proceedins of the Thirteenth World Conference on Operations Research IFORS 93, Lisbon.
- *Shorrocks, B.* (1978), **The Genetics of Diversity**, Hodder ang Stoughton, London.
- *Syslo, M. M.; Deo, N. & Kowalik, J. S.* (1983), **Discrete Optimization Algorithms**, Prentice-Hall, Inc

Exact and Approximate Algorithms for the Generalized Crew Scheduling Problem

Matteo Fischetti
DEI, Università di Padova

Silvano Martello
Dipartimento di Informatica, Università di Torino

Paolo Toth
DEIS, Università di Bologna

EXTENDED ABSTRACT

Given n *tasks* T_j $(j = 1, \ldots, n)$, requiring uninterrupted processing from a given *release time* r_j to a given *deadline* d_j, and an unlimited number of identical parallel *processors* $P_1, P_2, \ldots$, each of which can perform at most one task at a time, the *Fixed Job Schedule Problem (FJS)* is to determine a feasible assignment of tasks to processors such that the number of required processors is a minimum.

Two generalizations of FJS have already been introduced by the authors. Let $T_{f(i)}$ and $T_{l(i)}$ be, respectively, the earliest and latest task assigned to processor P_i. Imposing an upper bound s (*spread time*) on the difference $d_{l(i)} - r_{f(i)}$ for each P_i, produces the *Fixed Job Schedule Problem with Spread-Time Constraints (FJSS)*. The *processing time* of task T_j is defined as $d_j - r_j$. Imposing an upper bound w (*working time*) on the

sum of the processing times of the tasks assigned to each processor, produces the *Fixed Job Schedule Problem with Working-Time Constraints (FJSW)*. FJS, FJSS and FJSW can arise as subproblems in the more general *Crew Scheduling Problem (CSP)*, in which additional constraints laid down by the union contracts and company regulations are imposed, and different objective functions are considered.

FJS can be solved in polynomial time. $O(n \log n)$ algorithms have been given by Gupta, Lee and Leung (1979) and Nakajima, Hakimi and Lenstra (1982).

Both FJSS and FJSW are NP-hard. Fischetti, Martello and Toth (1987, 1989) have presented effective branch-and-bound algorithms for both problems, and polynomial-time algorithms for the exact solution of their preemptive relaxations. Fischetti, Martello and Toth (1992) have introduced polynomial-time algorithms for the approximate solution of both problems and have evaluated their worst-case and average behavior.

The crew scheduling problem (typically encountered in airline and transit companies) can be described as follows. The schedule for a prefixed period is defined as a set of n tasks T_j requiring a crew working without interruption in time interval $(r_j, d_j](j = 1, \ldots, n)$. Let us assume, without loss of generality, that all the numerical data are positive integers and that

$$r_1 \leq r_2 \leq \ldots \leq r_n.$$

The cost incurred if the same crew performs task T_j right after task T_i is c_{ij}, with $c_{ij} = \infty$ if $j \leq i$ or if T_i and T_j *overlap*, i.e., $r_j < d_i$. All the crews are initially located at a depot (denoted as a task T_0, having zero duration) and must return to the depot at the end of their duty. The times needed for going from the depot to task T_j and vice-versa are t'_j and t''_j, respectively; the respective costs are c_{0j} and c_{j0}.

Given an upper bound m on the number of crews, the problem is to find a set of crews of minimum total cost to perform all tasks by satisfying a set of additional constraints.

The additional constraints can vary considerably in the different practical contests. We will consider a *Generalised Fixed Job Schedule Problem* (GFJS) in which two of them, which are extremely common in practice, are explicitly included. Namely we impose, for each crew P_i, a limit w on the total working time (sum of the durations of the performed tasks) and a *spread-time* limit s on the total duration of the duty (time interval between the departure from the depot and the return to the depot). Observe that the model

allows inclusion of other constraints through the costs c_{ij}. For example, the regulations frequently impose a minimum and/or maximum rest period between two consecutive tasks: this is easily included by setting $c_{ij} = \infty$ if the time interval $r_j - d_i$ violates such limits.

GFJS is clearly an NP-hard problem, as it contains both FJSS and FJSW as special cases (arising when $c_{0j} = 1$ for all j, $c_{ij} \in \{0, \infty\}$ for all $i, j = 1, ..., n$, and when only one of the two additional constraints is imposed).

Beasley and Cao (1993) considered a special case of GFJS in which the working time constraint is not imposed. For this problem a bounding procedure based on dynamic programming and Lagrangian relaxation is proposed, and imbedded into a branch-and-bound scheme for the exact solution of the problem.

GFJS can be viewed as a network flow problem with additional constraints. To this end we define a directed graph $G = (V, A)$ having a node j for each T_j $(j = 0, ..., n)$. Each arc $(i, j) \in A$ represents the possibility for a crew of performing T_j right after T_i, and has cost c_{ij}. Notice that the ordered pairs (T_i, T_j) with $c_{ij} = \infty$ do not correspond to arcs in G. Hence the subgraph of G induced by $V \setminus \{0\}$ is acyclic, i.e., all the circuits of G pass through the depot node 0. We call a simple circuit *feasible* if it corresponds to a feasible duty for a crew, i.e., when it covers a subset of nodes in $V \setminus \{0\}$, say $\{j_1, ..., j_k\}$ (with $j_1 \leq ... \leq j_k$), such that: (i) $(d_{j_k} + t''_{j_k}) - (r_{j_1} - t'_{j_1}) \leq s$, and (ii) $\sum_{i=1}^{k} (d_{j_i} - r_{j_i}) \leq w$. GFJS then consists of finding a minimum-cost family of, at most, m feasible circuits covering once all the nodes of $V \setminus \{0\}$.

We next describe an Integer Linear Programming model for GFJS. For all $(i, j) \in A$, let $x_{ij} = 1$ if arc (i, j) belongs to the optimal solution (i.e., the same crew performs T_j right after T_i); $x_{ij} = 0$ otherwise.

$$v(GFJS) := \min \sum_{(i,j) \in A} c_{ij} x_{ij} \tag{1}$$

$$\sum_{i:(i,k) \in A} x_{ik} = \sum_{j:(k,j) \in A} x_{kj} \begin{cases} = 1 & \text{for } k = 1, \ldots, n \\ \leq m & \text{for } k = 0 \end{cases} \tag{2}$$

$$\sum_{(i,j) \in C} x_{ij} \leq |C| - 1 \quad \text{for all infeasible circuits } C \tag{3}$$

$$x_{ij} \in \{0,1\} \qquad \text{for } (i,j) \in A \tag{4}$$

Constraints (2) and (4) define a transportation problem, and impose the condition that all nodes in $V \setminus \{0\}$ must be covered exactly once, and that no more than m circuits are allowed. The exponentially many constraints (3) break possible infeasible circuits.

We present a branch-and-cut algorithm for the optimal solution of GFJS, in which the LP relaxation of model (1)-(4) is used to produce lower bounds on the optimal solution value. The quality of the bound is improved considerably by defining several strengthened versions of inequalities (3), based on the structure of the infeasible circuits. We also introduce additional families of valid cuts. For each family of cuts, exact and/or heuristic separation procedures are described which detect, in polynomial time, the most violated inequalities to be added to the current LP. Heuristic procedures for GFJS are also discussed, which exploit the fractional point available after each LP solution.

Finally, we report extensive computational experience on several classes of test problems from the literature, involving hundreds of tasks.

References

J.E. Beasley and B. Cao, 1993, "An Algorithm for the Crew Scheduling Problem", Research Report of the Management School, Imperial College, London, April 1993.

M. Fischetti, S. Martello and P. Toth, 1987, "The Fixed Job Schedule Problem with Spread-Time Constraints", *Operations Research* 35, 849 - 858.

M. Fischetti, S. Martello and P. Toth, 1989, "The Fixed Job Schedule Problem with Working-Time Constraints", *Operations Research* 37, 395 - 403.

M. Fischetti, S. Martello and P. Toth, 1992, "Approximation Algorithms for Fixed Job Schedule Problems", *Operations Research* 40, S96 - S108.

U.I. Gupta, D.T. Lee and J.Y.-T. Leung, 1979, "An Optimal Solution for the Channel Assignment Problem", *IEEE Transactions on Computers* C-28, 807 - 810.

K. Nakajima, S.L. Hakimi and J.K. Lenstra, 1982, "Complexity Results for Scheduling Tasks in Fixed Intervals on Two Types of Machines", *SIAM Journal on Computing* 11, 512 - 520.

Hybrid genetic algorithms for bus driver scheduling

Raymond S K Kwan, Anthony Wren

School of Computer Studies, University of Leeds, Leeds LS2 9JT, UK.

Extended Abstract

This paper describes a hybrid approach to bus driver scheduling that incorporates Genetic Algorithms (GAs), a rule-based driver duties estimator, and the IMPACS system which is based on set covering integer linear programming (ILP). The resulting scheduling system captures the complementary strengths of its components and has potential for solving large problems optimally and quickly.

IMPACS has been successfully used by many bus companies since mid 1970s. Being a mathematical programming approach, it has the advantage of producing optimal schedules. However it may encounter difficulty when the problem is large, or when there are difficult work rule constraints. A scheduling run might take hours to run, or the problem cannot be solved in one go and has to be decomposed into sub-problems.

Genetic algorithms are a class of search strategies that emulate the process of natural selection and evolution. Analogy is made between potential solution schedules and individual members of an evolving population. The fitness of an individual is mainly measured by the number of duties, the fewer the better, it is consisted of. Through processes of mating and mutation, generations of the population evolve with survival of the fittest. GAs are strategies designed to influence the evolution path such that the search space is effectively reduced in a controlled manner, and that the chance of finding the optimal solution is maximised.

Experimental results indicate that while progressive improvements in the design of "direct" GAs for driver scheduling could be made, so that the driver schedules yielded are getting closer to the optimal ones as found by IMPACS, there is a general characteristic that such GAs

would arrive very quickly at the best sub-optimal solutions they can yield and then fail to get any closer to optimality in further generations. Their performance does not degrade with increased problem sizes though. The conclusion is that more fundamental changes in this approach are needed.

In the new method, GAs take on an indirect role. They do not form schedules at all, but instead would yield a population of vastly condensed versions of the problem. The basic principle is that the optimal schedule will only involve a small number of points, called relief opportunities, where and when changing of drivers would occur. However the total number of relief opportunities to choose from is usually very large. The GAs in the new method are therefore aimed at obtaining good subsets of relief opportunities, which are small in sizes but do not leave out the critical relief opportunities. Then, IMPACS can be applied to yield the optimal schedules almost trivially.

Each individual of the evolving population in the new scheme now represents a subset of the original relief opportunities. The fitness of an individual is measured by two factors: its size, the smaller the better; and an estimation of the number of duties required, the nearer to the optimal the better. This implies that a fast and accurate driver duties estimator is needed for evaluating the latter. The initial estimate for the full set of relief opportunities sets the target.

IMPACS gives a consistently accurate estimate derived from solutions of the relaxed LP without integer constraints, but is impractical for our purpose here because it takes too much computational time. A rule-based estimator is used instead. The estimate is based on an idealised lower bound, which takes into account the pattern of work to be covered and some standard types of duty. A bus driver generally can only work continuously for some hours and then must be relieved by another driver in order to take a break. The estimator has a rule-base for analysing critical times when it may be difficult to do so. There are also rules to determine

if such difficulties can be overcome. If not, additional drivers are needed and the lower bound would be tightened, thereby producing a more accurate estimate. Experimental results of the estimator match almost exactly with the actual results of IMPACS in terms of the total number of duties and the spread of different types of duty in a schedule.

The biggest advantage of this hybrid approach is that it does not require any of its three major components to be perfect to produce superior results. IMPACS would only have difficulty with really large problems, so it would be quite acceptable for the final population members to be slightly larger than the absolutely minimal subset of relief opportunities.

IMPACS has already been thoroughly described elsewhere. This paper concentrates on the GA and the estimator components. Experimental results will be described.

Polyhedral Methods for Set-Covering/Partitioning Problems

Sebastian Ceria, University of Columbia and University of Rome "La Sapienza"
Carlo Mannino, University of Rome "La Sapienza", Italy
Paolo Nobili, IASI-CNR, Rome, Italy
Antonio Sassano, University of Rome "La Sapienza", Italy

Abstract: There are two major components to the solution of crew-scheduling problems. The first one is the generation of feasible rotations while the second is the selection of the best collection of rotations such that each task is covered by at least one rotation. This last phase can be formulated as a Set-Covering/Partitioning problem. We show how general cutting planes can be used within a Branch and Cut framework to efficiently solve large scale Set-Covering/Partitioning problems arising from crew-scheduling applications. We use two different classes of cutting planes: the first one is that of "lift and project" cuts which do not exploit the special structure of the Set-Covering polytope while the second one consists of general rank inequalities. We present separation algorithms for both classes of inequalities and discuss issues related to making these inequalities globally valid. We present extensive computational results comparing these two methods with state of the art codes.

CALIBRATING INCIDENT DETECTION ALGORITHMS BY SIMULATED ANNEALING

Simon COHEN

Institut National de Recherche sur les Transports et leur Sécurité

(INRETS, France)

ABSTRACT

Automatic Incident Detection (AID) plays a leading role in motorway traffic management. The AID process is based on computer algorithms using real-time traffic data. In its simplest form, a detection algorithm consists of comparing a traffic feature to a threshold. The operational efficiency of the algorithms depends closely on the quality of the calibration of these thresholds. For any algorithm and for any AID system there is always a tradeoff between the false alarms and the ability to detect incidents. So, it is up to the calibration process to search an optimal compromise between these two main criteria. Conventional optimization methods often lead to local optimum solutions. They are less efficient when the number of thresholds increases. The paper illustrates the interest the simulated annealing approach for this type of calibration. An experimental test is performed on the south part of the inner ring road of Paris. Simulated annealing produces generally better solutions than classical methods for all the classes of AID algorithms tested.

1. A NEW CONTEXT FOR A.I.D

The earliest explorations of Automatic Incident Detection (AID) techniques were performed in the USA more than twenty five years ago and their fortunes have been varied. Most AID activities involve inductive loop technology and were quickly confronted by various types of problems : methodologies which were still in the process of gestation, inadequate calibration tools and data, patchy comparative evaluations, practical installation problems, etc. In the course of implementation these monitoring techniques have furthermore generated rates of false alarms which are completely incompatible with the operational constraints of traffic management. Such impediments have tended to cast a degree of doubt on these techniques, which are nevertheless extremely useful. A considerable number of road operators have thus decided not to include traffic monitoring among their traffic management measures.

However, in the last few years there has been a marked revival of interest in AID techniques. One of the main reasons for this is the promising development of driver information systems. The credibility of information systems depends mainly on a good knowledge of traffic conditions as perceived through the various types of media, such as Variable Message Signs (VMS), for example. Traffic monitoring therefore becomes an indispensable traffic management measure. Its effectiveness greatly affects the quality and reliability of information systems.

This new context has acted as a catalyst for research into traffic monitoring. The DRIVE programme in Europe and the IVHS programme in North America have provided the framework for a large amount of work which focuses on the topic. Furthermore, AID techniques are currently benefiting from progress in the application of

new information and communications technology to traffic management. Recently, other technologies have been investigated, as for instance video image processing systems (5). Furthermore, new theories are now also explored in the context of incident detection. Examples of these are artificial neural networks (6), new filtering techniques (7) and catastrophe theory (3).

This statement underlines the multiplicity of all the possible monitoring systems. Choosing an adapted system/algorithm becomes a complex task. Thus, a serious investigation of the strengths and limitations of conventional and new algorithms appears necessary. The paper focuses on this aspect.

This paper describes a part of the research on motorway traffic surveillance currently undertaken at INRETS. More precisely, it examines the problems of incident detection on networks of expressways and motorways equipped with inductive loops. The paper focuses the calibration aspects of AID algorithms. It examines the potential of two optimisation techniques for a reliable assessment of the performances of the AID algorithms.

The potential of these different optimisation techniques is illustrated by experimental tests carried out in the Ile-de-France Region, on the Paris ring road. The conclusions enable us to improve the assessment of the real efficiency of the different existing approaches.

2. THE DIFFERENT APPROACHES

Automatic Incident Detection algorithms originated in the 1970s for detecting incidents using loop data sensors. While the earlier were simpler in theory, they had certain performance limitations which necessitated further enhancements and development of newer methods or approaches. This presented opportunities for exploring new theories to be applied and tested in the context of incident detection. Examples of these were several methods that fit into the category of statistical type. Of late, concepts such as catastrophe theory and neural networks are being explored for application in incident detection algorithms. While the superiority of one type of algorithms in all respects have to be demonstrated, each method possesses its own distinctness and advantages (2).

Pattern recognition algorithms were the first to be developed. They rely on recognizing and differentiating unusual pattern of traffic from the normal conditions. In this group the California algorithm is the most widely known and is often used as a standard for measuring the performance of others algorithms. There are several variations of this algorithm with different features characteristics. Algorithm #7 and Algorithm #8 are powerful versions of the California type developed by Payne et al (4).

Statistical techniques model the random traffic patterns using time series detector data. Statistical models include the Standard Normal Deviate (SND), Bayesian, time series (e.g exponential smoothing), dynamic algorithms... Most statistical techniques suffer from the static nature of the models. While the performance levels of statistical models vary under different conditions and environments, they all have the commonalty of comparing observed variables (e.g volume, speed, occupancy) with predicted model variables. This implies that a good prediction model is a precondition for efficcient results. Recently, a new set of filter algorithms have been developed mixing statistical technique to smooth raw data and pattern recognition approach to discriminate congestion and incident patterns (7).

Catastrophe theory is now proposed for incident detection. The logic defined by Hall et al (3) in the McMaster algorithm utilizes a template diagram on a volume-occupancy plane to identify an incident. The diagram is divided into four regions. Each region represents a specific traffic profile like for instance uncongested flow or congested flow patterns. Based on investigation of the difference in the congested traffic flow patterns caused

by an incident and recurrent congestion, a logic is proposed to determine the cause of the congestion. Initial application of catastrophe theory has shown promising results.

3. SELECTION FOR ASSESSMENT AND COMPARISON

A review of incident detection strategies currently implemented reveals that algorithms based on macroscopic traffic data are the most likely to be implemented for motorway control. Three AID algorithms, based on the same aggregate traffic data (over 1 mn), are selected for assessment and comparison. According to reported results, these algorithms seem particularly suited during recurring congestion periods.

In the first class of pattern recognition techniques, Algorithm #7 is considered as a reference. This algorithm is defined by three thresholds.
Algorithm #8 is also selected for assessment. The main advantage of this structure, with five thresholds, is its ability to suppress compression waves, principal source of false alarms in heavy traffic (4).

The third algorithm chosen for assessment belongs to the Delos family, recently developed by Stephanedes and Chassiakos (7). It corresponds to the Delos 1.1 version with two thresholds. The proposed logic uses filters that smooth raw data over large windows to eliminate short duration traffic disturbances such as random fluctuations, traffic pulses and compression waves. Comparisons of smoothed spatial difference on occupancy over time allows the discrimination between reccurrent congestion and incidents.

4. THE CALIBRATION PROBLEM

In its simplest form, a detection algorithm consists of comparing a traffic feature to a threshold. An incident is signaled if, for example, the value of the feature is greater than the chosen threshold.

More generally, the development of AID methods proceeds according to the following steps :

- identification of traffic features which take values unique to incident conditions ;
- combination of the above features (structure of the algorithm) ; and
- selection of the right thresholds (calibration).

In operational conditions, this approach requires the application of calibration procedures based on real data, contrary to the methods having recourse to simulation models. The operational efficiency of the algorithms depends closely on the quality of the calibration phase :

- too loose thresholds generate an intolerable number of false alarms ;
- too strict thresholds reduce detection capability.

For any algorithm and for any AID system there is always this tradeoff between the false alarms and the ability to detect incidents. So, it is up to the calibration process to search an "optimal" compromise between the two main criteria, the detection rate and the false alarm rate.

Each AID algorithm is characterized by a structure and a set of thresholds $T = \{T_i, i=1,...,n\}$. For each value of T, it is possible to calculate, for a given data set, the values of the usual performance criteria :

- the false alarm rate FAR(T), defined as the fraction of alarm signals generated for non-incident situations ;

- the detection rate DR(T), defined as the fraction of incidents detected; and

- the mean time-to-detect.

The choice of a threshold set T is optimal, when it corresponds to the lowest possible false alarm rate, for a given detection rate. Stated formally, the optimal T is the solution to a non-linear optimization problem under constraints, i.e., minimization of the false alarm rate FAR(T), subject to a detection rate DR(T) not inferior to a desired value y :

$$P(y) : \text{Min}_T \{FAR(T) \mid DR(T) > y\}$$

Algorithm performance is then assessed by means of the characteristics curves. These curves depict detection and false alarm rates performed by an algorithm across optimal thresholds values. The numerical process described by each P(y) program produces a performance point on the characteristic curve.

The main advantage of this evaluation approach is its independence from the algorithm structure.

5. OPTIMIZATION BY ERROR AND TRIALS

An optimization routine based on an error and trial approach is tested as a reference. It attempts to minimize a function it receives as an argument. The parameter of this function is the vector of threshodls wich is modified during this optimization process.

The initial value of the vector of thresholds is used to compute the initial fonction value of the false alarm rate. A gaussian random number generator is used to get a random increment in each dimension of the vector of thresholds. These random increments are normalized, yielding a uniformly distributed random unit vector , which, when multiplied by the specified stepsizes and added to the current value of the vector of thresholds produces a test value of this vector. The corresponding value of the false alarm is compared to the current minimum value. A test value larger than the current minimum is termed a failure, while a smaller value is called a success.

If a test value fails, a new random unit vector is generated and the test repats. After a specified number of consecutive failures, the stepsize is reduced by a specified factor and a new random vector is generated.

If the test value is a success, the search continues is the same direction until a failure occurs. This process terminates when the stepsize becomes too small, the maximum number of iterations exceeded or the desired minimum function value is reached.

Thresholds for operational algorithms are generally calibrated by this trial and error technique (1), (4).

6. THE SIMULATED ANNEALING APPROACH

Although deterministic algorithms may have efficient implementations and converge to solutions, they are not guaranteed to find optimal global minimum of a given optimization function. This is especially true if the function has distinct local minima. This statement leads us to test a simulated annealing approach to perform such optimization problems for AID.

Simulated annealing is a general purpose combinatorial optimization technique. It differs from a conventional iterative improvement algorithm in that, it not only accepts a solution with a better objective function value but also accepts a solution with worse objective function values probabilistically. The basic ideas of this method

can be extended to continuous minimization problems (8). Annealing based optimization methods employs noise to choose new parameters values (thresholds to calibrate). New points are chosen at various distances from their current point. Each new point is generated probabilistically according to a given distribution. The function value is calculated and probabilistically accepted or rejected. If accepted, the new point becomes the current point. The new point may be accepted even if it is worse and has a larger function value than the current point. The criteria for acceptance is determined by an acceptance function, a "temperature" parameter and the difference in the function values of the two points. Initially, the temperature is large, and a new point is accepted roughly half the time. As the numerical process progresses, this temperature parameter is reduced, thus lowering the probability that the acceptance function will accept a new point if it's functional value is greater than that of the current point. Because the algorithm occasionally chooses points uphill from its current point, it can escape from local minima and more effectively search the function space to find the global minimum. Thus, this approach is well suited to solving constrained nonlinear optimization problems like AID calibration. As previously, the false alarm rate is naturally the function to minimize.

The four elements required by the Metropolis procedure are as follows :

- the value of the false alarm rate is the objective function ;
- the system state is the point x defined by the thresholds ;
- the control parameter T is, something like a temperature, with an annealing schedule by which it is gradually reduced ;
- there is a generator of random changes in the configuration, that is, a procedure for taking a random step from x to $x + \Delta x$.

In practice, the last of these elements is the most problematical. Several different schemes for choosing Δx can be considered. The problem is one of efficiency. A generator of random changes is inefficient if, when local downhill moves exist, it nevertheless almost always proposes an uphill move. A good generator, should not become inefficient in narrow valleys ; nor should it become more and more inefficient as convergence to a minimum is approached.

A classical way of doing simulated annealing minimization on continuous control spaces is to use a modification of the downhill simplex method. This amounts to replacing the single point x as a description of the system state by a simplex of (N+1) points. The implementation of the Metropolis procedure is slightly subtle : We add a positive, logarithmically distributed random variable from the function value of every point that is tried as a replacement point. Like the ordinary Metropolis procedure, this method always accepts a true downhill step, but sometimes accepts an uphill one.

In the limit $T \mapsto 0$, this algorithm reduces exactly to the downhill simplex method and converges to a local minimum.

At a finite value of T, the simplex expands to a scale that approximates the size of the region that can be reached at this temperature, and then executes a stochastic tumbling Brownian motion within that region, sampling new, approximately random, points as it does so. The efficiency with which a region is explored is independent of its narrowness and orientation. If the temperature is reduced sufficiently slowly, it becomes highly likely that the simplex will shrink into that region containing the lowest relative minimum encountered.

7. A DATABASE FOR THE OPTIMIZATION PROCESS

The method of threshold optimization based on real data requires a significant number of observed incidents. In order to evaluate or calibrate an AID algorithm using real data from a motorway system, several categories of

data sets should be collected. More specifically, a relational database should be developed consisting of a large number of incident and non-incident traffic measures, supplemented by geographic information and incident characteristics.

Using the above data sets and by means of any classical DBMS tool, a relational incident database can be developed. The structure of the database is built around four basic tables : "station" ; "incident" ; "trafnorm" and "trafinc".

The "station" table represents the geometry of the roadway system. Each detection station is described in terms of :

- station name ;
- downstream station ;
- number of lanes of the roadway ; and
- position of the station with regard to a fixed point on the roadway.

The "incident" table describes each incident of the database in terms of :

- incident identification number ;
- date and estimated time of incident occurrence ;
- time of clearance ;
- upstream station of the incident site;
- site description ;
- kilometer-point of the incident site, as referred to a fixed point on the roadway;
- incident type ;
- number of vehicles involved ; and
- number of trucks involved.

Among these characteristics, several are not directly used during the calibration process. Nevertheless, they reveal to be useful for the analysis of the performance of the algorithms for various traffic and environmental conditions.

The nature of the two following tables depends closely upon the data acquisition system. Most of these systems use classical inductive loops. Macroscopic traffic parameters like flow, occupancy and mean speed are derived, agregated for instance every minute, and strored.

The normal trafic table or "normtraf", contains one-minute values of volume and occupancy for the normal traffic (incident-free) periods. Each period is 45 minutes long (but it could be of any length), beginning 15 minutes before the time of incident occurrence. The columns of the table include :

- period number ;
- date and time of measurement ;
- station name ;
- volume ; and
- occupancy rate.

Volume is defined as the one-minute count, summed over all lanes and expressed as vehicles per minute. Occupancy is defined as the one-minute occupancy, expressed as a percentage, at a given station, averaged across all lanes.

The incident traffic table or "inctraf" is similar to the previous table, but it refers to the incident counts. It contains the following columns :

- incident number ;
- date and time of measurement ;
- station name ;
- volume ; and
- occupancy rate.

It must be stated that, if speed is available, it could naturally be taken into account as an other column of the tables "normtraf" and "inctraf".

8. A TEST SITE ON THE PARIS RING ROAD

The Paris Ringway is a two-way circular motorway system around the city. Its length is 35.17 km and its particularity is that it lacks shoulders on both sides. Due to this, any incident, even a stall, causes a capacity reduction, especially critical during peak hours. On average, thirty incidents are reported every day.

The pilot section on the south part of the inner ring road is 6 km length. The section test has 3 lanes and includes 11 on-ramps. The surveillance of the section is achieved by detection stations positioned every 500 m. Traffic data - flow and occupancy - measured every 20 seconds are agregated over 1 minute.

In order to evaluate the performances of the AID algorithms using real data from a motorway system, several categories of data sets are collected. A relational database is developed consisting of a large number of incident and non-incident traffic measures, supplemented by geographic information and incident characteristics.

Informations related to incidents, like the beginning or the location, are surveyed by operators. Data concerning the exact instant of occurrence of the incidents appear inaccurate. Finally, the exploited database includes 177 incidents recorded on the test section, of which 125 occurred during peak-hours.

9. EXPERIMENTAL RESULTS

Several optimization runs of each numerical optimization process are performed from the same initial state, and with an equal number of iterations. In all the cases tested, the superiority of the simultated annealing approach is clearly stated. The interest of this annealing approach is particularly high when the number of parameters to calibrate increases.

9.1 Calibration of the AID algorithm Dclos 1.1(2, 5) with two thresholds

Figure 1 illustrates the final result of the different runs. The characteristic curves derived from each of the two numerical process are very closed at almost all the detection levels. Nevertheless, the characteristic curve obtained by the simulated annealing optimization is slightly above the one given by the error and trial process.

An example illustrates this difference. For the same detection rate of 69.5%, the false alarm rate reached - function to minimize - corresponds to 1.34% with the error and trial process and 1.20% with the simulated annealing optimization.

9.2 Calibration of the AID Algorithm #7 with three thresholds

For levels of detection above 50%, the differences between the two approaches become significant. The phenomenon is illustrated in Figure 2, showing the characteristic curves derived in each case.

At 80.2% of detection, the false alarm rate is 2.40% with the error and trial process and 1.29% with the simulated annealing optimization. Furthermore, the vectors of thresholds are largely distinct : (2.527 ; 0.323 ; 21.825) and (19.340 ; 0.145 ; 25.808) respectively.

9.3 Calibration of the AID Algorithm #8 with five thresholds

The difference between the two approaches increases strongly with the level of detection, as shown in Figure 3.

At 86.44% of detection, the minimum is 3.00% with the error and trial process and only 0.78% with the annealing approach. Once again, the vectors of thresholds obtained at the end of the optimization process are largely different : (5.272 ; - 0.418 ; 0.190 ; 29.908 ; 33.529) and (15.932 ; - 1.148 ; - 4.951 ; 26.431 ; 29.249) respectively.

10. RECOMMANDATION

Operational implementation of a surveillance system begins to take shape only once suitable and accurate calibration has been achieved. The assessment of the AID algorithms according to their potential performances remains a complex task. Conclusions depend of several factors such as the characteristics of the site and its environment, the nature of the traffic, the quality and the reliability of the data, and particularly the calibration technique, ... On this last point, experimental tests performed on the Paris ring road demonstrate clearly the efficiency of the simulated annealing approach for the calibration of many AID algorithms. This optimization technique is strongly recommanded when the number of thresholds to calibrate increases. It allows a reliable assessment of the real performances of each surveillance system.

REFERENCES

(1) Cohen S, Ketselidou Z 1992 "A calibration process for automatic incident detection algorithms", Fourth International Conference on Microcomputers in Transportation, Baltimore, Maryland, USA

(2) Cohen S (1994) "Comparative assessment of conventional and new incident dectection algorithms", seventh international conference on Road Traffic Monitoring and Control, Conference Publication n°391, IEE, London, pp. 156-159.

(3) Hall F.L, Shi Y, Atala G, 1993, Transp Research Record, 1394, 1-7

(4) Payne H.J, Helfenbein E.D and Knobel H.C, 1976, "Development and testing of incident detection algorithms", Vol 2,3 and 4, Report n° FHWA-RD 76-20, prepared for the Federal Highway Administration, Offices of Research and Development, Washington, DC, USA

(5) Michalopoulos P. G et al,1993, TEC, pp. 66-75

(6) Ritchie S. G, Cheu R. L ,1993, Transp Research, C.1, n°3, 203-217

(7) Stephanedes Y.J, Chassiakos A.P,1993, Transp Research, C. 1, n°3, 219-233

(8) Vanderbilt D, Louie S. G 1984, Journal of Computational Physics 56, 259-271

Figure 1 : Characteristic curves for 2 thresholds

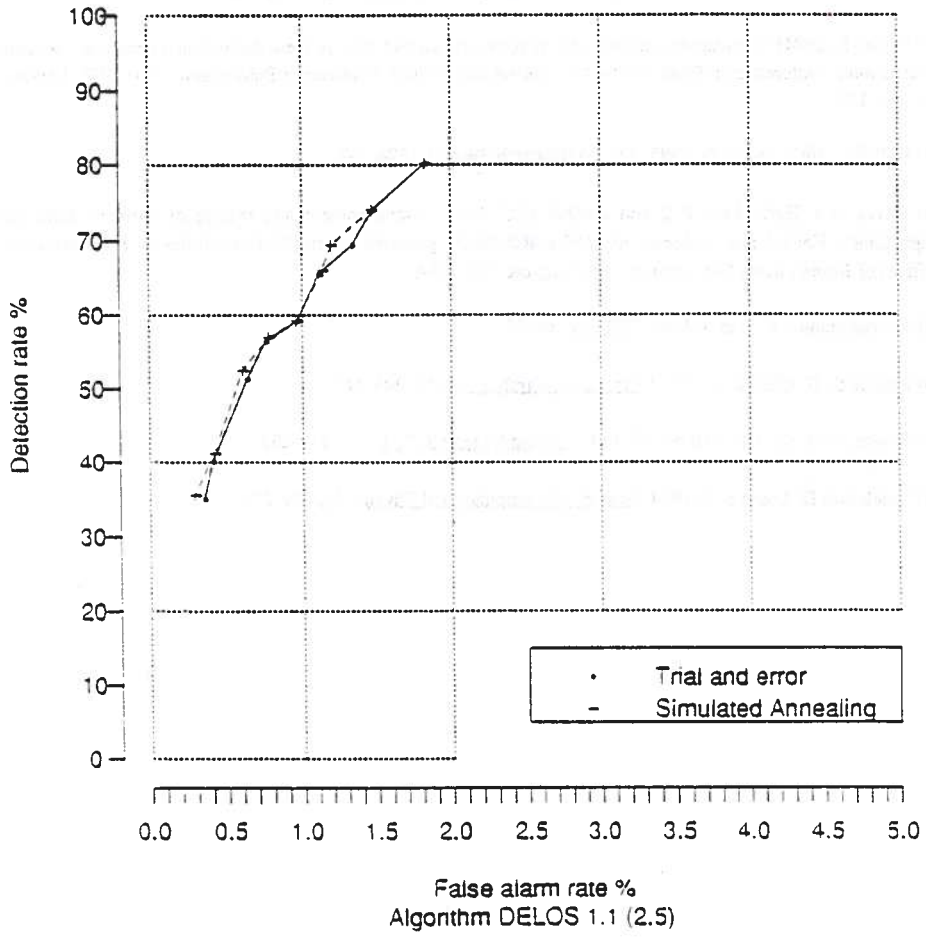


Figure 2 : Characteristic curves for 3 thresholds

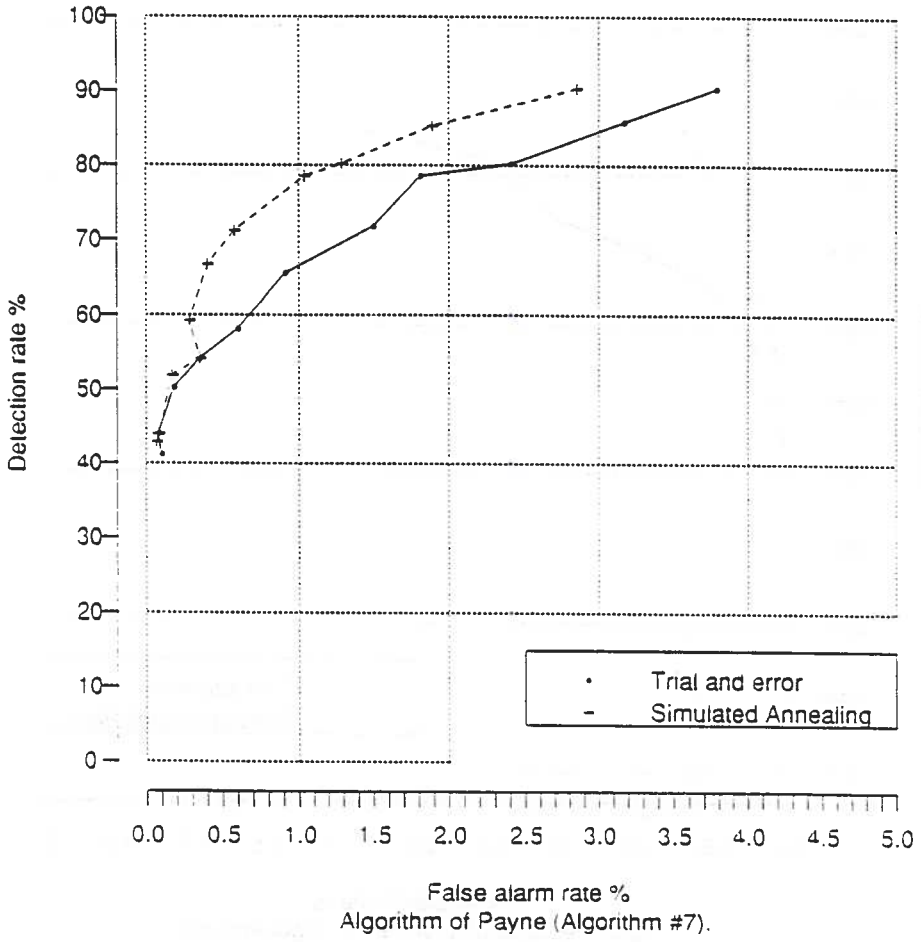
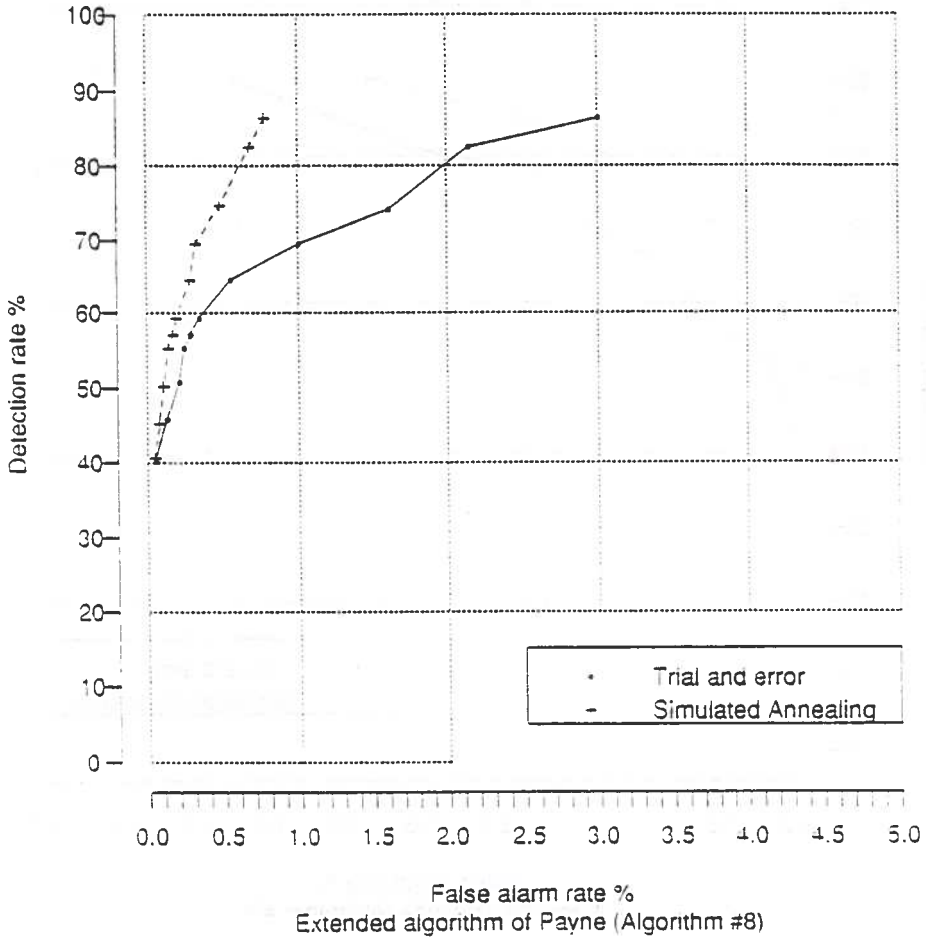


Figure 3 : Characteristic curves for 5 thresholds



NEURAL NETWORKS FOR THE FEEDBACK OPTIMAL CONTROL OF FREEWAY SYSTEMS

A. Di Febbraro §, D. Di Noto ‡, T. Parisini §, V. Recagno †, S. Sacone §and R. Zoppoli §

§Department of Communications, Computer and System Sciences
DIST-University of Genoa, Via Opera Pia 13, 16145 Genova, Italy

‡Institute of Transportation
University of Genoa, Via Montallegro 1, 16145 Genova, Italy

†Department of Electrical Engineering, Via Opera Pia 11A, 16145 Genoa, Italy

Abstract. The problem of clearing congestion situations in freeway traffic is addressed. Traffic is controlled by regulating the vehicle access to the freeway and by limiting the vehicle speed by means of variable message signs. Even though the problem is stated within a deterministic context, the optimal control law is sought in feedback form in order to react to unpredictable events. The resulting functional optimization is reduced to a nonlinear programming problem by constraining the control law to take on a fixed structure in which free parameters have to be optimized. For such a structure, a multilayer feedforward neural mapping is chosen. As the resulting neural control law approximates the optimal one, bounds to the approximation errors are established. The problem is solved for both an N-stage control horizon and a receding-horizon control mechanism. Simulation results for this latter control scheme are presented.

1. INTRODUCTION

Modelling and control of freeway traffic has been raising the interest of many researchers for some decades. In fact, freeways can be viewed as complex large-scale systems with such peculiar characteristics as to require dedicated models and control methods.

Freeway traffic models are essentially divided into two main classes, called microscopic and macroscopic models. Such classes differ from each other fundamentally in the way the vehicles are described. Microscopic models represent the behaviour of each single vehicle in the stream of interacting vehicles, whereas macroscopic models describe the behaviour of the whole traffic flow. The latter models exploit the analogy between the dynamics of fluids and of traffic flow, which results in describing such a flow in terms of aggregate state variables (traffic density, volume, and mean speed). Freeway control systems are usually specialized to monitor only one of such aspects as safety, performances, travel time estimation, energy consumption, pollution (Klijnhout, 1991). In this paper, the problem of managing traffic congestion on freeways is faced. One can find in the literature two major classes of control strategies to prevent and solve congestion situations on freeway systems: i) strategies intended to monitor the freeway access in order not to make traffic volume exceed road capacity; ii) strategies that aim at constraining drivers' behaviour by suggesting them what they should do to avoid congestion situations. In this paper, both classes of control policies will be considered.

The freeway system dealt with in the paper is represented by a model proposed by Payne (1971). Such a model has been used in a number of works about freeway traffic control (see, for instance, Isaksen and Payne, 1973; Looze *et al.*, 1978; Papageorgiou, 1983). After defining the traffic model, an open-loop optimal control problem aimed at clearing congestion situations is stated. The exact solution to such

a problem is an extremely difficult task because of both the strong nonlinearities and the high order of the model chosen. In (Papageorgiou, 1983), for example, a suboptimal solution is found based on the application of a multilayer control structure. Applications of various suboptimal control schemes based on the same principle can be found in (Isaksen and Payne, 1973; Payne *et al.*, 1973; Goldstein and Kumar, 1982).

The basic idea of the proposed approach consists in giving up "exact" control strategies and in searching for approximate solutions by constraining control strategies to take on a fixed structure in which a certain number of parameters have to be determined. Even though the problem is stated within a deterministic context, we shall determine a feedback control law and assume that the initial state may take *any* value from a given compact set. This enables the control system to react instantaneously to any possible congestion situation. As a control structure, we choose nonlinear approximators based on multilayer feedforward neural networks, since such approximators are easy to handle and, above all, for their powerful approximating capabilities. A nonlinear programming problem has then to be solved and a gradient technique is applied. To compute the gradient components, recursive equations are introduced, which generalize the classical adjoint equations of N-stage optimal control theory.

2. DESCRIPTION OF THE MODEL

We use a macroscopic model first proposed by Payne (Payne, 1971; Papageorgiou, 1983) based on the discretization in both space and time of the matter equation, from which the following equation is derived

$$\rho_k(i+1) = \rho_k(i) + \frac{T}{\Delta_k}\left[q_{k-1}(i) - q_k(i) + r_k(i) - s_k(i)\right], \quad i = 0, 1, \ldots, N-1;\ k = 1, \ldots, K \tag{1}$$

The freeway has been divided into K sections of length Δ_k, $k = 1, \ldots, K$ of 500 to 1000m, and N control stages have been defined. T is the sample time interval. $\rho_k(i)$ is the traffic density (number of vehicles in section k at time iT divided by the section length), $q_k(i)$ is the traffic volume (number of vehicles leaving section k during the time interval $[iT, (i+1)T]$ divided by T), $r_k(i)$ and $s_k(i)$ are the on-ramp and off-ramp traffic volumes of section k, respectively. Actually, $r_k(i)$ and $s_k(i)$ are different from zero only for those sections which contain an on/off ramp. The indexes of such sections make up set $\mathcal{I}_r$. The third aggregate variable of interest is the mean traffic speed, $v_k(i)$, whose dynamics is described by the equation

$$v_k(i+1) = v_k(i) + \frac{T}{\tau}\left\{V[\rho_k(i), b_k(i)] - v_k(i)\right\} + \frac{T}{\Delta_k} v_k(i)\left[v_{k-1}(i) - v_k(i)\right] + \frac{\nu T\left[\rho_{k+1}(i) - \rho_k(i)\right]}{\tau \Delta_k \left[\rho_k(i) + \chi\right]}$$
$$-\delta_{\text{on}} \frac{T}{\Delta_k} v_k(i) \frac{r_k(i)}{\rho_k(i) + \chi}, \quad i = 0, 1, \ldots, N-1;\ k = 1, \ldots, K \tag{2}$$

where τ, ν, χ, and δ_{on} are parameters determined experimentally. $V(\rho_k, b_k)$ is the steady-state speed-density characteristic, which in the described model takes on the form

$$V(\rho_k, b_k) = V_f\, b_k \left[1 - (\rho_k/\rho_{\max})^{m(3-2b_k)}\right]^l \tag{3}$$

where V_f denotes the "free" speed (in the absence of speed limits), $\rho_{\max}$ is the so-called "jam density", m and l are real-valued parameters, and $b_k(i)$ is a variable which models speed limits ($b_k = 1$ means that there is no speed limit in section k). Both $r_k(i)$ and $b_k(i)$ can be modified over time, thus playing the role of control variables. Variables $r_k(i)$ are imposed by monitoring the vehicle access via traffic lights placed on the freeway on-ramps. Speed limitations are imposed by variables $b_k(i)$, which are displayed on *variable message signs* properly placed along the road.

The three aggregate state variables so far defined are related to each other by means of the following weighted sum

$$q_k(i) = \alpha\, \rho_k(i)\, v_k(i) + (1-\alpha)\, \rho_{k+1}(i) v_{k+1}(i),\ i = 0, \ldots, N-1 \tag{4}$$

where $0 \leq \alpha \leq 1$ is an appropriate weighting factor. The off-ramp volumes $s_k(i)$ are related to the traffic volumes $q_{k-1}(i)$ through the relationship $s_k(i) = \gamma_k\, q_{k-1}(i)$, $k \in \mathcal{I}_r$ (Papageorgiou and Mayr, 1982), where $0 < \gamma_k < 1$. Substitution of this relation and of (4) into (1) yields

$$\rho_k(i+1) = \rho_k(i) + \frac{T}{\Delta_k}\left[\alpha\,(1-\gamma_k)\,\rho_{k-1}(i)v_{k-1}(i) + (1-2\alpha+\gamma_k\alpha-\gamma_k)\,\rho_k(i)\,v_k(i)\right.$$
$$\left.-(1-\alpha)\rho_{k+1}(i)v_{k+1}(i) + r_k(i)\right], \quad i = 0,1,\ldots,N-1;\ k = 1,\ldots,K \tag{5}$$

Equations (2) and (5) must be modified for both the first and the last sections of the freeway. To complete the model, we have to describe the dynamics of the queues on the on-ramps. Then, define the state variable $l_k(i)$ as the queue on the on-ramp of section $k \in \mathcal{I}_r$ at time iT, and $d_k(i)$ as the demand for access to the same ramp at the same time. Clearly, these quantities are related by the equation

$$l_k(i+1) = l_k(i) + T\,[d_k(i) - r_k(i)],\ i = 0,1,\ldots,N-1\,;\ k \in \mathcal{I}_r \tag{6}$$

To sum up, the model of the freeway turns out to be a discrete-time nonlinear dynamic system, whose state equations are Eqs. (2),(5) and (6). The state variables are the components of the following vectors: $\underline{\rho}_i \triangleq \mathrm{col}\,[\rho_k(i),\ k = 1,\ldots,K]$, $\underline{v}_i \triangleq \mathrm{col}\,[v_k(i),\ k = 1,\ldots,K]$, $\underline{l}_i \triangleq \mathrm{col}\,[l_k(i),\ k \in \mathcal{I}_r]$. In a more compact form, we define the state vector as $\underline{x}_i \triangleq \mathrm{col}\,(\underline{\rho}_i, \underline{v}_i, \underline{l}_i)$. In the same way, we define the control vector as $\underline{u}_i \triangleq \mathrm{col}\,(\underline{r}_i, \underline{b}_i)$, where $\underline{r}_i \triangleq \mathrm{col}\,[r_k(i),\ k \in \mathcal{I}_r]$, $\underline{b}_i \triangleq \mathrm{col}\,[b_k(i),\ k = 1,\ldots,K]$. Then, to use the standard notation of control theory, we write (2),(5),(6) in the form of a state equation

$$\underline{x}_{i+1} = \underline{f}\,(\underline{x}_i, \underline{u}_i), \quad i = 0,1,\ldots,N-1 \tag{7}$$

3. STATEMENT OF THE OPEN-LOOP AND CLOSED-LOOP OPTIMAL CONTROL PROBLEMS

The control objective is to minimize the total time spent by vehicles in the freeway system. Such a time is the sum of two terms. The first term represents the total travel time and the second one the total waiting time on the on-ramps. Let us now define the demand vector $\underline{d} \triangleq \mathrm{col}\,[d_k(i),\ k \in \mathcal{I}_r,\ i = 0,1,\ldots,N-1]$. Then, we can state the following (Papageorgiou, 1983)

Problem 1. Given the demand vector $\underline{d}$ and the initial condition vectors $\underline{\rho}_0$, $\underline{v}_0$, and $\underline{l}_0$, find the optimal control vectors $\{\underline{r}_i^\circ, \underline{b}_i^\circ, i = 0,1,\ldots,N-1\}$ that minimize the cost function

$$J = \sum_{i=0}^{N} T\,\underline{\rho}_i^T\,\underline{\Delta} + \sum_{i=0}^{N} T\,\underline{l}_i^T\,\underline{e} \tag{8}$$

where $\underline{\Delta} \triangleq \mathrm{col}\,(\Delta_k, k = 1,\ldots,K)$, $\underline{e} \triangleq \mathrm{col}\,[e_k = 1, k \in \mathcal{I}_r]$, subject to constraints (2),(5), (6) and

$$0 \leq \rho_k(i) \leq \rho_{k\,\max}, \quad i = 0,1,\ldots,N-1;\ k = 1,\ldots,K \tag{9}$$

$$0 \leq v_k(i) \leq v_{k\,\max}, \quad i = 0,1,\ldots,N-1;\ k = 1,\ldots,K \tag{10}$$

$$0 \leq l_k(i) \leq l_{k\max},\ i = 0,\ldots,N-1,\ k \in \mathcal{I}_r \tag{11}$$

$$0.7 \leq b_k(i) \leq 1,\ i = 0,1,\ldots,N-1;\ k = 1,\ldots,K \tag{12}$$

$$r_{k\,\min}(i) \leq r_k(i) \leq r_{k\,\max}(i), \quad i = 0,1,\ldots,N-1;\ k \in \mathcal{I}_r \tag{13}$$

where

$$r_{k\,\min}(i) = \max\left\{r_{\min}, d_k(i) + \frac{1}{T}\,[l_{k\max} - l_k(i)]\right\} \tag{14}$$

$$r_{k\,\max}(i) = \min\left\{r_{\max}, d_k(i) + \frac{1}{T}\,l_k(i)\right\} \tag{15}$$

$r_{\min}$ and $r_{\max}$ are fixed parameters dependent on the road characteristics.

□

We have assumed that the variables $b_k(i)$ take their values over the interval specified by (12). Actually, such variables can take only discrete values. After solving Problem 1, a suitable discretization of these speed limits seems reasonable. In addition, the frequency of changes in speed limits should be low, which makes the variables $b_k(i)$ keep the same values for a fixed number of subsequent time intervals (Papageorgiou, 1983). By using equations (5) and (6), it is possible to rewrite cost function (8) as

$$J = J_0 + T \sum_{i=0}^{N-1} \left\{ \sum_{k=1}^{K} \Delta_k \rho_k(i) + \sum_{k \in \mathcal{I}_r} \{ l_k(i) - T\gamma_k [\alpha \rho_{k-1}(i) v_{k-1}(i) + (1-\alpha)\rho_k(i) v_k(i)] \} \right.$$
$$\left. + T(1-\alpha)\rho_1(i) v_1(i) - \alpha T \rho_K(i) v_K(i) \right\} \tag{16}$$

where J_0 is a constant term not involved in the minimization of J, depending on the demand vector $\underline{d}$, on upstream traffic volumes, and on initial conditions.

We want to remark that Problem 1 is a complex constrained nonlinear programming problem. Moreover, its solution yields the optimal control law in open-loop form for a specific value of the initial state vector $\underline{x}_0$. This constitutes a serious limitation to the practical use of such a control law, since: i) the open-loop structure does not enable the control system to react to unpredictable disturbances; ii) even though no perturbation occurs, the computation time to determine the optimal control sequence $\underline{u}_0^o, \ldots, \underline{u}_{N-1}^o$ is in general too long to clear a congestion situation within an acceptable time. Then, we need to derive a feedback optimal control law (we assume that all the state variables can be measured exactly) for *any* initial state $\underline{x}_0 \in \mathcal{A}_0$, where $\mathcal{A}_0$ is the set of states $\underline{x}_0$ describing all possible initial congestion situations. We also introduce a first approximation in order to remove constraints (9)–(13). This approximation consists in adding suitable penalty functions to cost (8). As to constraints (9), we add the terms $g_\rho[\rho_k(i)] = K_\rho \left\{ [\max(0, -\rho_k(i))]^2 + [\max(0, \rho_k(i) - \rho_{k\max})]^2 \right\}$, where K_ρ is a positive scalar. Similar terms are added for the other constraints. Clearly, penalty functions of other types may also be used. The only condition to be satisfied is that the penalty functions be derivable with respect to their arguments. Then, by using a more compact notation, the cost function (8) can be rewritten as follows (the term J_0 is omitted)

$$J = \sum_{i=0}^{N-1} h(\underline{x}_i, \underline{u}_i) + h_N(\underline{x}_N) \tag{17}$$

where terms $h(\underline{x}_i, \underline{u}_i)$ have clear meanings ($h_N(\underline{x}_N)$ does not contain penalizing terms on the control variables). Then we can state the following

Problem 2. Find the optimal feedback control law $\{\underline{u}_i^o = \underline{\gamma}_i^o(\underline{x}_i),\ i = 0, 1, \ldots, N-1\}$ that minimizes cost (17) for any initial state vector $\underline{x}_0 \in \mathcal{A}_0$.

□

Note that Problem 2 is much more complex than Problem 1, as it is a functional optimization problem rather than a nonlinear programming one.

4. STATEMENT OF THE RECEDING-HORIZON OPTIMAL CONTROL PROBLEM

Problems 1 and 2 address the case in which the control actions on the freeway are exerted over a finite horizon of N decisional stages. This appears reasonable if a congestion on a freeway (due, for example, to the occurrence of an accident) has to be cleared within a finite time. However, if the traffic has to be regulated over an infinite time horizon, Problems 1 and 2 loose their meanings. On the other hand, it is well known that no tool is in general available for solving an infinite-horizon non-LQ optimal control

problem. An interesting approach to solve approximately such a problem consists in stating a suitable receding-horizon (RH) optimal control problem. Such a scheme can be described as follows. When the controlled system is in the state $\underline{x}_t$ at stage t, a finite horizon (FH) N-stage optimal control problem is solved, thus the sequence of optimal control vectors $\underline{u}_t^\circ, \ldots, \underline{u}_{t+N-1}^\circ$ is derived, and the first control of this sequence becomes the control action $\underline{u}_t^{RH^\bullet}$ generated by the RH regulator at stage t, that is $\underline{u}_t^{RH^\bullet} \triangleq \underline{u}_t^\circ$ (we shall use indexes t for the RH control problem, and we shall go on using indexes i for the FH case). Then, a *feedback* control law is obtained, as the control vector $\underline{u}_t^\circ$ depends on the current state $\underline{x}_t$.

In (Parisini and Zoppoli, 1994b) a possible technique is proposed for deriving the RH stabilizing optimal regulator without imposing either the "exact" constraint that the equilibrium point is reached in a finite time (as in (Keerthi and Gilbert, 1988; Mayne and Michalska, 1990)) or the condition of reaching the neighbourhood W of the equilibrium point (as in (Michalska and Mayne, 1993)). Unlike the previous approaches, the RH control law is computed "a priori", and not on line. Define now the following cost function

$$J_{FH}(\underline{x}_t, \underline{u}_{t,t+N-1}) = \sum_{i=t}^{t+N-1} h(\underline{x}_i, \underline{u}_i) + h_N(\underline{x}_{t+N}) \tag{18}$$

where $t \geq 0$, $\underline{u}_{t\tau} \triangleq \mathrm{col}(\underline{u}_t, \ldots, \underline{u}_\tau)$. Now N denotes the length of the control horizon. Then we can state the following

Problem 3. At every time instant $t \geq 0$, find the RH optimal control law $\underline{u}_t^{RH^\bullet} = \gamma_{RH}^\circ(\underline{x}_t)$, where $\underline{u}_t^{RH^\bullet}$ is the first vector of the control sequence $\underline{u}_t^{FH^\bullet}, \ldots, \underline{u}_{t+N-1}^{FH^\bullet}$ (i.e., $\underline{u}_t^{RH^\bullet} \triangleq \underline{u}_t^{FH^\bullet}$), that minimizes cost (18) for the state $\underline{x}_t$.

□

The statement of Problem 3 does not impose any particular way of computing the control vector $\underline{u}_t^{RH^\bullet}$ as a function of $\underline{x}_t$. Actually, we have two possibilities.

1) *On-line computation.* When the state $\underline{x}_t$ is reached at time t, cost (18) must be minimized at this instant. Problem 3 is then an open-loop optimal control problem and may be regarded as a nonlinear programming one. This problem can be solved on line by considering the vectors $\underline{u}_t, \ldots, \underline{u}_{t+N-1}, \underline{x}_{t+1}, \ldots, \underline{x}_{t+N}$ as independent variables. The main advantage of this approach is that many well-established nonlinear programming techniques are available to solve Problem 3. On the other hand, the approach involves a huge computational load for the regulator. If the dynamics of the controlled system is not sufficiently slow, as compared with the speed of the regulator's computing system, a practical application of the RH control mechanism turns out to be unfeasible.

2) *Off-line computation.* By following this approach, the regulator must be able to generate *instantaneously* $\underline{u}_t^{RH^\bullet}$ for *any* state $\underline{x}_t$ that may be reached at stage t. In practice, this implies that the control law $\gamma_{RH}^\circ(\underline{x}_t)$ has to be computed "a priori", i.e., off line, and stored in the regulator's memory. Clearly, the off-line computation has advantages and disadvantages that are opposite to the ones of the on-line approach. No on-line computational effort is requested by the regulator, but an excessive amount of computer memory may be required to store the closed-loop control law. Moreover, an N-stage functional optimization problem has to be solved instead of a nonlinear programming one. As is well known, such a functional optimization problem can be solved analytically only in very few cases, typically under LQ assumptions. As we are looking for feedback optimal control laws, dynamic programming seems to be the most efficient tool. This implies that the control function $\gamma_{RH}^\circ(\underline{x}_t)$ has to be computed when the backward phase of the dynamic programming procedure, starting from the final stage $t + N - 1$, has come back to the initial stage t. Unfortunately, dynamic programming exhibits computational drawbacks that, in general, are very difficult to overcome. In the next section, we shall return to the off-line solution of Problem 3 and present a neural approximation method to solve this problem.

5. REDUCTION OF PROBLEM 3 TO A NONLINEAR PROGRAMMING PROBLEM

To avoid the computational drawbacks (described in the previous section) resulting from the use of dynamic programming, we shall move toward an approximating technique that consists in assigning the control laws a given structure in which the values of a certain number of parameters have to be determined in order to minimize the cost function. Because of time-invariance of the dynamic system (7) and of the cost function (18), the closed-loop control functions $\underline{u}_i^{FH} = \underline{\gamma}_{FH}(\underline{x}_i, i)$, $i = t, \ldots, t+N-1$, can be written as $\underline{u}_i^{FH} = \underline{\gamma}_{FH}(\underline{x}_i, i)$, $i = 0, \ldots, N-1$. Then, the approximating control law takes on the form

$$\underline{u}_i = \hat{\underline{\gamma}}_i(\underline{x}_i, \underline{w}_i), \quad i = 0, 1, \ldots, N-1 \tag{19}$$

where $\hat{\underline{\gamma}}_i$ is a known function of its arguments and $\underline{w}_0, \ldots, \underline{w}_{N-1}$ are vectors of parameters to be optimized. Of course, the choice of the function $\hat{\underline{\gamma}}_i$ is quite arbitrary and, in any case, the control law (19), after the optimization of the vectors $\underline{w}_i$, constitutes an approximation for the solution of Problem 2. Among various possible approximating functions, we choose nonlinear mappings based on multilayer feedforward neural networks. This choice is suggested by both computational reasons and the approximation properties that characterize such neural mappings, as will be discussed later on. Now, we want to remark that the method of approximating the control laws by means of a preassigned control structure is not new in control theory; actually, it was proposed in the '60s (see, for example, the so-called Specific Optimal Control considered by Eisenberg and Sage (1966) for the general nonlinear nonquadratic case). However, it does not seem that such a method met with great success, probably because the selected structures were characterized by too small a number of free parameters to attain satisfactory approximation properties.

Let us now describe in some detail the N neural networks that implement the control functions (19). We assume that each of the N neural networks is composed of L layers, and that, in the generic layer s, n_s neural units are active. The input/output mapping of the q-th neural unit of the s-th layer is given by

$$y_q^i(s) = g\left[z_q^i(s)\right], \; s = 1, \ldots, L; \; q = 1, \ldots, n_s \tag{20}$$

$$z_q^i(s) = \sum_{p=1}^{n_{s-1}} w_{pq}^i(s) y_p^i(s-1) + w_{0q}^i(s) \tag{21}$$

where $y_q^i(s)$ is the output variable of the neural unit, $g(x) = 1/(1+e^{-x})$ is a sigmoidal activation function, and $w_{pq}^i(s)$ and $w_{0q}^i(s)$ are the weight and bias coefficients, respectively. All these coefficients are the components of the vector $\underline{w}_i$ appearing in the control function $\underline{u}_i = \hat{\underline{\gamma}}_i(\underline{x}_i, \underline{w}_i)$; the variables $y_q^i(0)$ are the components of $\underline{x}_i$ and the variables $y_q^i(L)$ are the components of $\underline{u}_i$. One can deduce that the control scheme gives rise to a chain of N neural networks, each followed by the dynamic system. This chain is related to the control scheme proposed in (Nguyen and Widrow, 1990). Our structure differs in the fact that the neural networks are allowed to be time-dependent. If we now substitute (19) into (7) and (17), and use repeatedly the state equation, thus eliminating the control and state vectors, the cost function takes on the form $J(\underline{w}, \underline{x}_0)$, where $\underline{w} \triangleq \mathrm{col}(\underline{w}_i, i = 0, 1, \ldots, N-1)$. Since the cost function $J(\underline{w}, \underline{x}_0)$ depends both on the vector $\underline{w}$ to be determined and on $\underline{x}_0$, we eliminate the dependence on $\underline{x}_0$ by assuming $\underline{x}_0$ to be a random vector uniformly distributed on $\mathcal{A}_0$, and by averaging $J(\underline{w}, \underline{x}_0)$ with respect to this vector. It is worth noting that, though this procedure contains a certain degree of arbitrariness, it is not unusual. For example, it has been applied to solve parametric LQ optimal control problems (see for instance, Kleinman and Athans (1968)). To sum up, we have to solve the following

Problem 4. Find the vector $\underline{w}^\circ$ that minimizes the expected cost $\underset{\underline{x}_0}{\mathrm{E}}\left[J(\underline{w}, \underline{x}_0)\right]$.

□

It follows that the functional Problem 2 has been reduced to an unconstrained nonlinear programming problem. Of course the RH neural optimal control law takes on the form $\hat{\underline{u}}_t^{RH^\circ} = \hat{\underline{\gamma}}_{RH}^\circ(\underline{x}_t, \underline{w}_0^\circ) \triangleq \hat{\underline{\gamma}}_{FH}^\circ(\underline{x}_t, \underline{w}_0^\circ)$, $\forall \underline{x}_t \in \mathcal{A}_0$, $t \geq 0$.

6. APPROXIMATING PROPERTIES OF THE NEURAL CONTROL LAW

In this section, we address two basic points. The first is the degree of accuracy that can be attained by using neural regulators. The second concerns the complexity that the neural networks implementing the control functions must exhibit in order to achieve a given degree of accuracy.

Suppose now that the approximating control functions $\hat{\underline{\gamma}}_i$ contain only one hidden layer (i.e., $L = 2$) composed of ν_i neural units, and that the output layer is composed of linear activation units. Denote such control functions by $\hat{\underline{\gamma}}_i^{(\nu_i)}(\underline{x}_i, \underline{w}_i)$. Let us also introduce some useful definitions. Given the maps $g_i \triangleq \underline{f}\left[\underline{x}_i, \underline{\gamma}_i^\circ(\underline{x}_i)\right] : \Re^n \to \Re^n$, we define the following family of sets: $\mathcal{B}_0 \triangleq \mathcal{A}_0; \mathcal{B}_i \triangleq g_{i-1}[\mathcal{B}_{i-1}]$, for $i = 1, 2, \ldots, N-1$. Now we can state the following proposition (Parisini and Zoppoli, 1994a)

Proposition 1. Assume that Problem 2 has only one solution $\underline{\gamma}_i^\circ(\underline{x}_i) \in C[\mathcal{B}_i, \Re^m]$, $i = 0, 1, \ldots, N-1$. Then, for every $\varepsilon \in \Re, \varepsilon > 0$ and every i with $0 \le i \le N-1$, there exist an integer ν_i and a weight vector $\underline{w}_i$ (i.e., a neural control function $\hat{\underline{\gamma}}_i^{(\nu_i)}(\underline{x}_i, \underline{w}_i)$) such that

$$\left\| \underline{\gamma}_i^\circ(\underline{x}_i) - \hat{\underline{\gamma}}_i^{(\nu_i)}(\underline{x}_i, \underline{w}_i) \right\| < \varepsilon, \quad \forall \underline{x}_i \in \mathcal{B}_i \tag{22}$$

□

Proposition 1 has been derived directly from the results reported in the literature on neural approximators (see, for instance, Hornik *et al.*, 1989). It is important to note that the results presented in Proposition 1 do not necessarily involve the need for using a feedforward neural network to approximate the optimal control function. Actually, results of this type are very common in approximation theory, and hold true even under rather weak assumptions about the functions to be approximated. Moreover, such results do not provide any information on the rate at which the approximation error decreases, as the number of parameters of the approximating structure increases.

To address this very important issue, we now apply Barron's results on neural approximation (Barron, 1991). To this end, let us introduce an approximating network that differs slightly from the one previously described to state Proposition 1. Each network is the parallel of m single-output neural networks of the type described above. Denote by $\hat{\gamma}_{ij}^{(\nu_{ij})}(\underline{x}_i, \underline{w}_{ij})$ the input-output mapping of such networks, where ν_{ij} is the number of neural units in the hidden layer and $\underline{w}_{ij}$ is the weight vector. Define also as $\gamma_{ij}^\circ(\underline{x}_i)$ the j-th component of the vector function $\underline{\gamma}_i^\circ$. We now need to introduce some smoothness assumptions on the optimal control functions γ_{ij}° to be approximated. Following (Barron, 1991), we define the class of functions

$$G_{c_{ij}}^i \triangleq \left\{ \overline{\gamma}_{ij} \quad \text{such that} \quad \int_{\Re^n} |\underline{\omega}| \, |\Gamma_{ij}(\underline{\omega})| \, \mathrm{d}\underline{\omega} \le c_{ij} \right\} \tag{23}$$

where $\Gamma_{ij}(\underline{\omega})$ is the Fourier transform of $\overline{\gamma}_{ij}$ and $\overline{\gamma}_{ij}(\underline{x}_i)$ is a smooth "extension" of the function $\gamma_{ij}(\underline{x}_i)$ from the set $\mathcal{B}_i$, where it coincides with $\gamma_{ij}(\underline{x}_i)$, to the space $\Re^n$. c_{ij} is any finite positive constant. Then, we state the following proposition (proved in (Parisini and Zoppoli, 1994a))

Proposition 2. Assume that Problem 2 has only one solution $\underline{\gamma}_i^\circ(\underline{x}_i) : \mathcal{B}_i \longrightarrow \Re^m$, $i = 0, 1, \ldots, N-1$, such that $\overline{\gamma}_{ij}^\circ \in G_{\tilde{c}_{ij}}^i$ for some finite positive scalar $\tilde{c}_{ij}$, for every j with $1 \le j \le m$. Then, for every i with $0 \le i \le N-1$, for every j with $1 \le j \le m$, for every probability measure σ, and for every $\nu_{ij} \ge 1$, there exist a weight vector $\underline{w}_{ij}$ (i.e., a neural strategy $\hat{\gamma}_{ij}^{(\nu_{ij})}(\underline{x}_i, \underline{w}_{ij})$) and a positive scalar c'_{ij} such that

$$\int_{\mathcal{B}_i} \left| \gamma_{ij}^\circ(\underline{x}_i) - \hat{\gamma}_{ij}^{(\nu_{ij})}(\underline{x}_i, \underline{w}_{ij}) \right|^2 \sigma\left[\mathrm{d}(\underline{x}_i)\right] \le \frac{c'_{ij}}{\nu_{ij}} \tag{24}$$

where $c'_{ij} = (2 r_i \tilde{c}_{ij})^2$. r_i is the radius of the smallest sphere (centered in the origin) that contains $\mathcal{B}_i$.

□

Proposition 2 specifies that, for any control function $\gamma_{ij}^{\circ}(\underline{x}_i)$, the number of parameters required to achieve an integrated square error of order $O(1/\nu_{ij})$ is $O(\nu_{ij} n)$, which grows linearly with the dimension n of $\underline{x}_i$. We now compare such a property with the ones of traditional linear approximation schemes (like polynomial and trigonometric expansions) as well as of other classes of nonlinear approximators.

Let us first address the case of linear approximators, that is, linear combinations of a number ν_{ij} of preassigned basis functions. In (Barron, 1991), it is shown "that there is no choice of ν_{ij} fixed basis functions such that linear combinations of them achieve integrated square error of smaller order than $(1/\nu_{ij})^{2/n}$". This applies to functions to be approximated that belong to the above-defined class $G^i_{c_{ij}}$. The presence of $2/n$ instead of 1 in the exponent of $1/\nu_{ij}$ may then give rise to the phenomenon of the "curse of dimensionality". However, if we assume a higher degree of smoothness for the functions γ_{ij}° by requiring them to have square-integrable partial derivatives of order up to s (then γ_{ij}° belong to the Sobolev space $W_2^{(s)}$), where s is the least integer greater than $1 + \frac{n}{2}$, two results can be established: 1) there exists a scalar c_{ij}^* such that $G^i_{c^*_{ij}} \supset W_2^{(s)}$ (i.e., $W_2^{(s)}$ is a proper subset of $G^i_{c^*_{ij}}$) (Barron, 1991), and 2) the linear schemes used to approximate functions belonging to Sobolev spaces do not suffer the "curse of dimensionality". It follows that neural approximators should behave better than linear ones in the difference set $G^i_{c^*_{ij}} \setminus W_2^{(s)}$.

As to the nonlinear approximators, similar approximation properties hold for Radial Basis Functions (Girosi and Anzellotti, 1992) (for which the centers and the weighting matrices of the radial activation functions can be tuned), or for linear combinations of trigonometric basis functions (for which the frequencies are adaptable parameters). In general, it is important that free parameters should not appear linearly, as is the case with the coefficients of linear combinations of fixed basis functions. It is also worth noting that the approximation bound of order $O(1/\nu_{ij})$ is achieved under smoothness assumptions that depend on the chosen nonlinear approximation schemes. The wider diffusion of neural approximators is probably to be ascribed to the simplicity of the tuning algorithms, to their robustness, and to other practical aspects.

7. SOLUTION OF THE NONLINEAR PROGRAMMING PROBLEM BY THE GRADIENT METHOD

The unconstrained nonlinear programming Problem 4 can be solved by means of some descent algorithm. We focus our attention on methods of the gradient type, as, when applied to neural networks, they are simple and well suited to distributed computation. However, due to the general statement of the problem, we are unable to express the gradient of the average cost $\underset{\underline{x}_0}{\mathrm{E}}[J(\underline{w}, \underline{x}_0)]$ in explicit form. This leads us to compute the "*realization*" $\nabla_{\underline{w}} J[\underline{w}(k), \underline{x}_0(k)]$ instead of $\nabla_{\underline{w}} \underset{\underline{x}_0}{\mathrm{E}} J[\underline{w}(k), \underline{x}_0]$. k denotes the iteration step of the descent procedure. The sequence $\{[\underline{x}_0(k)], k = 0, 1, \ldots\}$ is generated by randomly selecting the vectors $\underline{x}_0(k)$ from $\mathcal{A}_0$. Then, the updating algorithm is the following

$$\underline{w}(k+1) = \underline{w}(k) - \alpha(k) \nabla_{\underline{w}} J[\underline{w}(k), \underline{x}_0(k)], \; k = 0, 1, \ldots \tag{25}$$

The probabilistic algorithm (25) is related to the concept of "*stochastic approximation*" (see, for instance, Polyak and Tsypkin, 1973, for a description of the algorithm as well as for its convergence properties). The step-size $\alpha(k)$ must suitably decrease to meet one of such convergence requirements. In the examples given in the following, we take $\alpha(k) = c_1/(c_2 + k)$, $c_1, c_2 > 0$, and add a "momentum" $\eta[\underline{w}(k) - \underline{w}(k-1)]$ to (25), as is usually done in training neural networks.

The components of $\nabla_{\underline{w}} J[\underline{w}(k), \underline{x}_0(k)]$ can be derived by using the classical backpropagation procedure through each neural network implementing the control functions $\hat{\gamma}_i$. To initialize such a procedure, we have to compute the partial derivatives $\frac{\partial J}{\partial y_q^i(L)}$ (to simplify the notation, we drop the index k), which, as shown in (Zoppoli and Parisini, 1992; Parisini and Zoppoli 1994a), can be easily determined

once the vectors $\underline{\lambda}_i \triangleq \nabla_{\underline{x}_i} J\ (\underline{w}, \underline{x}_0)$, $i = 0, 1, \ldots, N-1$, have been determined. Such vectors can be computed as follows

$$\underline{\lambda}_i^T = \frac{\partial}{\partial \underline{x}_i} h(\underline{x}_i, \underline{u}_i) + \underline{\lambda}_{i+1}^T \frac{\partial}{\partial \underline{x}_i} \underline{f}(\underline{x}_i, \underline{u}_i) + \frac{\partial J}{\partial \underline{y}^i(0)} \tag{26}$$

for $i = 1, \ldots, N-1$, with $\underline{\lambda}_N^T = \frac{\partial}{\partial \underline{x}_N} h_N(\underline{x}_N)$. It is worth noting that (26) are the classical adjoint equations of N-stage optimal control theory, with the addition of a term (the third) to take into account the introduction of neural networks. The expressions of $\frac{\partial}{\partial \underline{x}_i} h(\underline{x}_i, \underline{u}_i)$, $\frac{\partial}{\partial \underline{u}_i} h(\underline{x}_i, \underline{u}_i)$, and of the matrices $\frac{\partial}{\partial \underline{u}_i} \underline{f}(\underline{x}_i, \underline{u}_i)$, $\frac{\partial}{\partial \underline{x}_i} \underline{f}(\underline{x}_i, \underline{u}_i)$, are detailed in (Di Febbraro *et al.*, 1994b). It can be seen that the algorithm consists of the following two "passes":

Forward pass . The state vector $\underline{x}_0\,(k)$ is randomly generated from $\mathcal{A}_0$. Then, the control sequence and the state trajectory are computed on the basis of these vectors and of $\underline{w}\,(k)$.

Backward pass . All the variables $\delta_q^i(s)$ and $\underline{\lambda}_i$ are computed, and the gradient $\nabla_{\underline{w}} J\ [\underline{w}(k), \underline{x}_0\,(k)]$ is determined. Then, the new weight vector $\underline{w}(k+1)$ is generated by means of (25).

It is worth noting that, instead of using a chain of N neural networks, as proposed for solving Problem 3 in approximate way, the neural optimal regulator $\hat{\underline{\gamma}}_{RH}(\underline{x}_t, \underline{w}^\circ)$ can be determined by minimizing the approximation error

$$E(\underline{w}) \triangleq \int_{\mathcal{A}_0} \left\| \underline{\gamma}_{RH}^\circ(\underline{x}_t) - \hat{\underline{\gamma}}_{RH}(\underline{x}_t, \underline{w}) \right\|^2 \mathrm{d}\,\underline{x} \tag{27}$$

with respect to $\underline{w}$. Define now the function $D(\underline{w}, \underline{x}_t) \triangleq \left\| \underline{\gamma}_{RH}^\circ(\underline{x}_t) - \hat{\underline{\gamma}}_{RH}(\underline{x}_t, \underline{w}) \right\|^2$, and note that we are able to evaluate $\underline{\gamma}_{RH}^\circ(\underline{x}_t)$ only pointwise, that is, by solving Problem 4 for specific values of $\underline{x}_t$. It follows that we are unable to compute the gradient $\nabla_{\underline{w}} E(\underline{w})$ in explicit form. Then, we interpret $E\,(\underline{w})$ as the expected value of the function $D(\underline{w}, \underline{x}_t)$ by considering $\underline{x}_t$ as a random vector uniformly distributed on $\mathcal{A}_0$. This leads us to compute the "realization" $\nabla_{\underline{w}} D[\underline{w}, \underline{x}_t(k)]$, instead of the gradient $\nabla_{\underline{w}} E[\,\underline{w}\,(k)\,]$, and to use a stochastic approximation approach, as done in Section 6. Of course, thanks to the time-invariances of the dynamic system and of the cost function, we can determine off line $\underline{\gamma}_{RH}^\circ[\underline{x}_0(k)]$ (instead of $\underline{\gamma}_{RH}^\circ[\underline{x}_t(k)]$) and the realization $\nabla_{\underline{w}} D[\underline{w}, \underline{x}_0(k)]$. In the next section, minimization of (27) will be used. The technique based on a chain of N neural networks is described in (Di Febbraro *et al.*, 1994a)

8. SIMULATION RESULTS

In this section, an example will be presented to point out the effectiveness of the receding-horizon neural control law described in the previous sections.

The freeway stretch considered consists of 30 sections of 1km length. On-ramps and off-ramps are present in sections 3,9,15,21, and 27. There are five control variables for speed limitations and each of them acts on six subsequent sections. In particular, VMS are placed in sections 1,7,13,19, and 25. It follows that $\dim(\underline{x}_i) = 65$ and $\dim(\underline{u}_i) = 10$. The model describing the dynamic behaviour of the freeway traffic is characterized by the following parameters: $\alpha = 0.9, V_f = 123\,\mathrm{km/h}, l = 4, m = 1.4, \tau = 0.01\,\mathrm{h}, \nu = 21.6\,\mathrm{km^2/h}, \chi = 20\,\mathrm{veh/km}, \delta_{\mathrm{on}} = 0.1, \rho_{\mathrm{kmax}} = 200\mathrm{veh/km}$ and $v_{\mathrm{kmax}} = 200\,\mathrm{km/h}$ for $k = 1, \ldots, K$, $l_{\mathrm{kmax}} = 200\,\mathrm{veh}$, $k \in \mathcal{I}_r$, $r_{\mathrm{min}} = 0\,\mathrm{veh/h}, r_{\mathrm{max}} = 20000\,\mathrm{veh/h}$, $\gamma_k = 0.1$ for each off-ramp. The demand value is $1000\,\mathrm{veh/h}$ for each on-ramp and for each time interval. The FH cost function takes on the form (16), where the number of temporal stages is $N = 5$ and the sampling time interval is 15 seconds.

The control function $\underline{u}_i = \hat{\underline{\gamma}}_{FH}(\underline{x}_i, \underline{w})$ considered to derive the neural receding-horizon regulator (NRH regulator) has been implemented by means of a neural network containing one hidden layer of 100 units. The constants of the step-size $\alpha(k)$ are $\eta = 0.9$, $c_1 = 1$, and $c_2 = 10^5$. These parameters, along with the number of neural units, were established experimentally. The state vector $\underline{x}_i$ has been considered as a random vector uniformly distributed on the set $\mathcal{A}_0$ of the possible initial congestion situations. The learning algorithm converged after $5 \cdot 10^6$ to $8 \cdot 10^6$ iterations.

In order to evaluate the effectiveness of the proposed method, the same problem has been solved by means of a classical RH control algorithm, that is, a standard RH regulator (in the following referred to as the ORH regulator) has been found using an on-line computed control law by means of a gradient technique. In the following figures , a comparison between the behaviour of the system under the application of the ORH regulator and the application of the approximate NRH solution, is shown.

The traffic situation considered shows a severe congestion in section 11 and, correspondingly, low values of traffic density in the subsequent section. More specifically, Figs. 1 and 2 show the traffic density and traffic mean speed in the sections of the freeway considered under the application of the ORH regulator. The same variables obtained with the use of the NRH control law are reported in Figs. 3 and 4. In Fig. 5, the values of the traffic density in the sections considered after 15 minutes under the application of the ORH and the NRH regulators are compared. In Fig. 6, the same comparison is done for traffic mean speed. It can be noticed that the NRH control law turns out to be very effective since it yields performances almost identical to those obtained by using the ORH control law (there is only about a 0.5% increase in the RH cost on the 15 minutes time interval considered).

As further results, Figs. 7 and 8 report the behaviours of traffic density and traffic mean speed starting from a different initial state. This state represents a typical traffic situation in which the congestion affects several subsequent sections. In particular, high values of traffic density and, consequently, low values of traffic mean speed, are found in sections from 11 to 21. Simulation results are quite satisfactory, as the NRH control law drives the system to the same state values reached in the previously considered situation.

9. FURTHER DEVELOPMENTS

The problem dealt with in the paper constitutes a first attempt to control a very complex large-scale system. More realistic control schemes are currently under investigation. We want to mention two important cases: i) the assumption that all state variables are measured exactly is relaxed and neural state estimators are proposed, and ii) the assumption that a single control center receives all traffic data and sends the control variables to the traffic control devices is relaxed, too. In place of such a centralized scheme, a team organization is considered, in which the measurement/control stations (the "members" of the team) cooperate on the minimization of the common cost function by exploiting local measures and some messages received from the other team members. An example of such a decentralized control structure, based on neural approximations, is given in (Parisini and Zoppoli, 1993), where the dynamic routing of a traffic or communication network is addressed.

10. REFERENCES

Barron , A. R. (1991) Universal approximation bounds for superpositions of a sigmoidal function. Technical Report No. 58, Department of Statistics, University of Illinois, 1991, to appear in the *IEEE Trans. on Information Theory.*

Di Febbraro , A. , D. Di Noto, T. Parisini, V. Recagno, S. Sacone and R. Zoppoli (1994a) Neural networks for the minimization of the travelling time on freeway systems. To be presented at the *7th IFAC Symposium on Transportation Systems* , Tianjin, August 1994.

Di Febbraro , A., V. Recagno, and S. Sacone (1994b) Gradient techniques for the optimal control of freeway systems. DIST Tech. Report No. 94/2.

Eisenberg , B. R. and A. P. Sage (1966) Closed-loop optimization of fixed configuration systems. *International Journal of Control*, **3**, 183-194.

Girosi , F. and G. Anzellotti (1992) Convergence rates of approximations by translates. *A.I. Memo No. 1288*, M.I.T., Cambridge, MA.

Goldstein , N. B. and K. S. P. Kumar (1982) A decentralized control strategy for freeway regulation. *Transportation Research B*, **16**, 279-290.

Hornik , K., M. Stinchombe, and H. White (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359-366.

Isaksen , L., and H. J. Payne (1973) Freeway traffic surveillance and control. *Proc. IEEE*, **61**, 526-536.

Keerthi , S. S. and E. G. Gilbert (1988) Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems: stability and moving-horizon approximations. *Journal of Optimization Theory and Applications*, **57**, 265-293.

Kleinman , D. L. and M. Athans (1968) The design of suboptimal linear time-varying systems. *IEEE Trans. Automatic Control*, **13**, 150-159.

Klijnhout , J. J. (1991) Freeway Control: an Overview. *Concise Encyclopedia of Traffic and Transportation Systems*, M. Papageorgiou, (Ed.), Pergamon Press, 152-156.

Looze , D. P., P. K. Houpt, N. R. Sandell, and M. Athans (1978) On decentralized estimation and control with application to freeway ramp metering. *IEEE Trans. on Automatic Control*, **23**, 268-275.

Mayne , D. Q. and H. Michalska (1990) Receding horizon control of nonlinear systems. *IEEE Trans. on Automatic Control*, **35**, 814-824.

Michalska , H. and D. Q. Mayne (1993) Robust receding horizon control of constrained nonlinear systems. *IEEE Trans. on Automatic Control*, **38**, 1623-1633.

Nguyen , D. H. and B. Widrow (1990) Neural networks for self-learning control systems. *IEEE Control System Magazine*, **10**, 18-23.

Papageorgiou , M. (1983) *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control.* Lecture Notes in Control and Information Sciencies, Springer-Verlag, **50**.

Papageorgiou , M. and R. Mayr (1982) Optimal decomposition methods applied to motorway traffic control. *Int. J. Control*, Vol. 35, No. 2, 269-280.

Parisini , T. and Zoppoli R. (1994a) Neural networks for feedback feedforward nonlinear control systems. *IEEE Transactions on Neural Networks*, Vol. 5, No. 3.

Parisini , T. and Zoppoli R. (1994b) A receding-horizon regulator for nonlinear systems and a neural approximation. To be presented at the *American Control Conference*, Baltimora, June 1994.

Payne , H. J. (1971) Models of Freeway Traffic and Control. *Simulation Council Proc.*, **1**, 51-61.

Payne , H. J., W. A. Thompson, and L. Isaksen (1973) Design of a traffic-responsive control system for a Los Angeles freeway. *IEEE Trans. on Systems, Man, and Cybernetics*, **3**, 213-224.

Polyak , B. T. and Tsypkin Ya. Z. (1973) Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, **12**, 377-397.

Zoppoli , R. and T. Parisini (1992) Learning techniques and neural networks for the solution of N-stage nonlinear nonquadratic optimal control problems. *Systems, Models and Feedback: Theory and Applications*, A. Isidori and T. J. Tarn (Eds.), Birkhäuser, Boston, 193-210.

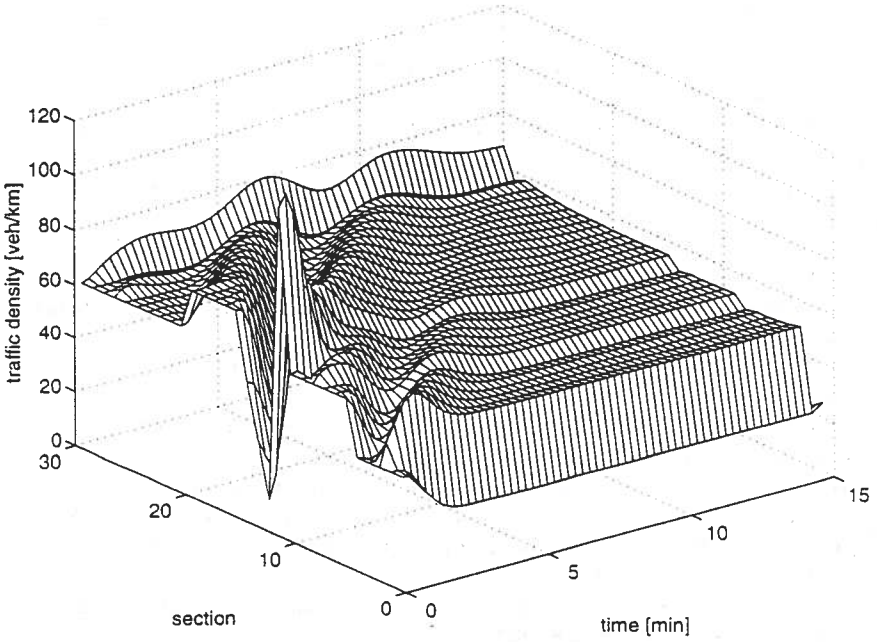


Fig. 1. Traffic density behaviour in clearing a severe congestion in section 11 under the application of the *on-line* computed RH regulator (ORH).

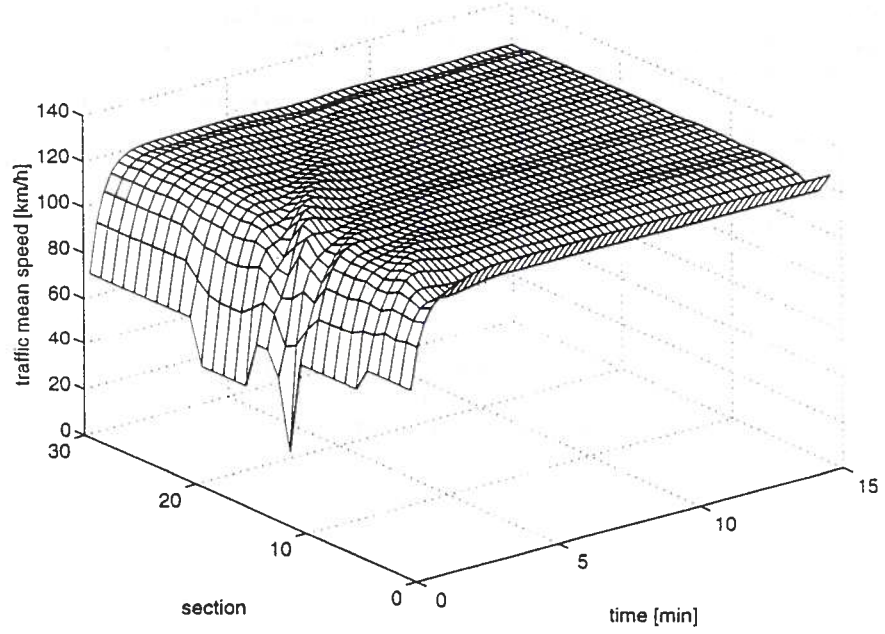


Fig. 2. Behaviour of the mean traffic speed for the example shown in Fig. 1.

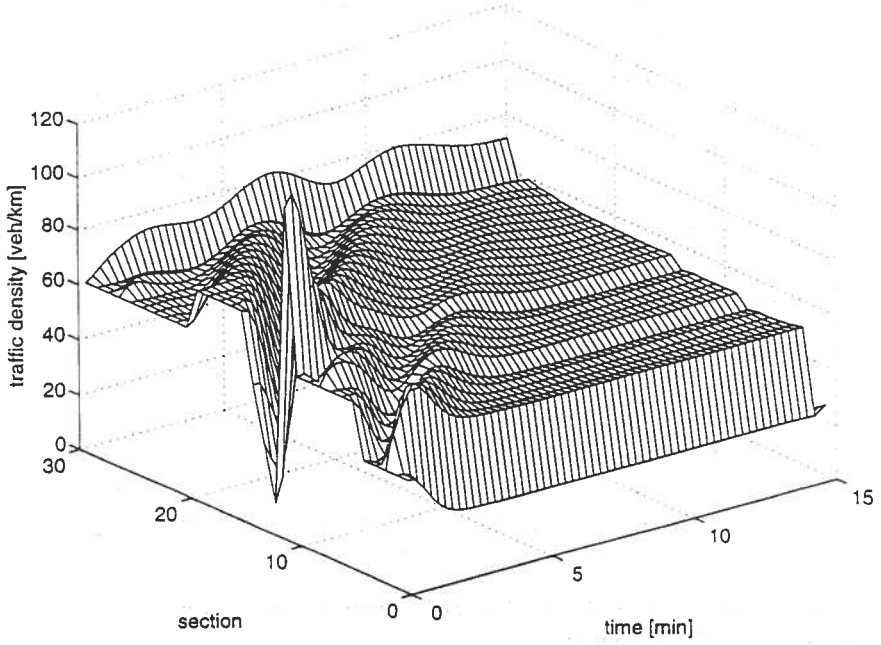


Fig. 3. Traffic density behaviour in clearing a severe congestion in section 11 under the application of the neural RH regulator (NRH).

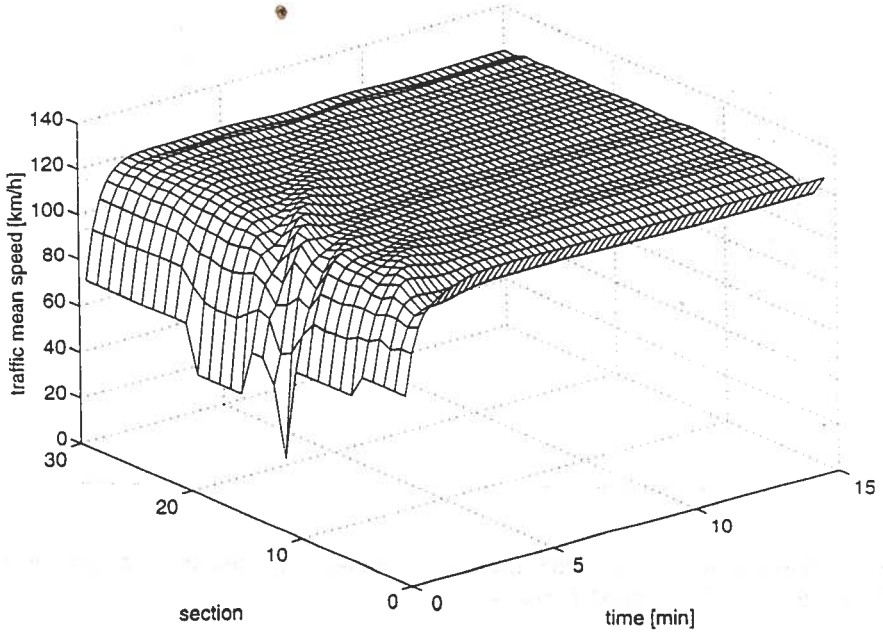


Fig. 4. Behaviour of the mean traffic speed for the example shown in Fig. 3.

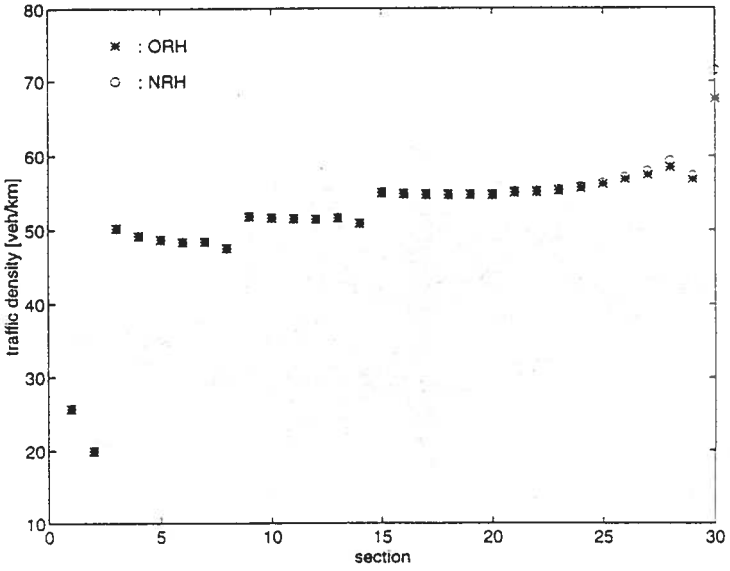


Fig. 5. Comparison of the traffic density values generated after 15 minutes by the ORH regulator and by the NRH regulator.

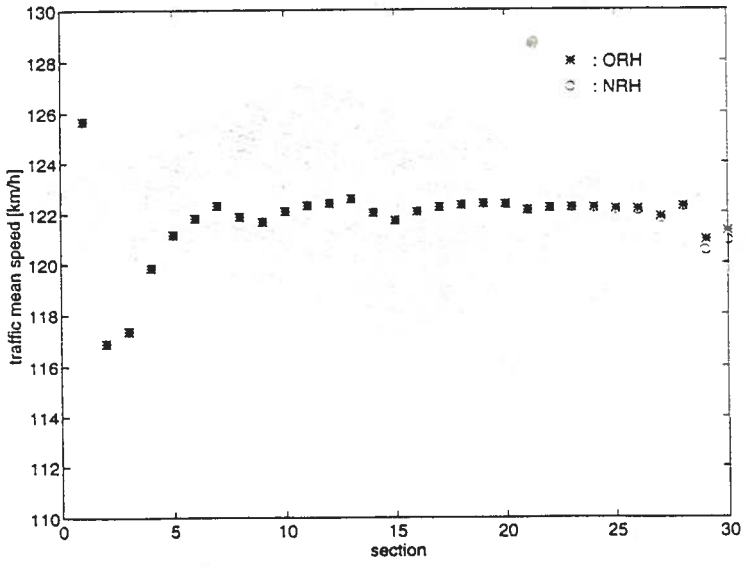


Fig. 6. Comparison of the traffic mean speed values generated after 15 minutes by the ORH regulator and by the NRH regulator.

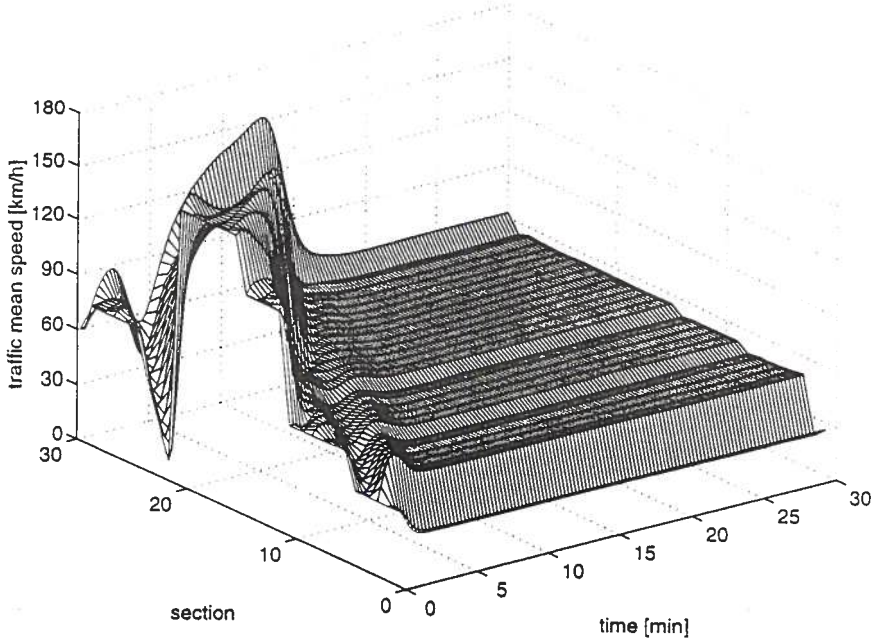


Fig. 7. Traffic density behaviour in clearing a congestion which lasts from section 11 to section 21.

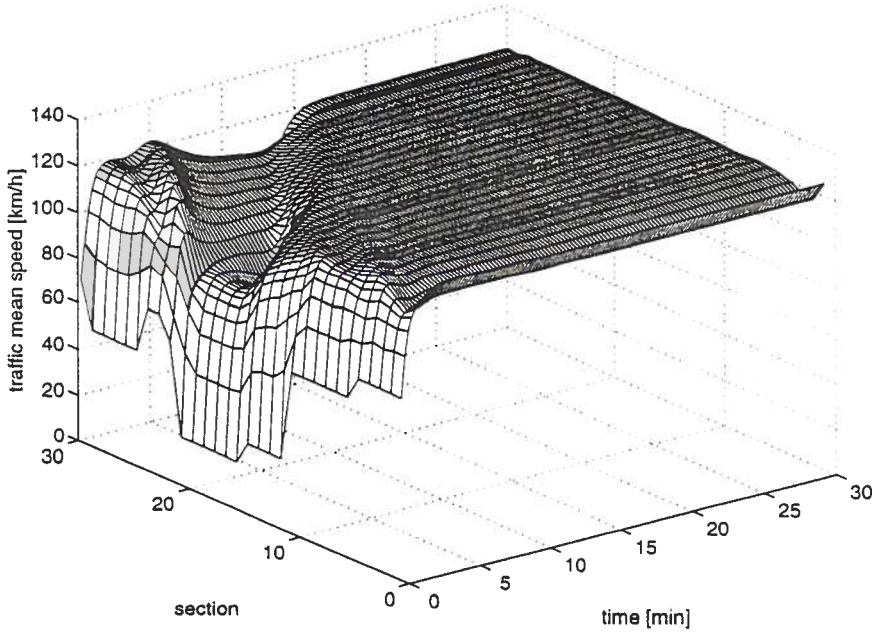


Fig. 8. Behaviour of the mean traffic speed for the example shown in Fig. 7.

An on-line dynamic traffic prediction model for an inter-urban motorway network

Ben-Akiva M., MIT, Hague Consulting Group, Netherlands
Cascetta E., University of Naples, Italy
Gunn H., Hague Consulting Group, Netherlands

Abstract: This paper specifies the overall structure of an on-line Dynamic Traffic Model System for an inter-urban motorway network. The model predicts traffic conditions on a motorway network in real time, and can be used to provide information on *expected(congestion to a motorway traffic control center.

The model receives traffic data such as traffic flow, average speed and average detector occupancy in real-time from a motorway monitoring system. Additionally, a database of historic traffic information will be used.

The model system contains several submodels: a statistical model for filtering noisy traffic data, and for very short term predictions *prediction horizon: 5-15 min.); a real-time O/D matrix prediction model to provide the input flows to the statistical and assignment models. Predictions from the two models are merged where their respective time horizons overlap. The present and predicted traffic conditions will be displayed for evaluation by the traffic operator. The real-time aspect of the system is reflected in the need for timing and reliable synchronisation of the various model components.

Acknowledgement

The work reported in this paper was undertaken as part of the DYNA project supported by the European Commission within the DRIVE II programme. The partners in the project are Hague Consulting Group, Centro Studi sui Sistemi di Trasporto, Elasis, the University of Lancaster, the University of Naples, the University Libre de Bruxelles and the Rijkswaterstaat.

The Impact of Predictive Information on Guidance Efficiency: An Analytical Approach

by

Moshe Ben-Akiva
Massachusetts Institute of Technology

André de Palma
University of Geneva

and

Isam Kaysi
American University of Beirut

Draft - May 1994

Abstract - The acceptance of route guidance advice by motorists is expected to occur only whenever such advice is experienced to be valid and reliable. Three major factors may cause route guidance systems to provide motorists with unreliable advice: (1) the traffic information constituting the basis for guidance advice is inaccurate; (2) the impact of a large fraction of motorists responding to the guidance, and the subsequent overreaction that occurs, is ignored whenever the guidance advice is being set; and, (3) concentration effects may occur and induce a higher level of (unpriced) congestion when drivers are provided with better information.

In this paper, the impact of the above factors is investigated for a small prototypical network. We provide numerical and analytical results in order to identify the critical factors involved in the design of driver information systems. The analysis presented in the paper involves (a) estimating the benefits of accurate predictions of traffic conditions; (b) assessing the sensitivity of guidance provision to inaccuracies and imperfections in the traffic predictions; and (c) comparing the behavior of uninformed drivers with that of drivers receiving myopic or predictive. Finally, conclusions and policy implications regarding the design features of advanced traveler information systems are presented.

1. Introduction

Advanced Traveler Information Systems (ATIS) based on modern information technology are expected to play an important role in reducing traffic congestion and improving traffic flows and safety. ATIS have the potential to reduce delays due to both incident and recurrent congestion by providing information to motorists with respect to alternative paths to their destinations and to actual traffic conditions on links of interest using a combination of roadside signals and onboard systems.

Potential Adverse Impacts of Guidance - In considering the potential benefits of alternative traveler information systems, it is also essential to evaluate the potential adverse effects that information may induce. Several behavioral phenomena may negate some of the beneficial effects of improved information, prime among them being overreaction and concentration [2].

ATIS Design Issues - The actual benefits realized from traveler information systems will depend heavily on the quality of the traffic information or guidance provided to drivers [3]. For instance, adverse impacts may not be avoided if the driver information system is not well-designed (see [5]). Moreover, inaccurate traffic information may result in poor guidance directives being provided to drivers. The likely impact would be a gradual loss of confidence in the system and its eventual abandonment.

Desirable Properties of Route Guidance - Guidance provided to drivers has to possess a number of *desirable properties* if it is to be useful for dynamic traffic management and if it is to be adopted by drivers [4]. To be useful for dynamic traffic management, guidance has to be able to *respond in real-time* to emerging traffic conditions and *anticipated* demand levels in such a way that it *reduces the overall extent of congestion*. In addition, since traveler information may prove to be counterproductive in some cases, guidance should be provided in such a way that it *reduces the potential adverse impacts of information*.

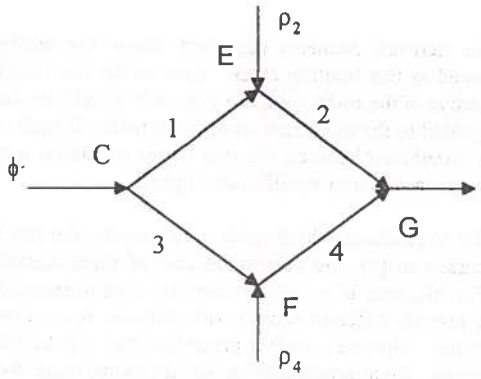
From the drivers' point of view, guidance should be *valid* in the sense that the travel time a driver experiences on the route he was guided to should be shorter than that on alternate routes. Moreover, guidance has to improve the *reliability* of travel times experienced by drivers by reducing variability and eliminating unusually long travel times. If drivers receive guidance which is not characterized by these desirable properties, they will not have confidence in the information system. As a result, drivers may not comply with the guidance being provided, in which case the dynamic traffic management potential of the information system would not materialize.

The objective of this paper is to conduct prototypical analyses of ATIS design elements using analytical approaches. A previous paper by the authors [6] provided a similar evaluation analysis which was, however, based on simulation tools.

2. The Sample Network

Network Structure - We now introduce the network that was used in the analytical study. The network was chosen to be simple but at the same time to include features that will enable an analysis of the factors involved in the design of traveler information systems. The network is presented in Figure 1 and consists of 4 links (labeled 1 to 4). Each of these four links has a potential bottleneck at its end.

Figure 1. Prototypical Network



O-D Flows - Associated with the network are 3 O-D flows all of which are destined to node G, as follows:

- $\Phi(t)$ rate of flow entering at node C at time t
- $\rho_2(t)$ rate of flow entering at node E at time t
- $\rho_4(t)$ rate of flow entering at node F at time t

Flow $\Phi(t)$ has two possible routes to the destination. Flows $\rho_2(t)$ and $\rho_4(t)$ have no route choice as they have to follow links 2 and 4, respectively. These latter flows represent "external demands" that are loaded onto the network and affect its traffic conditions but are outside the scope of the guidance system in the sense that they can not be controlled or diverted.

Guidance Provision - The network consists of two routes in parallel. Drivers receive route guidance at point C concerning which of the two routes they should follow to their destination, point G.

$\Phi(t)$ represents the incoming flow which is to be guided. $\Phi_1(t)$ and $\Phi_3(t)$ are the flows departing from C at time t and following routes A (links 1 and 2) and B (links 3 and 4), respectively ($\Phi(t) = \Phi_1(t) + \Phi_3(t)$).

If traffic to be guided constituted the only flow on the two routes (no side flows; i.e., $\rho_2(t)=\rho_4(t)=0$) then information on the current traffic conditions would be sufficient to compute future traffic conditions and to guide traffic properly, in the absence of accidents or any other factors that could vary route capacity. However, for the network adopted here the side flows need to be taken into consideration when providing guidance at C since future traffic conditions will depend on the future arrival of side flows in the network.

Network Performance - The link performance functions used to determine the travel times on links, or the time a vehicle spends on a specific link, assume that each link consists of a running section that has a constant free flow travel time and a bottleneck at the end. Without any loss of generality, the travel time on the running sections of the four links is set to zero. As such, we will consider delays at bottlenecks as the only measure of travel disutility faced by traffic. Moreover, each bottleneck is modeled as a deterministic queue which possesses a specific discharge capacity. The network has four bottlenecks in total (of capacities s_1, s_2, s_3, and s_4), two on each of the parallel routes. Queues in the bottlenecks build up or dissipate based on the level of the incoming flows relative to the bottleneck discharge capacity.

Network Topology - The adopted network represents an extreme case of network topology as it relates to the spatial frequency of guidance points. This network represents the case in which traffic that is guided

at point C cannot switch from the route selected at C. As such, *no possibilities* exist for adapting to evolving traffic conditions or for recovering from mistakes that were made when guiding vehicles at C.

3. Information Regimes

The Two Limiting Cases - Given the network elements described above, the analysis considers two information systems which can be viewed as two limiting cases. Each of the two information regimes is discussed next. In both cases, the objective of the route guidance system is to split the traffic at C between routes A and B such that drivers are guided to the minimum travel time route. If both routes A and B are used at time t, then traffic has to be *distributed* between the two routes in such a way that their travel times are equal. We refer to this condition as the user equilibrium objective.

Route Distributive Guidance - We refer to guidance which splits traffic among the two routes A and B as route distributive guidance. As discussed in [5], the implementation of route distributive guidance in reality may not be straightforward. For instance, if in-vehicle units are used to transmit guidance advice to drivers, it is technically possible to provide different drivers with different route directives at the same time to split them among alternate routes. However, such a procedure may not be politically or legally acceptable due to its inequity implications. An alternate scheme for achieving route distributive guidance is likely to be required. One possibility, which can also be implemented with variable message signs, is to change the guidance message or sign repetitively so that the time average of the route directives would correspond to the splitting fractions to be achieved for each route.

3.1 Instantaneous Information

The first information system uses *instantaneous* information to determine how many drivers should be diverted to routes A and B. Travel time on a route at time t is the sum of the (instantaneous) travel time on different links of that route as computed at time t and given the prevailing traffic conditions. The traffic control system aims here to achieve user equilibrium in the sense of equating the travel times on routes A and B as determined using instantaneous link travel times. This is, of course, an ex-ante user equilibrium objective since the travel time experienced by the drivers will be different in general from the instantaneous or *myopic* travel time .

Let $Q_i(t)$ be the queue length at time t at the entrance of bottleneck i, i=1...4. The travel time for a driver joining the queue before bottleneck i at time t is $Q_i(t)/s_i$. The instantaneous travel time for a driver arriving at C at time t and following route A is:

$$T^M_A(t) = Q_1(t)/s_1 + Q_2(t)/s_2 \qquad (1a)$$

and likewise for route B it is

$$T^M_B(t) = Q_3(t)/s_3 + Q_4(t)/s_4 \qquad (1b)$$

Equations (1a) and (1b) reflect the fact that, with the instantaneous information regime, the travel time on each link of the network is calculated at time t, and not at the time the link will be used, as should be the case.

The evolution of the queue length is provided by the following flow conservation relationships (assuming that all the bottlenecks are used at full capacity)

$$dQ_1(t)/dt = \Phi_1(t) - s_1 \qquad (2a)$$
$$dQ_2(t)/dt = s_1 + \rho_2(t) - s_2 \qquad (2b)$$
$$dQ_3(t)/dt = \Phi_3(t) - s_3 \qquad (2c)$$
$$dQ_4(t)/dt = s_3 + \rho_4(t) - s_4 \qquad (2d)$$

The user equilibrium objective based on instantaneous travel times is $T_A(t) = T_B(t)$. Differentiating this equation and using (1) and (2) one gets

$$\Phi_1(t) = \frac{s_1 s_3}{s_1 + s_3}\left[\frac{\dot{\Phi}(t)}{s_3} + \frac{s_3 + \rho_4(t)}{s_4} - \frac{s_1 + \rho_2(t)}{s_2}\right] \qquad (3)$$

All the comparative statistics are obvious and intuitively sound except for the effects of s_1 and s_3. In the absence of side flows ($\rho_2(t) = \rho_4(t) = 0$) the capacities s_1 and s_3 should not play any role in determining the relative split of traffic between the two parallel routes. The reason is that, considering route A, when the two bottlenecks 1 and 2 are in series with $s_1 > s_2$ (see equation (2b)), s_1 should not affect travel time on route A. In the case of equation (3), the splitting fraction depends on s_1 and s_3 even if the side flows are zero. This is a consequence of the fact that instantaneous rather than actual travel times have been used to achieve the user equilibrium objective. Comparative statistics as a function of s_1 and s_3 are thus ambiguous and the result of a mathematical artifact.

It is a matter of simple algebra to show that in general the travel time experienced by the drivers will not be the same on the two routes. The exact expressions of the travel times are complicated even for simple patterns of arrivals and are not discussed here.

3.2 Predictive Information

The second information system is based on *predictive information*. Here it is recognized that the time to clear bottleneck 2 on route A is not $Q_2(t)/s_2$ (the travel time computed at time t) but rather that Q_2 has to be computed at time $t+Q_1(t)/s_1$ which represents the arrival time at the second bottleneck of a car joining the queue of bottleneck 1 at time t. This compensates for the fact that a vehicle will not reach the second bottleneck instantaneously - at time t - but will do so only after all cars in front of it in bottleneck 1 have cleared this bottleneck. Therefore, in the case of perfect foresight, the travel times for a driver arriving at C at time t are computed using the correct expressions, as follows:

$$T^P{}_A(t) = Q_1(t)/s_1 + Q_2[t+Q_1(t)/s_1]/s_2 \qquad (4a)$$

$$T^P{}_B(t) = Q_3(t)/s_3 + Q_4[t+Q_3(t)/s_3]/s_4 \qquad (4b)$$

In practice, Q_1 and Q_3 will be based on the latest measurements of traffic conditions and Q_2 and Q_4 need to be predicted. The user equilibrium objective, based on foresight, is achieved by requiring that $T^P{}_A(t) = T^P{}_B(t)$. Again, differentiating this equation and using (2) and (4), one obtains:

$$\Phi_1(t) = \Phi(t)\ \frac{\Psi_2(t)^{-1}}{\Psi_2(t)^{-1} + \Psi_4(t)^{-1}} \qquad (5)$$

where

$$\Psi_2(t) = \frac{s_1 + \rho_2^F(t)}{s_1 s_2}; \quad \Psi_4(t) = \frac{s_3 + \rho_4^F(t)}{s_3 s_4}$$

$$\rho_2^F(t) = \rho_2\left[t + \frac{Q_1(t)}{s_1}\right]; \quad \rho_4^F(t) = \rho_4\left[t + \frac{Q_3(t)}{s_3}\right]$$

Note that the splitting fraction at time t depends on the observed queues on links 1 and 3 at time t and on the side flows at later times. Thus, the implementation of this route guidance objective requires measurements of the queues at 1 and 3 and predictions of the side flows. If traffic patterns are repetitive, information concerning future traffic conditions may be approximated from data on historical travel patterns.

In this case all the comparative statistics give intuitive results. It can also be checked that when the side flow on either (or both) route is null, only the capacity of the downstream bottleneck matters.

The splitting pattern defined by equation (5) guarantees that the travel time experienced by the drivers are the same at each instant on the two routes (although, of course, they generally change as a function of t).

3.3 Uninformed Drivers

The above two cases reflect the situation whereby drivers are split between the two routes based on either myopic or predictive information. The split is implemented by providing drivers with information regarding the preferred route to their destination. The information will either reflect instantaneous traffic conditions at the time drivers reach point C or traffic conditions predicted to materialize by the time drivers reach the downstream bottlenecks.

The travel times experienced based on the two information regimes also need to be compared with the case where some drivers are uninformed. In such a case, the behavior of drivers is assumed to reflect their response to the "average" traffic conditions they are likely to experience on the downstream bottlenecks. That is, such drivers formulate a "predictive" view of the traffic conditions on an average day and split themselves in such a manner that the travel time on the two routes will be equal if "average" traffic conditions materialize. Obviously, on any given day, traffic conditions differ from the average, and therefore uninformed drivers will be unable to adjust their behavior since they have no information regarding actual or future traffic conditions when they make their route choice at C. In general, their route choices will be suboptimal.

4. Evaluation of Myopic and Predictive Information Regimes

In this section we consider the situation whereby all travelers receive either myopic or predictive information and compare the two information regimes. Two basic cases will be addressed, namely, constant flows and day-to-day flow variability.

4.1 Constant Flows

4.1.1 Formulation

Here we discuss the case whereby all the arrival flows (upstream and side flows) are constant as a function of time. In such a case the splitting factors are constant over time and can be computed explicitly from equations (3) and (5). The actual travel times on the two routes can be computed using equations (4a) and (4b) and by integrating equations (2a)-(2d). It can be checked that when using the instantaneous information splitting factor, the actual travel times on the two routes are generally different and consequently that a fraction of users have ex-post been misguided. Alternatively, when using the perfect foresight splitting factors, the travel time on the two routes are the same and are given by:

$$T_A(t) = T_B(t) = \left(\Phi \frac{\Psi_2 \Psi_4}{\Psi_2 + \Psi_4} - 1 \right) t \qquad (6)$$

4.1.2 Equivalence Between Information Regimes

An obvious question concerns the conditions under which the two information regimes would provide the same splitting ratios and, consequently, the same guidance. In other words, it may be the case that traffic evolves in such a way that errors in the instantaneous case wash out and the two formulas (3) and (5) become equivalent. This can happen in two situations: (i) if the numerical values of the factors included in the two expressions just happen to produce the same splitting factors; and, (ii) for a more systematic case which occurs under the following conditions:

$$\frac{s_1 + \rho_2^F(t)}{s_1 s_2} = \frac{s_3 + \rho_4^F(t)}{s_3 s_4} \qquad (7a) \quad \textit{and}$$

$$\frac{s_1 + \rho_2(t)}{s_1 s_2} = \frac{s_3 + \rho_4(t)}{s_3 s_4} \qquad (7b)$$

For the current case of constant flows, the above two conditions simplify to

$$\frac{s_1 + \rho_2}{s_1 s_2} = \frac{s_3 + \rho_4}{s_3 s_4} \qquad (7c)$$

A sufficient condition for the above two equalities (i.e., (7a) and (7b)) to hold simultaneously is that the downstream queue length computed by the instantaneous case (at time t) and by the predictive case (taking into account the expected delays at the upstream bottlenecks) be the same, even though these may be time dependent. Another interpretation is to assume a "naive" prediction method such that the predicted value for $\rho_i^F(t)$ is $\rho_i(t)$, i=2, 4. In this case, although the instantaneous model does not provide correct estimates of travel times, the prediction errors on the two routes are such that the splitting ratios are unaffected. Note finally that when the downstream capacities are "infinite" (that is, practically, when they are not binding), the two information approaches are equivalent as expected.

4.1.3 Ranking Predictive and Myopic Information Regimes

For the network under consideration, and with constant flows, it is possible to make a definite statement concerning the benefits of predictive information compared to instantaneous information. For this purpose, we first compute the total travel time experienced by flow incoming to C at time t. The expression for this total travel time can be written as:

$$TT(t) = T_A(t).\Phi_1 + T_B(t).\Phi_3 \qquad (8)$$

TT is computed for the predictive case based on equations (4) and (5). For the myopic case, the splitting ratio is obtained from equation (3) while $T_A(t)$ and $T_B(t)$ to be used in equation (8) relate to the *actual* travel times experienced on the two routes, based on the designated splitting ration. Comparing the total travel time for flow incoming at time t for the two cases of predictive (P) and myopic (M) information, the following expression is obtained:

$$\frac{TT^M(t) - TT^P(t)}{t} = \frac{\left(\Phi_1^M - \Phi_1^P\right)^2}{\Psi_2 + \Psi_4} \qquad (9)$$

In other words, it can be concluded that $TT^M \geq TT^P$; that is, the total travel time experienced by flow incoming to C and split based on predictive information is at all times less than or equal to the value experienced if traffic is split according to myopic information. The equality holds for the condition stated in equation (7c). In addition, point equalities occur whenever $\Phi_1^M = \Phi_1^P$.

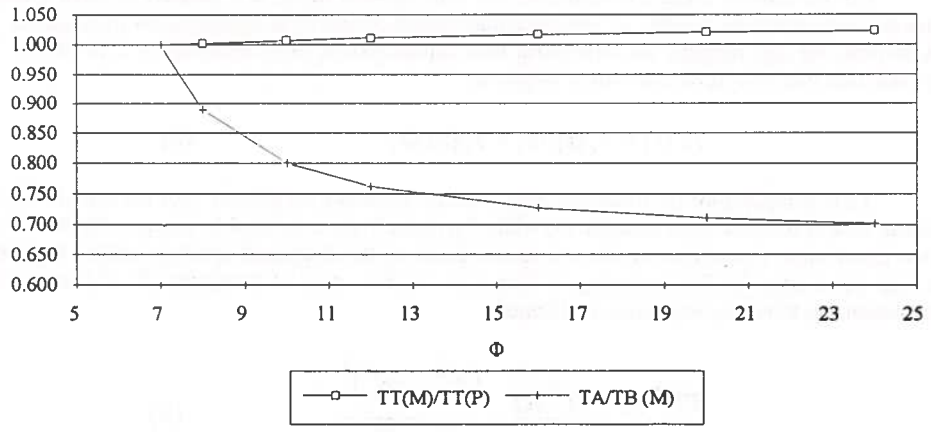
It should be noted, however, that the result presented above is not obvious since perfect information could leave traffic conditions worse off than with imperfect information. For instance, Arnott, de Palma, and Lindsey [1] concluded that instances exist whereby zero information can be more desirable from the social point of view than perfect information.

4.1.4 Sample Results

To illustrate some of the results for the network being considered, a simple numerical example will be considered. The following values of the input parameters will be utilized: $\Phi = 10$; $s_1 = 2$; $s_2 = s_3 = 4$; $s_4 = 6$; $\rho_2 = 2$; $\rho_4 = 5$. It should be noted that the basic assumption that all bottlenecks are used up to capacity at all times introduces restrictions on the feasible combinations of these input parameters. Although the results presented below are associated with specific input parameters, the conclusions are robust with respect to variations in the values of these parameters.

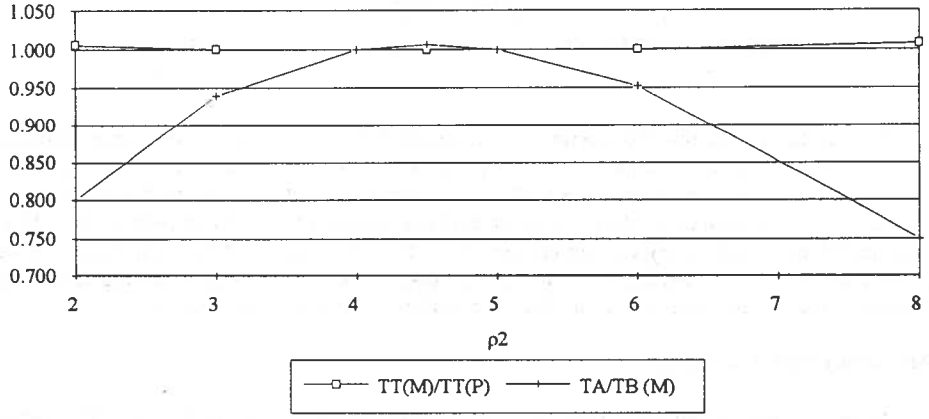
Level of Incoming Flows - Figure 2 indicates that the total travel time incurred with instantaneous information regarding traffic conditions is somewhat larger than that incurred by traffic which is split based on perfect foresight information. However, a more critical conclusion relates to the observation that the ratio of travel time on the two alternate routes when traffic is split according to instantaneous information becomes quite different from 1.00 as Φ increases. This indicates that confidence in such information, and therefore a high level of compliance, are unlikely to be sustained as drivers using the longer route are bound to find out that they are being misguided ex-post. In other words, while a slightly lower guidance efficiency is to be expected from using instantaneous information, the compliance rate is likely to be quite low. The figure also shows that the inefficiency introduced by the use of myopic information increases as the incoming flow at C becomes larger since the inaccuracy introduced by the use of such information increases.

Figure 2. Effect of Incoming Flow Level



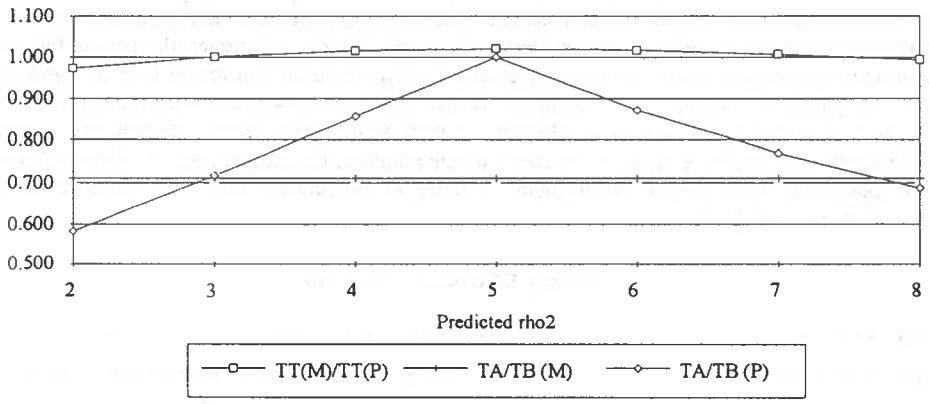
Level of Side Flows - Figure 3 illustrates the variation of TT^M/TT^P and T_A/T_B for the myopic case as the side flow ρ_2 varies between 2 and 8 units, representing the feasible interval for ρ_2. This figure indicates that the ratio of TT^M/TT^P changes from 1.006 (at ρ_2=2), to 1 (at ρ_2=4 and ρ_2=5), and to 1.008 (at ρ_2=8). As such, this ratio is always greater than or equal to 1, and predictive information is always more effective than instantaneous information, as guaranteed by equation (9). Moreover, the ratio of the travel times on the two alternate routes for the myopic case, T_A/T_B, varies from a low of 0.75 (at ρ_2=8) to a high of 1.01 (at ρ_2=4.5) while T_A=T_B occurs at the two values of ρ_2=4 and 5 since the expression for T_A=T_B is quadratic in ρ_2. These last results indicate that certain values of the side flows can generate wide discrepancies in the travel times on the two alternate routes for the myopic case. As discussed above, this can potentially have adverse implications in terms of compliance with and confidence in the information system being adopted.

Figure 3. Effect of Side Flow Levels



Prediction Accuracy - It is clear that perfect foresight constitutes an ideal situation given that it requires totally accurate predictions of the levels of the future side flows. Predictions of future side flows can be obtained in general but are unlikely to be perfect. Figure 4 illustrates how the ratio of the total travel times as well as the ratio of travel times on the two alternate routes for the predictive and myopic cases vary as the (imperfectly) predicted level of ρ_2 varies around the actual level which is equal to 5 units. It should be noted that the splitting ratio for the myopic case, as determined by equation (3), is based on the correct value of ρ_2 which is observed, and as such no inaccuracies are introduced in the evaluation of the ratio. The results presented below are associated with a value Φ=20.

Figure 4. Effect of Prediction Accuracy
(actual rho2=5)



The results indicate that the superiority of predictive information over instantaneous information is quite robust with respect to the quality of flow prediction. Even though the total travel cost for the predictive case TT^P increases as the quality of the flow prediction deteriorates, predictive information remains superior to instantaneous information even with a prediction error in ρ_2 as large as 40% (that is, with the imperfectly predicted ρ_2 varying between the values of 3 and 7). Moreover, even with quite inaccurate predictions of ρ_2, predictive information still does a much better job at achieving almost equal travel times on the two alternate routes, and thus user equilibrium is almost maintained.

4.2 Day-to-Day Flow Variability

In the previous section we considered a case whereby flow levels are constant over the course of the period of analysis. We now expand our analysis to consider situations in which flow levels remain constant over any day but vary from day to day. In specific, we consider three different types of days, each with a different combination of (constant) flow levels and a probability of occurring, as follows:

Day Type	Probability	Φ Level	ρ_2 Level	ρ_4 Level
1	0.25	Φ	$\theta\rho_2$	ρ_4
2 ("normal")	0.50	Φ	ρ_2	ρ_4
3	0.25	Φ	ρ_2	$\theta\rho_4$

The parameter θ reflects the extent to which side flows vary on days with heavy side traffic from levels experienced on "normal" days. In the above table, a "normal" day has a 0.5 probability of occurring while days whereby each of the side flows is heavy ($\theta \geq 1$) have a probability 0.25 each of occurring. Figure 5 presents average values of the total travel times (TT) over the different day types for drivers affected by each of three different information availability scenarios: predictive information, instantaneous information, and no information. The first two cases are familiar by now with the only difference being that splitting ratios are set based on traffic information that will reflect (in a myopic or predictive sense) the type of day being experienced. On the other hand, the splitting ratios for the uninformed drivers remain the same from day to day and are simply based on achieving equal travel times on the two routes if a normal day is experienced. That is, such drivers are assumed to adopt a behavior which responds to

normal traffic conditions and which is not expected to perform well if there is wide variability in traffic conditions.

Figure 5. Day-to-Day Flow Variability

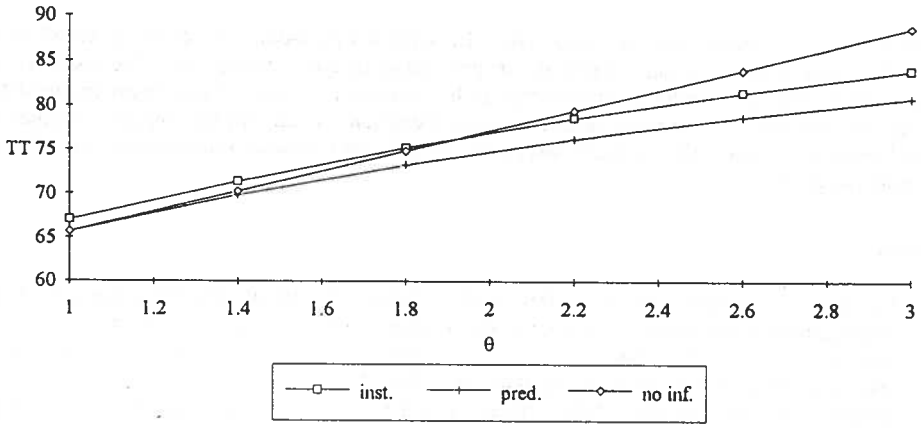


Figure 5 indicates that for the day types being considered, and for values of θ ranging from 1 to 2, total travel time for uninformed drivers is less than that experienced by drivers responding to myopic information. On the other hand, predictive information provides benefits, especially as θ increases, reflecting greater variability in traffic conditions. The implication here is that in case traffic conditions do not display large variability from day to day ($\theta < 1.8$) *and* uninformed drivers are able to respond to "normal" traffic conditions properly, then uninformed drivers are better off than drivers provided with guidance based on myopic information. Moreover, for values of $\theta < 1.2$, indicating very little variability in traffic conditions from day to day, predictive information offers only marginal improvement over the no information case.

5. Conclusions and Policy Implications

Previous Research - Previous research by the authors (see [6]) addressed ATIS design issues using simulation tools and concluded that one of the critical characteristics of successful guidance system designs consists of basing guidance on predictive traffic information which accounts for driver reaction. Simulation results indicated that guidance based on very good flow predictions provide minor reductions in average delay but significant improvements in guidance validity levels. These same conclusions were confirmed by the analytical derivations and numerical results presented in this paper, as described below.

Conclusions and Policy Implications - The analytical study described in this paper has demonstrated, for a small prototypical network, the value of predictive information as a basis of route guidance. Three main conclusions can be made, namely: (1) While total travel time is only reduced slightly by basing guidance on predictive rather than instantaneous information, the latter type of information results in large discrepancies in travel time between the two alternate routes; (2) The superiority of predictive information over instantaneous information is quite robust with respect to the quality of required flow prediction; and, (3) When day to day variations in traffic flows are not large, uninformed drivers who behave according to "normal" traffic conditions may be better off than drivers who receive guidance based on myopic or instantaneous information. Moreover, if the day to day variability is very limited, guidance based on predictive information is not called for as it offers only marginal improvements over the no information case.

Since the success of any ATIS implementation will clearly depend on the quality of the guidance it provides to drivers, guidance validity represents one of the major considerations at the ATIS design stage. The first conclusion described above indicates that a guidance system based on instantaneous information is not likely to sustain operation due to the large discrepancies in travel times between the two alternate routes which result in low guidance validity levels and, as a result, a loss of confidence in the system.

Further Research - Ongoing research is addressing the issues of within-day flow variability as well as the impact of market penetration and value of information using analytical techniques. The results of this research will be compared with conclusions reached by investigating similar issues based on simulation tools, as was undertaken in previous research. Other extensions which will be considered include the issues of endogenous flow arrival rates and the possible coordination between traffic control measures and route guidance systems.

References

[1] Arnott, R., A. de Palma, and R. Lindsey (1994). "Dynamic Model of Peak Period Congestion with Heterogeneous Commuters", *Journal of Transportation Economics and Policy*, forthcoming.

[2] Ben-Akiva, M., A. de Palma, and I. Kaysi (1991). "Dynamic Network Models and Driver Information Systems", *Transportation Research*, Vol. 25A, No. 5.

[3] Bonsall, P.W. and A.D. May (1989). "Evaluation of In-Vehicle Route Guidance Systems", paper presented at the 5th WCTR, Yokohama, Japan.

[4] Kaysi, I. (1991). "Framework and Models for the Provision of Real-Time Driver Information", Unpublished Ph.D. Dissertation, Department of Civil Engineering, MIT, November 1991.

[5] Kaysi, I., M. Ben-Akiva, and H. Koutsopoulos (1993). "An Integrated Approach to Vehicle Routing and Congestion Prediction for Real-Time Driver Guidance", *Transportation Research Record*, No. 1408.

[6] Kaysi, I., M. Ben-Akiva, and A. de Palma (1994)."Design Aspects of Advanced Traveler Information Systems", *Compendium of Papers*, 1992 Capri Seminar on Urban Traffic Networks.

APPLICATION OF CONTEXTUAL TECHNOLOGY FOR SUPPORTING DECISION MAKING IN URBAN TRANSPORTATION NETWORKS

L. BIANCO
I.A.S.I. - C.N.R. - Viale Manzoni, 30 - I-00185 Rome
Dept. of Electronic Eng. - Univ. of Rome "Tor Vergata"

P. DELL'OLMO
I.A.S.I. - C.N.R. - Viale Manzoni, 30 - I-00185 Rome

Irina V. EZHKOVA
Computing Center - Russian Academy of Sciences,
40 Vavilova Str.117333 - Moscow

Abstract

In this paper we examine the application of a new methodology to the analysis and the evaluation of transportation systems. In general, a transportation system can be described by a limited number of attributes, like mode and geographical area characteristics, efficiency, etc., including also, not easy to formalize, valuations of costs-benefits affecting etherogeneous categories of users. Data are collected objectively with reference to the same scale, but often it turns out to be more useful to have different views of data, in which the decision maker measures attributes values in relation to a given "scenario" (or "context"). In order to support complex analysis, by providing adequate information in understandable form, we present a new advanced paradigm based on a contextual technology for flexible decision making. Contextual systems are designed for formation, updating, and processing of knowledge with a view to decision making in fuzzy environments. The paper discuss several examples in which the basics tools of the system, like quantitative comparison of etherogeneous attributes, determination of proximity of each reference notion to an active context, are applied to the evaluation of the state of an urban network.

Key Words. Artificial intelligence, Contextual systems, Decision theory, Multicontextual decision making, Fuzzy sets theory, Road Traffic, Transportation

1. Introduction

Strategic planning and decision making in transportation systems regard a number of different aspects, like economy, social attitudes, environment, impact, efficiency, and costs. As shown by Bianco (1989), especially in urban transportation the decision maker has to evaluate the traffic network state having simultaneously in mind technical-economical issues, and preferences of several categories of people whose living, working, and travelling conditions can strongly depend on the network performances. While standard traffic engineering parameters (like flow, speed travel times, etc) are quantitatively measurable for every link of the network, social preferences express in a

qualitative manner the perception of the road state by a given class of users. Therefore, a correct description of the state of the traffic network requires a set of attributes, both quantitative and qualitative, whose values the decision maker has to compare with some reference measures. These latter are often obtained as average values on the network or from predetermined standards. In general, current tools for supporting evaluation of the state of transportation systems present several limitations as discussed, for instance in Takyi, (1993). Furthermore, the evaluation of a road can be different accordingly to the point of view under consideration. If the points of view are in contrast (as could be for drivers and resident, for instance) it will be necessary a correct and rigourous trade-off analysis which cannot be achieved without a methodological support. In fact, a decision based only on empirical evaluation of data, may previlege in a no correct way one point of view with respect to the other. With respect to this problem, as the relative importance of all attributes to take into account changes with the considered framework, it is necessary to introduce a measure of the quantitative dependency from the reference context. The methodology presented in this paper proposes several features, like the establishment of an evaluation paradigm in which the value of each attribute is expressed at the background of a fixed context, in order to overcome the above problems.

2. Traffic Network State Evaluation

Current monitoring systems, based on surveillance detectors, video cameras and communication networks, can provide a set of data describing the on-line traffic state on network roads (flows of vehicles, queue lengths, speeds, etc.). Statistic techniques and databases provide a platform for the decision maker for the analysis of the network conditions. Of course, high-level political decisions cannot ignore how the roads conditions are perceived by large categories of users, including drivers, residents, and shopkeepers. At strategic level, any decision must take into account also the degree of satisfaction and the evaluation of these users' categories. These evaluations can be obtained by means of questionnaries as proposed by Hopkinson *et al.* (1992).

Let us consider, for example, a traffic network database whose attributes are showed in Table 1 and in which each entry is a network link (road). The values of time, speed, flow, flow/capacity can be average values on a fixed interval, for example one hour, while the social preferences for each category are assumed to be: good, normal, unacceptable.

Table1 Database Attributes

time (secs)
speed (km/h)
flow (vehicles/h)
flow/capacity
Drivers judgment (good, normal, unacceptable)
Resident judgment (good, normal, unacceptable)
Shopkeepers judgment (good,normal,unacceptable)
Experts judgment (good, normal, unacceptable)

The decision maker may face the problem of planning interventions to improve the condition of heavily conjested areas. His experience suggests to start from "busy roads". In the decision maker's mind, the notion "busy road" is an high level concept representative of a category of roads suffering from a given type of conjestion. A road X in the database is stored as a t-ple of attributes values like, X =<40 secs, 30 Km/h, 600 v/h, 0.8, good, normal, normal, good>. On the other hand, from the his experience, the concept "busy road" is not so directely related to the absolute values of these attributes. It rather depends on the comparison of these values against some *surrounding conditions.*

For example, a speed of 30 km/h in a middle size Italian city might be considered "normal" if viewed in the scenario of roads "good for drivers", and "high" in the scenario of roads "good for residents". In this sense a single or even multiattribute analysis would not be satisfactory without a reference context. The decision maker should be able to define *concepts*, like "busy road", representative of specific roads conditions, and *contexts*, like "good for drivers", which can act as different points of view for the same problem area. In such a way, he could build its own framework for evaluating roads condition and formulate a correct problem statement for strategic difficult to formalize decision making problems.

3. Contextual Systems

Contextual Systems have been applied in different areas (see Ezhkova (1989), (1992), (1993), Bianco *et al.* (1994)).

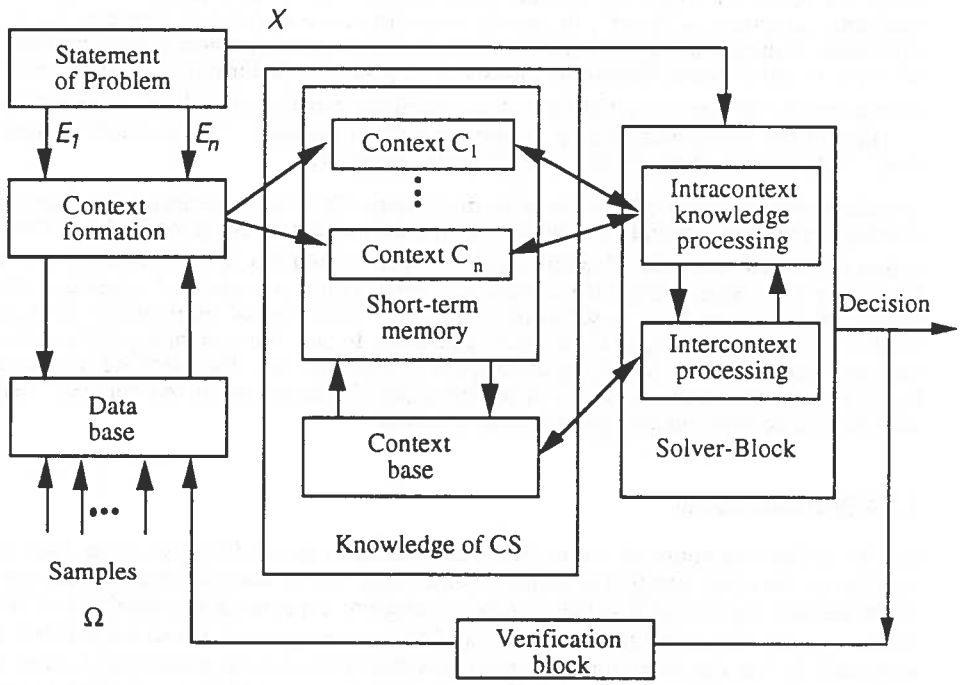


Fig. 1 Contextual system blocks

3.1 General view

The scheme of the major Contextual System blocks is shown in Fig. 1. The Contextual System has two types of memory: *long-term* and *short-term*. The former consists of a Primary Data Base and a Base of Contexts. The Primary Data Base stores a variety of samples originating from statistics, standards, experts, etc.. The Base of Contexts serves for storing contexts specific of some problem areas. A short-term (active) memory stores a variety of contexts handled by the Contextual System in solving a current problem. The

problem statement block contains a procedure for the generation of a set of decision alternatives E for the current local problem. In a partial case, a set of alternatives E is defined by a user. The key functional Contextual System blocks are: Context Formation Block and Solver. The latter comprises two blocks: a block for Intracontext Knowledge processing and a block for Intercontext Knowledge processing. The basic principles of blocks functioning are considered below. Note also, that Contextual System has a Verification Block. This verification is exercised by the expert. This kind of feedback permits Contextual System to adapt its experience and behaviour.

3.2 Context Formation Block

Context Formation Block is responsible for forming a description of a context. For this purpose more than one method can be applied (see Ezhkova (1992)). In the analysis of complex transportation systems it turns out to be effective to define a suitable context as a result of a *selection* on the Primary Database. As an example, the decision maker can select the roads for which the Drivers judgement is "good" and those for which the Residents judgment is "good". In general more articulated selection formulae can be expressed. In particular, following the traditional statement of decision making problem, the input to the Context Formation Block can be given in the form of a set of decision alternatives E, viewed as a set of basic concepts of the considered Problem Area $E \in \{E_1, \ldots, E_n\}$. In the transportation area, E may be a set of concepts {"busy road", "normal road", "free road"}, or any other particular set of alternatives.

The Context Formation Block draws from Primary Database a variety of samples E relating to the selection and then, while forming the context of the given Problem Area, makes use of this subset Ω^E of samples only. They are stored in a short-term memory as a Secondary Data Base. Behind the method is a verbalization procedure of mapping fuzzy values of attributes from a semantic space S to their verbal equivalents such as "high","very heavy", "frequent", "normal" and so on. In fact, it is a mapping of the initial semantic space to the so called cognitive space of a current task. The mapping is defined by an experience function, which is the principal characteristic of the context. Next sections will describe main steps of context formation.

3.3 Experience function

Let D_k^E be the distribution of values of the attribute p_k on the set Ω^E. In so doing, take the samples for the entire group of concepts E considering it as an integral whole. An analysis of the derived function D_k^E suggests some conclusions concerning the suitability of the attribute p_k for describing the given context. Most advantageous is the case when D_k^E is unimodal, i.e. has one maximum. The latter sets the standard as the most typical value of p_k for the given group of concepts E, for the given context. Only when there is one *maximum* D_k^E and, respectively, one *norm* with respect to which a verbal estimate of the type "high", "fast", etc. could be made, it is possible to be sure that all prerequisites for verbal estimate application are met. Otherwise, the attribute p_k must be broken down into a variety of attributes $p^1_k, \ldots, p_k^{mk}$, to which function D_k^E is unimodal. As it can be seen from experience, this approach is well consistent with the experts' ideas about a given attribute.

As an example, let consider the values of flow in the central area of the city showed in Table 2.

Table 2 Distribution of flows in the central area of the network in the pick hour

Flow (v/h)	Number of links
20	7
40	55
60	17
80	6
100	15
200	14
300	18
400	21
500	24
600	60
700	45
800	32

As a matter of fact, the two distribution maximum correspond to different cases: flow for *local* traffic and flow for *major* traffic. This results into two attributes instead of one: *local flow* and *major flow.*

The approaches similar to the aforementioned one make it possible to refine the semantic space of attributes S for the specific set of concepts E. In the new revised space S^* the given context is characterized by unimodal distribution of attributes values. In Ezhkova (1989) it is shown that the attribute measurement with a verbal estimate of the type "frequent", "high", "fast", etc., requires an experience function of the subject with respect to the frequency of occurence of attributes values. The same function can be used for mapping a concrete value from a particular semantic scale, measuring the attributes in meters, vehicles/hour, seconds, etc. into a so called *cognitive scale*. On the cognitive scale, the verbal estimates may be formalized as fuzzy subsets on interval [0,1] independent of the subject and semantics of the estimated attribute.

Here, it is the Contextual System that acts as a subject generating estimates. Its experience in assesing p_k within the framework of context C is determined by the experience function $F_k{}^E$ derived from the distribution $D_k{}^E$ as a result of the following transformation:

$$(1) \qquad F_k^E = \begin{cases} \dfrac{D_k^E(x)}{2\max\limits_x D_k^E(x)}, & x < p_{norm} \\ 1 - \dfrac{D_k^E(x)}{2\max\limits_x D_k^E(x)}, & x \geq p_{norm} \end{cases}$$

where p_{norm} is a "normal road" attribute value (norm) for which $D_k{}^E$ is maximum.

For example let the distribution of the "speed" of roads in a *good for drivers* context and a *good for residents* context be like those shown in Figure 2.

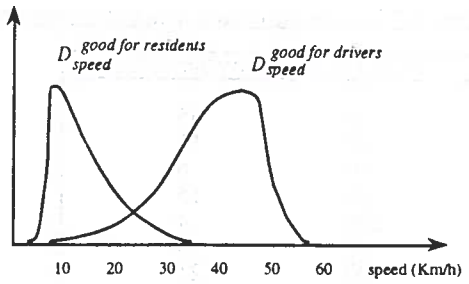


Fig. 2 Distribution of values for "speed" in a *good for drivers* context and in a *good for residents* context

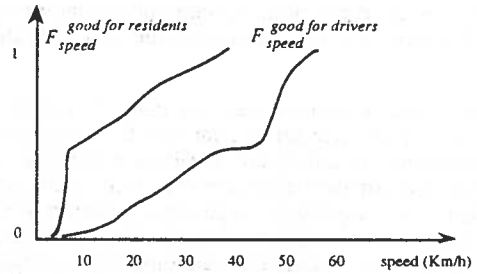


Fig. 3 Experience functions relative to "speed" in a *good for drivers* context and in a *good for residents* context

The function built on the basis of the context makes it possible to transfer the attribute value from the semantic scale to verbal estimates on cognitive scale. The vector of experience functions $F^E = <F_1^E, \ldots F_n^E>$ built for the attributes of the updated space S^*, will be referred to as an experience vector. The latter assigns the transference (mapping) of the semantic space S^* onto the cognitive space U. The latter's scales are for measuring the attributes with verbal estimates. The transfer is context dependent, hence the experience vector shapes the context.

3.4 Verbalization Procedure

Given the semantic attribute values, the verbalization procedure makes it possible to produce verbal estimates with regard to the Problem Area context. This procedure is applicable for example to the following problem.

Let X be a value of attribute p on the semantic scale. Consider a general case were X is a fuzzy set, i.e. is defined with a membership function $\mu(x)$, $x \in \Delta_p$ where Δ_p is a set of possible values for p measured on the specified semantic scale. In a partial case, X may be a fuzzy set of specific simplified form, a fuzzy number, a simple number, or any value of the ordinal scale.Let F be an experience function with respect to a certain attribute in the given context; let V be a set of verbal estimates used for measuring p; each verbal estimate from V is represented in the form of fuzzy subset built as a result of psycological experiments or analyses (special interval form, triangular form, trapezoidal form, etc.).

Verbalization procedure boils down to the following. Map the fuzzy set X with the experience function F onto cognitive scale. Let R be a fuzzy subset from the interval [0,1] resulting from this mapping: the attribute x value is converted from semantic scale to the value $y=F(x)$ in the interval [0,1] of cognitive scale, and is differentiated by $\mu(x)$. Denote the membership function of the fuzzy subset R in terms of $r(y)$, $y \in [0,1]$. Select the verbal estimate L, from the set V, such that the associated fuzzy subset describes in the best possible way the fuzzy subset R.

For example, let $F_{speed}^{good\ for\ drivers}$ and $F_{speed}^{good\ for\ residents}$, be the experience functions about the "speed" of roads in a *good for drivers* context and in a *good for residents* context, respectively. Figure 4 illustrates the mapping from initial semantic scale for measuring attribute "speed" (in Km/h) to the cognitive scale for measuring "speed" within the above defined contexts.

Verbalization Procedure VERB

Step 1. Map the value X from semantic scale onto the value R of cognitive scale with experience function $F: X \rightarrow R$,

$$r(F(x)) = \mu(x),\ x \in \Delta_p$$

Step 2. Selection of a verbal estimate M for the verbal description of the fuzzy subset R

$$M \overset{V}{\approx} R$$

$$m(y) \overset{V}{\approx} r(y), \qquad y \in [0,1]$$

Here M is a verbal estimate from the set V such that the relevant fuzzy subset with a membership function $m(y)$ is the best approximation (out of V) of the fuzzy subset R. The verbalization procedure results in a verbal estimate M derived with regard to the prescribed context.

For example, the value of speed 20 Km/h will be reflected by $F_{speed}^{good\ for\ drivers}$ to the value about 0.1 on the interval [0,1] on cognitive scale. By means of verbalization procedure it will be approximated as "very slow" road relatively to a *good for drivers* context. The same value 20 Km/h will be evaluated as enough "fast" by means of verbalization procedure relative to experience function for *good for residents* context (see Figure 4). So the same value of the speed of a road will be evaluated as "very slow" and "fast" in different contexts.

The interpretation procedure, which is reverse to the verbalization, permits to interpret verbal estimates on the background of a given context. Figure 4 illustrates also the procedures of verbalization and translation for different contexts. For example, by means of the interpretation procedure, the same verbal estimate "fast" of "speed" will be interpreted in different ways in different contexts: as about 20 Km/h m in a *good for resident* context, and as about 50 Km/h in a *good for drivers* context. It is possible, by using of verbalization and interpretation procedures, to translate verbal estimates from one context to another one.

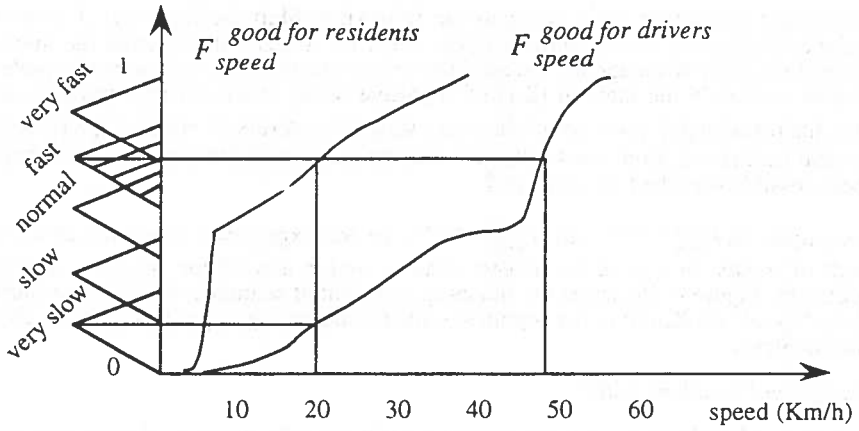


Fig. 4 Cognitive scale showing "very slow" and "fast" verbal estimates in a *good for drivers* and *good for resident* context

3.5 Verbalization description of basic concepts

The Contextual System is able of verbal description of the basic concepts $e \in E$ with verbalization procedure. First, describe the basic concepts in the update semantic space S^*. Let us fix attribute p_k and concept e from the set E (e.g. p_k = "speed", e= "busy road"). Let Σ^e be a set of samples in the KB relative to the concept e and let $f^e{}_k$ be the distribution function of these examples with respect to the values of attribute p_k. Function $\mu^e{}_k$, derived from $f^e{}_k$ by normalizing to 1 can be viewed as a membership function of a fuzzy value $T^e{}_k$ of attribute p_k characterizing concept e. Then each e basic concepts can be described in an updated semantic space S^* as a vector

$$e \overset{S^*}{=} <T^e{}_1,\ldots,T^e{}_m>$$

Let $F=<F_1, \ldots, F_m>$ be an experience function vector with respect to attributes $p_1,\ldots,p_m$. Let $V_1, \ldots, V_m$ be sets of verbal estimates ordinarily applicable to attributes $p_1,\ldots,p_m$ respectively. By applying verbalization procedure to each attribute separately, the Contextual System can define verbal estimates $<L^e{}_1,\ldots,L^e{}_m>$ for attributes $p_1,\ldots,p_m$, respectively. With regard to the given Problem Area context:

$$L^e{}_i = VERB\ (T^e{}_i,\ F_i,\ V_i)$$

Here, the input parameters of the VERB procedure are:

- $T^e{}_i$: value of attribute p_i on concept e,
- F_i: experience function with respect to attribute p_i in the given context,
- V_i: a set of verbal estimates traditionally employed for measuring attribute p_i.

As a result, the Contextual System describes concepts e in terms of linguistic estimates of attribute $p_1,\ldots,p_m$ in cognitive space U:

$$e \overset{U}{=} <L^e{}_1,\ldots,L^e{}_m>$$

Figure 5 illustrates the procedure of description of concepts "busy road", "normal road", and "free road" in the frame both *good for drivers* context and *good for resident* context for the attribute "speed" of a road.

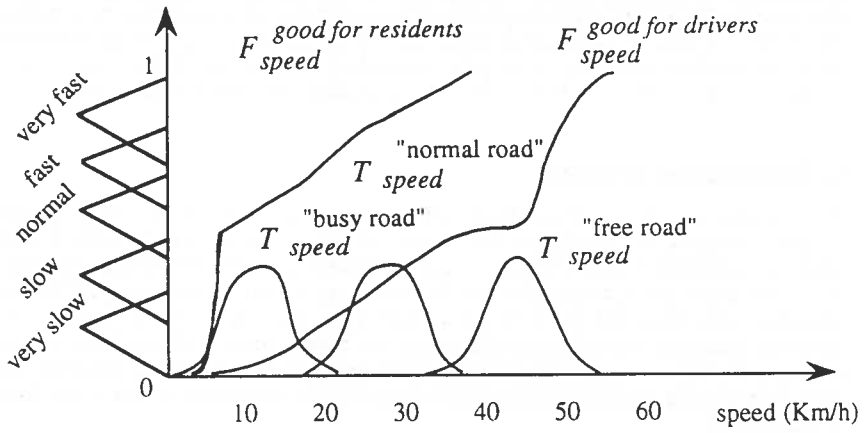


Fig. 5 Description of basic concepts "busy road", "normal road", and "free road", through the attribute "speed" in a *good for drivers* context and for the *good for residents* context.

Here $T_{speed}^{"busy\ road"}$, $T_{speed}^{"normal\ road"}$, $T_{speed}^{"free\ road"}$ are distribution values of concepts "busy road" "normal road", and "free road" respectively, which are treated as fuzzy sets on semantic scale. In this case appropriate verbal estimates for the concepts will be described at the background of a *good for drivers* context as "very slow", "normal" and "fast" respectively, namely:

"busy road" = < "very slow" "speed" , ..., >
"normal road" = < "normal " "speed", ..., >
"free road" = < "normal" "speed", ..., >

A different description is obtained at the background of a *good for resident* context:

"busy road" = < "normal " "speed" , ..., >
"normal road" = < "high" "speed", ..., >
"free road" = < "high" "speed", ..., >

In this manner, the Contextual System describes all Problem Area concepts. One should recall herewith that, following changes in the Problem Area context, another set of basic concepts is assigned, a new experience vector is built, hence new verbal description of concepts are made. One and the same concept will be described differently at the background of different contexts.

4. Solver Block

It is common knowledge that in decision making man can intesively draws knowledge from other areas shaping metaphors, analogies, and associations. This type of knowledge processing requires the development of a meta-view at the body of problem area knowledge, i.e. formalization of the problem area context. The context formalization is also necessary for defining the more valid procedures of knowledge processing within the context of each problem area separately. Note that the proximity of concepts behind decision making schemes is largely dependent on the problem area context. Two concepts close to one context may become diametrically opposite in the other.

4.1 Intracontextual processing

The introduction of contextual space makes it possible to determine distances between the context concepts. Given context variations, distances between the concepts change too, which is consistent with the human concept of the real world. The cognitive scale enables one to measure the distance between the attribute values of the context. The traditional semantic scale does not provide this opportunity. As was shown in Ezhkova (1989) to compute distances between the estimates at the background of the context, it is necessary to map them onto a congnitive scale and then compute the difference between them. The cognitive scale is actually absolute for measuring the attributes values at the background of the context.

Let $e_1 = <L^{e1}_1, \ldots, L^{e1}_n>$ and $e_2 = <L^{e2}_1, \ldots, L^{e2}_n>$ be descriptions of concepts e_1 and e_2 in the conceptual space. The linguistic estimates L^{e1}_1 and L^{e2}_1 are in the form of fuzzy subset.

Let ρ_i be a distance between the linguistic estimates of concepts e_1 and e_2 with respect to the i_{th} attribute. The distances between the concepts e_1 and e_2 in the context space is measured in the same manners in the euclidean space in terms of coordinate distances ρ_i, $i=1,\ldots, n$:

$$(2) \qquad \rho = \sqrt{\frac{\sum_i \rho_i^2}{n}}$$

The use of distances in the context space provides a foundation for devising intracontext schemes of knowledge processing decision making.

Let $X=<X_1, \ldots, X_k>$ be a case under evaluation in the local problem, related to the context C.The Solver Block computes distances in the contextual space from vector X to each basic concept $e_j \subseteq E$:

$$\rho^{x_{e1}}, \ldots, \rho^{x_{en}}$$

where distance $\rho^{x_{ej}}$, $j = 1, \ldots, t$ was calculated by means of procedure (2) from distances $\rho_i^{X_{ej}}$ for each attribute p_i, , $i = 1 \ldots, n$.

For example, the computed distances from the road X under evaluation to each of concepts "busy road", "normal road", and "free road" for the attribute "speed" at the background of a *good for drivers* context are:

$$\rho^{X,busy}_{speed} = 0.751, \ \rho^{X,normal}_{speed} = 0.382, \ \rho^{X,free}_{speed} = 0.101,$$

By using procedure (2) we calculate the distances from X to concepts "busy road", "normal road", and "free road":

$$\rho^{X,busy} = 0.694,\ \rho^{X,normal} = 0.513,\ \rho^{X,free} = 0.141$$

By using experience function for the distances in the current context, it is possible to verbalize the degrees of possibilities. For example, in the case under consideration the result of evaluating road X at the background of *good for drivers* context could be:

- the distance from X to the concept "free road" is *"small"*
- the distance from X to the concept "normal road" is *"large"*
- the distance from X to the concept "busy road" is *"very large"*

This set of distances can be treated as distribution of possibilities of alternatives e_i, $i = 1, \ldots, t$, $e_i \in E$, in the local decision making problem. So as a result, the block produces a set of values $\{ \rho^{x_q} \}$ determining the possibility of each alternative $e_i \in E$.

4.2 Intercontextual processing

Intercontext knowledge processing block works concurrently with several contexts in the process of decision making. Contexts can be formed by request of the block itself, or be selected both from context base and active memory. It is Algebra and logic of contexts that underlie the intercontext schemes of knowledge processing. Algebra of contexts involves contraction, extension, immersion, coupling and intersection of contexts. For example context C immersion in context C_1 is used by the Contextual System for extended learning, adaptation, introduction of new notions, etc. Knowledge immersion in a broader context is often used, in practice, in solving of interpretation problems.

The *logic of constext* is based on mapping one context into the other: $C \xrightarrow{\varphi} C_1$, where φ, in a general case, is a train of mappings $\varphi = < \varphi_E, \varphi_A, \varphi_F, \varphi_Q, \varphi_K, >$, each one influencing its own element. In handling each fixed context separately, the given block addresses the intracontext knowledge processing block. In general solver results in a set of estimates of decision alternatives.

5. Conclusions

In the paper is presented a new advanced paradigm based on contextual technology for flexible decision making in transportation. Complex analysis of etherogeneous data describing a transportation system can be performed in relation to a given scenario or context. This seems to provide adequate information in understandable form for supporting strategic decision making in traffic networks.

References

Bianco, L. (1989), The Role of Quantitative Methods in Urban Transportation Planning, *Syst. Anal. Model. Simul.*, Akademie-Verlag Berlin, 6, 29-59.

Bianco, L., P. Dell'Olmo, and, I.V. Ezhkova (1994), Context-dependent Decision-making in the Transportation Management Process, *IASI-CNR Tech. Rep.*, (in press).

Ezhkova, I.V. (1989), Knowledge Formation Through Context Formalization, *Computers and Artificial Intelligence*, 8, 305-22.

Ezhkova, I.V. (1992), Contextual Technology for Supporting Decision Making, *Proc. of the Eleventh Eur. Meeting on Cybernetics and Systems Research*, Vienna 1992, 503-510.

Ezhkova, I.V. (1993), Contextual Approach for AI Systems Development, Fuzzy Logic in Artificial Intelligence, E.P. Klement, W. Slany (Eds), *Lecture Notes in Artificial Intelligence*, 695, Springer Verlag.

Hopkinson, P.G., C.A. Nash and N. Sheehy, (1992), How much do people value the environment?, *Transportation*, 19, 97-115.

Takyi, I.K., (1993), A Multidimensional Methodology for Evaluating Public Transportation Services, *Transp. Res. A.*, 27A, 395-405.

INTERMODALITY IN URBAN TRANSPORTATION : MODELLING AND ANALYSIS

Angela Di Febbraro, Valerio Recagno, and Simona Sacone

Department of Communications, Computer, and System Sciences

University of Genova

Via Opera Pia 13, I-16145 Genova, Italy
Tel.: +39-10-3532983 Fax: +39-10-3532948

Abstract

This paper deals with modelling, evaluating the performances, and controlling an intermodal transportation network. Such system is represented as an oriented graph whose nodes can represent single-mode stations or intermodal stations. For the system considered, a discrete event model is presented, integrating public and private transportation means. Some disturbances are included to model the stochastic nature of the system. Based on such a model, an urban traffic simulator (INTRANET) has been on-purpose designed. INTRANET includes two major modules. The first one, the Traffic Simulation Kernel, allows to study the dynamic behaviour of the transportation system, analyzing its performances and applying control strategies to optimize them. The second one, the Passenger Information Service, gives the system users at any time updated information about the different intermodal paths between any pair of origin/destination nodes. Some experimental results relevant to a case study, showing the effectiveness of the proposed control strategies, are presented and discussed.

1. INTRODUCTION

The progressive concentration of economic activities within cities has resulted in an ever-increasing need of efficient transportation facilities inside metropolitan areas. Such a need is strongly connected with the limitation of the 'private' traffic in the city centers, so as to prevent situations of severe traffic congestion. This is feasible only in a city where suitably working public transportation services exist. Such services should be widespread, efficient, and sensitive to the changes in users' demand.

This is the reason why in recent years, the integration between public and private transport services, with the aim of increasing the 'attractiveness' of public transportation means with respect to private ones, has become a research topic of great interest. Such an integration leads to realize intermodality in transportation. *Intermodal transportation* can be defined as the serial use of different means of transport to move passengers and/or freight from a place to another. From the above definition two major research topics result, relevant to the movement of passengers and to the movement of goods (Kondratowicz, 1990; Guelat, et al., 1990). In this paper, the problem of integrating passenger transportation services in a urban area is dealt with. The realization of efficient intermodal transportation systems is especially intended to lead people not to use their cars to move in the cities, which would result in both an optimization of the travelling times and a decrease in air pollution.

The two fundamental components of a transportation system are the *demand* system, which represents the needs and the behaviour of the users, and the *supply* system, which includes all what concerns the service production from the infrastructures to the planning rules of the system (Kanafani, 1983). It usually happens that some aspects of the demand depend on the supply characteristics and vice versa. Then, a model of a transportation system can be seen as the integration of three different models: the demand model, the supply

model, and the demand/supply interaction model (Ben Akiva and Lerman, 1987; Cascetta and Cantarella, 1993). This paper deals with a supply model relevant to an intermodal transportation network, proposed with the aim of analyzing the performances of the system and designing control strategies to optimize them. The network realizes the integration among urban public transportation services themselves and private traffic. The proposed model can be viewed as an extension of the one presented in (Di Febbraro, Recagno and Sacone, 1994). The system is represented as an oriented graph, in which the nodes are stations where it is possible to switch from a mode of transport to another one.

For the peculiar characteristics that an intermodal transportation network presents, it seems suitable to model it as a Discrete Event Dynamic System (DEDS). In fact, discrete event modelling is suitable for those systems in which significant changes in their state only occur at discrete time instants. In general, DEDS respond to the following characteristics: they are discrete in time and space, asynchronous, and modular. Moreover, they may include control strategies and communication systems. Ruling communication and control strategies, it is possible to make the event flow satisfy the project requirements (Ho and Cao, 1991; Cassandras, 1993). For the transportation system under study, a peculiar discrete event model is proposed. Due to the stochastic characteristics of the transportation system considered, some disturbances are also included. The behaviour of the discrete event system modelling the transportation network is studied by means of an on-purpose designed simulation tool, which is the kernel of the urban traffic simulation program INTRANET (INtegrated TRAnsportation NETwork).

INTRANET is designed to perform two major functions, almost independent of each other. The first objective is the validation of integrated timetables for the different transportation modes, so as to consider the various transportation services to be parts of a whole intermodal transportation system. In this sense, as will be clear in the following, INTRANET can evaluate the performances of the system, supporting the user in modifying adaptively a given timetable until reaching the desired performances.

The second objective pursued in designing INTRANET is to give the users of the intermodal transportation network some real-time updated information about the state of the network itself. The optimal situation to realize is that in which a user can choose, when approaching the area covered by the transportation network, the best path to reach his destination. Each user could enter the information system in the stations of the network, or, in a more advanced realization, even from his car. In particular, if a user approaches the network by car, he can evaluate where it is possible (and convenient) to leave his car and carry on with his journey by public transportation means. To this end, some intermodal stations have a parking area associated with. More specifically, the available information regard essentially : i) the travelling times along all the possible paths, intermodal or not, between any pair of nodes in the network; ii) the availability of places in the parking areas associated with some macronodes; iii) the occurrence of irregular traffic conditions in some part of the network. Such an on-line information system could lead to a fair distribution of the traffic on the different (multimodal) paths and transportation modes, resulting in an improvement of the system performances.

The paper has the following structure. Next section is devoted to the description of the discrete event model designed for the intermodal transportation network considered. Section 3 describes the main modules of the simulation program INTRANET, which are : the Input/Output Interface, the Traffic Simulation Kernel (T.S.K.), and the Passenger Information Service (P.I.S.). As regards the T.S.K., it is essentially divided into two modules: the Discrete Event Model (D.E.M.), described in Section 2, and the Control Module. Some control strategies, based on the measure of the delays of the transport means, are proposed with the objective of guaranteeing the regularity of the transportation services, i.e., the fulfilment of the timetables for the considered transportation modes. Then, a case study is presented for which simulation results are available, showing the differences between the performances of the transportation system with and without the application of control strategies. Some conclusions end the paper, also giving some insight into possible further developments of the presented work.

2. MODELLING AN INTERMODAL TRANSPORTATION NETWORK

In this section, the model of an intermodal transportation network is presented. First, the general characteristics of the model are described, and its main components defined. Then, the discrete events selected to represent the functioning of the network are introduced. On the whole, this section is a description of the module of the T.S.K. mentioned in the Introduction as D.E.M.. The considered intermodal transportation

network is modelled as an oriented graph, whose fundamental elements are nodes, macronodes, and links. A *node* is a station for a single transportation mode, and it can only exist as a part of a macronode. A *macronode* is an intermodal station, that is, a place where people can change mode of transport, or simply enter the transportation system. Therefore, a macronode is composed of one or more nodes. A *link* is a unidirectional transportation path between two adjacent nodes/macronodes, and which can have two different functions: there are links which connect two macronodes and are devoted to a single transportation mode, and links which connect two nodes in a macronode and can be traversed only on foot. In the following, the second kind of links will be referred to as *inner links*.

It is assumed that users can approach the urban integrated transportation system using any means of transport. To model the arrivals of passengers by private transport means, parking areas are considered to be located near some macronodes of the network, where passengers can leave their transport means, if room is available. The existence of a parking area at a macronode and the availability of places in it play an important role in modelling the integration between public and private transport means.

It is assumed that there are three different kinds of public transport means, which are underground, railway, and bus lines. In this framework, the links which connect the N macronodes in the network are divided into three classes: a) bus links (*B-links*); b) railway links (*R-links*); c) underground links (*U-links*). For each macronode n, n=1, .., N, the types of its input and output links (U, R, or B) are specified. Clearly, each macronode can have at most six links in common with another macronode.

Macronodes n, n=1,...,N-1, represent the real intermodal stations of the transportation system under study, while the last macronode, macronode N, is a special macronode, i.e., a 'virtual' intermodal station which represents the outer world with respect to the network. In other words, each macronode which connects the urban area with its surroundings has a link towards macronode N. The presence of such a virtual macronode allows to model the considered transportation system as a closed network. This yields a major advantage consisting in the possibility of applying particular storage methods for closed graphs, which reduce sensibly the computational load. Moreover, the presence of macronode N also allows to model the movement of those means of transport which go, or come from outside the considered urban transportation network.

2.1. Nodes, macronodes, links, and transportation means

In this section, the main entities of the system to be modelled are described. An *entity* is a component of the system which requires explicit representation in the model (Banks and Carson, 1984). In this case, the entities of the model consist of those objects defined before as network entities (macronodes, nodes, and links) plus the transportation means. With each class of entities, static quantities (*parameters*) and dynamic quantities (*state variables*) are associated.

Nodes. A node is a part of a macronode which models a station for a single transportation mode. The parameters associated with a node are :

a) a numerical identifier;
b) the transportation mode to which it refers;
c) its capacity (the maximum number of means that can be in the node at the same time);
d) the minimum stopping time of transport means;
e) its typology (station or terminus) with respect to the different lines which include the node. This parameter is necessary since the terminal node has the following features: i) its capacity is typically larger than that of an ordinary station; ii) vehicles can be parked; iii) transportation means can change their line numbers;
f) the time necessary to a vehicle to travel to the node from the nearest depot. This is necessary to evaluate, in the control phase, the possibility of getting a new vehicle from a depot. The influence of depots on the normal functioning of the system stands in the fact that transportation means which are travelling from a depot to a macronode of the network can affect the traffic in the links they are passing through. For bus lines, the interference can be neglected since the capacity assigned to a bus link is seldom equal to the maximum physical capacity of the road. As far as it regards railway and underground train depots, it has been assumed that there exist dedicated links for the vehicles which move from depots to macronodes and vice versa, so as not to affect the normal functioning of the system.

The state variables of a node are:

α) the number of active transport means in it;
β) the number of out-of-service transport means in it. In each railway and underground station a place to recover broken vehicles exists. In particular, each node has a dedicated platform where the broken vehicle can stop and passengers can leave the vehicle to proceed with their travel in the network.

Macronodes. The parameters associated with a macronode, which is the entity modelling an intermodal station, are:
a) a numerical identifier;
b) the transportation modes of the nodes in it;
c) the existence of a parking place close to the station;
d) the travelling times on the inner links; obviously, these times depend on the structure of the intermodal station.
The state variable associated with a macronode is the availability of parking places in the area close to it, if such an area exists.

Links. A link, i.e., the physical connection between two macronodes, is characterized by the following parameters:
a) the transportation mode to which it refers;
b) the pair of macronodes which it connects;
c) its length;
d) its minimum travelling time;
e) its timetabling travelling time;
f) its capacity, i.e., the number of transportation means which can travel along the link (only for R and U-links). The capacity is a parameter for railway and underground links, since in these cases it cannot vary during simulation, whereas it is a state variable for B-links, where the real capacity is strongly dependent on the traffic conditions.

The state variables of a link are:
α) the number of vehicles travelling along it;
β) the real travelling time, which can differ from the minimum and the timetabling times because of the stochastic disturbances and the occurrence of some particular event;
γ) its capacity, only for B-links (see point (f) above).

Transportation means. Each transportation means is characterized by an alphanumerical parameter also indicating the relevant transportation mode. The state variables of a vehicle are:
α) the number of line;
β) the state of the vehicle, which can be one of the following : i) travelling on a link; ii) stopped in a node with closed doors (passengers cannot leave the vehicle); iii) stopped in a node (passengers can leave the vehicle); iv) ready to leave a node; v) waiting for entering a node; vi) waiting for leaving a node; vii) broken in a node; viii) broken on a link; ix) inactive (no more present in the network);
γ) timetabling leaving time; this time is significant only when the transportation means is stopped in a node.

Besides to the main static and dynamic quantities associated with the fundamental entities, the model comprehends a further set of parameters and state variables used to describe particular behaviours of the considered transportation system. For example, there exist some variables which concern the synchronization of vehicle movement in the network. These synchronization mechanisms are quite necessary since nodes and links are limited resources shared by the transportation means. Furthermore, there are variables used to memorize break-downs of vehicles, in order to introduce these aspects in the analysis of the system behaviour.

2.2. Events and state transitions

An *event* can be defined as an instantaneous occurrence which may produce changes in the system state. The events that describe the evolution of the considered urban transportation network, and then rule the functioning of the INTRANET simulation tool, are divided into the following two classes:

i) events that describe the nominal traffic conditions;

ii) events that represent unpredictable conditions affecting the system.

The events belonging to the first class are:

i1) arrival of a transportation means at the end of a link;
i2) beginning of a stopping time;
i3) end of a stopping time;
i4) departure of a transportation means from a node through a link;
i5) car parking available;
i6) car parking not available;
i7) simulation end.

The second class consists of the events:

ii1) beginning of a failure in a node;
ii2) end of a severe failure in a node;
ii3) end of a light failure in a node;
ii4) beginning of a failure in a B-link;
ii5) end of a severe failure in a B-link;
ii6) end of a light failure in a B-link;
ii7) beginning of a failure in a R-link or in a U-link;
ii8) end of a severe failure in a R-link or in a U-link;
ii9) end of a light failure in a R-link or in a U-link;
ii10) beginning of a B-link congestion;
ii11) end of a B-link congestion.

Each type of event is associated with an operational procedure which performs the state transitions resulting from the event occurrence. How simulation proceeds on the basis of the event occurrences is explained in next section.

3. INTRANET : A URBAN TRAFFIC SIMULATOR

In this section, the structure of the INTRANET urban traffic simulator is described. It consists of different modules, as shown in the figure below.

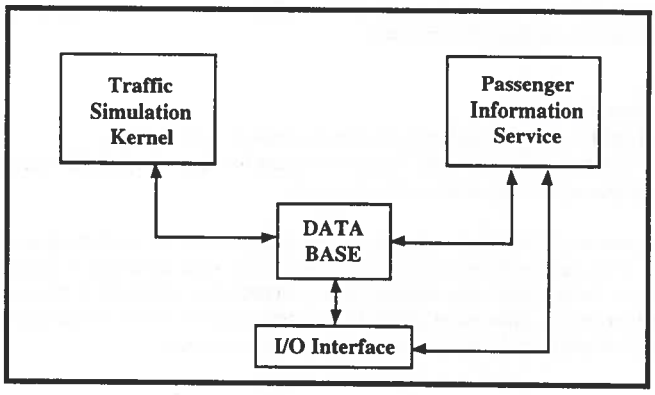


Fig. 1 - Structure of the INTRANET simulator.

For each module among the I/O Interface, the Traffic Simulation Kernel (T.S.K.), and the Passenger Information Service (P.I.S.), a detailed description is given in the relevant subsections. These three modules communicate by means of the data base, which contains both static and dynamic information. The static data consist of the inputs required by the simulation tool, as detailed in the following subsection. The dynamic information regard the state of the network, which is updated at each event occurrence. The P.I.S. has a bidirectional link to the I/O interface, so as to give it the possibility of receiving as an input the user's requests for shortest paths to move from a macronode to another one in the network, and sending back the answer. This mechanism is described in subsection 3.3.

3.1. The Input/Output Interface

As regards inputs, the I/O interface accepts data in a given format and makes up the static information of the data base to be used by the main blocks of INTRANET, i.e., the T.S.K. and the P.I.S.. The inputs required by the simulator are the following:

a) the topology of the network and the timetables for the three transportation modes;
b) the parameters of the probability distributions of the stochastic variables modelling the disturbances introduced in the model of the intermodal transportation network to fit its stochastic nature, which are :

   b1) perturbation of the timetabling leaving time of a transport means from a node; it consists in a time which is added to the timetabling leaving time, according to a Gaussian probability distribution;
   b2) perturbation of the timetabling travelling time of a link; again, this is a time which is added to the timetabling travelling time, according to a Gaussian probability distribution;
   b3) time interval between the arrival of a transportation means at the end of a link and the beginning of its stopping time in the node which the link enters;
   b4) time interval between the end of a stopping time of a transportation means in a node and its departure from a node;
   b5) duration of the different kinds of failures;
   b6) time interval between two subsequent failures.

Variables (b3)-(b6) are associated with different uniform probability distributions. Such distributions are generated by means of the inverse transform technique, while the Gaussian distributions are obtained by means of an approximation made using the convolution method (Banks and Carson, 1984).

In addition, the I/O interface manages the user's requests to the P.I.S., as mentioned above. As for outputs, the I/O interface gets from the data base information to make available the outputs of each simulation session. In particular, the output of a simulation run consists of:

a) a simulation trace; such a trace is essentially a list of all the events occurred during simulation, indicating the type and the occurrence time of each event;

b) statistics about:
   - link travelling times;
   - stopping times for each transportation mode and means at each node;
   - delays for each transportation mode, for each line, and for each transportation means;
   - duration of failures for each node and for each link.

The output statistics above can be used to analyze and optimize the performances of the transportation system. The choice of the performance indexes to evaluate can be made according to the particular purposes of the simulation analysis. In this paper, the objective is to guarantee the fulfilment of the timetables for the three different transportation modes. This can be achieved by monitoring the delays of the transportation means and implementing control strategies to be applied whenever a delay is recorded.

3.2. The Traffic Simulation Kernel

The Traffic Simulation Kernel consists of two fundamental blocks, which are the Discrete Event Model, and the Control Module, as shown in the figure below.

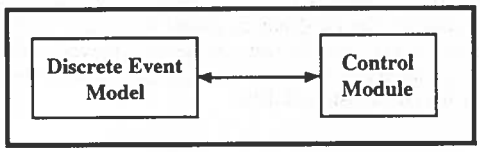


Fig. 2 - Structure of the Traffic Simulation Kernel.

The D.E.M., which represents the dynamic behaviour of the transportation network under study, has been described in section 2. The evolution of the system is reproduced by means of discrete event simulation, which can be viewed as the modelling over time of a system in which changes in the state only occur at discrete time instants, i.e., those istants when events occur. The mechanism to make simulation advance correctly, with all the events occurring in chronological order, is based on the *active event list* . Such a list is initialized before starting a simulation session, including those events which can be scheduled when the session starts. During the simulation, the active event list is examined and updated, by processing and eliminating, time after time, the event which occupies the head of the list (Banks and Carson, 1984).

Each type of event is associated with an operational procedure which, once invoked, modifies the state variables of the model in dependence of the event occurred. Such procedure has to insert in the active event list the events which will occur as a consequence of the current event, and cancel those events which the current event prevents from occurring. However, there are events which are not scheduled in dependence of any other event. These are the events gathered in class (ii) in subsection 2.2., i.e., those events which represent unpredictable conditions affecting the system. In fact, such events occur according to statistical distributions, a-priori given as inputs to the simulator. Whenever any nonrecurrent situation happens, the Control Module of the T.S.K. is invoked. This module can implement different control strategies to try to recover the regular functioning of the system, as explained in next section.

3.2.1. Control strategies

In such a transportation system as the one described in this paper, i.e., when one neglects all the modelling aspects relevant to the single passengers moving in the network, a sensible control objective is that of guaranteeing the regularity of transport services. This can be achieved by designing suitable control strategies in order to fulfil the timetable. The proposed criterion to analyze the performances of the intermodal transportation system is based on the measures of the delays of the transport means on the different lines associated with the three considered transportation modes.

As mentioned above, the model presented in this paper only represents the transportation supply. Then, it is supposed that the demand characteristics have been taken into account in the procedure to determine the timetables given as inputs to the simulation tool. The timetables for the different transportation modes vary during the day just to follow the demand dynamics, in the fulfilment of the budget constraints.

In this sense, two different control strategies to fulfil the timetable in the event of delay of some means of transport are proposed. The first, simple one is that of making the transport means which are late traverse the links in the minimum travelling time until they achieve again the fulfilment of the timetable. This 'acceleration' is feasible since the timetabling travelling times of the links are usually longer than the minimum ones, so as to make it possible to regain some short delay. Note that another mechanism to make up for short delays is embedded in the system. Consider a transportation means stopped in a node. In fact, its timetabling leaving time from that node usually allows a stopping time longer than the minimum one. So, following the same lines as for the travelling times, such a stopping time can be reduced until reaching the minimum one, in order to reduce the delay of the transportation means.

The described control strategy proved to be useful when delays occur for railway and underground transportation modes, but it presents some drawbacks while applied in case of bus delays. In fact, the acceleration proposed is useless whenever bus delays are caused by link congestions, as usually happens. This consideration leads to the definition of a dedicated control strategy, hereafter described. At the occurrence of a "beginning of a B-link congestion event", the congested link is identified. Thus, the set of the bus lines including the congested link is determined. For each of these lines, the arrival time at the node at the end of the congested link of the bus closest to such a link is estimated, so as to compute its delay as in the first control strategy. Then, the introduction of a new bus, coming from the depot the closest to the node which the congested link enters, is taken into account. Such a bus will leave from that node and head for the end of the considered line. The cost of such an operation is evaluated on the basis of maintenance, fuel, and staff costs. The action is intended to be realized if it appears less costly than the considered delay.

3.2.2. A case study

In order to analyze the behaviour of the transportation system described, simulation experiments have been performed on different intermodal networks. In this section, an applicative example is reported. The considered transportation network is depicted in Fig. 1. It is composed of 80 links and 21 macronodes, each of which consists of at least two nodes. In particular, in the network there are 21 bus stations, 13 railway stations, and 11 underground stations. As regards the transportation lines, the network comprehends 24 bus lines, 8 railway lines, and 4 underground lines; the timetables present a number of runs up to 20/hour during the peak hour for each transportation mode.

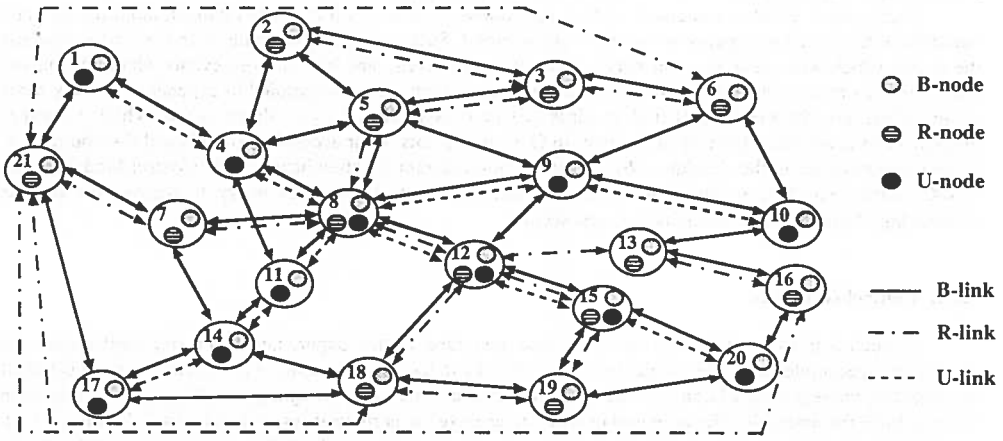


Fig. 3 - The considered intermodal transportation network.

Various simulation experiments with one-day duration have been run on a Sun Sparc 10 Station, and statistics about vehicle, line, and mode delays, as well as the other outputs mentioned in the subsection 3.1, have been gathered. Some significant experimental results are reported in the figures below. In particular, significant comparisons have been made between the behaviours of the transportation system with and without the application of control actions. In Fig. 4 the comparison between the mean bus delays during the day without control (plain line) and under the application of the first control strategy (dashed line) is reported. The same kind of comparison is made in Figg. 5 and 6 for railway and underground, respectively. Note that the scale is different for each diagram so as to make the variations in the delay behaviours as evident as possible.

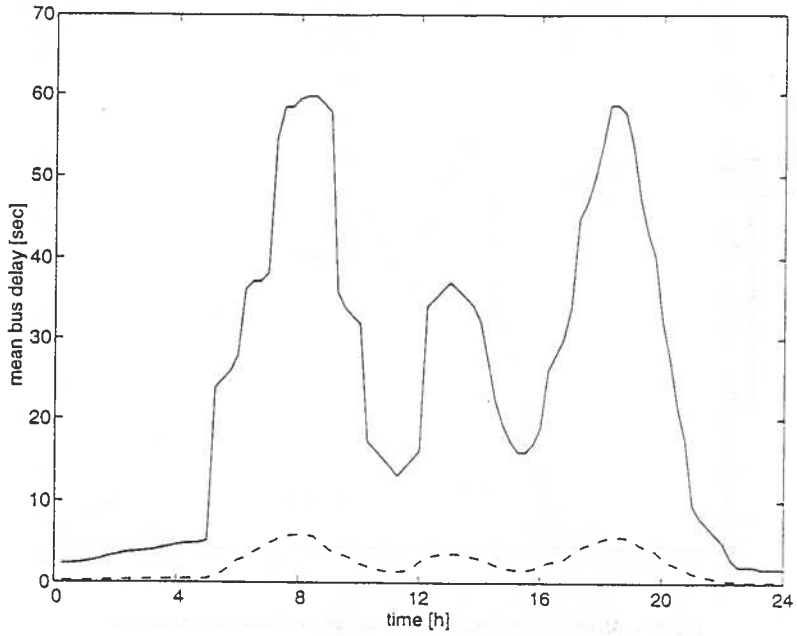


Fig. 4 - Mean bus delays with and without control action.

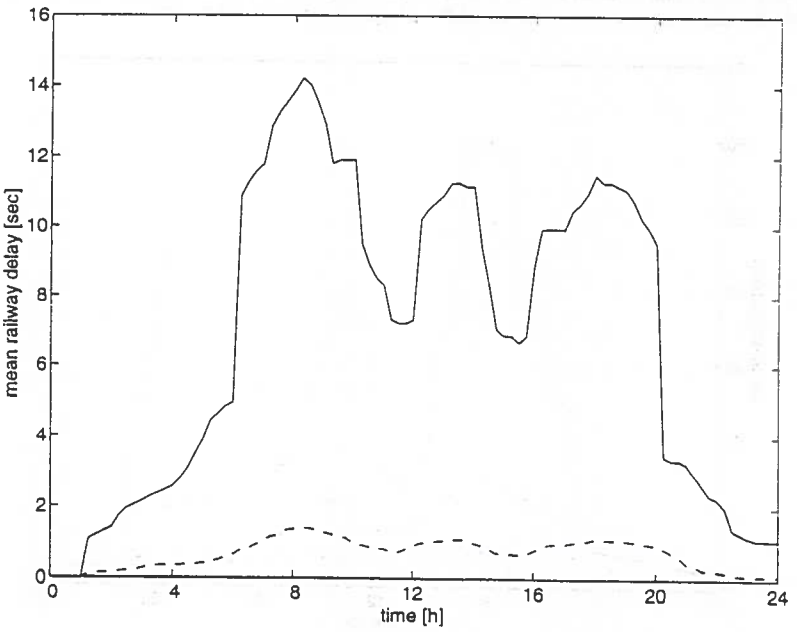


Fig. 5 - Mean railway delays with and without control action.

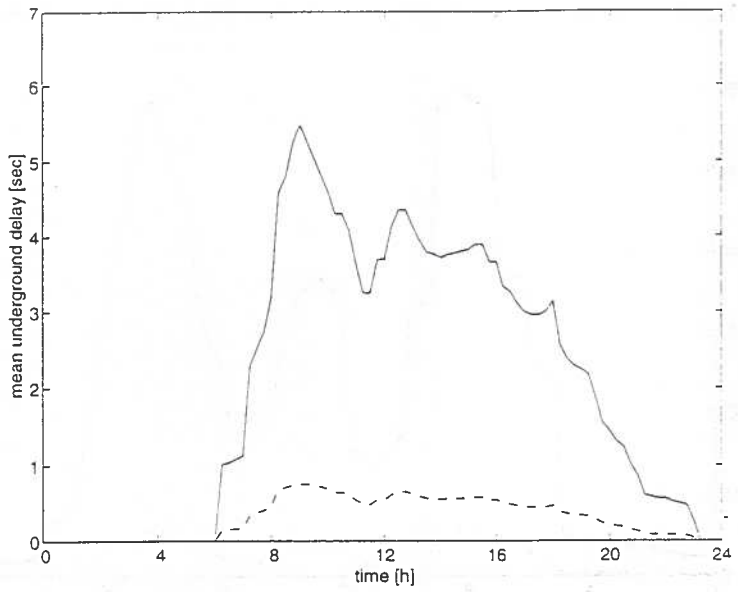


Fig. 6 - Mean underground delays with and without control action.

Fig. 7 reports in the same diagram the mean delays of the three transportation modes: the plain curve represents the bus delays, the dash-dotted one stands for the railway delays and the dashed one for the underground delays. This figure puts into evidence the differences between the behaviours of the three transportation modes and, in particular, the fact that bus delays are always significantly greater than the other ones.

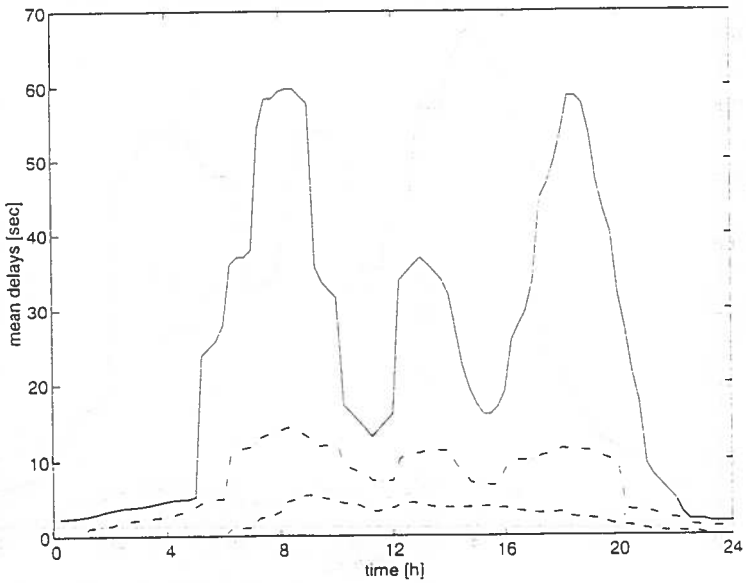


Fig. 7 - Mean delays of the three transportation modes.

The delays of the three transportation modes under the application of the first control strategy are shown in Fig. 8. It should be noted that the proportions among the delays of the three modes, as well as their 'shape', are similar to those shown in Fig. 7, but the control action reduces the values of such delays by a factor 10.

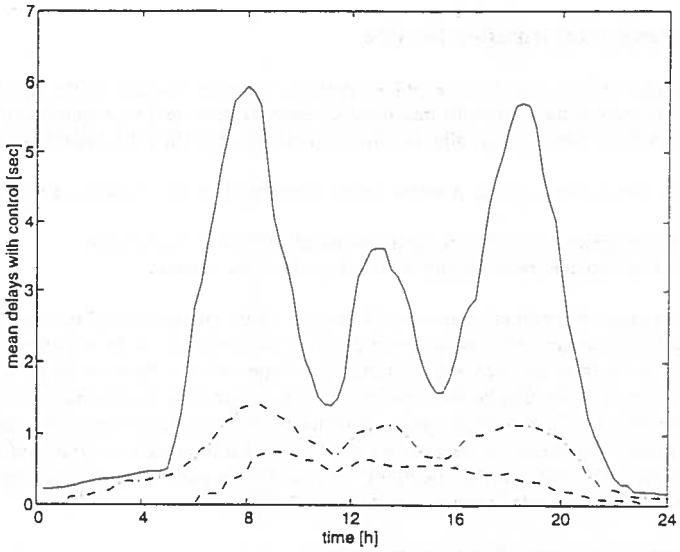


Fig. 8 - Mean delays of the three transportation modes with the application of the first control strategy.

Besides to the delays with respect to the timetable, there is another class of variables which is strongly dependent on the demand behaviour during the day, i.e., the stopping times at the nodes. This is especially true for the nodes in which more lines of different transportation modes intersect. For instance, in Fig. 9 the behaviour of the stopping times at macronode 8, which is one of the most important from a 'traffic' point of view, is reported.

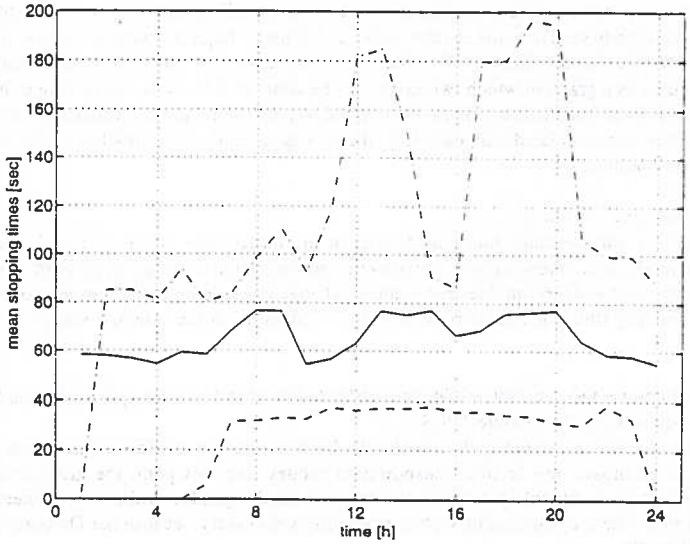


Fig. 9 - Mean stopping times at macronode 8 for the three transportation modes.

Here again, the plain curve represents the bus mean stopping times, the dash-dotted one stands for the railway mean stopping times and the dashed one for the underground mean stopping times.

3.3. The Passenger Information Service

The Passenger Information Service (P.I.S.) performs a major function of the INTRANET simulator, which is to give the users of the intermodal transportation network some real-time updated information about the state of the network itself. More specifically, the information offered by the P.I.S. regard essentially :

i) the travelling times along all the possible paths, intermodal or not, between any pair of nodes in the network;
ii) the availability of places in the parking areas associated with some macronodes;
iii) the occurrence of irregular traffic conditions in some part of the network.

Such an on-line information system could lead to a fair distribution of the traffic on the different (multimodal) paths and transportation modes, resulting in an improvement of the system performances. All the paths, intermodal or not, between a couple of origin/destination nodes in the network are a-priori determined. The real travelling times of the links between two macronodes are updated in real-time depending on the state of the network. Each time, the P.I.S. receives via the I/O interface a request from a user who needs to move from a macronode to another macronode of the network and, using real-time data, provides information about the shortest multimodal paths between the couple of macronodes. In particular, the passenger's request which constitutes the input to this module, is composed of the following data:

i) an origin macronode n_o and a destination macronode n_d;
ii) the transportation modes neglected by the user (possibly none); in fact, the user can choose which modes he prefers to use to reach its destination;
iii) the maximum travelling time t_{max} of the path from n_o to n_d (possibly ∞); this represents another degree of freedom of the user, who can fix the maximum time he wants to spend to reach its destination;
iv) the number k of alternative paths requested (possibly 1).

The P.I.S. retrieves both the static information about the travelling times on inner links and about the topology of the network, and such dynamic quantities as the link travelling times and the expected waiting times at the nodes from the data base. Then, the module satisfies the users' requests using an ad-hoc modified version of the *Dijkstra algorithm* (Christofides, 1975). This algorithm takes into account the fact that the considered network is represented as a graph in which two nodes can be connected by more than a single link (in particular there could be at most three links connecting two macronodes in each direction). Furthermore, another feature of the method used is the cost associated with each link; the cost associated with a link l of mode m is a function of the following three quantities:

i) the real travelling time on link l;
ii) the travelling time on the inner link connecting, in the macronode origin of link l, the station of the transportation mode used to reach such a macronode and the station of mode m; this time can be zero, if the path found so far by the algorithm does not require a change of mode in the current macronode;
iii) the expected waiting time for transportation means of mode m in the relevant station of the macronode origin of link l.

The times in (ii) are known and constant, while those indicated in (i) and (iii) are updated in real time, depending on the state of the network, by the module T.K.S..

The degrees of freedom offered to the user by the P.I.S. can result in different problems. In particular, if the user chooses to eliminate one or two transportation modes from his path, the graph could become not connected. For this reason, a suitable procedure to remove from the graph possibly not connected macronodes has been implemented. The introduction of such a procedure is necessary because the Dijkstra algorithm works only for connected graphs.

Another complication arises from the fact that the user can require to know the k shortest multimodal paths to reach its destination. In the event of k>1, once the shortest path between the origin and destination macronodes is determined, the P.I.S. passes to a second phase of its work. It consists in finding the shortest (k-1) alternative paths besides to the best path already found. In doing so, the method proposed by Yen (1971) is applied.

The output of the P.I.S. is a list containing the k shortest paths with total travelling time less than t_{max}, connecting macronodes n_o and n_d, ordered by increasing travelling times. This list contains the information about the nodes which compose each path and about the changes of transportation modes to be made to follow the path. Furthermore, the P.I.S. associates with each path further information, consisting of:

- the total travelling time on the path;
- the number of changes of transportation modes to be made;
- indication of possibly not granted connections; this occurs when the connection time to perform a change of mode at a macronode is tight.

All the above information are passed to the I/O interface, which makes them available to the user, so as to make him able to choose at any time the best path to move from a macronode to another one in the network.

4. CONCLUSIONS

In this paper, the problems related to modelling, evaluating the performances, and controlling an intermodal transportation system have been dealt with. Such system is represented as an oriented graph whose nodes can represent single-mode stations or intermodal stations. The considered transportation network, integrating public and private transportation means, has been modelled as a discrete event system, whose behaviour has been studied by means of an on-purpose designed discrete event simulator, called INTRANET (INtegrated TRAnsportation NETwork).

INTRANET includes four modules: the Traffic Simulation Kernel (T.S.K.), the Passenger Information Service (P.I.S.), the Input/Output Interface, and the Data Base. The first two modules are the most important, since they perform the main functions of INTRANET. In particular, the T.S.K. allows to study the dynamic behaviour of the transportation system, analyzing its performances and applying control strategies to guarantee the regularity of the transportation services, i.e., the fulfilment of the timetables. The second one, the P.I.S., gives the system users at any time updated information about the state of the transportation network. It receives requests from users who need to move from a macronode to another macronode in the network and, using real-time data, provides information about the shortest intermodal paths between the selected pair of macronodes. Some experimental results relevant to a case study, showing the effectiveness of the control strategies proposed to guarantee the fulfilment of the timetables, have been presented.

In an advanced configuration, the described urban traffic simulator can be considered to be a module of a more complex structure, in which an optimization module is connected to the simulator. In this sense, the optimization module could find, in a deterministic framework, the optimal timetables for the three transportation modes to be passed as inputs to the simulator. Work is in progress to realize such an optimization module, and then the overall structure.

5. REFERENCES

Banks, J., and J.S.Carson (1984). *Discrete-Event System Simulation*. Prentice-Hall, Englewoods Cliffs, NJ.

Ben Akiva, M., and H.Lerman (1987). *Discrete Choice Analysis*. MIT Press, Cambridge, MA.

Casalino, G., A.Di Febbraro, A.Ferrara, R.Minciardi, and D.Nicoletti (1991). Underground modelling and control: discrete event approach. *Concise Encyclopedia of Traffic and Transportation Systems*. Pergamon Press, Oxford, 585-588.

Cascetta, E., and G.E.Cantarella (1993). Modelling dynamics in transportation networks: State of the art and future developments. *Simulation Practice and Theory*, 65-91.

Cassandras, C. G. (1993). *Discrete Event Systems*. Irwin and Aksen Ass., Boston, MA.

Christofides, N. (1975). *Graph Theory*. Academic Press, London, UK.

Di Febbraro, A., V. Recagno, and S. Sacone (1994). Synchronization and control of a multimodal transportation system. *Proc. IEEE Melecon '94*, Antalya, Turkey.

Guelat, J., M.Florian, and T.G.Crainic (1990). Multimode Multiproduct Network Assignment Model for Strategic Planning of Freight Flows. *Transportation Science*, **24**, .25-39.

Ho, Y.C., and X.R. Cao (1991). *Perturbation Analysis of Discrete Event Dynamic Systems*. Kluwer Academic, Boston, MA.

Kanafani, A. (1983). *Transportation Demand Analysis*. McGraw-Hill, New York.

Kondratowicz, L.J. (1990). Simulation Methodology for Intermodal Freight Transportation Terminals. *Simulation*, **55**, 49-58.

Yen, J.Y. (1971). Finding the k-shortest, loopless paths in a network. *Man. Sci.*, **17**, 712.

A Reliability Model for Urban Transportation Planning

Brunilde Sansó
and
Luc Milot

Ecole Polytechnique de Montreal
Department of Applied Mathematics
Centre for Research on Transportation and GERAD

Extended Abstract

1 Introduction

Traditional reliability measures have not been widely studied in urban transportation networks. This is due in part to the difficulty of finding good reliability measures for such networks. In fact, a traditional approach to Network Reliability has been to evaluate connectivity measures which are mainly related to the network structure and do not necessarily reflect the network capability to transmit flow, which is the ultimate goal of any flow network. Therefore, connectivity measures cannot be seen as suitable measures for evaluating urban transportation network performance. Moreover, most connectivity measures have been proven or are suspected to give rise to NP-hard problems see Ball(1979). Thus, it would be very hard to calculate exact connectivity measures for real-size urban transportation networks. For a critical survey on the applicability of

different types of measures to urban transportation networks, see Sansó (1993).

Despite the difficulty in finding suitable reliability measures for real size transportation networks, some authors have tried different approaches. Jordan and Turnquist (1979) have studied the reliability from the customer point of view, evaluating the influence of delimiting some service zones in order to increase service reliability. Their work is one of the few articles presenting a mathematical model to incorporate the notion of reliability into transportation planning. Turnquist and Bowman (1980) present a set of simulation experiences to study the effect of the structure of the urban network into the service reliability. For the first time the underlying network structure is related to transportation service reliability. Hagstrom (1983) presented a decomposition algorithm to find several connectivity measures of a small transportation network. The paper concentrated on the algorithm rather than the transportation application. In fact, since the reliability measures treated were connectivity measures, they could only be found for a small network example. Sansó and Soumis (1991) proposed a general framework to evaluate a performance measure for different types of flow networks. The method is based in a partial enumeration of the more probable states of the system and on the evaluation of a performance measure for every one of those states. In order to be able to asses the performance measure for every state, a dynamic assignment problem should be solved. In fact, the failure states are either accidental states or states for which the arc capacity has been reduced.

Even though Sansó and Soumis' approach combines the notions of accident, dynamic assignment and network reliability, that is not the first attempt at incorporating accidents into transportation planning. In fact, in the transportation literature two major field of research have related to this issue. They can be resumed into the following two questions. How does the user behave under the knowledge of the possibility of accidents? And, given an accident, how is the traffic to be assigned during the temporal period of the study?

The first problem is treated in the literature on stochastic equilibrium models (see, for instance, Dial (1977), Sheffi and Powell (1981,1982), Mirchandani and Soroush (1987) (and the references cited therein)). The second problem relates to the literature on dynamic traffic assignment (see, for instance Drissi and Hameda (1992) and Chang,Ho and Wui (1993)).

In this paper we present a model to incorporate the notion of accidents into the traffic assignment process. The model was designed so that it is possible to evaluate a measure of network performance following the general approach proposed by Sansó and Soumis (1991). The next section is devoted to briefly revised the more relevant points described

by Sansó and Soumis. The model and a discussion are presented in section 3.

2 Preliminaries

Let Z be a parameter of network performance in a system. For urban transportation networks, a suitable measure of Z could be the mean travelling time for all the origin-destinations considered in the system. In urban transportation networks, a state can be described by a vector given the arc capacity of all the network links. An accidental state is thus one in which at least one link presents, due to an accident, a reduced capacity.

It is almost impossible to enumerate all possible accidental states of the system. Moreover, even if one considered only a fixed number of capacity reductions, r, there will be $(r+1)^n$ states to be taken into account, n being the number of links in the network. This situation, although more critical in urban transportation networks, given the important number of links and levels of link operation, it is quite common in most types of flow networks. Li and Silvester (1984), Chiou and Li (1986) Lam and Li (1986) and Shier (1988) have all proposed different algorithms based in a partial state enumeration approach. The idea being to enumerate only the most probable states of the system and, for each state, evaluate a suitable performance measure, next, upper and lower bounds on the expected measure of performance could be computed. These type of measures have been named "performability measures" in the reliability literature.

Sansó and Soumis (1991) proposed that, for urban transportation networks, rather than evaluating the most probable states, the contribution of the most important states to the expected performance measure be considered. The most important states could be any state a planner would consider of importance. For instance, in a city island like Montreal, accidents in the bridges are major concern, thus making the states considering such accidents to be among the most important ones.

Within every important state i, it is nevertheless necessary to make an accurate evaluation of Z_i, the performance measure for the state. At the beginning of the section we mentioned that a suitable performability measure would be the mean travelling time for all network origin-destinations. For the perfect state, such a measure can be found using a classical traffic equilibrium package. Nevertheless, the accidental states would ideally need a dynamic traffic assignment model or, at least, an approximate model that would consider the dynamic behavior of the traffic taking into account also the quantity

and time of information given to the users. The next section is devoted to a description of the most basic features of the model.

3 The three-T's model

The objective of the this model is to use a static tool, such as a traffic equilibrium package, to model the highly dynamic behavior of drivers in an accidental state. We consider that before an accident takes place a state of equilibrium exists. This is what we call the first "time", $T1$, of the model.

Immediately after the accident, the beginning of the second "time",$T2$, of the model takes place. $T2$ will itself be discretized into several intervals, called "intervals of information"of equal length δ. The intervals represent how often a fixed percentage of the user get information about the state of the system. Let t mark the beginning of $T2$. Then, information will be provided to the users at discrete times $\tau = t, t + \delta, t + 2\delta, ...$ until we enter a new equilibrium state $T3$. The intervals of information can also be seen as how often the manager of the system get a sample of the state of the system.

The percentage of people informed of the accident for every discrete time τ is $\lambda[\tau]$. We consider that right after the accident nobody has been informed, therefore $\lambda[t] = 0$ and all the users get stuck in the accident. It is considered that the percentage of users informed of the accident increases as time elapses. However, not all informed users will decide to change routes: some will have no choice or will prefer to stay in the accidented route, depending on the information they have gotten on the state of the system from the previous interval . This process is repeated until time $T3$ is reached.

4 Discussion

The three-T's model has been implemented using the traffic assignment package EMME2. Input/output features of the software have been adapted to calculate the performance measure for every T and for every interval of information. The travelling times during the accidental periods are calculated by a simple queuing model developed for the accidented link.

The network used in the project depicted some of the major features of the urban network linking the island of Montreal to the South Shore. Accidents on the four bridges were considered and performance measures were taken. We varied the $T's$ length, the importance and location of the accident, the discretization time, the number of informed people for every interval and the percentage of people deciding to stay in their routes despite the accident. Results were obtained in seconds. Once those results are ponderated by every important state probability, a compound measure of network performance is given.

The model we have presented is approximate, however, we believe that it provides a good tool for the planner to evaluate the importance of accidents in the network taking into account users information and behavior, network structure, level of demand, accident location, and accident probability and duration. This is particularly important in the context of information systems and IVHS.

References

[1] Ball, M.O. 1979. "Computing Network Reliability". *Operations Research* 27(4), 823-838.

[2] Chang, G.L., Ho, P.K. and Wui, C.H. 1993. "A Dynamic System-Optimum Control Model for Commuting Traffic Corridors", *Transportation Research C*, vol1C, no.1, 3-22.

[3] Chiou, S. N. and Li, V.O.K. 1986. "Reliability Analysis of Communication Network with Multimode Components", IEEE Transactions on Selected Areas of Communication, SAC-4, 1156-1161.

[4] Dial, R.B. 1971. "A Probabilistic Multipath Traffic Assignment Model Which Obviates Path Enumeration". *Transportation Research* 5, 83-111.

[5] Drissi-Kaitoumi, O. and Hameda-Benchekroun, A. 1992. "A Dynamic Traffic Assignment Model and Solution Algorithm", *Transportation Science*, 26, 119-128.

[6] Hagstrom, J.N. 1983. "Using the Decomposition-Tree of a Network in Reliability Computation". *IEEE Transactions on Reliability* 32, 71-78.

[7] Jordan, W.C. and M.A. Turnquist. 1979. "Zone Scheduling of Bus Routes to Improve Service Reliability". *Transportation Science*, 13(3), 242-268.

[8] Lam, Y. and Li, V.O.K. 1986. "An Improved Algorithm for Performance Analysis of Networks with Unreliable Components", *IEEE Transactions on Communications*, 34, 496-497.

[9] Li, V.O.K. and Silvester, J.A. 1984. "Performance Analysis of Networks with Unreliable Components", *IEEE Transactions on Communications*, 32, 1105-1110.

[10] Mirchandani, P. and H. Soroush. 1987. "Generalized Traffic Equilibrium With Probabilistic Travel Times and Perceptions". *Transportation Science* 21(3), 133-151.

[11] Sansó, B. and F.Soumis. 1991. "Communication and Transportation Networks Reliability Using Routing Models", *IEEE Transactions on Reliability*, 40(1), 29-38.

[12] Sansó, B. 1992. "Planification des Réseaux de Transport Routier en Fonction de la Notion de Fiabilité", Centre de Recherche sur les Transports, publication CRT847.

[13] Sansó, B. 1993. "Fiabilité et Planification de Réseaux de Transport Routier: l'État de l'Art", Centre de Recherche sur les Transports, publication CRT-930.

[14] Sheffi, Y. and W. Powell. 1981. "A Comparison of Stochastic and Deterministic Traffic Assignment Over Congested Networks". *Transportation Research-B* 15B, 53-64.

[15] Sheffi, Y. and W. Powell. 1982. "An Algorithm For the Equilibrium Assignment Problem with Random Link Times". *Networks* 12, 191-207.

[16] Shier, D.R. 1988. "A New Algorithm for Performance Analysis of Communication Systems" *IEEE Transactions on Communications*, 36, 516-519.

[17] Turnquist, M.A. and L.A. Bowman. 1980. "The effect of Network Structure on Reliability of Transit Service". *Transportation Research-B* 14B, 79-86.

RELATIONSHIP BETWEEN PARKING LOCALIZATION AND TRAFFIC FLOWS IN URBAN AREA

Stefano Carrese*, Stefano Gori** and Tommaso Picano*

ABSTRACT

In the last twenty years, most studies on parking demand distribution have been analyzing user choice criteria, and after determining the utilization level of proposed parking facilities, they individuated optimal pattern of the parking plan, based on optimisation of only two parameters: parking cost and walking distance. Searching for optimal allocation criteria, it has been always done a design verification of different scenarios for dwell demand satisfaction.
The proposed model develops this subject of research, considering a car volume equilibrium on the network. It allows to simulate the choices on a multi-modal network of different user classes, everyone characterized by a typical value of the associated utility to every transportation alternative. The model deals with the simulation problem of the transportation system operation, starting, for several time intervals in which the day is divided, from so many demand matrices O/D as the user classes are, that it intends modeling.
In order to verify the performance of the proposed model, a model application was conducted in the town of Viterbo, the first one whose a usable data base could be obtained. Anyway it wasn't available a realistic demand function and so a fixed demand assignment was only possible.

INTRODUCTION

In the last twenty years, most studies on parking demand distribution have been analyzing user choice criteria, and after determining the utilization level of proposed parking facilities, they individuated optimal pattern of the parking plan. Searching for optimal allocation criteria, it has been always done a design verification of different scenarios for dwell demand satisfaction.

The main criterion adopted for establishing facilities land position, is then their maximum utilization and consequent investment effectiveness. This means maximum personal utility for the user himself. Most studies refer to walking distance and parking cost as main variables in different forms and, although they provide different models, results show the position of the facilities where attraction is higher. User choice is generally taken as to optimize a disutility function, e. g. generalized cost function or walking speed total sum.

From the urbanization point of view, in monocentric structure towns (as most ancient Italian towns) the focus of highest attraction lays in the center, because of the presence of commercial and administrative activities but tourism.

In most European urban areas of wide dimension, the more recent parking policies sustain use of public transport and try to reduce private traffic through fare and law measures. In particular main objectives are:

- to reduce traffic congestion in the historical center, penalizing the use of private cars through limitations to on-street and long-time dwell;
- to avoid the increase of trips in peripheral areas near the urban centers, through the use of Park & Ride;
- to develop public transport competitivety, increasing the level of service (bus lane, traffic management, signal preemption systems, etc.)
- to satisfy parking needs of householders and commercial activities.

From these considerations about the exceedingly sectoral nature of the approaches founded in the state of the art, started the idea of a model capable to face parking problems in urban area, in terms of mobility as a whole system.

CLASSIFICATION OF EXISTING MODELS

The localization problem of a public service, as a hospital, a stadium or a parking, consists mainly in the search for where it will be possible to create a structure that responds to specific predetermined objectives (Tadei 92). It includes also the following components:

* Dipartimento Idraulica Trasporti e Strade - Università di Roma "La Sapienza".
** Dipartimento di Scienze dell'Ingegneria Civile - Terza Università di Roma.

- a single decision maker, usually public, which controls the spatial distribution of the service;
- a demand, based on some user categories , which moves to places where service is provided, to use it;
- an evaluation criterion for service distribution, which usually stands on some measures of user benefit;
- costs of equipment installation and management with budget constraints.

The general problem splits into two sub problems, which provide a first important separation among existing models:

- the allocation sub problem: it means the definition of the trips scheme between user households and localization of service equipment;
- the localization sub problem of service equipment. It means the choice of where equipment will be installed.

Obviously the two sub problems are interrelated and can't be solved separately (fig. 1)

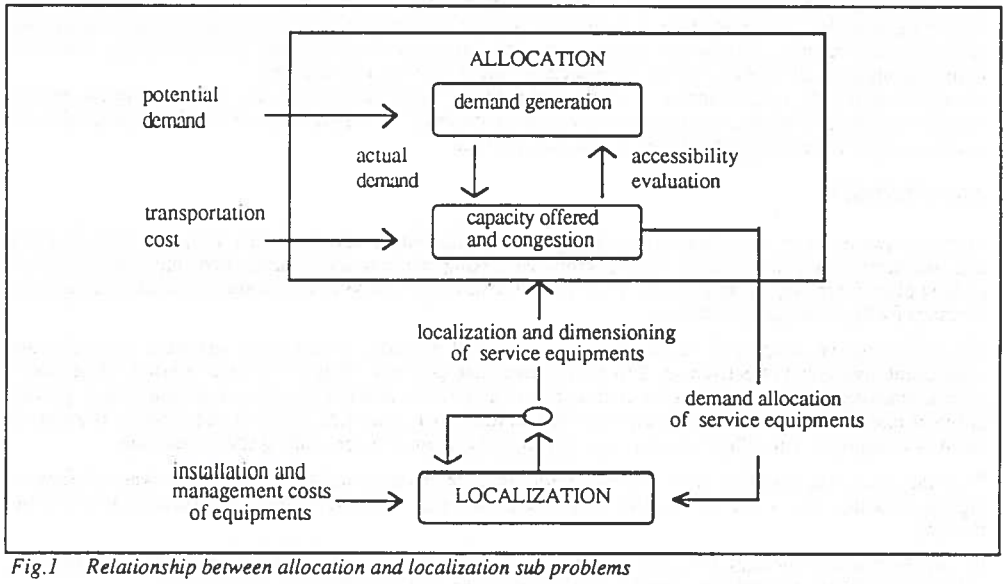


Fig.1 Relationship between allocation and localization sub problems

In the allocation phase, main decision makers are the users (demand). Decision steps are:

- comparison of transportation cost and capacity offered by different service equipment and consequent evaluation of accessibility;
- choice of equipment to reach, and relative trips scheme.

In the localization phase instead, there is usually just one decision maker, an authority that executes the following operation:

- comparison of localization and dimensions of existing service equipment with their allocated demand;
- evaluation of costs and benefits about the effective state of the system and to its possible modification;
- generation of a new plan of localization and dimensions of equipment, including new realizations, changing of management strategy or dismissing of old ones.

Applying this settlement to the parking problem, a classification of existing models has been conducted according to the following criteria:

- optimal localization of parking infrastructure;
- management of attraction related to dwell disposability, which conditions the flow distribution on the road network;
- strategies of fare management and duration limitation.

Two main classes of models have been individuated:

I) localization models, based on procedures of the Operational Research and Regional Sciences;

II) allocation models, which represent user behavior in parking.

The first class is composed by models that, stating the supply and the demand locations on the network, choose the positions of service points as to optimize an objective function, subject to spatial and economic constraints (Odoni & Larson, 1981). The most used in parking studies are:

- the *p-center* models, which minimize the maximum user cost and are mainly used for emergencies services localization, as ambulances and hospitals. These models permit to localize p supply points on a network as to minimize the maximum distance between demand and supply points, as to favorite the more disadvantaged demand point;
- the *p-medians* models, which minimize the average user cost and result to be the more applicable to dwell problems. These models in fact permit to localize p supply points on a network as to minimize the average distance between demand and supply points. Problems which have been solved with these algorithms deal with the localization of equipment for non-emergency service distribution, as schools, transit terminals and public administration offices;
- the *requirements* models deal with the solution of problems in which some service efficiency standards are a priori fixed, and the localization and the number of equipment, necessary to reach the fixed level of service, is found. This kind of models is applicable to p-center and p-median problem, extending their functional capacity.

These models had low degree of development in dwell problems because of the difficulty to describe particular features of parking demand, as fare and law strategies, and their interdependence with traffic mobility. Anyway an application of the p-median problem (Perl & Ho, 1990) keeps in account the elastic nature of the demand, and has the maximization of consumer surplus, as localization's objective: different demand functions represent user behavior in the choice of more convenient service.

Referring to the second class, much more experiments in parking simulation have been conducted, using allocation models.

ALLOCATION MODELS FOR DISTRIBUTION DEMAND

There can be individuated three main categories of models for parking facilities allocation:

1) *behavioral* models, that provide the optimum parking plan among all the possible ones, through an objective function representing the user choice;
2) *gravity* models, that provide the optimal plan through a model of gravitational type;
3) *logit-based* models, that simulate user behavior through the modal split.

Models belonging to the first class represent user criteria of parking choice and differ by the calculation method: some use equilibrium assignment (Beinborn 84, Gray 72, Lablanche 72); other authors use linear programming (Ellis 72, Whitlock 73); others use disaggregated utility function (Austin 73).

All of them analyze user behavior through a disutility function, and so individuate underutilized or insufficient parking. In this class final choice of parking localization is taken after the comparison of the disutility function values assumed by each scenario. This function changes expression depending by the author; this means different sensitivity toward parameters considered, although variables generally considered representing of dwell demand distribution in different parking, are mainly the following:

C_k Parking cost in the k facility;

d_{kj} Walking distance between parking k and destination j;

s_k Capacity of parking k.

These models for a real representation of the phenomenon of parking distribution, are first calibrated by existing situations through analytical regression, then they are applied to the study of effectiveness of parking proposed.

For example Ellis presents a disutility function subject to two constraints:

$$\text{Min} \quad \sum_j \sum_k \sum_q \sum_d Z(j,k,q,d) \times X(j,k,q,d)$$

$$\text{s.t.:} \quad \sum_j \sum_q \sum_d X(j,k,q,d) \le s(a,k) \qquad \forall k$$

$$\sum_k X(j,k,q,d) = T(j,q,d) \qquad \forall (j,q,d)$$

where:

j= index of the final destination zone;

a= index of the arrival period;

d= index of the starting period (d>=a);

$X(j,k,q,d)$= number of user arrived in period a, starting in period d, belonging to group q, destined in zone j and parked in k;

$Z(j,k,q,d)$= disutility for every user;

$s(a,k)$= supply (number of spaces) available in period a in parking k;

$T(j,q,d)$= number of user arrived in period a, starting in period d, belonging to group q, destined in zone j.

Then three different kind of relationship between costs and walking distances, according to user sensitivity, are considered and calibrated by observed data, in order to explicit disutility function.

The use of d_{kj} as main variable brings all these models to localize parking facility near destinations and don't consider the systemic vision of transportation and their effect's analysis on mobility.

The major contribution of these models is their simulating capacity of user behavior, when looking for a free space near his destination: more over they can represent "how" and "how much" user choice is influenced by in force rules and fare measures in urban area. Referring to these, the probabilistic approach by Austin (73) must be outlined for his capability to describe user behavior in parking, defining the value assigned to walking time from parking to destination, according to travel purpose and dwell duration. In Beinborn's model (84) instead, the analysis of in force rules and fare measures and the effects on user distribution among parking facilities is really interesting: it is even considered the phenomenon of illegal dwell, taking in account the following variables:

m_k Expected fee for illegal dwell in parking k;

P_{kt} Probability that a user pays a fee for the illegal dwell of duration t in parking k. It depends by frequency of police controls;

m_{kt} $=P_{kt}*m_k$ Expected fee for the illegal dwell of duration t in parking k.

In the second class of models (Bates 72, Pratt 84) the choice of optimum parking plan is obtained by a zone comparison for the different plans in which a lack or a surplus of supply versus demand has been found. In other words after an appropriate zone split, the values of requested and disposable parking spaces can be calculated and every zone is characterized by a parking state of lack or surplus. Repeat the procedure for all the feasible parking plans, a complete framework of all possible situations of every zone can be obtained.

Bates for example consider that parking demand in a zone i P_i will be equal to the sum of expected (O-D matrix) user demand in zone i D_i , plus demand parking in zone i coming from other zone j A_i , minus parking demand in zone i satisfied in other zones j $\sum_j A_{ij}$.

In formula:
$$P_i = D_i + A_i - \sum_{j=1}^{n} A_{ij}$$

where
$$A_i = \sum_{j=1}^{n} A_{ji} - A_{ii}$$

and
$$A_{ij} = R_i \left[S f_{ij} \Big/ \sum S f_{ij} \right]$$

where:

A_{ij}= number of vehicles that although would like to park in i, go parking in j;

S_j = parking spaces supply in zone j;

R_i = residual demand of zone i = D_i-S_i;

f_{ij}= distribution factor between zone i and j, proportional to the inverse of their distance.

Must be outlined the validity of this model only when parking supply in the whole town is greater than demand.

The third class of models, the so called logit-based (Bifulco 93, Gattuso 92, Hunt & Teply 93) provides a general representation of transport system and then examine different kind of traffic measure. Objective of these models is the identification of a homogeneous class of parking users through the casual utility theory. In fact a nested logit model has been used to calculate the share of private traffic that choose a specific parking facility according to a minimization criterion of times or costs.

In fact in the urban centers with elevated levels of congestion it is almost impossible fulfilling the demand of dwell placing parking in the immediate neighborhood of their destinations, because of the capacity constraints of the network and the lack of available space for the realization of parking.

It needs remembering that an increase of the dwell supply can affect the demand producing an increment of this one (driven demand) that can make inefficient the executed interventions. The dwell supply increase takes back the mobility to a congestion condition with no project situation (Papola 93). As commonly recognized the interventions, provoking modifications of the supply, influence the demand. They cause substantial variations of the existing equilibrium between various modal-splits and transportation alternatives. They also lead to undesired modifications of the mobility that influence the competitivity, the performances and the level of service of the whole transportation system.

In Gattuso's model information required to demand representation are really specific. Although they regard only private users, the knowledge of the following data is requested: destination zone, interval during which the trip is executed, access door to urban center, dwell purpose and the desired transfer opportunity between parking and destination. The specification of the model is further outlined by a logit-based modal split model, which doesn't care of all the user and the transportation modes of the area, but only car users that have to choose the trip mode between parking and destination. In fact the utilization of a generic parking facility means a user total cost equal to the sum of the following terms: the fare (constant value), the access cost (variable with the parking load factor), the parking approach cost (constant value, independent from car volumes on the network) and the transfer cost from parking to destination (constant value, depending on the "as the crow flies" distance). For the representation of changing in access cost, the model executes a phenomenon simulation, based on discreet intervals time progress, which describes the evolution of parking load, that is the number of vehicles inside the parking in the n-th temporal interval. Last, this model, applied only to the verification of alternatives parking plans, designed out of the model, presents a strongly operative characteristic, which more over appears to be too much specific for a good reproducibility of a multi modal context (present in most case studies).

In Bifulco's model the context area is the complete transport network. The consequent demand and supply description consider also information necessary to satisfy transit simulation. He uses a logit model for modal split too, but the aim is the definition of volumes of every user category for every transport mode. Regarding dwell supply modeling, every parking facility is represented by a link cost function characterized by:

- a fixed cost, caused by possible dwell limitation and parking fare structure;
- a variable cost, that considers the search time for a free space.
- a cost of access links to parking, that has a constant time depending on the type of parking link considered (on-street and off-street).

In this model assignment procedure aims to calculate links volumes on the complete network and to assess different mode alternatives, even as integration of them (Park & Ride is explicitly considered). Finally Bifulco's model, through its multi-user and multi-class representation of the transport system, better interprets the need to analyze mobility as a whole system. In particular all that measures that change and influence parking capacity, because of its powerful attraction, condition car volumes on network links. Thus, although this model seems a complete tool for the representation of actual mobility in urban area, for its possible application, even to simplified cases, a heuristic algorithm has been used, whose convergence capacity is elsewhere described.

The model in the followings proposed, tries to be a simulation tool, capable to overpass limits of the before described models: it integrates them, outlining some interesting role of dwell analysis in mobility management in urban area. Major contribution is identified in the refined representation of transport supply (Bifulco 93), in the algorithmic effectiveness (Gattuso 92) and in the utilization of most significant variables (Ellis 72, Beinborn 84).

A MODEL FOR PARKING LOCALIZATION WITH MOBILITY VERIFICATION

An innovative methodology is proposed (Carrese 93, Carrese Gori 93) to consider two macro effect on traffic flows, in the localization test phase of an off-street parking facility.

The first effect, to be considered in the hypothesis of parking demand (steady) smaller than available supply, presents a phenomenon of traffic flows redistribution in the exam area with possible vehicular congestion in

parking neighborhood. Here the local accessibility to the parking facility should be considered, taking care of maintaining the level of service in the access links over a predetermined lower bound (e.g. level of service not worst of the no project situation).

In the second effect the parking demand exceeds available supply: new parking construction allows then to increase the level of satisfied demand. The relaxation of the steady demand constraint increases the number of degrees of freedom: extending the flow redistribution problem to the determination of the increasing demand.

In both cases, a new parking infrastructure is considered as an attraction element (Avarello Carrese 92), like a school or a shopping or business center. In practice, in a regime of weak illegal dwell control, the experience shows an increasing zone dwell capacity due to a new parking construction.

DEMAND MODELING

The proposed model develops this subject of research. It allows to simulate the choices on a multi-modal network of different user classes, everyone characterized by a typical value of the associated utility to every transportation alternative. The model deals with the simulation problem of the transportation system operation starting, for several time intervals in which the day is divided, from so many demand matrices O/D as the user classes are that it intends modeling.

The user classes are characterized not only, by purpose trip, but mostly by the required duration of the dwell after the trip and before executing the possible following trip. This required dwell time influences the utility perceived by the users in decisive way. Especially regarding to auto alternatives because it produces the value of parking cost, that the explicit modeling wants to deal with.

It supposes that the origin/destination matrices will be noted for several classes and will contain the total information of the demand on the all transportation modes. It remains to evaluate in the assignment to the network the aliquot of the demand that will use the different transportation modes available on every pair O/D, according to the total performances of the transportation system.

SUPPLY MODELING

To define transport supply, a detailed description of the multi-modal individual and public networks is required. In particular:

- main viability
- centroids-network connection
- parking links (on and off street)
- pedestrian parking-destination links
- transit links
- transfer links
- waiting links
- parking-transit network connection links
- transit network- destination connection links

The function, representing the different branches described, assumes different pattern that can be generally taken back to a constant component and a variable (t) with flows ($f_{l,m}$ with l = branches and m = modes):

$$C_{l,m} = C^0_{l,m} + a_{lm} * (f_{lm}).$$

For better explanations on these type of functions it refers to other studies (Branston 76, Matsoukis 80, Bifulco 93), here it pays attention on the modeling of the parking phenomena on road and in a plant.

Here parking time-flow functions , that have been used in the simulation model are shown (Fig. 2). It can be seen that for low volume of parking, user has t_0 minutes of time parking anyhow. For values over the parking capacity the user has t_c minutes of time parking obtained adding to the default value t_0, the value of the in force fine M, multiplying its probability to be fined P (Bates 72). In formulas:

$$t_c = t_0 + (M * P) / r,$$

where r is the conversion rate Lit/mins that in this application has been fixed in £1000=2 minutes. The realization of a parking is represented by a step F (fare), starting from the default value of t_0. Different parking capacities are shown as growing steps on volumes scale.

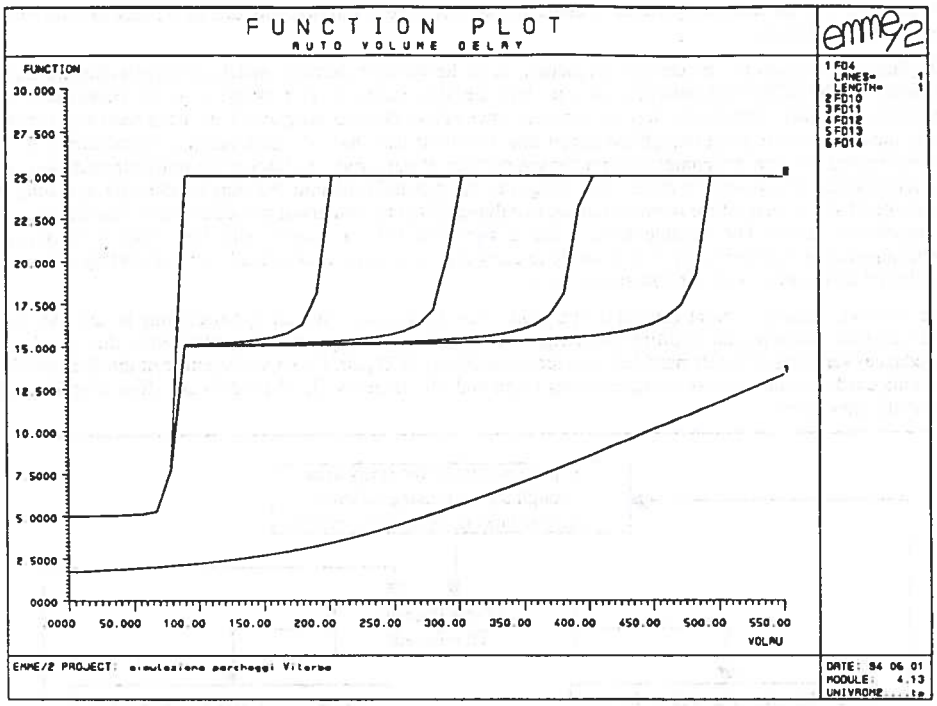


Fig.2 Parking time-flow functions

MULTI-MODAL EQUILIBRIUM

The demand and supply so schematized are inputs of the multi-modal equilibrium assignment model. It is composed by:

- a transit assignment model, based on the concept of optimal (Spiess and Florian 89). In this model the traveler's choice set consists of all feasible strategies. It is evident that, due to the asymmetry inherent to the transit trip, this model can only be solved efficiently by assigning the demands from all origins to a single destination. A non negative time (or cost) is usually used to quantify each trip component, with the exception of "waiting for a vehicle", which is quantified by using the statistical distribution of waiting times for the arrival of the first vehicle of a given transit line at a given stop. The model may be solved by an algorithm which is composed of two parts. In a first pass, from the destination node to all origins, the optimal strategy and the expected total travel times from each node to the destination node are computed. In a second pass, from all origins to the destination, the demand is assigned to the network according to the optimal strategy.
- a demand variable auto assignment model. This model computes the equilibrium auto demand, link flows and travel times by the Partial Linear Approximation Method - PLAM - (Florian 86), which is an improvement of the Linear Approximation Method - LAM - (Frank and Wolfe 56) used for Fixed Demand. The LAM finds, given a current feasible solution, a descent direction by solving a sub problem where the objective function is a linear approximation of the nonlinear objective function of the problem to be solved. Moreover in the PLAM the algorithm may be implemented by requiring only evaluations of the inverse demand function. As it is only required the auto demand function, the inverse demand is obtained numerically, by using the secant method. The PLAM finds, given a current solution, a descent direction, by solving a sub problem where the delay functions are linearized, but the inverse demand functions are not.

When the demand function is not available, it will be possible to use the fixed demand auto assignment, simulating only the above mentioned first effect of car volumes redistribution. In fact, in the following model

application in a small town, only the fixed demand auto assignment will perform, due to the lack of information about demand function.

If the demand function is a mode choice function, as in the variable demand model, the transit demand can be computed as the difference between the total and the auto matrices, and assigned to the transit network. Frequently the transit impedances may be computed once before the auto assignment and have been kept constant during the assignment , even though the transit time functions used have the auto times as parameter. It is also possible to end the auto assignment after a chosen number of iterations, recompute the transit impedances, and then readjust the auto-matrix to reflect the changes in the demand function; this may be done by obtaining the expression of the inverse of the demand function analytically (or by numerical procedure) and then recomputing the impedance values. The variable demand auto assignment is then resumed with the effect of the adjusted transit impedances accounted for. It is possible to carry out these steps automatically, and setting up a bi-modal equilibrium assignment with variable transit times.

If the demand function is more general than a mode choice function, a similar approach may be used to set up more complex multi-modal equilibrium assignment models. This approach is valid since the travel time (impedance) variables of the demand function are separable by O-D pair. Given the structure of the different flow functions used, including those representatives of on and off street dwell, all model's activities (Fig.3) can be seen in the flow chart.

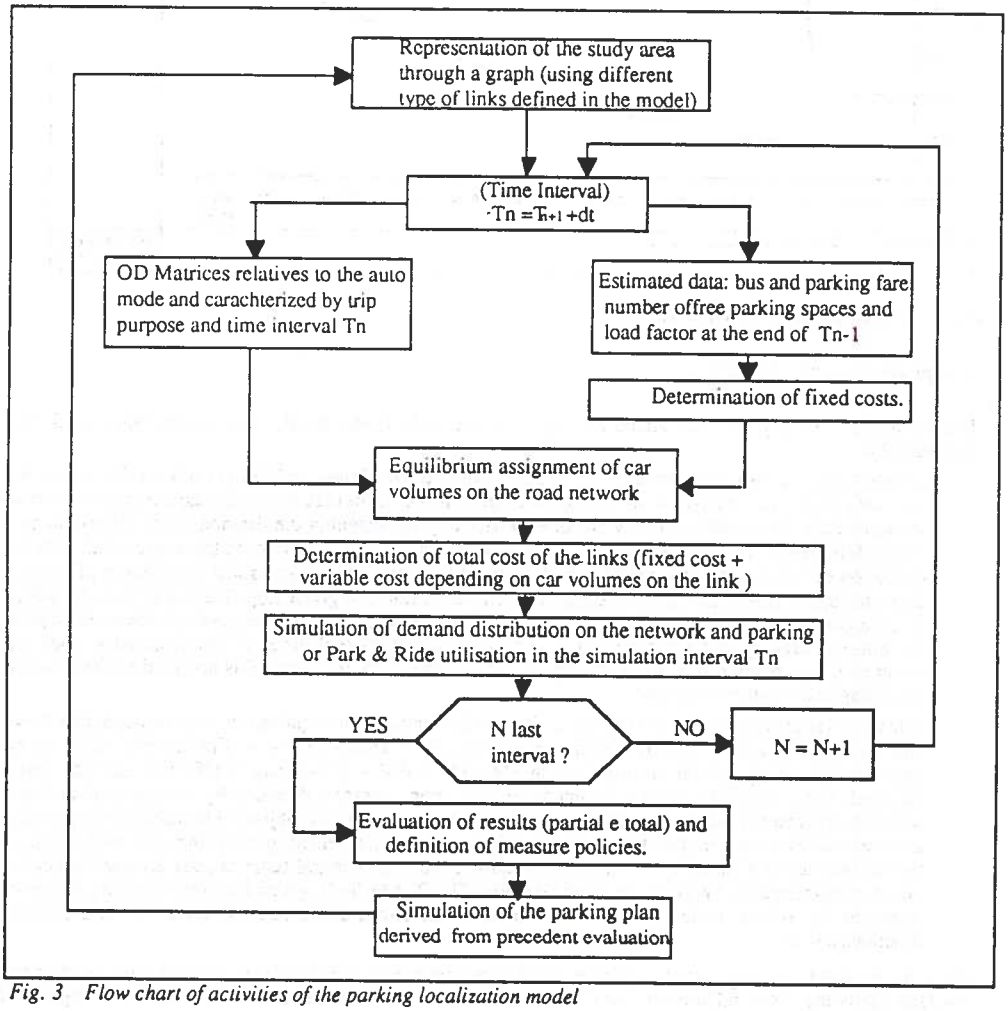


Fig. 3 Flow chart of activities of the parking localization model

From the operative point of view the multi-user equilibrium model proposed, representing explicitly the dwell, gives interesting reliable results. Then it is true that the warranty, on the equilibrium singleness, is only ensured under particular conditions (Daganzo 93, Bifulco 93) for one mode assignment.

Regarding multi-modal models, they haven't great possibilities. For a multi-modal model, with e.g. only one user class, the generalized cost of two transportation modes on the same pair (o,d) must be the same. Besides it needs that the contribution of the congestion cost (transit on board time) on road branches to the generalized cost is the same for all modes. This situation is not realistic because it has been proved that the contributions of cost concerning different transportation modes are weighed differently in the determination of the generalized transportation cost. Further in multi-modal models, the generalized volume on links, is not the algebraic sum of every mode volumes.

In order to execute an application of the model is only necessary to fix a fare policy (changing the coefficient of the parking curves), the budget allocation and the available localization of all parking installations. In fact in all existing cities, it is not possible to realize parking facilities everywhere and of all dimensions. The parking plan design (Fig.4) can be set initially in terms of available areas and maximum dimensions allowed in the areas.

The output of the model is the occupation status of all available parking facilities (on and off street), and car volumes on all links of the road network from which it is easy to find times, costs and all the usual performance indicators. Changing fare policy (modifying coefficients of the flow-costs curves), substantial differences in occupation status and car volumes on network links can be obtained. There are two choice problems concerning:

- which parking to realize;
- optimal fare policy to introduce.

In fact, often it is not possible to realize all the feasible parking facilities, but some priority problems must be faced. However it has to be considered how the proposed model can simulate in correct terms the development of all the traffic system and how it can provide outputs (occupation status, rotation, dwell accumulation, congested links) that allow effective choices.

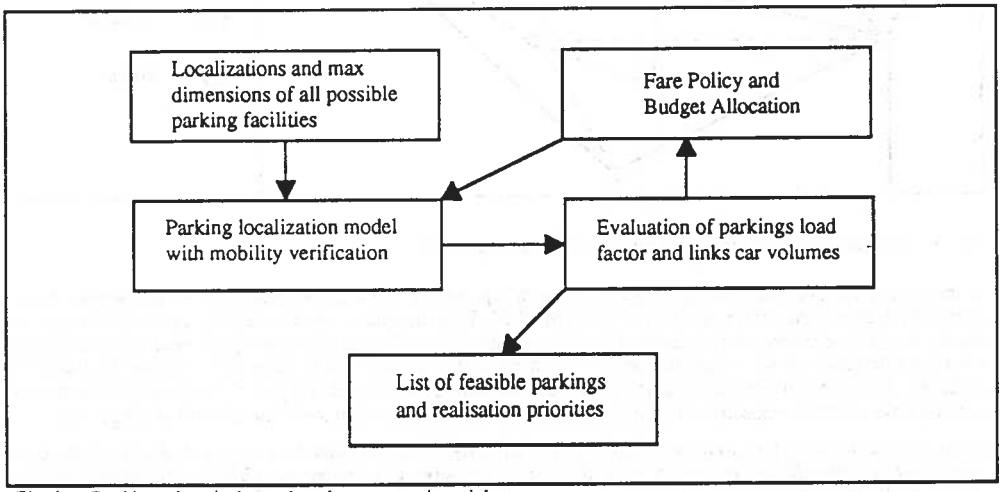


Fig.4 Parking plan design using the proposed model

CASE STUDY OF THE MODEL APPLICATION

In order to verify the performance of the proposed model, some application to real cases of different dimension are foreseen:

1) small size town;
2) medium size town;
3) metropolitan urban area

Next it is presented the results of a study conducted in the town of Viterbo (an example of case 1), the first one whose a usable data base could be obtained. Anyway it wasn't available a realistic demand function and so a fixed

demand assignment was only possible.

The existing public transportation system presents headways greater then 30 minutes, and this condition is absolutely not compatible with an hypothesis of Park & Ride. In fact in this size of town, a realistic hypothesis of P&R exists only in case that It will be possible to realize a specific high frequency system (e.g. minibus), which joint the P&R with the main attractive poles.

Viterbo has about 60,000 inhabitants. The study area, including the historical center inside the ancient walls and the next surroundings, is divided in 23 zones. The transportation network is represented at different levels of aggregation (Fig.5). The network consists of 41 nodes and 266 links. First of all a main viability is characterized by 86 links and a secondary viability is formed by 180 links. This last one also contains the "parking links", that are defined in number of 42. These parking links represented the share of the capacity of the existing or planning dwell in the single zones. These last links represent more than one access to the park from the network, further pedestrian links allow the reach destinations even from the neighboring zones.

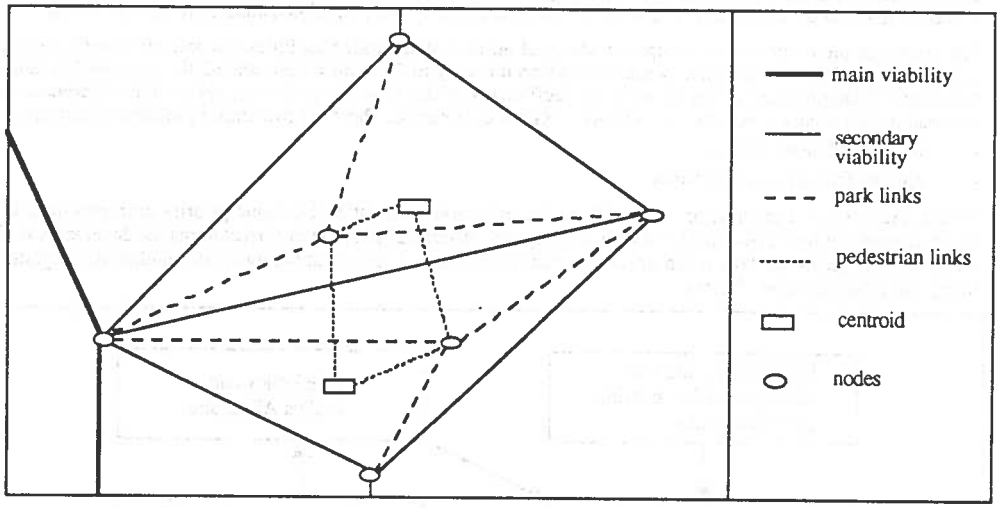


Fig. 5 Network representation: car, parking and pedestrian links.

In the first simulation the existing situation has been represented, through the peak hour matrix which, due to access limitation to the center, takes from 11.20 to 12.20. This simulation regards parking user behavior and the supply function represents the situation of parking availability at the starting of the peak hour, so as it springs from the calculation of the accumulation in every zone of the previous hours flows O/D, divided for the dwell duration.. Then the equilibrium assignment has been performed and results (Figg. 6, 7, 8) show that the most of parking links reach the capacity, while only 3 links exceed it. It means that some users dwell in illegal way.

In the second simulation a parking of 1000 places of capacity, has been added in a very central zone (it has been extracted from the effective parking plan of the town). This parking is connected to the network by two links. It can be easily observed (Figg.9, 10, 11) that there is a volume redistribution of the flows in some links.

Evaluation of results can be conducted only on the basis of car, parking and pedestrian volumes. In fact due to the availability of only accumulation and rotation parking data and due to the lack of a parking duration cluster analysis, a comparison between times wouldn't be interesting without the simulation of a realistic fare policy.

The comparison between the two simulations show the following results: concerning the main viability links, in the project simulation, there are decreases from 2% to 7,5%, with an average decrease of 4,9%. Concerning the parking links, there are four of them (near and over capacity) that show decrease from 1,5% to 10,7% with an average decrease of 4,8%. At last about the pedestrian links the decrease goes from 7,1% to 30,5% and the average decrease is 24,5%.

Finally the model has provided satisfying differences in results according to the different parking policies which can be simulated.

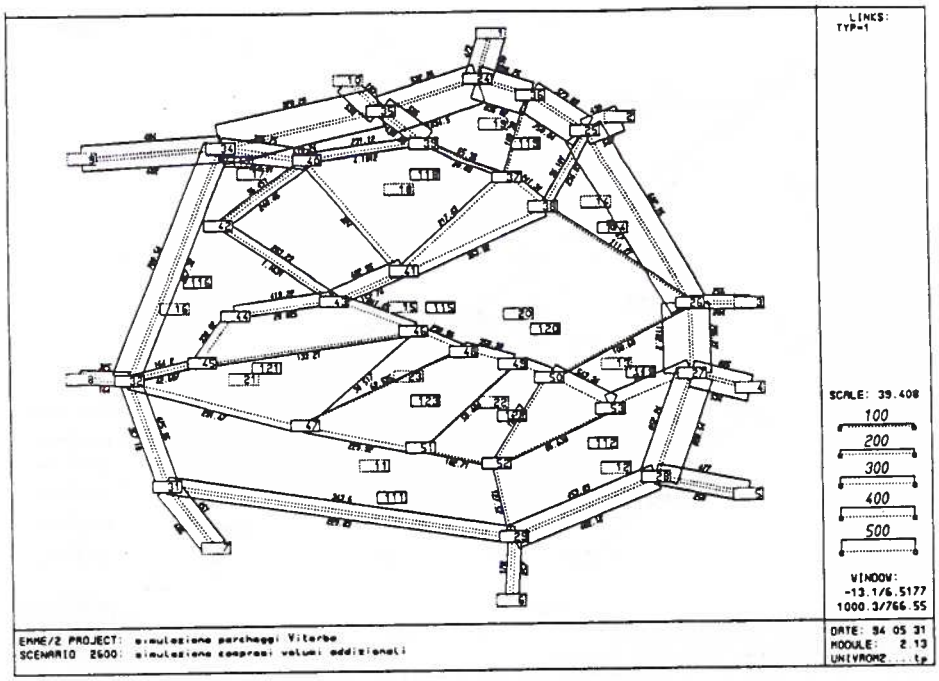


Fig.6 First simulation: actual volumes distribution

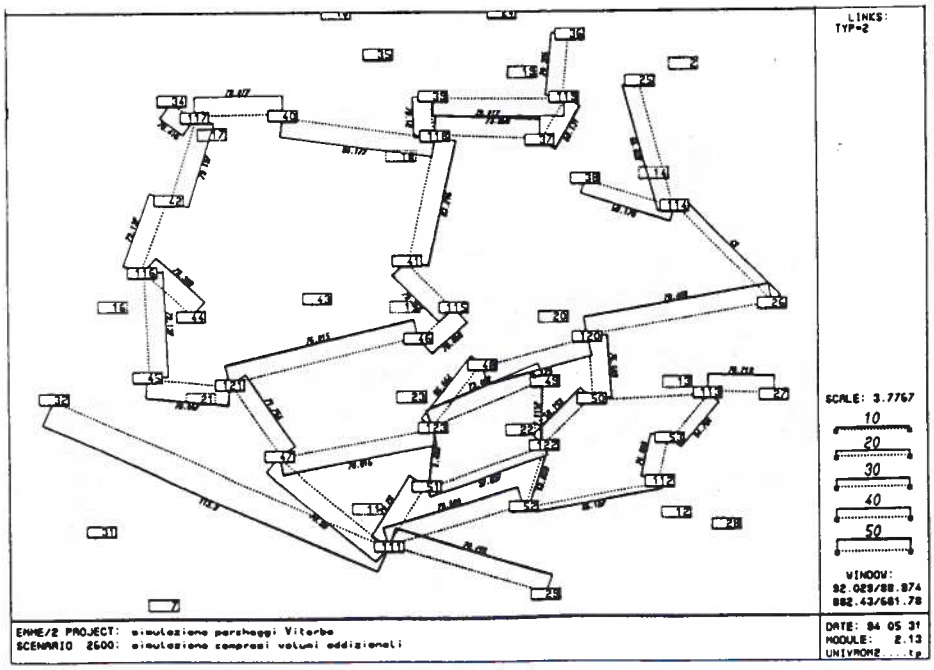


Fig.7 First simulation: actual parking volumes distribution

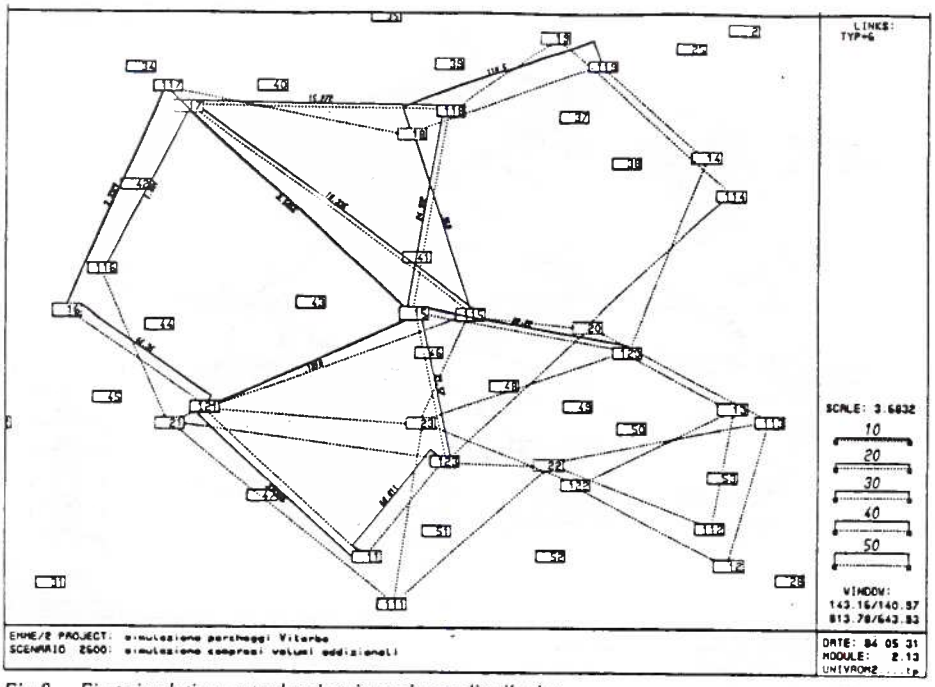


Fig.8 First simulation: actual pedestrian volumes distribution

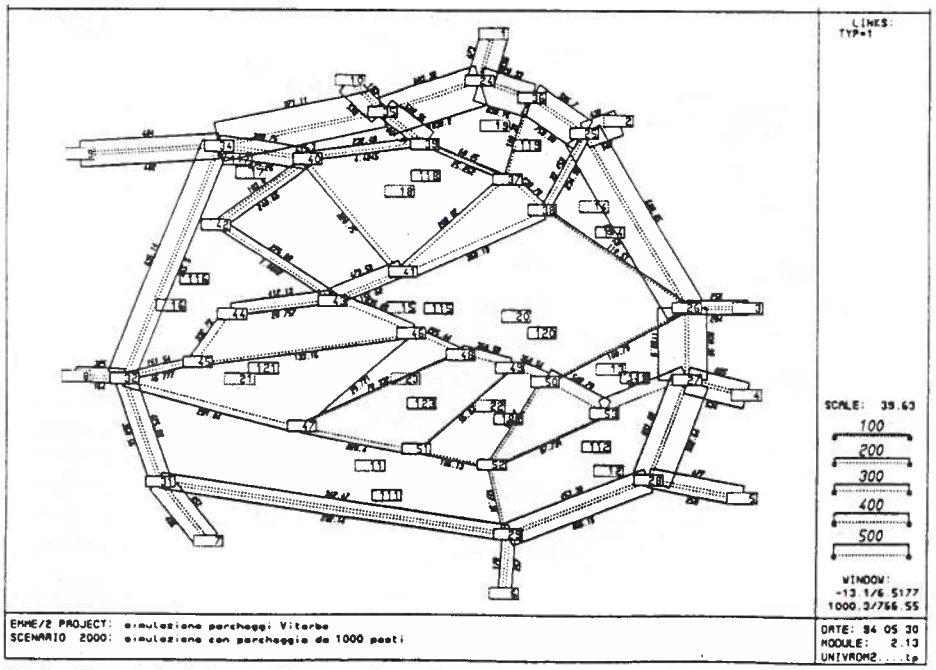


Fig.9 Second simulation: project car volumes distribution

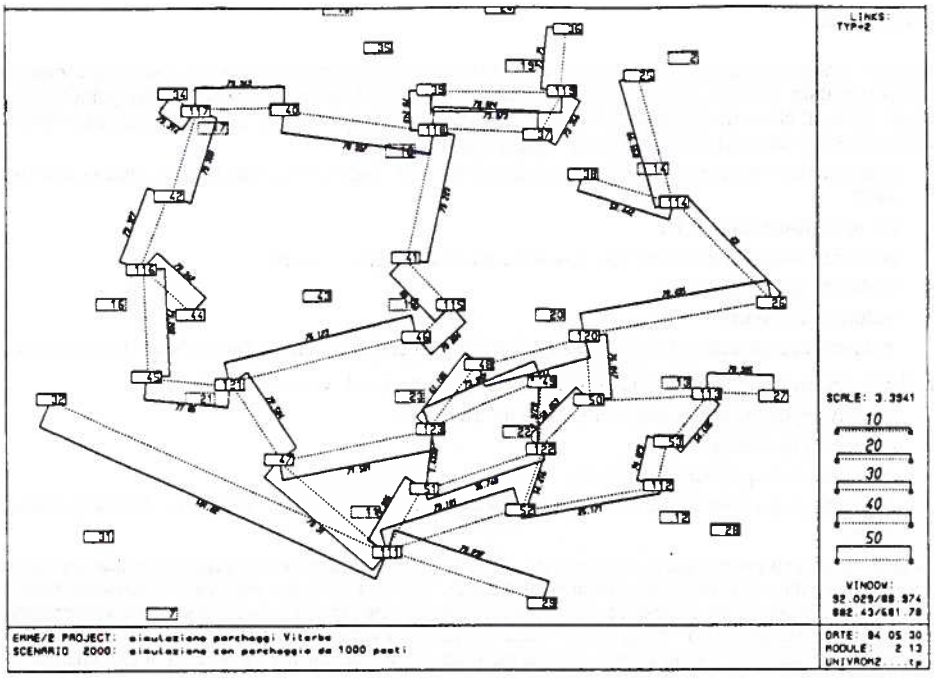


Fig.10 Second simulation: project parking volumes distribution

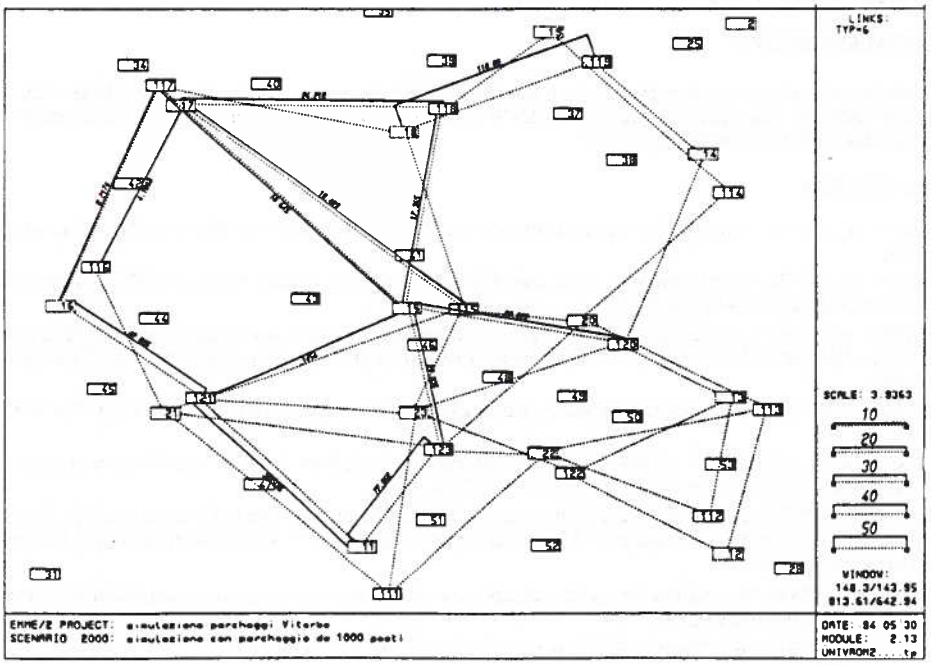


Fig.11 Second simulation: project pedestrian volumes distribution

CONCLUSIONS

This study limits the analysis of the remarkable effects only for those concerning to the system performances of the transportation supply and to the mobility demand. A general agreement exists to this point among the analysts (Bonsall et al. 82, Polak et al. 90), that the models, used traditionally, are often inadequate and partial fairly and that the characteristics of an "ideal" model would have to be:

- mathematics congruence in the representation of the route-choices and of the location choices and kind of dwell;
- use of available data easily;
- explicit modeling of the interaction between demand and supply of dwell;
- explicit temporal approach;
- multi-user approach;
- multi-modal approach and modeling of the demand elasticity at least regarding to the modal choice level .

The "key" variables, considered in the explicit simulation of the dwell, would have to be:

- capacity, typology and location of the dwell infrastructures;
- structure of dwell fare;
- restriction and regulation of the dwell;
- modeling of the illegal dwell and of the parking that lasts more than policies of the dwell limitation allow.

It must be stated that the proposed model is not capable to provide directly which parking facilities to build, but only to describe the occupation status of each, also the car volumes of all the integrated transportation network components. It is interesting to observe that the model contains the ideal "key variables" and also it presents all the principle characteristics of the so called "ideal" model. Only the explicit modeling capability of demand elasticity is not included. A limit of the model is the need of a further clustering in relation to human behavior when dwell duration is the same.

The execution of further model application is foreseen.

AKNOWLEDGEMENTS

The application of the model in the town of Viterbo wouldn't be possible, without the data base relative to demand modeling, which has been provided by ENEA. Authors hope to get more information about demand and parking data base for further developments.

BIBLIOGRAPHY

Austin T. W., (1973). "Allocation of parking demand in a CBD", *Transportation Research Record*, n. 444, pp. 1-8.

Austin T. W., (1973). "Estimation of potential use of peripheral parking for Los Angeles CBD", *Transportation Research Record*, n. 444, pp. 21-26.

Avarello L. e Carrese S. (1992) "Integrazione tra pacchetti esistenti e modelli di simulazione per nuove necessità: il caso del Park & Ride". *13° Corso in: Comunicazioni, Trasporti e Governo della Mobilità*. Capri, giugno 1992.

Bates J., (1972). "A gravity allocation model for parking demand". *Transportation Research Record*, n. 395, pp. 1-5.

Beinborn E. A. e Gur Y. J., (1984). "Analysis of Parking in Urban Centers: Equilibrium Assignment Approach", *Transportation Research Record*, n. 957, pp. 55-62.

Bifulco G.N., (1992-93). "Un approccio Multimodale ai Modelli Matematici per la Pianificazione dei Trasporti con Simulazione esplicita della Sosta", *Università di Napoli "Federico II" Dottorato di Ricerca in Ingegneria Trasporti-VII ciclo.*

Bonsall et al., (1982). "A model of the combined parking location mode split/journey timing decision", *paper of I.T.S. Leeds University* (UK).

Branston D., (1976). "Link Capacity Functions: a review". Transp. Res., Vol 10.

Carrese S. (1993) "La modellistica della domanda di sosta e l'influenza dei parcheggi sui flussi di traffico". *Il*

conferenza italiana degli utenti EMME/2. Roma, giugno 1993.

Carrese S. e Gori S. (1993) "Elementi per un piano parcheggi compatibile con i problemi di congestione urbana". *Seminario su rilievi e modellizzazione del traffico veicolare*. Università di Napoli, Giugno 1993

Daganzo C.F., (1993). "Stochastic Network Equilibrium with multiple vehicle types and asymmetric, indefinite link cost iacobians". Transp. Science, pp. 282-300

Ellis R. H. e Rassam P. R., (1970). "Structuring a system analysis of parking", *Transportation Research Record*, n. 317, pp. 1-13.

Ellis R. H. e Rassam P. R. e Bennet J. C., (1972). "Development and implementation of a parking allocation model", *Transportation Research Record*, n. 395, pp. 5-20.

Florian M., (1986). "Non linear Cost Network Models in Transportation Analysis". Mathematical Programming Study 26, pp.167-196.

Frank M. and Wolfe P., (1956). "An Algorithm for Quadratic Programming". Naval Res. Logist. Q., 3, pp.95-110

Gattuso D., (1992). "Un Modello di Assegnazione Dinamica della Domanda di Sosta. Applicazione per la valutazione del Programma Urbano Parcheggi della città di Reggio Calabria", *Modelli e Metodi per l'analisi ed il Controllo delle Reti di Trasporto*, a cura di *Cascetta E.* e *Salerno G.* S.I.D.T. 1992 Angeli F. Collana Trasporti.

Gray V.O., (1972). "Parking space allocation by computer model", *Transportation Research Record*, n. 395, pp. 21-30.

Hunt J.D. e Teply S., (1993). "A Nested Logit Model of Parking Location Choice", *Transportation Research*, Vol. 27B,n. 4, pp. 253-265.

Matsoukis E.C. (1980). "Speed/flow Relationships in the Central Glasgows Area". Transp. Planning and Tech., Vol. 6, pp. 21-26.

Odoni A.R. e Larson R.C., (1981). "Urban Operation Research", *Prentice-Hall, Inc., Englewood Cliffs, NJ.07632*, pp. 383-480.

Papola N., (1993). "Trasporti Urbani" Atti del dipartimento I.T.S. Trasporti n° 10, *ESAGRAFICA*, Roma.

Perl J. e Ho P.K. (1990). "Public Facilities Location under Elastic Demand", *Transportation Science*, Vol. 24, n. 2, pp. 117-136, (maggio 1990).

Polak J. et al., (1990). "The appllication of CLAMP to the analysis of parking policy in Birmingham City Centre." Proc. of the PTRC Summer Annual Meetings

Spiess H. and Florian M. (1989). "Optimal Strategies: A New Assignment Model for Transit Networks". Transp. Res. B, Vil. 23B, pp. 83-102.

Tadei R., (1992). "Teorie e metodi di localizzazione", *13° Corso: Comunicazioni, Trasporti e Governo della Mobilità. Modelli, Metodi e Strumenti di Supporto alle Decisioni, Capri* (31 maggio - 6 giugno 1992). Politecnico di Torino, Istituto di Analisi dei Sistemi ed Informatica.

Whitlock M., (1973). "Use of linear programming to evaluate alternative parking sites", *Transportation* Research Record, n. 444, pp. 9-19.

Deterministic and Probabilistic Sales-Delivery Man Problems

by
Igor Averbakh*and Oded Berman†

1 Sales-Delivery Man Problem.

Suppose that customers who require some service are situated at the nodes of a network, and service is provided by a server who performs a service tour through all the nodes and returns back to his home location. On-scene service times are equal to zero. In the travelling salesman problem (TSP) the objective is to minimize the length of the service tour; in the delivery man problem (DMP) the objective is to minimize the total waiting time (w.t.) of all the customers. We consider the combination of the two criteria, assuming that the TSP criterion is of primary importance.

Problem 1 (routing sales-delivery man problem (SDMP)). Given home location of the server, find a service tour so as to minimize the total w.t. of customers subject to the constraint that the tour must be optimal for TSP.

Problem 2 (location -routing SDMP) The same as problem 1, but home location is not known and must be found as well.

*University of Toronto, Division of Management and Economics at Scarborough, 1265 Military Trail, Scarborough, Ontario M1C 1A4, Canada.
†University of Toronto, Faculty of Management, Toronto, Ontario M5S 1V4

SDMP is especially important for the types of networks where the TSP has many optimal tours and can be solved in polynomial time. The most important examples of such networks are trees and cactuses (connected networks where each node and each edge belong to at most one cycle).

2 SDMP on Trees and Cactuses.

Let the network be a tree T. For any node v connected components of set $T\backslash\{v\}$ are referred to as $v-$branches.

TSP on a tree is trivial and is solved by any depth-first tour. The number of depth-first tours can be an exponential function of n, where n is the number of nodes. Any depth-first tour can be defined by a rule describing a choice of a nontraversed $v-$branch after visiting any node v.

Let us define *density* of $v-$branch as a ratio between the total number of customers situated on this branch and the total length of the branch.

Proposition 1. Optimal solution to routing SDMP on a tree is defined by the following *density rule*: after visiting any node v the server chooses a nontraversed v-branch (if any) with maximal density.

Consequence. Optimal SDMP tour on a tree can be found in $O(n \log n)$ time for general trees and in $O(n)$ time for trees with bounded vertex degree.

Proposition 2. Problem 2 can be solved with the same complexity as problem 1.

In the case of general service times the problems can be reduced to SDMP with zero service times on an auxiliary tree.

The main difficulty of SDMP on a cactus in comparison with a tree is that for each cycle there are two possible directions of traversing. Straightforward enumeration of all variants is

.not a good approach, since the number of cycles can be $O(n)$ and the number of variants can be exponential. Nevertheless, it is possible to make a decision about direction of traversing for each cycle separately using an auxiliary SDMP subproblem for each cycle.

Proposition 3. Routing and locating-routing SDMP on a cactus can be solved in $O(n \log n)$ time ($O(n)$ for the case of bounded vertex degree).

We consider also multiserver variants of SDMP. Using reduction from scheduling problems it is proved that m-server DMP and $m-$server SDMP on a tree are NP-complete for any fixed $m \geq 2$ and NP-complete in the strong sense if m is variable.

Another class of problems arises if we consider models with probabilistic demand. Suppose that customers generate calls for service with known probabilities.

On any given day the server receives a list of calls from customers waiting for service and it immediately starts a service tour that includes all the calls in the list. A specific multistop location or routing problem is defined by a criterion to be optimized and by a strategy for the server to choose a service tour for each potential list of calls. There are two fundamental strategies: reoptimization strategy where for each realization of the list the server tries to optimize his tour, and a priori optimization strategy, where for each list of calls the sequence of serving customers is uniquely defined by some basic sequence of all potential customers (a priori tour) chosen in advance.

Consider first a prior optimization strategy; then probabilistic DMP and TSP can be formulated as follows (home location of the server is assumed to be known).

Problem 3. (routing probabilistic minisum DMP). Find a priori tour so as to minimize the expected of the total w.t.

Problem 4. (routing probabilistic TSP). Find a priori tour so as to minimize the expected total length of the tour.

Problem 5. (routing probabilistic minimax DMP). Find a priori tour so as to minimize the expectation of maximal w.t.of customers (w.t. of the customer which receives service last).

Probabilistic SDMP can be formulated as follows.

Problem 6 (routing probabilistic minisum SDMP). Find a priori tour optimal for problem 4 so as to minimize the expectation of the total w.t. of customers.

Problem 7 (routing probabilistic minimax SDMP). Find a priori tour optimal for problem 4 so as to minimize the expectation of maximal w.t. of customers.

All these problems are considered on a tree.

It can be proved that the set of optimal solutions for problem 4 on a tree is exactly the set of depth-first tours. Therefore, to define optimal solution to problem 6 or problem 7 one should define a rule how to choose a nontraversed v-branch in a priori tour after visiting any node v. We present such optimal rules for problems 6 and 7 (called respectively *generalized density rule* and *generalized depth rule.*) Algorithms for implementing these rules, which lead to complexity $O(n \log n)$ ($O(n)$ for the case of trees with bounded vertex degree) for both problems are developed. Also, we prove that location-routing versions of problems 6 and 7 can be solved with the same complexity as the routing problems.

If reoptimization strategy is used, then it is assumed that the server can choose an optimal SDMP tour for any realization of the list of calls. The problem is to find optimal home location. We consider two versions of this sales-delivery facility location problem: minisum version, where the expectation of the total w.t. is minimized, and minimax version, where the expectation of the maximal w.t. is minimized. For the minimax version on a tree an algorithm with complexity $O(n^2)$ is developed. Complexity of the minisum version on a tree is still open.

AN EXACT ALGORITHM FOR THE VEHICLE ROUTING PROBLEM BASED ON SET PARTITIONING

N. Christofides[1], E. Hadjiconstantinou[1], A. Mingozzi[2]

(1) *The Management School, Imperial College, University of London, London, UK*

(2) *Department of Mathematics, University of Bologna, Bologna, Italy*

Abstract

We consider the problem in which a fleet of M vehicles stationed at a central depot is to be optimally routed to supply customers with known demands subject to vehicle capacity constraints. This problem is referred as the Vehicle Routing Problem (VRP). In this paper we present an exact algorithm for solving the VRP based on a Set Partitioning formulation of the problem.

We describe a procedure for computing a valid lower bound to the cost of the optimal VRP solution that finds a heuristic solution of the dual of the LP-relaxation of the set partitioning formulation without generating the entire set partitioning matrix. The dual solution obtained is then used to limit the set of the feasible routes containing the optimal VRP solution. The resulting Set Partitioning problem is solved by using a simple branch and bound method.

Computational results are presented for a number of problems derived from the literature. The computational results show the effectiveness of the proposed method.

INCORPORATING FACET-INDUCING INEQUALITIES INTO GRAPHICAL-CONSTRUCT-BASED LAGRANGEAN RELAXATION METHODOLOGIES

AGHA IQBAL ALI

SHIRLEY SHMERLING

The University of Massachusetts at Amherst

May, 1993

Abstract. This paper shows that facet-inducing inequalities can be incorporated into graphical-construct-based Lagrangean relaxation methodologies by including them as dualized constraints. It develops an algorithm that successively identifies additional facet-inducing inequalities and incorporates them into the Lagrangean function. Computational experience shows that the algorithm can resolve duality gaps using relatively few facet-inducing inequalities.

Keywords. Integer Programming, Lagrangean Relaxation, Spanning Trees, Routing Problems, Travelling Salesman Problem, Hamiltonian Paths, Facet-inducing Inequalities.

1. Introduction.

This paper synthesizes polyhedral theory with Lagrangean relaxation in the context of graphical-construct-based models and, thereby, obtains a tight lower bound on a routing problem. Polyhedral theory seeks a representation of the convex hull of the polytope associated with a routing problem using facet-inducing inequalities [Dantzig, Fulkerson & Johnson 1954, 1959; Grötschel & Padberg, 1979; Grötschel & Padberg, 1985; Grötschel & Pulleyblank, 1986; Nadef & Rinaldi, 1991; Boyd & Pulleyblank, 1991]. It has been operationalized in LP-based methodologies [Padberg & Hong, 1980; Crowder & Padberg, 1980; Grötschel 1980; Padberg & Rinaldi, 1987; Grötschel & Holland, 1987; Grötschel & Holland, 1991; Padberg & Rinaldi, 1991]. We demonstrate that it can equivalently be operationalized in Lagrangean dual based methodologies. These latter methodologies have capitalized on computational characteristics of graphical structures that are closely related to a routing

problem. By incorporating facet-inducing inequalities into Lagrangean dual based methodologies, we obtain tight lower bounds while retaining the inherent advantages of such graphical structures.

Graphical-construct-based methodologies for the solution of routing problems use a related subgraph that is amenable to efficient solution. For example, a 1-tree [Held & Karp, 1970, 1971; Volgenant & Jonker, 1982; Malik & Fisher, 1990] retains most of the characteristics of a travelling salesman tour, an m-tree retains most of the characteristics of an m-tour [Ali & Kennington, 1986; Gavish & Srikanth, 1986], and similarly constructed relaxations are possible for other routing problems [Ali & Huang 1988]. A routing problem is modeled by imposing additional routing constraints on such graphical constructs, e.g. the travelling salesman problem is modeled as a 1-tree with additional node-degree constraints. The constructs provide an immediate relaxation obtained by dualizing all additional routing constraints. The Lagrangean relaxation then requires optimization over the subgraph structure, which is achieved using greedy (or quasi-greedy) algorithms. This inherent computational efficiency allows determination of a lower bound using the corresponding Lagrangean dual. Lower bounds so obtained have been held to be strong since the revealed duality gap is in the vicinity of 1 percent [Helbig Hansen & Krarup, 1974; Balas & Toth, 1985; Ali & Kennington, 1986; Gavish & Srikanth, 1986].

We demonstrate the incorporation of facet-inducing inequalities into graphical-construct-based Lagrangean duals in the context of the minimum cost Hamiltonian path problem. In section 2 this problem is defined and the application of polyhedral theory to the problem is stated. In section 3 we describe an algorithmic procedure that solves a sequence of Lagrangean duals each providing a tighter lower bound. The Lagrangean duals are obtained by dualizing node-degree constraints and facet-inducing inequalities. Section 4 reports computational experience with this algorithmic procedure and concluding comments are made in section 5.

2. Prototype for Discussion.

An instance of the minimum (s,t) Hamiltonian path problem (HPP) is given by an undirected graph $\mathbf{G} = (\mathbf{N}, \mathbf{E})$, a cost vector $\mathbf{c}$, and two vertices $s, t \in \mathbf{N} = \{1, \ldots, n\}$. The problem seeks a minimum cost Hamiltonian path between the end-point vertices s and t. We assume the problem has a feasible solution (i.e. that there exists at least one Hamiltonian path between vertices s and t) and to simplify notation let $s \equiv 1$, $t \equiv n$. Defining the problem from a graphical point of view, we want

to find a minimum cost spanning tree in which the degree of each of the vertices $\{i : i \in \mathbf{N}\backslash\{1,n\}\}$ is 2 and, the degrees of vertices 1 and n are, each, 1. The degree requirement for vertex i, denoted ρ_i, is defined as follows:

$$\rho_i = \begin{cases} 1, & \text{if } i = 1 \text{ or } i = n; \\ 2, & \text{if } i = 2,\ldots,n-1. \end{cases}$$

The minimum cost $(1,n)$ Hamiltonian path problem is as stated follows:

$$(HP_T) \quad Z_{HP_T} = \min \sum_{e \in \mathbf{E}} c_e x_e$$

$$s.t. \quad \sum_{e \in \delta(\{i\})} x_e = \rho_i, \qquad i = 1,\ldots,n \ (node\text{-}degree\ constraints)$$

$$\mathbf{x} \ is\ a\ spanning\ tree,$$

where, $\delta(\mathbf{Q})$ is the *cut* induced by $\mathbf{Q} \subset \mathbf{N}$, i.e., $\delta(\mathbf{Q}) = \{(i,j) \in \mathbf{E} : (i \in \mathbf{Q} \text{ and } j \notin \mathbf{Q}) \text{ or } (i \notin \mathbf{Q} \text{ and } j \in \mathbf{Q})\}$.

From the above statement it is easily seen that HPP is closely related to the symmetric travelling salesman problem (STSP) and, as such, embodies the characteristics of various routing problems. Consequently, we use it, as a prototype, to demonstrate our developments.

An alternative formulation of HPP is obtained using polyhedral theory [Grötschel & Padberg, 1985; Padberg & Rinaldi, 1991]. This formulation, which is LP-based, seeks a representation of the convex hull of Hamiltonian paths by using a set of facet-inducing inequalities, each of the form $\mathbf{f}\,\mathbf{x} \leq f_0$. As an example of a facet-inducing inequality consider the following variation of a comb inequality (which has been proven to be a facet-inducing inequality for STSP [Grötschel & Padberg, 1979]). Let $\mathbf{H} \subset \mathbf{N}$ and $\mathbf{W}_i \subset \mathbf{N}$ for $i = 1,\ldots,k$. A comb is a subgraph generated by a set of vertices: $\mathbf{C} = \{\mathbf{H}, \mathbf{W}_1, \ldots, \mathbf{W}_k\}$ with the following properties:

(a) $|\mathbf{H} \cap \mathbf{W}_i| \geq 1$, $i = 1,\ldots,k$,

(b) $|\mathbf{W}_i \backslash \mathbf{H}| \geq 1$, $i = 1,\ldots,k$,

(c) $2 \leq |\mathbf{W}_i| \leq n-2$, $i = 1,\ldots,k$,

(d) $\mathbf{W}_i \cap \mathbf{W}_j = \emptyset$, $i \neq j$,

(e) If $\varphi(\mathbf{H})$ is even (odd) then k is odd (even) and at least 3 (2),

where, $\varphi(\mathbf{H})$ denotes the number of end-point vertices that are in $\mathbf{H}$ and is defined as follows:

$$\varphi(\mathbf{H}) = \begin{cases} 0, & \text{if } 1,n \notin \mathbf{H}; \\ 1, & \text{if } 1 \in \mathbf{H}, n \notin \mathbf{H} \text{ or } 1 \notin \mathbf{H}, n \in \mathbf{H}; \\ 2, & \text{if } 1,n \in \mathbf{H}. \end{cases}$$

For a given comb $\mathbf{C} = \{\mathbf{H}, \mathbf{W}_1, \ldots, \mathbf{W}_k\}$ the following comb inequality represents a facet of the minimum cost Hamiltonian path polytope:

$$\sum_{e \in \mathbf{E}(\mathbf{H})} x_e + \sum_{i=1}^{k} \sum_{e \in \mathbf{E}(\mathbf{W}_i)} x_e \leq |\mathbf{H}| + \sum_{i=1}^{k} (|\mathbf{W}_i| - 1) - \left\lceil \frac{k + \varphi(\mathbf{H})}{2} \right\rceil = s(\mathbf{C}),$$

where, $s(\mathbf{C})$ is referred to as the size of the comb $\mathbf{C}$, and $\mathbf{E}(\mathbf{Q})$ is the set of clique edges of $\mathbf{Q} \subset \mathbf{N}$, i.e. $\mathbf{E}(\mathbf{Q}) = \{(i,j) \ : \ (i \in \mathbf{Q} \ and \ j \in \mathbf{Q})\}$. Expressions for other inequalities (subtour-elimination constraints, and variations of 2-matching and clique-tree inequalities) are similarly obtained and can be shown to be facet-inducing inequalities for HPP as done for STSP in Grötschel & Padberg [1979] and Grötschel & Pulleyblank [1986]. Denoting the set of all facet-inducing inequalities by F_n, HPP is modeled as the following linear program:

$$\begin{aligned}
(HP_C) \quad & Z_{HP_C} = \min \sum_{e \in \mathbf{E}} c_e x_e \\
s.t. \quad & \sum_{e \in \delta(\{i\})} x_e = \rho_i, \quad i = 1, \ldots, n \quad (node\text{-}degree\ constraints) \\
& \mathbf{f}\,\mathbf{x} \leq f_0, \quad \forall (\mathbf{f}, f_0) \in F_n \quad (facets\text{-}inducing\ inequalities) \\
& 0 \leq \mathbf{x} \leq 1 \\
& \mathbf{x} \in \Re^{|\mathbf{E}|}
\end{aligned}$$

In the next section we show how solution methodologies based on the graphical-construct-based model, HP_T, are enhanced using facet-inducing inequalities.

3. Synthesis of Facet-inducing Inequalities and Graphical-construct-based Lagrangean Relaxation.

Graphical-construct-based methodologies have, hitherto, not explicitly utilized facet-inducing inequalities. The set of all subtour-elimination constraints, denoted $\mathcal{S}_n$, is implicit in the related subgraph structure (i.e. spanning tree). However, other types of facet-inducing inequalities, $\mathcal{F}_n = F_n \backslash \mathcal{S}_n$, such as comb and clique-tree inequalities, have not been incorporated into graphical-construct-based methodologies.† The proposed incorporation is based on the following, inherently

† Facet-inducing inequalities have been applied within an LP-based Lagrangean relaxation framework [Smith, Meyer & Thompson,1990].

redundant, formulation:

$$(HP_G) \quad Z_{HP_G} = \min \sum_{e \in \mathbf{E}} c_e x_e$$

$$s.t. \quad \sum_{e \in \delta(\{i\})} x_e = \rho_i, \quad i = 1, \ldots, n \quad (node\text{-}degree\ constraints)$$

$$\mathbf{f}\,\mathbf{x} \leq f_0, \quad \forall (\mathbf{f}, f_0) \in \mathcal{F}_n \quad (facet\text{-}inducing\ inequalities)$$

$$\mathbf{x} \ is\ a\ spanning\ tree.$$

By dualizing both the node-degree constraints and the set of facet-inducing inequalities, $\mathcal{F}_n$, the resulting Lagrangean subproblem is a minimum spanning tree problem; a problem that has the integrality property [Nemhauser & Wolsey, 1988], i.e. all extreme points of the convex hull of the set of spanning trees are integral. Associating the Lagrange multiplier $\phi_r \geq 0$, with the rth facet-inducing inequality, $\mathbf{f}^r\mathbf{x} \leq f_0^r$, and the Lagrange multiplier u_i with the ith node-degree constraint, the Lagrangean relaxation is stated as follows:

$$(\overline{L}(\mathbf{u}, \Phi)) \qquad \overline{Z}_L(\mathbf{u}, \Phi) = \min \sum_{e \in \mathbf{E}} c_e x_e + \sum_{i=1}^{n} u_i \left(\sum_{e \in \delta(\{i\})} x_e - \rho_i \right) + \sum_{\{r\,:\,(\mathbf{f}^r, f_0^r) \in \mathcal{F}_n\}} \phi_r (\mathbf{f}^r \mathbf{x} - f_0^r)$$

$$s.t. \quad \mathbf{x} \ is\ a\ spanning\ tree$$

The Lagrangean dual is stated as:

$$(\ \overline{LD}\ ) \qquad \overline{Z}_{LD} = \max_{\Phi \geq 0} \overline{Z}_L(\mathbf{u}, \Phi)$$

Due to the integrality property the optimal objective value of the Lagrangean dual obtained by dualizing both the node-degree constraints and all facet-inducing inequalities, $\mathcal{F}_n$, is equal to the optimal value of an LP relaxation that includes all facet-inducing inequalities, F_n. That is, $\overline{Z}_{LD} = Z_{HP_C} = Z_{HP_T} = Z_{HP_G}$.

As in the LP context, it is not necessary to identify and include all facet-inducing inequalities associated with HPP; it suffices to explicitly define only a subset $\mathcal{F}^m \subset \mathcal{F}_n$. We develop an algorithm that is based on the solution of a sequence of Lagrangean duals, LD^0, $LD^1, \ldots, LD^m \ldots$, where the sets of corresponding dualized facet-inducing inequalities are $\mathcal{F}^0$, $\mathcal{F}^1, \ldots, \mathcal{F}^m, \ldots$. Initially, the set of facet-inducing inequalities is null (i.e. $\mathcal{F}^0 = \emptyset$ and the Lagrangean dual, LD^0, has only the node-degree constraints dualized). Each successive Lagrangean dual incorporates additional facet-inducing inequalities, that is $\mathcal{F}^m \subset \mathcal{F}^{m+1}$, and therefore yields a tighter bound, i.e. $Z_{LD}^0 \leq Z_{LD}^1 \leq \ldots \leq Z_{LD}^m \ldots \leq \overline{Z}_{LD}$. Obviously, the size of the vector Φ is different for successive Lagrangean duals. In particular, the number of Lagrange multipliers increases as the number of dualized facet-inducing inequalities increases.

To simplify the notation we denote the set of dualized facet-inducing inequalities for any Lagrangean dual by $\overline{\mathcal{F}}$. The cardinality of $\overline{\mathcal{F}}$ is denoted by α (i.e. $\alpha = |\overline{\mathcal{F}}|$) so that $\Phi \in \Re_+^{\alpha}$. Since a dualized facet-inducing inequality corresponds to a sub-structure, such as a comb or a clique-tree, it takes the form $\sum_{e \in \mathbf{E}_r} f_e^r \; x_e \leq s_r$, where $\mathbf{E}_r$ is the set of edges in the particular sub-structure. For example, for a comb $\mathbf{C} = \{\mathbf{H}, \mathbf{W}_1, \ldots, \mathbf{W}_k\}$, $\mathbf{E}_r = \mathbf{E}(\mathbf{H}) \cup \{\cup_{i=1}^{k} \mathbf{E}(\mathbf{W}_i)\}$ and s_r is the size of the comb. When $|\mathbf{H} \cap \mathbf{W}_i| = 1 \; \forall i = 1, \ldots, k$, $f_e^r = 1 \; \forall e \in \mathbf{E}_r$ and s_r is the maximum number of edges, from among $\mathbf{E}_r$, that may be present in a Hamiltonian path. The Lagrangean dual, LD^m, is then stated as follows:

$$(\; LD^m \;) \qquad Z_{LD}^m = \max_{\Phi \geq 0} Z_L^m(\mathbf{u}, \Phi)$$

where

$$Z_L^m(\mathbf{u}, \Phi) = \min \sum_{e \in \mathbf{E}} c_e x_e + \sum_{i=1}^{n} u_i \left(\sum_{e \in \delta(\{i\})} x_e - \rho_i \right) + \sum_{r=1}^{\alpha} \phi_r \left(\sum_{e \in \mathbf{E}_r} f_e^r \; x_e - s_r \right)$$

$$s.t. \qquad \mathbf{x} \textit{ is a spanning tree}$$

Central to the development of the sequence $\{LD^m\}$ is the ability to augment $\overline{\mathcal{F}}$ by additional facet-inducing inequalities, $\mathcal{F}'$. Such inequalities, e.g. comb and clique-tree, are based on a solution that satisfies all node-degree constraints. Therefore, the identification of $\mathcal{F}'$ is possible if we have a solution that satisfies all dualized facet-inducing inequalities, $\overline{\mathcal{F}}$, as well as all node-degree constraints. In a companion paper [Ali & Shmerling, 1993] it is shown that such a solution is obtained by solving the Lagrangean dual LD^m to optimality. This solution, denoted $\hat{\mathbf{x}} \in \Re_+^{|\mathbf{E}|}$, is a convex combination of minimum spanning trees at the optimal dual solution to LD^m, $(\overline{\mathbf{u}}^m, \overline{\Phi}^m)$; it is either a Hamiltonian path or else violates facet-inducing inequalities $\mathcal{F}'$ that can be identified using existing algorithms [Padberg & Rao, 1982; Padberg & Grötschel, 1985; Padberg & Rinaldi, 1990; Grötschel & Holland 1991]. At each iteration m of the algorithm given below, the Lagrangean dual is solved to optimality to yield $\hat{\mathbf{x}}$. If $\hat{\mathbf{x}}$ violates additional facet-inducing inequalities, $\mathcal{F}'$, then $\overline{\mathcal{F}}$ is augmented to define a new and tighter Lagrangean dual, LD^{m+1}.

Algorithm: Dualized Facets

Step 1: (Initialization) $m \leftarrow 0$, $\overline{\mathcal{F}} \leftarrow \emptyset$

Step 2: Optimally solve LD^m to obtain $(\overline{\mathbf{u}}^m, \overline{\Phi}^m)$ and $\hat{\mathbf{x}} \in \Re_+^{|\mathbf{E}|}$

Step 3: If $\hat{\mathbf{x}}$ is integral, stop - $\hat{\mathbf{x}}$ is the optimal Hamiltonian path.

Step 4: (facet identification)

4.1: If $\hat{\mathbf{x}}$ violates additional inequalities, $\mathcal{F}' \not\subset \overline{\mathcal{F}}$ then

$\overline{\mathcal{F}} \leftarrow \overline{\mathcal{F}} \cup \mathcal{F}'$,

$u_i \leftarrow \overline{u}_i^m, \quad \forall i = 1, \ldots, n,$

$\phi_r \leftarrow \overline{\phi}_r^m, \quad \forall r = 1, \ldots, \alpha,$

$\phi_r \leftarrow 0, \quad \forall r = (\alpha+1), \ldots, (\alpha + |\mathcal{F}'|),$

$\alpha \leftarrow \alpha + |\mathcal{F}'|,$

$m \leftarrow m + 1$, go to step 2.

4.2: Stop, no violated facet-inducing inequalities were found.

The above algorithm terminates when an optimal $(1, n)$ Hamiltonian path is obtained, or if a violated facet-inducing inequality cannot be identified. The latter can occur since current facet identification algorithms do not guarantee identification of all comb and clique-tree inequalities. Moreover, even if all comb and clique-tree inequalities are not violated, there may be some other, hitherto unknown, classes of facet-inducing inequalities that are violated [Fleischmann, 1988].

Previously used graphical-construct-based methodologies do not incorporate facet-inducing inequalities, $\mathcal{F}_n$; they attempt to solve LD^0 only for the purpose of obtaining a lower bound. In the above algorithm, progressively stronger relaxations are defined by the sequence of Lagrangean duals $\{LD^m\}$. The duality gap at iteration m is explained, at least partially, by the violated facet-inducing inequalities, $\mathcal{F}' \not\subset \overline{\mathcal{F}}$. In this sense, the algorithm iteratively resolves the duality gap, i.e., $Z_{HP_G} - Z_L^m(\mathbf{u}, \Phi) > Z_{HP_G} - Z_L^{m+1}(\mathbf{u}, \Phi)$.

4. Computational Experience.

The algorithmic procedure of the previous section is illustrated using a graph with 50 vertices and 677 edges. We seek a minimum cost Hamiltonian path between vertices 1 and 50. The initial Lagrangean relaxation (at $\mathbf{u} = 0$, $\overline{\mathcal{F}} = \emptyset$), yields a minimum spanning tree, illustrated in Figure 1, and has value $Z_L^0(0) = 531$. The optimal value of LD^0 is 634. The sub-graph given in Figure 2 corresponds to $\hat{\mathbf{x}}$, at optimality of LD^0 (the edges, e, have value $\hat{x}_e = 1.0$, except those labelled $\frac{1}{2}$ which have $\hat{x}_e = 0.5$). This solution, which satisfies all node-degree constraints, reveals the following

two comb inequalities to be violated:

$$\sum_{e \in \mathbf{E}(\{17,20,45\})} x_e + x_{19,20} + x_{14,17} + x_{18,45} \leq 4 \qquad (f1)$$

$$\sum_{e \in \mathbf{E}(\{1,25,43\})} x_e + x_{2,25} + x_{26,43} \leq 3 \qquad (f2)$$

These inequalities are incorporated into the Lagrangean function (i.e. $(f1)$, $(f2)$ are appended to $\overline{\mathcal{F}}$) to define the Lagrangean dual LD^1. The optimal value of LD^1 is 635.75. The sub-graph given in Figure 3 corresponds to $\hat{\mathbf{x}}$, at optimality of LD^1 (again, all edges, e, have value $\hat{x}_e = 1.0$, except those whose value appears beside the edge). This solution, obviously, satisfies all node-degree constraints as well as inequalities $(f1)$ and $(f2)$. It reveals the following two comb inequalities to be violated:

$$\sum_{e \in \mathbf{E}(\{9,10,14,15,17,18,20,34,35,37,40,45\})} x_e + x_{40,39} + x_{15,13} + x_{18,41} + x_{20,19} + x_{14,12} + x_{9,8} + x_{10,11} \leq 15 \qquad (f3)$$

$$\sum_{e \in \mathbf{E}(\{21,22,46\})} x_e + x_{21,47} + x_{19,46} + x_{22,23} \leq 4 \qquad (f4)$$

Comb inequalities, $(f3)$ and $(f4)$ are incorporated into the Lagrangean function (i.e. $(f3)$, $(f4)$ appended to $\overline{\mathcal{F}}$) to define the Lagrangean dual LD^2. The optimal value of LD^2 is 637. The obtained solution $\hat{\mathbf{x}}$, given in Figure 4, is integral and thus is an optimal solution to the primal problem. We see that the duality gap at LD^0 is entirely resolved by four comb inequalities.

Table 1 summarizes our computational experience with the developed algorithm.† For each problem, we report the problem ID, the number of vertices, the edge density and the optimization history. Test problems 1-4 have been derived from TSPs in the literature; each of the problems seeks a minimum cost Hamiltonian path between two cities that are adjacent in the optimal tour for the TSP. Problems 5-12 are randomly generated with 50 vertices and edge density about 55%.

For each iteration m, Table 1 reports the number of dualized facet-inducing inequalities and, the initial and final values of the Lagrangean function. It is evident, as established in the literature, that LD^0 obtains a tight bound on the primal problem. Problems 1, 3, 4, 5, 7, 8, 9, 11, and 12 terminate with an optimal path. Almost all of the identified facet-inducing inequalities are of the 2-matching type. Problems 1-4 obtain a bound that is equivalent to that reported in the literature [Grötschel & Holland, 1991]. Problem 1 does not require the incorporation of any facet-inducing inequality, problem 2 requires 18 and, problems 3 and 4 require 2 each.

† The algorithm has been implemented in FORTRAN on a 486-based PC under the 640K limitation.

TABLE 1. Summary of Computational Experience

PROBLEM CHARACTERISTICS							
No.	ID	Vertices	Density	m	$\lvert\overline{\mathcal{F}}\rvert$	Z_{LD}^{m-1}†	Z_{LD}^{m}
1	DF42[a]	42	1.00	0	0	591.00	670.00[e]
2	GH48[b]	48	1.00	0	0	4082.00	4889.00
				1	4	4889.00	4912.00
				2	8	4912.00	4943.00
				3	10	4943.00	4952.50
				4	11	4952.50	4955.25
				5	12	4955.25	4960.75
				6	14	4960.75	4967.50
				7	16	4967.50	4969.50
				8	18	4969.50	4971.00
3	HK48[c]	48	1.00	0	0	9905.00	11361.50
				1	2	11361.50	11378.00[e]
4	KT57[d]	57	1.00	0	0	10835.00	12734.00
				1	2	12734.00	12749.00[e]
5	R02	50	0.55	0	0	531.00	634.00
				1	2	634.00	635.75
				2	4	635.75	637.00[e]
6	R05	50	0.55	0	0	536.00	642.00
				1	4	642.00	643.50
				2	6	643.50	645.00
				3	8	645.00	646.00
				4	10	646.00	647.00
				5	12	647.00	647.13
7	R06	50	0.56	0	0	600.00	718.00
				1	2	718.00	719.33
				2	4	719.33	719.66
				3	6	719.66	720.00[e]
8	R08	50	0.56	0	0	541.00	625.00
				1	2	625.00	625.75
				2	4	625.75	626.00[e]
9	R11	50	0.55	0	0	5865.00	7038.50
				1	2	7038.50	7041.00[e]
10	R14	50	0.56	0	0	6342.00	7110.00
				1	2	7110.00	7117.00
				2	4	7117.00	7124.50
11	R16	50	0.56	0	0	6204.00	7727.80
				1	2	7727.80	7728.00[e]
12	R51	50	0.55	0	0	6825.00	8191.50
				1	2	8191.50	8195.00[e]

†For $m = 0$, $Z_{LD}^{m-1} \equiv Z_L^0(\mathbf{0})$.

[a] HPP between cities 23 and 24 ($c_{23,24} = 29$) of TSP from Dantzig, Fulkerson & Johnson, 1954.

[b] HPP between cities 1 and 29 ($c_{1,29} = 70$) of TSP from Grötschel & Holland, 1991.

[c] HPP between cities 34 and 36 ($c_{34,36} = 83$) of TSP from Held & Karp, 1962.

[d] HPP between cities 15 and 26 ($c_{15,26} = 206$) of TSP from Karg & Thompson, 1964.

[e] Optimal solution to primal is obtained.

5. Concluding Comments.

This paper enhances graphical-construct-based methodologies by successively identifying facet-inducing inequalities and incorporating them into the Lagrangean function. It extends the state-of-the-art of such methodologies in as far as the resolution of the duality gap is concerned. The primary significance of resolving the gap using facet-inducing inequalities is not only that the revealed gap is explained by (or due to) facet-inducing inequalities but also that optimal solutions can be obtained without resorting to enumeration. The inherent advantage afforded by retaining graphical structures, while incorporating the power of facet-inducing inequalities, is that relatively few such inequalities are required.

References

A.I. Ali and J.L. Kennington. 1986. The Asymmetric M- travelling Salesmen Problem: A Duality Based Branch-and-bound Algorithm. *Discrete Applied Mathematics*, **13**, pp. 259-276.

A.I. Ali and C.H. Huang. 1988. Graphical-Structure-Based Models for Routing Problems. *International Journal of Systems Science* **19**, pp. 1667-1686.

A.I. Ali and S. Shmerling. 1993. A graphical-construct-based algorithm for optimal solutions to a Lagrangean dual of the Hamiltonian path problem. Working paper. SOM, University of Massachusetts at Amherst, Amherst, MA.

E. Balas and P. Toth. 1985. Branch and bound methods, in the Travelling Salesman Problem. Lawler, et al., pp. 371-378.

S.C. Boyd and W.R. Pulleyblank. 1991. Optimizing over the subtour polytope of the travelling salesman problem. *Math. Programming* **49**, pp. 163-187.

H. Crowder and M.W. Padberg. 1980. Solving large-scale travelling salesman problem to optimality. *Management Science* **26**, pp. 495-509.

G.B. Dantzig, D.R. Fulkerson and S.M. Johnson. 1954. Solutions of a large-scale travelling salesman problem. *Operations Research* **2**, pp. 393-410.

G.B. Dantzig, D.R. Fulkerson and S.M. Johnson. 1959. On a linear-programming, combinatorial approach to the travelling salesman problem. *Operations Research* **7**, pp. 58-66.

B. Fleischmann. 1988. A new class of cutting planes for the symmetric travelling salesman problem. *Mathematical Programming* **40**, pp. 225-246.

B. Gavish and K. Srikanth. 1986. An Optimal Solution Method For Large-Scale Multiple Travelling Salesmen Problems. *Operations Research* **34** pp. 698-717.

M. Grötschel. 1980. On the symmetric travelling salesman problem: solution of a 120-city problem. *Math. Programming Study* **12**, pp. 61-77.

M. Grötschel and O. Holland. 1987. A cutting plane algorithm for minimum perfect 2-matchings. *Computing* **39**, pp. 327-344.

M. Grötschel and O. Holland. 1991. Solutions of large-scale travelling salesman problems. *Math. Programming* **51**, pp. 141-202.

M. Grötschel and M.W. Padberg. 1979. On the symmetric travelling salesman problem I: Inequalities. *Mathematical Programming* **16**, pp. 265-280.

M. Grötschel and M.W. Padberg. 1985. Polyhedral Theory, in the Travelling Salesman Problem, Lawler, et al., pp. 251-302.

M. Grötschel and W.R. Pulleyblank. 1986. Clique tree inequalities and the symmetric travelling salesman problem. *Mathematics of Operations Research* **11**, pp. 537-569.

K. Helbig Hansen and J. Krarup. 1974. Improvements of the Held-Karp algorithm for the symmetric travelling salesman problem. *Math. Programming* **7**, pp. 87-96.

M. Held and R.M. Karp. 1962. A dynamic programming approach to sequencing problems. *J. Soc. Indust. Appl. Math* **10**, pp. 196-210.

M. Held and R.M. Karp. 1970. The travelling salesman problem and minimum spanning trees. *Operations Research* **18**, pp. 1138-1162.

M. Held and R.M. Karp. 1971. The travelling salesman problem and minimum spanning trees: part II. *Math. Programming* **1**, pp. 6-25.

R.L. Karg and G.L. Thompson. 1964. A heuristic approach to solving travelling salesman problems. *Management Science* **10**, pp. 225-247.

K. Malik and M.L. Fisher. 1990. A dual Ascent algorithm for the 1-tree relaxation of the symmetric travelling salesman problem. *Operations Research Letters* **9**, pp. 1-7.

D. Nadef and G. Rinaldi. 1991. The symmetric travelling salesman polytope and its graphical relaxation: Composition of valid inequalities. *Math. Programming* **51**, pp. 359-400.

G.L. Nemhauser and L.A. Wolsey. 1988. Integer and combinatorial optimization, John Wiley. Chichester, UK.

M.W. Padberg and M. Grötschel. 1985. Polyhedral Computations, in the Travelling Salesman Problem, Lawler, et al., pp. 307-360.

M.W. Padberg and S. Hong. 1980. On the symmetric travelling salesman problem: a computational study. *Math. Programming Study* **12**, pp. 78-107.

M.W. Padberg and M.R. Rao. 1982. Odd minimum cut-sets and b-matchings. *Mathematics of Operations Research* **7**, pp. 67-80.

M.W. Padberg and G. Rinaldi. 1987. Optimization of a 532- city travelling salesman problem by branch and cut. *Operations Research Letters* **6**, pp. 1-8.

M.W. Padberg and G. Rinaldi. 1990. Facet identification for the travelling salesman polytope. *Math. Programming* **47**, pp. 219-257.

M.W. Padberg and G. Rinaldi. 1991. A branch-and-cut algorithm for the resolution of large-scale travelling salesman problems. *SIAM Review* **33**, pp. 60-100.

T.H.C. Smith, T.W.S. Meyer and G.L. Thompson. 1990. Lower bounds for the STSP for Lagrangean relaxation. *Discrete Applied Mathematics* **26**, pp. 209-217.

T. Volgenant and R. Jonker. 1982. A branch and bound algorithm for the symmetric travelling salesman problem based on the 1-tree relaxation. *European Journal of Operational Research* **9**. pp. 83-89.

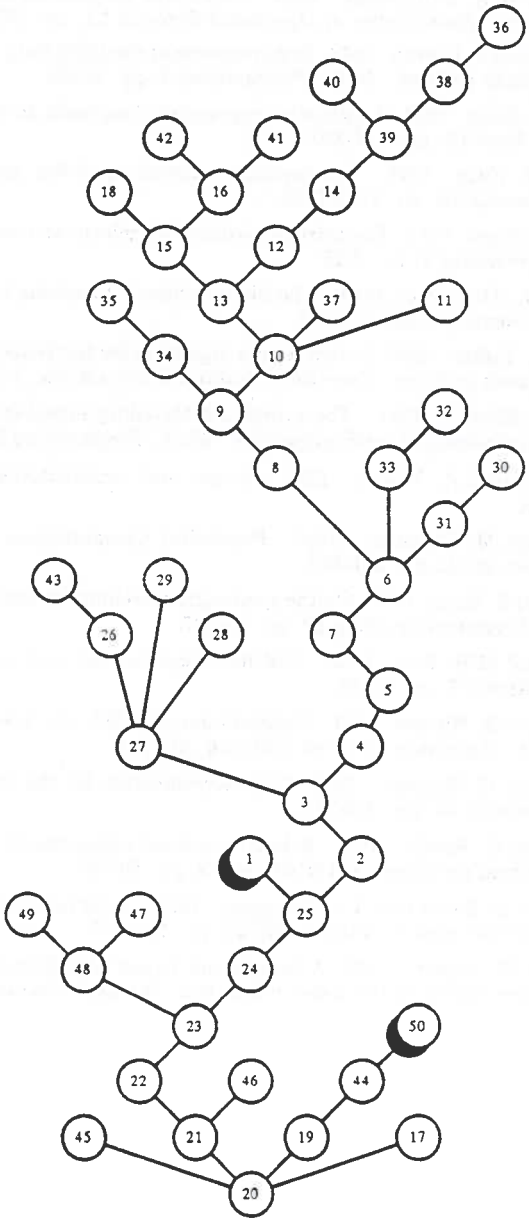


Figure 1. Minimum Spanning Tree at u = 0.

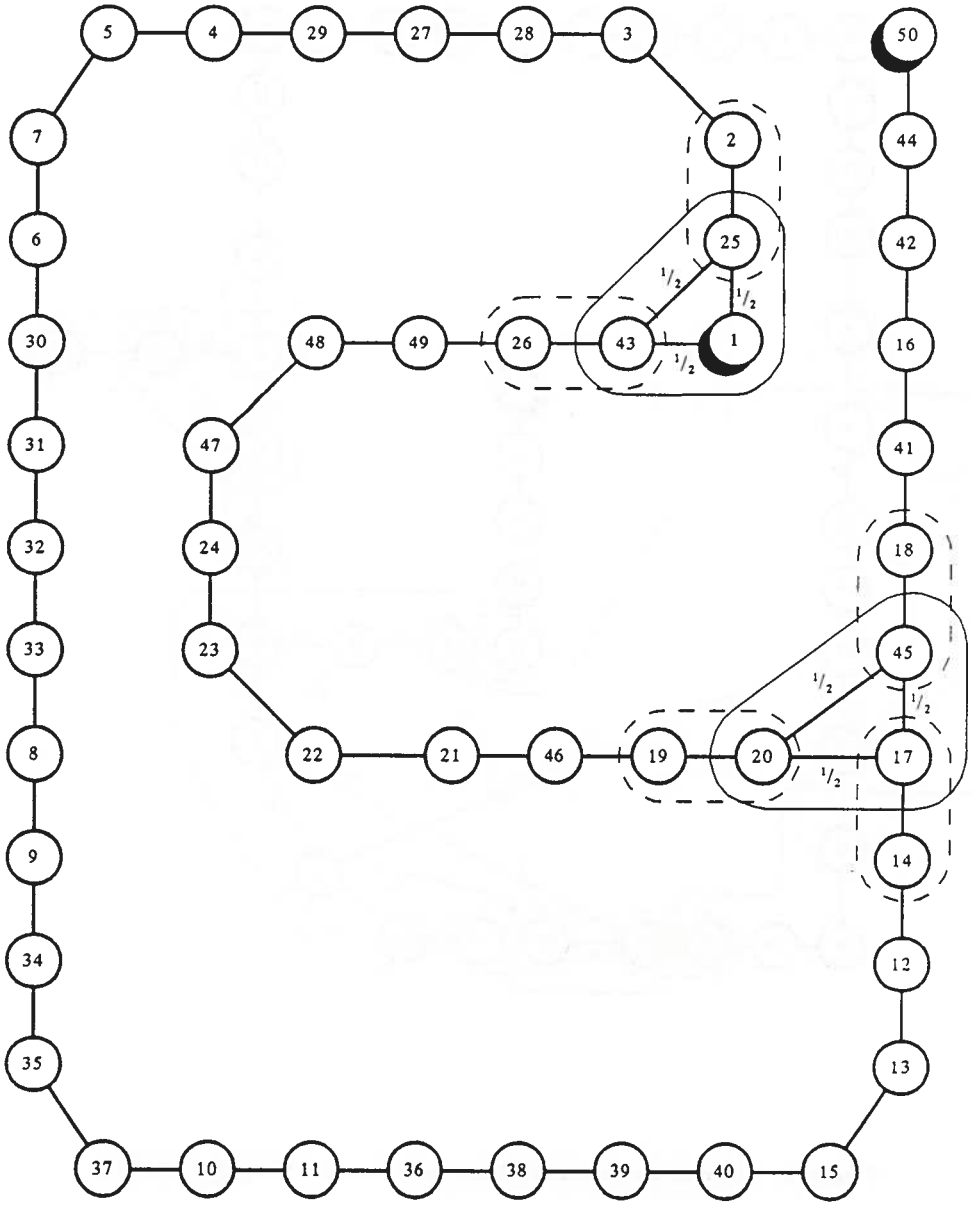


Figure 2. Violated Comb Inequalities and $\hat{x}$ at $\bar{u}^0$ ($Z^0_{LD} = 634$).

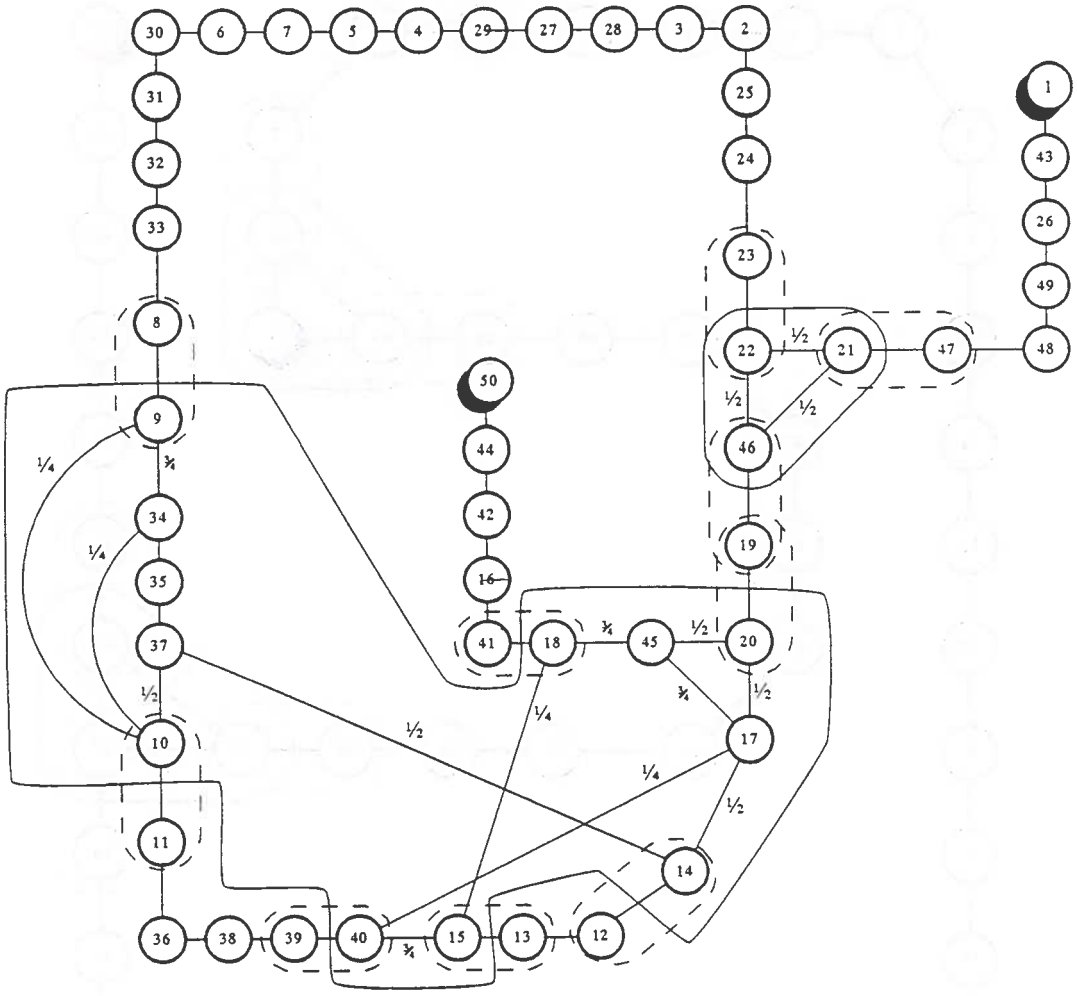


Figure 3. Violated Comb Inequalities and $\hat{\mathbf{x}}$ at $(\overline{\mathbf{u}}^1, \overline{\Phi}^1)$ ($Z^1_{LD} = 635.75$).

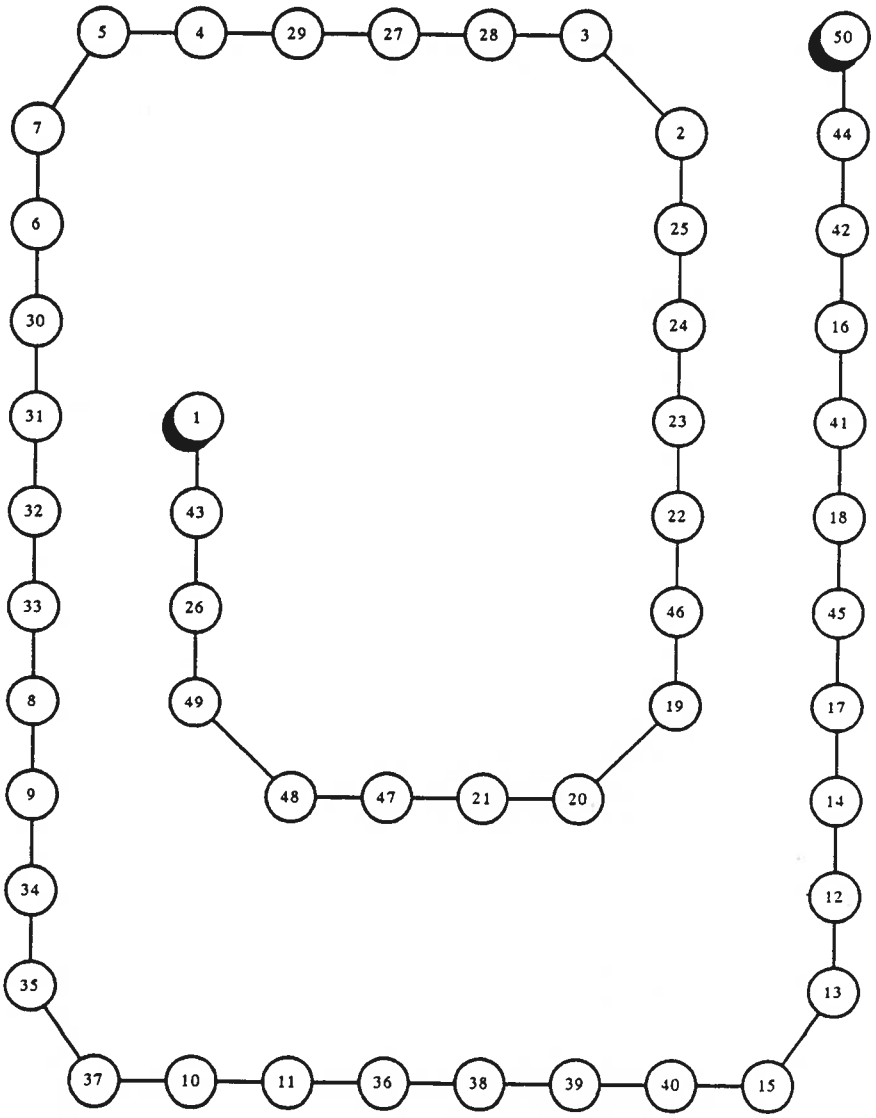


Figure 4. Optimal Solution Given by $\hat{\mathbf{x}}$ at $(\overline{\mathbf{u}}^2, \overline{\Phi}^2)$ $(Z_{LD}^2 = \overline{Z}_{LD} = 637)$.

Optimization in Air Traffic Flow Management

Dimitris Bertsimas
Amedeo R. Odoni
Massachusetts Institute of Technology

1. Introduction

The severe congestion problems increasingly experienced since the mid-1980's by Air Traffic Management (ATM) systems in the United States, Western and Central Europe and parts of the Pacific Rim region have directed much current attention toward Traffic Flow Management (TFM).

Congestion occurs whenever the capacity of one or more elements of the ATM system (airports, terminal-area or en-route sectors, various waypoints, etc.) is exceeded by demand over a period of time. Thus, congestion is mostly associated with peak traffic hours of the day, peak travel times of the year and, especially, periods of adverse weather conditions, when airport service rates may be significantly reduced or the flows of air traffic on some airways may be disrupted. In the long-run, i.e., over periods of 5-20 years, congestion may be alleviated by means of capacity improvements attained through the construction of additional runways and airports or through advances in ATM technology and procedures. In the medium-term, i.e., on a planning horizon of six months or a year, "demand management" measures, such as slot assignment at busy airports or use of congestion-pricing at such airports, may also be helpful. However, on a short-term basis, i.e., for any given level of demand and of ATM system capacity, TFM provides the only approach for reducing the *costs* or, more generally, the *impacts* of delays on the ATM system and its users. *TFM's objective is to "match" dynamically air traffic demand with the available capacity of airports and airspace sectors, on a day-to-day basis, in a way that minimizes delay costs or impacts.*

Should the growth in air traffic demand continue to outpace improvements in ATM system capacity in the future --as it well might-- the

importance of TFM will increase even beyond the central role that it already plays today. Major research and development (R&D) programs are currently under way, in he United States and in Europe aimed at designing and building the TFM systems of the future.

It is important to realize that the extensive use of flow management in ATM is only a quite recent development and that the TFM systems that operate today have evolved in an essentially *ad hoc* way over the last decade in response to the acute pressures of growing traffic. The event most responsible for advancing the widespread use of TFM was the strike of air traffic controllers in the United States in 1981. To reduce the workload on the ATM system, it was decided at the time that airborne delays would be avoided as much as possible by holding aircraft on the ground prior to take-off. An aircraft would not be allowed to take off unless there was reasonable assurance that, after departure, it would be able to proceed to its destination with a minimum of delay in the air. After the ATC system returned to normalcy, the use of ground-holding was preserved as an option and, in fact, has become the primary tool for dealing with the most serious instances of congestion.

Even more important, that experience resulted in a better appreciation of the potential of TFM. The Central Flow Control Facility's (CFCF) responsibilities and capabilities for co-ordinating TFM throughout the continental airspace of the United States have steadily expanded ever since. (CFCF is now called the Air Traffic Control System Command Center (ATCSCC) and is located in Washington, DC.) Traffic Management Units (TMU) are also operating at regional Air Route Traffic Control Centers (ARTCC) and at the major Terminal Radar Control (TRACON) facilities. By taking actions that may relieve a local problem without the need for ATCSCC intervention, the regional and terminal area TMUs help to de-centralize TFM decision-making. The TMUs also assist in implementing ATCSCC strategies and decisions whenever an intervention by ATCSCC becomes necessary.

In Western Europe, where the problem of air traffic delays has reached critical proportions, there is currently no counterpart to a facility such as the ATCSCC. EUROCONTROL, however, is leading the effort to develop an advanced

TFM system for Europe which would co-ordinate traffic planning on an inter-European level.

It is fair to state, in any event, that true long-term R&D on advanced TFM is still in its early phases. For this reason, many of the fundamental conceptual issues regarding TFM have not yet been resolved and overall "visions" of TFM's future are still in a formative stage. Moreover, the state-of-the-art when it comes to automation aids for TFM is now at an important juncture. The Enhanced Traffic Management System (ETMS) in the United States (Volpe National Transportation Systems Cente, 1992) has demonstrated the feasibility of developing a comprehensive database and a knowledge infrastructure that will support TFM activities. But when it comes to developing decision-support systems that would utilize the available database and would assist, in real time, TFM specialists in devising effective flow management strategies, the task is only beginning. In colloquial terms, there is much work to be done on developing some "brain power" for existing and future automation aids for TFM.

This paper attempts to provide, at the same time, an overview of the practical context within which decision-support tools (or "automation aids") for TFM are to be developed, as well as a brief review of state-of-the-art optimization models that might constitute the core of such tools. In Section 2, we define TFM's scope (what activities fall within its purview) and limitations (what types of user benefits TFM can and cannot provide). Section 3 identifies the fundamental technical characteristics of the environment within which TFM must operate and their implications for TFM automation aids. Section 4 argues that it is useful to classify TFM initiatives in a hierarchical way and to distinguish between *strategy generation* and *strategy implementation* activities. Each of these two types of activities calls for different combinations of methodological approaches. Ground-holding is a particularly critical TFM device and the difficulties in devising good ground-holding strategies in the presence of uncertainty are reviewed in Section 5. Section 6 contains a brief survey of optimization models for TFM.such models. Section 7 presents a specific example of a multi-airport, integer-programming problem that introduces the fundamental approach to formulating optimization models in this area. In Section 8, we describe a recently-developed, very generalized

formulation that can be used to model TFM environments in which capacity constraints exist both at airports and in en route sectors. The multi-airport model of Section 7 is shown to be just a special case of the generalized model. Through a judicious choice of decision variables and careful statement of objective function and constraints, this generalized model can be solved very efficiently, as discussed in Section 8 and could eventually become a truly "real time" TFM tool. Sections 9 and 10 contains brief discussions on two important current issues in TFM: the "fairness" of the distribution of delays among ATM users; and the possibility of increased user participation in TFM decision-making. Finally, conclusions are summarized in Section 11.

2. Scope and Limitations of TFM

Indicative of the fact that conceptual thinking about TFM is still in its formative stages is the existing range of views on the scope and role of TFM in air traffic management. For the purposes of this discussion, we shall concentrate only on the "core" activities that everyone agrees properly fall within TFM's scope. Specifically, we shall define TFM as comprising only those activities aimed at *strategic planning and implementation* of regional, national or local *traffic flows*. Stated differently, TFM "maps the contours" of solutions to problems of congestion and overloading and oversees the execution of these solutions. This definition parallels the one suggested by Pozesky (1989).

Since the focus is on strategic initiatives, three primary types of interventions can be considered as the primary "tools" that can be used by TFM: *ground-holding*, i.e., intentionally delaying an aircraft's take-off for a specified amount of time; *metering*, i.e., controlling the rate of traffic flow over time past specified spatial boundaries; and *re-routing*, i.e., altering the distribution of traffic flows over a set of alternative routes or even restructuring the routes themselves. The models to be presented below utilize only the first two (ground-holding and metering). Our ongoing research is also considering re-outing strategies. Other types of interventions (e.g., sequencing and spacing, speed control, vectoring, and high-altitude and low-altitude holding of individual aircraft) could also be included, but, from our

perspective, these are more "tactical" tools that provide the means for "fine-tuning" the implementation of a strategic flow management plan.

It is also useful to understand *a priori* in pragmatic terms the nature of the potential contributions of TFM. One may start by pointing to some differences between those TFM actions caused by "bottlenecks" in en route sectors, on the one hand, and those due to inadequate capacity in airports or in terminal-area airspace, on the other. In the former case, the bottlenecks can often be bypassed at little cost to ATM users, through TFM interventions of a usually local nature, such as re-routing of aircraft or re-structuring the flows of traffic in local airspace. In such cases, TFM essentially generates capacity which would not have been available in TFM's absence and can, thus, reduce the delays that would have been experienced by ATM users.

However, when a flight's airport of destination, for one reason or another, cannot provide adequate capacity, delay becomes *unavoidable* and TFM *cannot* reduce the total delay *time*. In fact, it is quite possible (indeed probable) that total delay time will increase as a result of TFM, because some of the available airport capacity may not be utilized, as will be discussed later in this paper. TFM *can*, nevertheless, accomplish two things in such situations: reduce the *cost* (or the impact) of the waiting time (for example, through ground-holding, so that most of the unavoidable delay can be taken without consuming fuel); and also affect, if desirable, the *distribution* of the total delay time among users of the ATM system.

3. The TFM Environment

To be successful, a TFM system must be designed to take into account some fundamental technical characteristics of the ATM environment. The most important of these, which largely define the feasible "solution space" that TFM seeks to explore, are four:

1. Information intensive: As a result of a combination of a very large data base on past and future events (e.g., flight plans on file, capacity-related data, statistics regarding non-scheduled flights, etc.) and of a highly-developed communications, surveillance and weather-monitoring and

forecasting infrastructure, the ATM system receives and processes enormous amounts of information every day. The TFM system must "harness" all this information, i.e., it needs to: identify what information is available where; decide what part of that information is relevant to support flow management; and automate the acquisition (from dispersed sources within the ATM system), the processing and the distribution and display of that information. TFM must also consider whether any additional data, which may not be readily available currently, would improve performance and, if so, seek to obtain such data, either directly through the ATM system or indirectly from such external sources as the airlines.

2. *Dynamic:* Information about current and short-term future conditions at the various parts of the ATM system is constantly being updated during the course of each day. For example (and of special relevance to flow management) weather forecasts --and, therefore, airport capacity predictions-- typically improve as any critical time period approaches. To take full advantage, TFM must adopt dynamic strategies, which may be continually revised as new information becomes available.

3. *Stochastic:* The future capacity of most ATM elements is inherently *stochastic*, often even on a time horizon of less than one hour, due to dependence on unpredictable or partly predictable variables, such as detailed weather characteristics, equipment outages and other probabilistic events. ATM demand is similarly stochastic, even in the short-run. Ideally, TFM strategies must be designed to take into consideration the level of uncertainty associated with such key parameters.

4. *Complex:* In many instances, particularly in the case of traffic flow-planning at a regional or national level, the number of alternative courses of action available to TFM may be enormous and may increase exponentially with the number of "decision variables". This is, for instance, the case for ground-holding programs that may involve several hundred flights, for each of which TFM must determine the amount of ground-delay, if any, that the flight will absorb. In such instances, even the best-trained human operators can consider only a very small subset of the alternatives and evaluate in detail even a smaller subset, in the absence of advanced automation aids. Computer-

based decision support systems (DSS) are required to assist TFM in identifying, evaluating, comparing and implementing such complex, "combinatorial" strategies in a real-time decision-making environment.

It is fair to state that major progress has been achieved to date with respect to the first of the above areas, automated information gathering, processing and display for TFM --although there is still a lot of room for improvement, especially outside the United States. Dynamic aspects of TFM decision-making are dealt with in an *ad hoc* manner in today's system, with the assistance of a few simple automation tools. More advanced automation aids are currently under development. As for items 3 and 4, the exploration of stochastic and combinatorial aspects is still in the realm of basic research, with some important insights emerging during the past few years. In the following sections we shall make some additional comments along these lines.

4. Hierarchical Decomposition

Hierarchical decomposition would seem to be a key to the generation and implementation of TFM strategies. Given the dynamic ATM environment and the prevalent uncertainty, the best approach to devising and carrying out TFM strategies is to work progressively at several different levels of aggregation. There are three dimensions to this process: level of detail, time-to-go and geographical scope. On any given day, the process should begin with an approximate determination (low level of detail) several hours in advance, of desirable targets for the various critical traffic flow rates on a national or regional scale; and continue by updating, refining and finally executing these initial plans at an increasingly local level as the time-to-go becomes shorter. The need for this type of approach has been recognized already by the current TFM system in the United States and is reflected in the existence of a national (ATCSCC) and many local (TMU) planning units, the allocation of roles and responsibilities among these units and the operating practices followed in developing a daily traffic plan.

An analogous approach is called for in developing automation aids for TFM. A natural and convenient way for "solving" the overall TFM problem is to decompose it into a set of hierarchical and nested subproblems. At the

highest level in this scheme is the *strategy generation (SG)* problem which can be stated as follows: given an approximate estimate of how much capacity is available over a time-horizon of several hours at capacity-constrained airports and on critical route segments of the ATM system, determine an initial ground-holding strategy and an initial allocation of airborne traffic among route segments. Answers to the SG problem need not be overly precise. For example, in view of the performance characteristics of today's ATM systems and of the typical level of accuracy in capacity and demand predictions, it is sufficient to subdivide the time axis into 10- or 15-minute-long intervals and specify which flights will be scheduled to land at and depart from each congested airport during each of these intervals. It makes little sense, at this high level of TFM planning, to try to be more precise than that.

At lower levels in this hierarchy, one is confronted with a continuum of *strategy implementation (SI)* problems involving progressively shorter time-horizons and aircraft which are either close to becoming available for departure or may already be airborne. Solutions to SI problems require much greater specificity regarding the timing of various events and the flows of traffic on route segments.

Clearly SG and SI problems are closely inter-related: the quality of top-level TFM strategies obtained by solving the SG problem largely determines whether a satisfactory strategy implementation can be developed later. Conversely, the SG will have to be re-solved if, during SI, it turns out that the traffic cannot be handled in an acceptable manner or that there is room for more traffic.

Three broad types of approaches are available for addressing SG and SI problems:

(a) *Manual solutions:* TFM specialists devise solutions with the aid of timely, well-displayed information and, possibly, with support from an easily accessible historical database that indicates how similar situations were handled in the past.

(b) *Knowledge-based expert systems:* offer recommendations to TFM specialists by working with a historical database such as the one mentioned under (a), automating recognition of "similar TFM situations" and, possibly, extrapolating knowledge to new situations.

(c) *Exact and heuristic algorithms:* offer recommendations to TFM specialists by solving (exactly or approximately) appropriate variations of optimization models in network design, network flows, mathematical programming, etc.

The distinction between SG and SI problems is very useful in selecting the right combination of the above approaches. In general, for most SI problems (and, especially, for those of a largely local scope) high-quality solutions can be devised manually by TFM specialists, with support from relatively simple tools based on approaches (b) and (c). Several such tools are currently available and many more are in various stages of development. They are concerned with such tasks as departure sequencing, "miles-in-trail" en route metering, arrival sequencing and spacing, development of alternative flow routes, merging of traffic streams, etc. The quality of solutions obtained in this way are probably comparable to the best that could be obtained through approaches that place additional emphasis on (b) and (c).

By contrast SG problems, especially those involving more than one or two contiguous ARTCC (or several different countries in Europe) may require considering and comparing a very large number of alternatives, each of which sets off a chain-sequence of other events. In such cases, approaches based on (b) and, particularly, on (c) hold promise for generating solutions which, on average, may be far superior to those generated by less advanced automation, under time pressure. We shall return to this subject in Sections 6 and 7, after first discussing the specific difficulties involved in generating strategies that include ground-holds.

6. Ground-Holding and Uncertainty

It was suggested earlier that ground-holding is the principal and, if done well, the most effective device that can be employed by TFM whenever

delays are anticipated to be severe. This is particularly true in light of the general desire in ATM to keep airborne delays within comfortable bounds. It is therefore useful to examine the ground-holding issue in more detail in this section.

Ground-holding is typically imposed on aircraft flying to congested airports or scheduled to traverse congested airspace. The motivation is that, as long as a delay is unavoidable, it is safer and less costly for the flight to absorb this delay on the ground before take-off, rather than in the air. Unfortunately, deciding how much ground-holding delay to assign to a flight is far from simple (Andrews, 1993). The reason is that *it is difficult to predict how much delay a flight will actually suffer*, because sector capacities and, especially, airport capacities are often highly variable and can change quickly over time, as weather changes or other events occur.

It is an often insufficiently appreciated fact that airport capacity is *not* a constant, but a *random variable*. For any given airport, capacity during any period of a day can assume one of several different values, depending on weather conditions (visibility, cloud ceiling, winds, precipitation), the mix of aircraft types, the mix of operations (landings vs. take-offs), the runways in use, equipment outages, human factors, etc. It is not unusual in the United States to encounter a 2:1 or even 3:1 ratio between the highest and lowest capacities of any given airport. This is illustrated in Figure 1 which refers to Boston's Logan International Airport, one of the busiest commercial airports in the United States and in the world with approximately 24,000,000 passengers and 480,000 aircraft operations in 1993. As can be readily seen, the airfield capacity of Logan Airport is a highly probabilistic quantity. It can range from a high of 130-120 operations (landings and take-offs) per hour in visual meteorological conditions (VMC, about 78% of the time), to 90-85 operations per hour under some intermediate weather conditions (about 8% of the time) to a low of approximately 60-50 operations per hour in strict instrument meteorological conditions (IMC, about 14% of the time). (Readers will undoubtedly realize that, although the capacity profile shown in Figure 1 is referred to as the "capacity coverage chart" by airport managers, it is, in fact, a plot of the *probability density function* of Logan's airfield capacity.) One of the principal reasons for these dramatic changes is that, while in VMC two

runways can often be used simultaneously for arrivals at Logan, only one runway can typically be used in IMC.

Small changes in visibility at ground level or in the cloud-cover may then translate into large differences in airport capacity. Meteorologists cannot yet predict such changes to this level of accuracy, even over a very short time-horizon of an hour or less. Thus, ground-holding decisions must be made under uncertainty and must consider the trade-off between "conservative" strategies that may at times assign excessive ground-holds and more "liberal" ones that may result in more expensive airborne delays. Airlines in the United States have been contending that current TFM tactics err excessively on the conservative side, i.e., that there may be too many instances in which airport landing capacity (at destination airports) is being wasted, while aircraft stand waiting on the ground at the airports of origin.

Unfortunately, no automation aid or other tool currently in use in support of TFM considers explicitly whatever probabilistic information may be available on any given day (e.g., the probability that a thunderstorm will occur at a particular airport in the afternoon of that day). Thus, TFM specialists are forced to deal with uncertainty on the basis of experience and in an *ad hoc* way. This may result in postponing excessively the initiation of ground-holding, while waiting for more definite information, or in initiating such programs unnecessarily. This is yet another important research area that requires urgent attention.

6. Optimization Models

It was noted at the end of Section 5 that certain types of mathematical optimization models probably offer the most promising methodologies for building decision support systems (DSS) for strategy generation in TFM planning. Extensive research has been performed recently on the application of mathematical modeling to SG problems in TFM. The objective is to achieve a better starting solution to the planning of air traffic flows and the assignment of ground-holds to flights than can be done today, either manually or through simple heuristic automation tools.

The problem statement is essentially the following version of the SG problem: if demand is expected to exceed capacity at one or more airports or in airspace for a certain period of time, generate a plan of air traffic flows (metering, re-routing, etc.) and of delay assignments to individual aircraft (ground holds, airborne delays) which maximizes (or minimizes) an aggregate measure of efficiency (or, respectively, of cost) to all ATM users, while respecting all airport and airspace capacity constraints.

The earliest models developed (Terrab and Odoni, 1993) address a single airport at a time and assume that there are no en route sector capacity constraints and that the forecasts of airport demand and capacity are "deterministic", i.e., there is no uncertainty about the level of aircraft demand and the available airport capacity over the time-horizon of interest. Increasing levels of detail and complexity have subsequently been added such as: dynamic updating of forecasts of capacity and demand (Richetta and Odoni, 1994; Vranas et al., 1994b); stochastic capacity and demand (Richetta and Odoni,1993); a multi-airport environment in which individual aircraft fly multi-leg itineraries, so that a delay at one "upstream" airport may result in further flight delays at "downstream" airports; and inclusion of airspace capacity considerations (Bertsimas and Stock, 1993; Helme, 1992; Lindsay, 1993; Lindsay et al., 1993). Other families of models have adopted similar approaches, but at different levels of aggregation (see, e.g., Booth and Harvey, 1983, and Zenios,1991).

Naturally, the most appropriate combination of model features to use depends on the TFM situation at hand. The most general version of a SG model (multi-airport, dynamic, stochastic, with en route capacity constraints) may not be appropriate for many (probably most) practical situations. For example, deterministic (rather than stochastic) versions will be preferable for locations where the airport capacities are stable enough to be approximated as perfectly predictable quantities. Similarly, static (rather than dynamic) versions may be more appropriate for environments where (i) there are significant lags in updating information concerning weather or capacities or (ii) a strategic TFM plan is prepared at a single point in time (typically at the beginning of the day) and that plan is revised only in a marginal way ("tactical" changes) from that point on.

As the complexity and realism of model assumptions grows so does the challenge of obtaining efficiently solutions to large problems of realistic size. The approaches taken have included the use of standard "off-the-shelf" optimization software, as well as of customized software that may take advantage of special characteristics of the model to speed up its solution.

7. A Multi-Airport Optimization Model

In this section, we shall describe briefly an example of an optimization model for strategic TFM planning. This particular model, presented in detail in Vranas et al. (1994a), considers traffic-flow bottlenecks at airports only, i.e., assumes that en route sectors are not primary congestion points. In this respect, the model is more appropriate for the current TFM environment in the United States (where the overwhelming majority of serious delays are due to airport capacity constraints) rather than in Europe. The model considers any given set of airports expected to experience some delays on a particular day and the entire set of flights scheduled to operate at these airports. It subdivides the day into time-periods of equal length (e.g., of 10 or 15 minutes). Given a forecast of the arrival capacity and the departure capacity at each of the airports for each of the time periods the model assigns to each flight, when necessary, a ground delay and/or an airborne delay, in a manner that satisfies all the airport capacity constraints, while minimizing a function that reflects total delay costs (or impacts of delay).

A feature of this model is that it takes into consideration the propagation of delays over the network of airports: because each aircraft typically performs more than one flight legs on any given day, whenever a specific aircraft is seriously delayed, the next flight leg by the same aircraft will also be delayed. Moreover, at "hub" airports, a late-arriving aircraft may delay the departure of several flights, given current airline scheduling practices which emphasize passenger transfers. We shall refer to this as the Multi-Airport (MA) model. The FAA is currently developing a DSS, called OPTIFLOW, for possible use at the ATCSCC in preparing daily national strategic traffic flow plans. OPTIFLOW uses a variation of the MA model as the basic tool for preparing this traffic plan.

Consider then a set of airports $\mathbf{K}=\{1,\ldots,K\}$ and an ordered set of time periods $\mathbf{T}=\{1,\ldots,T\}$. For instance, $\mathbf{K}$ might be the set of the 20 or so busiest U.S. airports (plus, if desired, an "imaginary" airport that represents all other airports in the world and acts as the "source" or the "sink" for all flights not traveling between two of the real airports in $\mathbf{K}$); and $\mathbf{T}$ might be a set of 64 time periods of 15 minutes each, amounting to a time horizon of 16 hours, i.e., the portion of a day from 07:00 to 23:00 (when most flights take place). Consider also a set of flights $\mathbf{F}=\{1,\ldots,F\}$. (Note that a single aircraft may perform several of these flights.) $\mathbf{F}$ is the set of all flights of interest, e.g., all flights departing from an airport in $\mathbf{K}$ and arriving to another airport in $\mathbf{K}$. (For brevity, we use here the term "flight" for what should more precisely be called a "flight leg".) The set of data defined in Table 1 is assumed to be known for each flight and each airport. Note that the data refer to flight characteristics and to airport capacities.

Consider finally the set $\mathbf{F}' \subset \mathbf{F}$ of those flights that are *continued*. A flight is said to be continued if the aircraft which is scheduled to perform it, is also scheduled to perform at least one more flight later in the day. For each flight $f' \in \mathbf{F}'$, we assume we know the next flight f scheduled to be performed by the same aircraft, and the ``slack'' or ``absorption'' time $s_{f'}$ such that, if f' arrives at its destination at most $s_{f'}$ time periods late, the departure of the next flight f will not be affected. $s_{f'}$ is obviously equal to the difference between (i) the time interval between the scheduled departure time of f and the scheduled arrival time of f', and (ii) the minimum ``turnaround'' time of the aircraft performing both flights.

The MA model is now presented in Table 2. The objective function is to minimize the total cost of delay (ground and airborne) for all flights in the set F. The decision variables of the MA model are u_{ft} and v_{ft}, where u_{ft} is 1 if flight f is assigned to take off at period t (i.e., if $d_f + g_f = t$) and 0 otherwise, and v_{ft} is 1 if flight f is assigned to land at period t (i.e., if $r_f + g_f + a_f = t$) and 0 otherwise. Constraints (3) and (4) ensure that, for every flight f, exactly one of the variables u_{ft} must be equal to 1 and the others must be equal to zero, and similarly for the variables v_{ft}. Given this fact, the number of time-periods of ground delay, g_f, assigned to flight f can be expressed in terms of the

assignment variables u_{ft} by means of constraint (7); similarly, a_f, the amount of airborne delay, can be expressed in terms of v_{ft} by means of (8).

Constraints (1) and (2) are the departure and arrival capacity constraints, respectively: the number of take-offs and landings at airport k in time period t cannot exceed the available departure and arrival capacities. Constraints (5) are the *flight coupling* constraints: they "transfer" any excessive delay of flight f' to its next flight f. In fact, the coupling constraints say that, if flight f' arrives at its destination with a total delay $g_{f'} + a_{f'}$ which is greater than $s_{f'}$ (the "slack" defined above), then the next flight f will have to be delayed on the ground at least $g_{f'} + a_{f'} - s_{f'}$ time periods; otherwise, the departure of the next flight f will not be affected. Note that, if the coupling constraints did not exist, the problem would be decomposable into K subproblems concerning one airport each. Note that nonnegativity of g_f is guaranteed by (5), whereas nonnegativity of a_f is not yet guaranteed; this is why constraints (6) are needed. A final remark concerning the coupling constraints (5) is that they can be interpreted in a more general way than the linking of successive flights scheduled to be performed by the same aircraft; i.e., they can be used to link any pair of flights f and f' such that f cannot be allowed to depart before f' has arrived (possibly because passengers in f' must transfer to f). In this interpretation, a flight f' may have more than one "next" flights f.

The following additional points are of interest: First, the MA model can be extended to include additional considerations. For example the departure and arrival capacities, $D_k(t)$ and $A_k(t)$, of a given airport k at a given time period t are often not independent, because they are determined by the way in which runway use is allocated among departing and arriving aircraft. In such cases, arrival capacity can be increased during period t at the expense of departure capacity, and vice versa. This can be accounted for in the MA model by considering $D_k(t)$ and $A_k(t)$ to be decision variables and adding to the model constraints that quantify the tradeoffs between arrival and departure capacities at airport k (Vranas et al., 1994b; Gilbo, 1993). Second, to achieve an equitable distribution of delays the MA model can be solved with additional constraints or other modifications which ensure that there will be no "discrimination" among flights by aircraft type, when it comes to the

assignment of airborne and ground delays. In fact, numerous hypothetical examples of the MA model have been solved with such restrictions to date. Finally, Bertsimas and Stock (1993) have developed an approach that solves the MA problem very efficiently. For example, an instance involving an ATM region with three major congested airports and approximately 1000 flights scheduled within a six hour period was solved in about 5 seconds on a fast workstation.

8. A Multi-Airport Model with Enroute Capacities

In this section we present an important extension of the model in the previous section that takes into account the capacities of the sectors of the National Airspace System (NAS). Each flight passes through contiguous sectors while it is en route to its destination. There is a restriction on the number of airplanes that may fly within a sector at a given time. This number is dependent on the number of aircraft that an airtraffic controller can manage at one time, the geographic location and the weather conditions. We will refer to the restrictions on the number of aircraft in a given sector at a given time as the en route sector capacities. There are several key sectors throughout the United States (and especially in Europe) that are often operated at full capacity. The issue of congestion at these sectors is as critical as congestion in the terminal areas, since the cost of holding an airborne aircraft is not dependent on the location of the aircraft. Thus, airborne delay costs could further be reduced if we could determine the optimaltime for a flight to traverse the capacitated sectors.

Bertsimas and Stock (1993) present an integer programming formulation for this general Air Traffic Flow Management Problem (TFMP), in the presence of possible restrictions on the en route sector capacities. As in the previous section we consider a set of flights $\mathbf{F} = \{1,\ldots,F\}$, a set of airports $\mathbf{K} = \{1,\ldots,K\}$, an ordered set of time periods $\mathbf{T} = \{1,\ldots,T\}$ and a set C of pairs of flights that are continued. The problem input data are given in Table 3.

The objective in the TFMP is to decide how much each flight is going to be held in the ground and in the air in order to minimize the total delay cost. In order to model the problem we introduce decision variables w_{ft}^{j} which is equal

to one if flight f arrives at sector j by time t and zero otherwise. This definition using *by* and not *at* is critical to the understanding of the formulation. Also recall that we have also defined for each flight a set Pf of sectors which includes the departure and arrival airports, so that the variable w^j_{ft} will only be defined for those sectors jin the set Pf. Furthermore, one variable per flight-sector pair can be eliminated from the formulation by setting $w^j_{fT}= 1$, where T is the last time period that flight f can reach sector j. Having defined the variables w^j_{ft} we can express several quantities of interest as linear functions of these variables as follows:

1. The variable $u^j_{ft}=1$ if flight f arrives at sector j at time t and 0 otherwise, can be expressed as follows: $u^j_{ft} = w^j_{ft} - w^j_{f,t-1}$

2. Noticing that the first sector for every flight represents the departing airport, then the total number of time units that flight f is held on the ground is the actual departure time minus the scheduled departure time, i.e.,

$$g_f = \sum_{t \in T^k_f, k=P(f,1)} t u^k_{ft} - d_f = \sum_{t \in T^k_f, k=P(f,1)} t(w^k_{ft} - w^k_{f,t-1}) - d_f.$$

3. Noticing that the last sector for every flight represents the destination airport, the total number of time units that flight f is held in the air can be expressed as the actual arrival time minus the scheduled arrival time minus the amount of time that the flight has been held on the ground, i.e.,

$$a_f = \sum_{t \in T^k_f, k=P(f,N_f)} t u^k_{ft} - r_f - g_f = \sum_{t \in T^k_f, k=P(f,N_f)} t(w^k_{ft} - w^k_{f,t-1}) - r_f - g_f.$$

With these observations we can now formulate the problem, which we present in Table 4. Constraints (9), (10) and (11) take into account the capacities of various aspects of the system. Constraints (9) ensure that the number of flights which may take off from airport k at time t, will not exceed the departure capacity of airport k at time t. Likewise, constraints (10) ensure that the number of flights which may arrive at airport k at time t, will not exceed the arrival capacity of airport k at time t. The third constraint ensures that the sum of all flights which may feasibly be in sector j at time t will not exceed the capacity of sector j at time t.

Constraints (12) represent connectivity between sectors. They stipulate that if a flight arrives at sector j' by time $t+l_{jj'}$, then it must have arrived at sector j by time t, where j and j' are contiguous sectors in flight f's path. In other words, a flight cannot enter the next sector on its path until it has spent l_{ij} time units traveling through sector j, the current sector in its path.

Constraints (13) represent connectivity between airports. They handle the cases in which a flight is continued, i.e., the flight's aircraft is scheduled to perform a later flight within some time interval.We will call the first flight f' and the following flight f. Constraints (13) state that if flight f departs from airport k by time t, then flight f' must have arrived at airport k by time $t-s_f$. The turnaround time, s_f, takes into account the time that is needed to clean, refuel, unload and load, and further prepare the aircraft for the next flight.

Constraints (14) represent connectivity in time. Thus, if a flight f has arrived by time t then w_{ft}^j has to have a value of 1 for all later time periods t'.

Important Remark:

The major reason we used the variables w_{ft}^j, as opposed to the variables u_{ft}^j is that the variables w_{ft}^j nicely capture the three types of connectivity in TFMP: connectivity among sectors, connectivity among airports and connectivity in time. Moreover, Bertsimas and Stock (1993) show that the connectivity constraints they define are facets of the convex hull of solutions. As a result, the linear programming relaxation of the formulation is almost always integral, i.e., the given formulation is a particularly strong one. Indeed when specialized to the problem of the previous section this formulation is stronger than the formulation of the previous section.

The previous formulations were tested extensively with very encouraging results. As an example Bertsimas and Stock (1993) performed experiments on a connected network of four airports: Boston Logan (BOS), NY LaGuardia (LGA), Washington National (DCA) and a node representing all other airports (X).
Three hypothetical sectors, surrounding LaGuardia, were also introduced into the model. Different flights would traverse these sectors while en route to LaGuardia depending on the origin of the flight. The three airports (BOS, LGA, DCA) and the three sectors were the only capacitated elements in the system.

The other sectors were allocated unlimited capacity. One set of experiments for 200 flights over a 24 hour period and another set for 1000 flights over varying time periods were performed.

The model was programmed in GAMS. For the set of 200 flights, the model was run on a Hewlett Packard Workstation, HP-UX/9000-710, and OSL was used as the solver. GAMS required 5.1 Mb of memory to store this information and 2923 seconds CPU time to generate the matrix to input to OSL. To solve the problem OSL only requires 15.6 seconds CPU time. The set of 1000 flights was tested on a Sun SPARCstation 10 model 41. CPLEXMIP 2.1 was used as the solver. The solution time for the 1000 flights over a 6 hour period was 4.267 seconds CPU time. In all cases the solution from the linear programming relaxation was integral, so there was no need for enumerative approaches and the computatiuon times were extremely fast.

Experiments were also performed using the specialization of the model for uncapacitated sectors, with impressive results. Again, solving the 1000 flight dataset on the SPARC workstation a solution was found considering an 8 hour time period in only 7.834 seconds of CPU time

Overall, the formulations given in this section gave rise to very encouraging results. An optimal integral solution was found for the very large scale instances of the problem within seconds. This is particularly important, because in practice we might want to re-optimize often as more information becomes available.

9. The Issue of "Fairness"

Questions about "fairness" will undoubtedly become more prominent, and possibly contentious, if the role of TFM continues to expand in the future. TFM strategies that aim exclusively at minimizing or maximizing some aggregate measure of performance, without concern for distributive effects, may result in very uneven allocation of delays among ATM users. As this may be unacceptable, models that treat all users as a single class (as the models presented above do) may have a better chance of being adopted in practice.

This does not mean that some deviations from a strictly uniform distribution of delays among ATM users may not be beneficial at times. A notable example is the allocation of ground-holds between long-range and short-range flights headed to the same (potentially congested) airport. Optimal dynamic TFM strategies favor long-range flights over short-range ones in such cases, in the sense that long-range flights are more likely to be allowed to take off without any ground-holding. Intuitively, good dynamic strategies tend to be more active with short-range flights (i.e., impose more ground-holds on them) in order to take advantage of the improved state of knowledge regarding the capacity of the destination airport at the time when short-range flights are scheduled to depart. Indeed current practice partly reflects this tendency. For instance, flights to the United States from Europe, or coast-to-coast domestic flights are generally exempt from ground-holding in all but exceptional circumstances.

Similar questions are also implicit in the ideas discussed in the next section.

10. A More "Participatory" System

TFM, even in its existing form, involves many decision-makers. The operators of ATM systems (the FAA and other national or international ATM organizations) must clearly play the central role as the co-ordinating agents in the process and the ultimate eventual decision-makers. But aircraft operators also have a role to play. For example, at the most obvious, an airline may decide to cancel a planned flight, should the delay assigned to it by TFM be deemed unacceptable. In the United States, each airline is also given the opportunity to re-assign to other flights of its own (subject to certain restrictions) the airport slots vacated by any other flights that the airline may decide to cancel on any given day ("slot substitution program").

Currently, many planners of TFM systems believe that user participation in TFM decision-making should increase in the future. For example, the TFM system's operator might determine how much total delay will be allocated to each individual airline on any given day, but then give

considerable flexibility to each specific airline to decide how this total delay will be allocated among that airline's individual flights. The OPTIFLOW DSS, mentioned earlier, includes a feature along these lines: an airline may be able to request that TFM assign to all members of a specified group of its flights (called a "flight bank") headed to the same airport, arrival times that all fall within a short span of time. This would make it possible for airlines operating airport "hubs" to transfer passengers efficiently between flights.

At an even more advanced level, a mechanism might be developed that would give airlines the opportunity to express, in real time, their preferences on how available capacity on any given day should be allocated and, possibly, even to enter "bids" of some kind for specific time slots. Some research is currently beginning (e.g., at the MITRE Corporation and at the Massachusetts Institute of Technology) on developing such a dynamic and interactive decision-making process, involving more closely the airlines and other ATC users. This research obviously must address difficult methodological and practical problems.

11. Conclusion

In this paper we have presented a structured overview of several major conceptual issues in TFM optimization. TFM can undoubtedly provide large benefits in terms of workload reduction and cost savings to the operators and users of ATM systems. These benefits increase, along with the importance of TFM, when ATM systems become more congested.

Major progress has been achieved to date in the area of data acquisition, processing and display for TFM. However, the process of developing advanced automation aids is only in its early stages. Varying combinations of methodological approaches are called for, depending on the type of TFM activity at hand. In this respect, the distinction between "strategy generation" and "strategy implementation" tasks is very useful. It was argued that the potential benefits from advanced automation tools in connection with strategy generation may be particularly significant. The development of decision support systems for TFM, based on optimization models, which explore and compare, in real time, very large numbers of alternative strategies seems

feasible. Thus, it may become feasible in the near future to deploy a prototype DSS for TFM that utilizes such optimization models. Open areas for TFM-related basic research were also identified. Prominent among those are the following: dealing with uncertainty, in general, and especially in assigning ground-holding delays; understanding the distributive effects of TFM decisions; and working toward a more participatory future TFM system.

12. References

Andrews, John W. (1993) "Impact of Weather Event Uncertainty upon an Optimum Ground-Holding Strategy, *ATC Quarterly, 1*, 59-84.

Bertsimas, Dimitris and S. Stock (1993) *The Multi-Airport Flow Management Problem with En Route Capacities*, Working Paper, Operations Research Center. Massachusetts Institute of Technology, Cambridge, MA 02139.

Booth, G. and C. Harvey (1983) "Minimizing Air Traffic Delays by Mathematical Programming", *Proceedings of the Conference on Safety Issues in Air Traffic Systems and Design*, Princeton University, Princeton, NJ, pp. 390-403.

Gilbo, Eugene P. (1993) "Dynamic Optimization of Airport Capacity Utilization in Air Traffic Management", Working Paper, Unisys Corporation, Cambridge, MA 02140, U.S.A.

Goranson, Jesse (1993) *Looking for Trouble: How Well the FAA's Enhanced Traffic Management System Predicts Aircraft Congestion*, S.M. Thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139.

Helme, Marcia P. (1992) "Reducing Air Traffic Delay in a Space-Time Network," in *Proceedings of the 1992 IEEE International Conference on Systems, Man, and Cybernetics*, Chicago, IL.

Lindsay, Kenneth S., September (1993) *Traffic Flow Management Modeling with the Time Assignment Model*, MTR-93W0000119, The MITRE Corporation, McLean, VA, to appear.

Lindsay, Kenneth S., E. Andrew Boyd, and Rusty Burlingame (1993) "Traffic Flow Management Modeling with the Time Assignment Model," in *Air Traffic Control Quarterly*, to appear.

Medeiros, Manuel F. (1989) "Advanced Traffic Management System Automation", *Proceedings of the IEEE, 77*, pp. 1643-1652.

Pozesky, Martin T. (1989) "The U.S. Air Traffic Control System Architecture", *Proceedings of the IEEE, 77*, pp. 1605-1617.

Richetta, O. and A. R. Odoni (1993) "Solving Optimally the Static Ground-Holding Policy Problem in Air Traffic Control", *Transportation Science, 27*, pp. 228-238.

Richetta, O. and A. R. Odoni (1994) "Dynamic Solution to the Ground-Holding Policy Problem in Air Traffic Control", in *Transportation Research, to* appear.

Terrab, M. and A.R. Odoni (1993) "Strategic Flow Control on an Air Traffic Network" *Operations Research, 41*, pp. 138-152.

Volpe National Transportation Systems Center (1992) *Enhanced Traffic Management System (ETMS): Functional Description (Version 4.2)*, Report VNTSC-DTS56-TMS-003, U.S. Dept. of Transportation, Kendall Square, Cambridge, MA 02142, U.S.A.

Vranas, Peter, D. Bertsimas and A. R. Odoni (1994a) "The Multi-Airport Ground-Holding Problem in Air Traffic Control", in *Operations Research*, to appear.

Vranas, Peter, D. Bertsimas and A. R. Odoni (1994b) "Dynamic Ground-Holding Policies for a Network of Airports", in *Transportation Science*, to appear.

Zenios, S. (1991) "Network Based Models for Air Traffic Control", *European Journal of Operational Research, 50*, pp. 166-178.

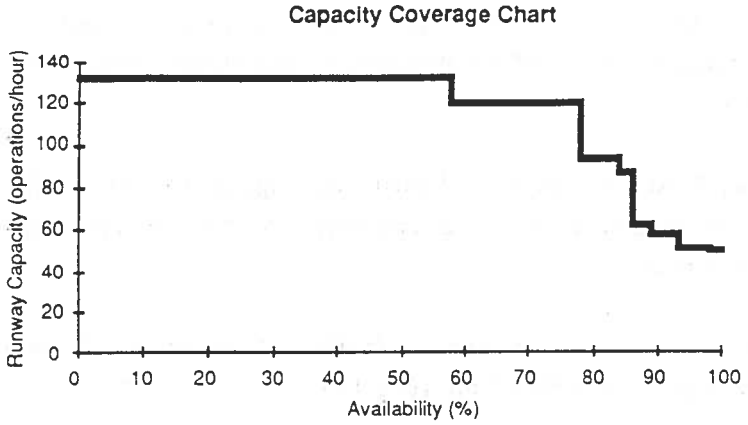


Figure 1: Capacity availability at Logan International Airport, Boston.

Symbol	Denotes
$\mathbf{K} = \{1,\ldots,K\}$	Set of airports k.
$\mathbf{T} = \{1,\ldots,T\}$	Set of time periods t.
$\mathbf{F} = \{1,\ldots,F\}$	Set of flights f.
$k_f^d \in \mathbf{K}$	Departure airport of flight f.
$k_f^a \in \mathbf{K}$	Arrival airport of flight f.
$d_f \in \mathbf{T}$	Scheduled departure time of flight f.
$r_f \in \mathbf{T}$	Scheduled arrival time of flight f.
c_f^g	Ground delay cost per unit of time of flight f.
c_f^a	Airborne delay cost per unit of time of flight f.
$D_k(t)$	Departure capacity at airport k at period t.
$R_k(t)$	Arrival capacity at airport k at period t.
g_f	Ground delay assigned to flight f.
a_f	Airborne delay assigned to flight f.
u_{ft}	Departure assignment decision variable for flight f (see text).
v_{ft}	Arrival assignment decision variable for flight f (see text).
G_f	Upper bound on the ground delay of flight f.
A_f	Upper bound on the airborne delay of flight f.
$\mathbf{T}_f^d$	Set of allowable departure periods for flight f; $\left(\mathbf{T}_f^d = \left\{t \in \mathbf{T}: d_f \le t \le \min(d_f + G_f, T)\right\}\right)$
$\mathbf{T}_f^a$	Set of allowable arrival periods for flight f; $\left(\mathbf{T}_f^a = \left\{t \in \mathbf{T}: r_f \le t \le \min(r_f + A_f, T)\right\}\right)$
$\mathbf{F}' \subset \mathbf{F}$	Set of continued flights f'.
$s_{f'}$	Slack time of continued flight f'.

Table 1: Notation for the MA model.

$$\min \quad \sum_{f=1}^{F} c_f^g g_f + c_f^a a_f$$

$$\text{s.t.} \quad \sum_{f:k_f^d=k} u_{ft} \le D_k(t), \qquad (k,t) \in \mathbf{K} \times \mathbf{T}; \qquad (1)$$

$$\sum_{f:k_f^a=k} v_{ft} \le R_k(t), \qquad (k,t) \in \mathbf{K} \times \mathbf{T}; \qquad (2)$$

$$\sum_{t \in \mathcal{T}_f^d} u_{ft} = 1 \qquad f \in \mathbf{F}; \qquad (3)$$

$$\sum_{t \in \mathcal{T}_f^a} v_{ft} = 1 \qquad f \in \mathbf{F}; \qquad (4)$$

$$g_{f'} + a_{f'} - s_{f'} \le g_{f''} \qquad f' \in \mathbf{F}'; \qquad (5)$$

$$a_f \ge 0 \qquad f \in \mathbf{F}; \qquad (6)$$

$$g_f = \sum_{t \in \mathcal{T}_f^d} t u_{ft} - d_f, \qquad f \in \mathbf{F}, \qquad (7)$$

$$a_f = \sum_{t \in \mathcal{T}_f^a} t v_{ft} - r_f - g_f, \qquad f \in \mathbf{F}, \qquad (8)$$

$$u_{ft}, v_{ft} \in \{0,1\};$$

Table 2: Formulation of the Multi-Airport (MA) model.

N_f	=	number of sectors in flight f's path
$P(f,i)$	=	the ith sector in flight f's path
P_f	=	$\{P(f,i) : 1 \leq i \leq N_f\}$
$D_k(t)$	=	departure capacity of airport k at time t
$A_k(t)$	=	arrival capacity of airport k at time t
$S_j(t)$	=	capacity of sector j at time t
d_f	=	scheduled departure time of flight f
r_f	=	scheduled arrival time of flight f
s_f	=	turnaround time of an airplane after flight f
c_f^g	=	cost of holding flight f on the ground for one unit of time
c_f^a	=	cost of holding flight f in the air for one unit of time
L_{fj}	=	max amount of time that flight f can arrive early to sector j
U_{fj}	=	max amount of time that flight f can arrive late to sector j
l_{fj}	=	number of time units that flight f must spend in sector j
T_f^j	=	set of feasible times for flight f to be in sector j

Table 3: Data needed for the TFMP with enroute capacities.

$$(TFMP)\quad Min \sum_{f\in F}[(c_f^g - c_f^a) \sum_{t\in T_f^k, k=P(f,1)} t(w_{ft}^k - w_{f,t-1}^k) + c_f^a \sum_{t\in T_f^k, k=P(f,N_f)} t(w_{ft}^k - w_{f,t-1}^k)]$$
$$-(c_f^a - c_f^g)d_f + c_f^a r_f$$

$$\text{subject to} \sum_{f:P(f,1)=k} (w_{ft}^k - w_{f,t-1}^k) \leq D_k(t), \qquad \forall k \in \mathcal{K}, t \in \mathcal{T} \tag{9}$$

$$\sum_{f:P(f,N_f)=k} (w_{ft}^k - w_{f,t-1}^k) \leq A_k(t), \qquad \forall k \in \mathcal{K}, t \in \mathcal{T} \tag{10}$$

$$\sum_{f:P(f,i)=j,P(f,i+1)=j',i<N_f} (w_{ft}^j - w_{ft}^{j'}) \leq S_j(t), \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{11}$$

$$w_{f,t+t_{f,j}}^{j+1} - w_{ft}^j \leq 0, \qquad \forall f \in \mathcal{F}, j \in P_f, t \in T_f^j \tag{12}$$

$$w_{f,t}^k - w_{f',t-s_{f'}}^k \leq 0, \qquad \begin{cases} \forall (f',f) \in \mathcal{C}, t \in T_f^k, \\ k = P(f,1) = P(f',N_f) \end{cases} \tag{13}$$

$$w_{f,t}^j - w_{f,t-1}^j \geq 0, \qquad \forall f \in \mathcal{F}, j \in P_f, t \in T_f^j \tag{14}$$

$$w_{ft}^j \in \{0,1\}, \qquad \forall f \in \mathcal{F}, j \in P_f, t \in T_f^j \tag{15}$$

Table 4: Formulation of the TFMP with enroute capapcities.

EXPERIMENTS WITH COMPLEX INTEGER PROGRAMMING MODELS FOR THE AIRPORT GROUND DELAY PROBLEM

by

Tom Mifflin
Federal Aviation Administration

Arjang Assad and Michael Ball
DISTINCT Management Consultants and U. of Maryland

Larry Stone and Terry Thompson
METRON

ABSTRACT

In this talk we will present the models and analyses that have led to the development of a prototype system, OPTI-FLOW. The purpose of the OPTI-FLOW system is to support the operational decisions made by the controllers in the FAA Air Traffic Control System Command Center (ATCSCC -- also known as the Central Flow Control Facility). ATCSCC has general responsibility for the coordination of the activities of the Air Route Traffic Control Centers. Thus, the ATCSCC specialists are involved in decisions regarding the implementation of a variety of controls, such as ground delays and ground stops issued to aircraft on the ground, and airborne holds and reroutings. There is a division of responsibility between ATCSCC and local controllers with respect to many of the above controls. In particular, in some cases the ATCSCC has direct responsibility and in other cases it serves as an advisor or intermediary. ATCSCC has direct responsibility for implementing Estimated Departure Clearance Time (EDCT) Programs, also known as ground delay programs. An EDCT Program is a revision to the original scheduled takeoff times of aircraft destined for a particular airport. EDCTs are typically implemented when the arrival capacity at an airport is reduced.

Bad weather is the most common reason for a capacity reduction. Ultimately, the OPTI-FLOW system is intended to support a wide range

of the activities of the central flow specialists. The current system addresses the single airport ground delay problem.

The starting point for our work is a set of papers by Bertsimas, Odoni and their students. In the modeling paradigm used, time is broken down into discrete intervals (typical interval lengths might be 5 or 10 minutes). Activity scheduling consists of assigning an activity to one such interval. Any more refined scheduling would be carried out in a post-processing step. The ground delay problem would be solved for airport A when the maximum arrival capacity at airport A was decreased below its normal level. The decrease would have to be such that the airport could not accept all of the planes at their normally scheduled arrival times. The airport's maximum arrival capacity would be converted into arrival time interval capacities at airport A. For example, if the arrival rate was 30 planes per hour and the time interval length was 10 minutes then the arrival interval capacity would be 5. The output of the ground delay problem is the assignment of flights destined to airport A to arrival time intervals at airport A. The simplest version of this problem can be formulated as a transportation problem. We enhanced the model in several ways based on an extensive analysis of the requirements of the ATCSCC specialists, various airline interests, including the issue of fairness to airlines, and general public welfare. Some of the specific enhancements we included are:

1.) We found it necessary to add so-called "bank preservation" constraints to this basic model. These constraints insure that a "bank of flights" associated with a particular airline's hubbing operation remain close to one another in the revised schedule.

We considered three alternate approaches for handling these constraints and found that the performance of general purpose integer programming solvers varied substantially depending on which formulation was used.

2.) We devised a complex function for generating the model's costs. This cost function considered the length of the delay received by a flight, the type of flight and fairness among airlines.

3.) The actual problem has certain important dynamic and stochastic aspects. In particular, information regarding the arrival airport's capacity depends on the weather and is thus stochastic. Furthermore, this information is updated regularly.

At the same time, decisions on when a flight can take off can be revised until a certain amount of time before the flight's scheduled departure. We present results of experiments with a number of sample scenarios in which the model was dynamically resolved over time in response to varying conditions.

4.) Although the ultimate decision regarding ground delays rests with the ATCSCC, the airlines can cause substantial revisions to the final program through the use of a "cancellations and substitutions" mechanism. This mechanism allows airlines to cancel certain flights and to substitute one flight into another flight's arrival time slot. It is typical for the airlines to use this mechanism quite extensively, which in turn can have a major impact on the effectiveness of a ground delay program. We present results of experiments which show how our model can be used to respond to airline actions in order to produce new schedules that improve the overall schedule quality.
The results we present are based on a prototype system we built on a Unix workstation using OSL as the integer programming solver. The system design and the models we use reflect substantial interaction with ATCSCC specialists, other FAA personnel and airline industry experts.

We also plan to present preliminary results on an advanced model that more explicitly handles certain stochastic problem aspects.

SOME MODELS AND ALGORITHMS FOR EN ROUTE AIR TRAFFIC FLOW MANAGEMENT

V. Tošić, O. Babić, M. Čangalović and Đ. Hohlačov

Faculty of Transport and Traffic Engineering
University of Belgrade
Vojvode Stepe 305, 11000 Belgrade, Yugoslavia

ABSTRACT: The paper considers the air traffic Flow Management Problem (FMP) for the airspace case (bottlenecks are not only airports but could also be e.g. sectors). Only temporal solutions are allowed which means that no alternative routes can be used, thus holds are the management instruments to be used. The formulation of a general model is first proposed and then a more precise and realistic BIP-based model is developed. Solutions are obtained using LP supplemented by a heuristic, or a heuristic only. Some of the results of numerous experiments using real-life data are presented for illustrative purposes.

1. INTRODUCTION

Early research efforts in the area of en route flow management were reported by Bianco et al. (1981), Bielli et al. (1982) and Bailey and Smith (1983). The FMP (Flow Management Problem) was appropriately discussed and defined by Odoni (1987). FMP in the airspace or on the air route network has been more specifically discussed by Tosic and Babic (1992) along with a review of the existing efforts to solve the problem (Tosic, 1993). Some results in solving the problem are presented here.

Generally the problem to be solved (FMP), as presented here, can be defined in the following way:

The system supplying service to customers consists of a finite number of elements. Each element can handle several customers simultaneously but has a finite capacity which may vary in time. Individual customers demand service from one (rarely) or more elements sequentially. Customer demands are initially independent. It can clearly happen that total demand exceeds some elements' capacity at certain periods. The assumption is that after such an overloaded period there is a lower-demand-than-capacity period when the overflow of demand from the previous period can be served completely. The question to answer is: how to manage demand so as to minimize the delay cost to customers, or more precisely, which customers should be delayed and for how long, in order to produce minimum total cost.

2. GENERAL MODEL

One can try to formalize in a very general way the above-described problem as applied to the Air Traffic Flow Management Problem. In the first step, time will be considered a continuous variable, while the system elements as well as customers will be discrete (which corresponds to their nature, anyway). Now the problem can be stated as:

Find delay t_i for each flight (customer) i to minimize

$$\sum_i c_i(t_i) \tag{1}$$

where

$c_i(t)$ nonnegative delay cost function for flight i and $c_i(0)=0$,

subject to

$$\sum_i q_{ij}(t) \le Q_j(t), \quad \forall j, \forall t \tag{2}$$

where $q_{ij}(t)$, $Q_j(t)$ are non-negative functions:

$q_{ij}(t)$ - capacity of element j used by flight i at time t
$Q_j(t)$ - total capacity of element j at time t

The capacity of element j used by flight i at time t, $q_{ij}(t)$, can be expressed as the demand of flight i for that capacity at time $(t-t_i)$, $d_{ij}(t-t_i)$, where $d_{ij}(t)$ is given demand, i.e.

$q_{ij}(t) = d_{ij}(t-t_i)$ for all t such that $t-t_i \geq 0$,

so that the constraints (2) become

$$\sum_{i:t-t_i\geq 0} d_{ij}(t-t_i) \leq Q_j(t), \quad \forall j, \forall t \tag{3}$$

Namely, one should try to find the optimum (minimum) delay vector by minimizing (1) so that capacity constraints (3) are satisfied at all times.

As mentioned earlier, the FMP is always considered over a finite time interval during which there are subintervals when demand exceeds capacity. All the demand not served at the time it occurs will be served later during the considered interval, or at the latest immediately after the interval. Consequently, flight delay cannot be greater than the length of this interval. This means that the interval should be extended long enough until cumulative surplus capacity is sufficient to serve demand spillover from the bottlenecks in that period.

Formally, this means that variables t_i and parameters t in the general model belong to the finite time interval [0, T] where

$$\sum_i d_{ij}(t) > Q_j(t)$$

for at least one j and at least one t from [0, T].

3. DISCUSSION OF THE OBJECTIVES

In their previous work (Tosic and Babic, 1992) the authors discussed several metrics that can be used in delay measurements when solving the FMP. It turns out that the proposed general model and the objective function (1) include and formalize all such possible metrics which will be discussed here one by one.

1. If all flights have the same delay cost c which does not depend on delay length, then the delay function is

$$c_i(t_i) = \begin{cases} c, & t_i > 0 \\ 0, & t_i = 0 \end{cases} \tag{4}$$

where c = const. > 0.
Now the objective function (1) is reduced to the total number of delayed flights, not taking into account aircraft and flight type or delay length. This would be a rather simplistic egalitarian form of the model.

2. If the cost of delay is linear with time and all flights have equal delay cost c per time unit, then $c_i(t_i) = c\, t_i$ and the objective function (1) can be reduced to

$$\sum_i t_i \tag{5}$$

which represents the total delay.

3. If the delay cost functions are linear with time, and delay cost per time unit differs from flight to flight (or is fixed for a group of flights) then $c_i(t_i) = c_i t_i$ and the objective function (1) is

$$\sum_i c_i \cdot t_i \tag{6}$$

4. Sometimes longer delay can be penalized by setting an appropriate (non-linear) form of the delay function keeping it the same for all flights, i.e. $c_i(t_i) = c(t_i)$ for all i. The objective function is now

$$\sum_i c(t_i) \tag{7}$$

5. Finally, in the most general case the delay cost can be nonlinear with different forms from flight to flight. So the objective function is

$$\sum_i c_i(t_i) \tag{8}$$

Thus, depending on the concrete form of the objective function, i.e. the chosen metric for delay measurement, the proposed general model can represent different cases of the FMP.

4. DISCRETIZATION OF THE MODEL

It was found that a general (conceptual) model in which time is a continuous variable is difficult to make operational, i.e. so that it gives a solution, a vector of optimal delays. Even in the case of discrete time (when such delays should be integers) an adequate model has not been found at this stage of research. A solution is proposed herein based on a discrete case model and previous work by Babic et al. (1992a, 1992b). This model in FMP terminology could be classified as multi-element, multi-period, deterministic and static (Odoni, 1987) and can be formulated in the following way:

In practice, flow control deals with a time interval divided into a finite number of periods rather than with a continuous time variable. On the other hand, the capacity of the elements is often given as the number of flights handled by an element of the system per unit of time. Therefore, starting from the general model, time can be treated as discrete by dividing the considered interval [0, T] into periods k of equal length, $k = 1,2,...,K$, where K is the total number of such periods. In this case, variables t_i take integer values, $t_i \in \{0, 1, ..., K\}$, while the constraints (3) become

$$\sum_{t:k-t_i>0} d_{ij,k-t_i} \le Q_{j,k} \quad , \forall j, \forall k \tag{9}$$

Thus, d_{ijk} is 1 if flight i demands service at element j during period k, and 0 otherwise, and Q_{jk} is the maximum number of flights that can be served by element j during period k.

Since in such discrete models variable t_i appears in constraints (9) as an index, this makes it impossible to solve the model directly for t_i. Therefore, it is proposed that a new binary variable $x_{i,l}$ be introduced for each flight i and each of its possible delays l. This variable equals 1 if flight i is delayed l periods, and 0 otherwise. Obviously

$$\sum_{l=0}^{l_{max}} x_{i,l} = 1 , \quad t_i = \sum_{l=1}^{l_{max}} l \cdot x_{i,l}$$

and

$$c_i(t_i) = \sum_{l=0}^{l_{max}} c_i(l) \cdot x_{i,l}$$

Value ${}_il_{max}$ is the maximum possible delay of flight i, ${}_il_{max} = K + 1 - {}_iP_a$, where ${}_iP_a$ is the first period during which flight i requires service in the system (arrival period).

Now the discrete model derived from the general one becomes:

$$(\min) \sum_{i=1}^{M} \sum_{l=0}^{l_{max}} c_i(l) \cdot x_{i,l} \tag{10}$$

subject to

$$\sum_{l=0}^{l_{max}} x_{i,l} = 1, \quad i=\overline{1,M} \tag{11}$$

$$\sum_{i=1}^{M} \sum_{l:k>l} d_{i,j,k-l} \cdot x_{i,l} \le Q_{j,k}, \quad j=\overline{1,N} , \; k=\overline{1,K} , \tag{12}$$

$$x_{i,l} \in \{0,1\}$$

where M is the total number of flights, N the total number of elements and the second sum in the constraints (12) is established over all possible delays l of flight i, $l \in [0, {}_il_{max}]$, less than period k.

The discrete model developed is a problem of linear Binary Integer Programming (BIP). The BIP model as formulated here always has a feasible solution.

Some very important features of the proposed model are:

First, it covers all different cases of the FMP defined by objective functions (4)-(8).

Second, the model stays linear even when the delay cost is not linear, because the nonlinear dependence is expressed here through the objective function coefficients.

Solutions to the defined cases, based on the BIP model, will be presented and discussed further in this paper. Special attention will be paid to minimizing the total number of delayed flights and minimizing the total delay.

5. MINIMIZING THE NUMBER OF DELAYED FLIGHTS

If the objective function has the form (4), the BIP model is reduced to minimizing

$$(\min) \quad \sum_{i=1}^{M} \sum_{l=1}^{{}_i l_{max}} x_{i,l}$$

subject to (11) and (12). This model is, obviously, equivalent to maximizing the total number of flights without delay, i.e.

$$(\max) \quad \sum_{i=1}^{M} x_{i,0} \tag{13}$$

subject to (11) and (12).

In order to solve the problem (13), several heuristic algorithms have been developed. They are all based on the following lexicographic approach:

Each algorithm first sets as many flights as possible without delay. Then it tries to sequentially maximize the number of remaining flights served with 1,2, ..., K-1 periods of delay.

Formally, at each step l (l=0,1, ...) the algorithm tries to solve the following linear BIP model:

$$(\max) \quad \sum_{i \in F_l} x_{i,l} \tag{14}$$

subject to

$$\sum_{i \in F_l} d_{i,j,k-l} \cdot x_{i,l} \le Q^{l}_{j,k}, \quad j=\overline{1,N}, \quad k=\overline{l+1,K}$$

$$x_{i,l} \in \{0,1\}$$

F_l is the set of all flights i still unserved after l steps, with a delay of at least l periods, i.e. their maximum possible delays ${}_i l_{max}$ are greater then l. Obviously, $F_0=\{1,2,...,M\}$. $Q^{l}_{j,k}$ is the available capacity of element j during period k, remaining after l steps, and $Q^{0}_{j,k}=Q_{j,k}$.

Then, all flights $i \in F_l$ for which ${}_i l_{max}=l+1$ and $x_{il}=0$ in the solution to the model (14), are served with delay ${}_i l_{max}$. This means that they are all served after the considered K periods.

The algorithm is completed after step l if $F_{l+1}=\varnothing$, i.e. all flights have been already served.

Depending on the ways of solving the problem (14), two types of heuristics are proposed: heuristics based on bottlenecks and heuristics based on linear programming (LP) relaxation.

A short description of both types is presented here.

5.1 Heuristics based on bottlenecks

A pair (j,k) is a bottleneck if during period k the total demand in element j exceeds its capacity, i.e.

$$\sum_{l=1}^{M} d_{ljk} > Q_{j,k} .$$

The heuristics proposed here treat only flights which appear in bottlenecks and, consequently, consider only elements and periods used by these flights.

In the initial step of such heuristics, the size of the problem could be reduced by applying the following reduction procedure:

Form a set of all bottlenecks. Remove from set {1,2,...,M} all flights which do not request service in any bottleneck and consider the reduced set as F_0. For each element/period (j,k) update the available capacity subtracting from it the total number of removed flights which request service in (j,k). Consider the new capacity to be $Q_{j,k}$. Remove from the problem all such element/periods in which no flights from F_0 request service.

Flights removed by the reduction procedure are served without delay.

Further, heuristics solve the sequence of problems (14), using at each step one of two principles: either load the system until its capacities are reached ("INFMIN" heuristic) or decrease the number of flights requesting service in bottlenecks ("MMBFR" heuristic).

The "Maxmin pipe cross section" (INFMIN) algorithm considers flights as particle fluid flows through pipes of different cross sections. Namely, at step l, each element/period (j,k) is considered to be a segment of a pipe with a cross section equal to the available capacity $Q^l_{j,k}$.

Thus, each flight from F_l can be represented as a flow of fluid being considered for service along a pipe that has several segments with different cross sections. These segments correspond to elements/periods where the flight is considered for service, with delay l. Now, any flight from F_l that passes through the pipe with the maximum minimum cross section is served (with delay l). After that, the cross section in each segment of this pipe is updated, i.e. reduced by one, and the next flight is served based on the same criterion. The procedure continues until the minimum cross sections for all remaining flights from F_l are 0, i.e. these flights cannot be served by delay l. Then, the procedure goes to step $l+1$.

The "Maxmax bottleneck flow reduction" (MMBFR) algorithm, at step l, finds the overload

$$\sum_{i \in F_l} d_{ijk-l} - Q^l_{j,k}$$

for each element/period (j,k), $k>l$, that represents a bottleneck at this step. Then, it tries to remove from F_l as few flights as possible such that all overloads are eliminated. Namely, each flight that is considered for service with delay l in at least one bottleneck, is also considered as a flow of fluid through a pipe with several segments of different cross sections. But here the segment represents such a bottleneck and its cross section is equal to the corresponding overload. Now, the flight that passes through the pipe with the maximum cross section is removed from F_l. After that, cross sections in all segments of this pipe are reduced by one and the next flight is removed based on the same criterion. The procedure is stopped when the cross section (overload) is equal to 0 in all bottlenecks. Then all remaining flights from F_l are served (with delay l) and the procedure goes to step $l+1$.

A detailed description of previous heuristics is presented in Babic et. al. (1992).

5.2 Heuristics based on LP relaxation

As usual, the linear BIP model (14) can be solved by applying at each step l a standard LP package to its Linear programming relaxation in which $0 \le x_{il} \le 1$.

If such a solution is integer, then all flights of F_l are served according to this solution. But if it is partially non-integer, then an additional heuristic procedure is used: the integer part of the solution is accepted and, consequently, the corresponding available capacities are updated. Now the described procedure, applied at step l in one of the heuristics INFMIN or MMBFR, is used on the set of all flights i with non-integer values of the variables x_{il}. In that way a complete integer solution to the model (14) is found.

Obviously, if the LP solution is integer at the first step $l=0$, the corresponding value of the function (14) represents the maximum number of flights without delay, while a complete schedule of all flights, obtained by such heuristics, is an optimal solution to the model (13).

6. MINIMIZING TOTAL DELAY

When objective function (5) is used, the BIP model expresses minimization of the total delay, having the form

$$(\min) \quad \sum_{i=1}^{M} \sum_{l=1}^{l_{max}} l \cdot x_{i,l} \tag{15}$$

subject to (11) and (12).

The LP relaxation approach can be also used for solving the model (15):

LP relaxation of the model (15) is solved using a standard LP package. If the solution is integer, it is an optimal one and the value of the objective function (15) represents the minimum total delay. Otherwise, each flight i with value 1 for corresponding variable $x_{i,l}$ is served with delay l. Then, all flights with non-integer values for some of the corresponding variables are served by applying one of the two heuristic approaches:

In the first one, the minimum delay $l(i)$ for which $x_{il(i)}$ is non-integer is determined for each flight i. In the second approach $l(i)$ is obtained such that the value of $x_{il(i)}$ is maximum among the values of all variables corresponding to flight i. In both approaches each flight i is not allowed to be served with a delay less than $l(i)$.

Then flights are completely served by using the slightly modified INFMIN or MMBFR algorithm: At each step of the algorithm, set F_l contains only flights which are allowed to be delayed l or more periods (i.e. for $i \in F_l$, $l(i) \le l$) and, therefore, this set can be empty. Consequently, the algorithm starts from the first step l for which F_l is non-empty and omits all further steps with empty sets of unserved flights.

Obviously, all other cases of the FMP, expressed by the BIP model with concrete forms of the functions $c_i(l)$ in the objective (10), can be solved by the proposed heuristics based on LP relaxation. Of course, in each of these cases, some new heuristic procedures might be developed for the non-integer part of an LP solution. They should take care of the concrete objective function. But, at this moment it was decided to use the heuristic approach already designed for the objective (4), and then verify it through numerical experiments. Actually, for the practical problems considered in Section 7, the number of flights with non-integer values in an LP solution was shown to be rather small and, consequently, it seems that applying this heuristic to them, did not have too much influence on the final integer solution.

All types of heuristics proposed in Sections 5 and 6 will be applied to some real-life problems. The numerical results obtained for different cases of the criteria will be discussed in detail.

7. EXPERIMENTS

Parallel with the development of the models, numerical experiments were conducted to learn more about the modelled process as well as the models under development.

The main objectives of the experiments were to:

- learn more about the structure of the problem
- compare different models developed regarding:
  - the "optimum solutions" they offer
  - the structure of the "optimum solutions"
  - computing time
  - utilization of the available resources (capacity)
- test the models in real-life situations dealing with small-medium size problems which could be managed using the available computing facility (PC)

In the beginning phase of the research numerical examples were hypothetical and smaller in size. Usually examples with the following general characteristics were used:

- three Air Traffic Control (ATC) sectors
- three time periods
- 30 flights
- one bottleneck per sector (in the second period)

Experiments with the above numerical example size were very useful primarily to become more familiar with the structure of both the problem and the solutions offered using different models under development. The size of the problem allowed for (visual) inspection and hand calculations to be used easily. This helped to better understand both the problem and the solutions and to proceed more easily during the modelling process. It was particularly interesting to see the distribution of delays and even more which particular flights were penalized based on different solutions. It is well known that the problem of different "slot assignment" procedures is rather delicate and controversial in real life. Any kind of automation of this assignment even if it pretends to be an "optimization" should be carefully examined.

At a later phase in research some real-life size problems were considered during the experiments. Available real data were used for these experiments.

A problem was noted in the upper airspace area consisting of four sectors handling nearly only fly overs. The peak periods, when the bottlenecks occurred, were 2-3 hours long. Numerous experiments were conducted and a typical example will be presented and discussed in this paper.

A sketch of the spatial pattern of the considered airspace with four sectors is shown in Fig. 1. Sectors are labeled due to their geographical position. The Figure also includes the traffic demand pattern during the considered four-hour time interval. Lines show the flow direction through the four sectors, while associated numbers, at the exit point, show total flow intensities during the considered time interval (e.g. flow N-E-S consists of 15 flights during the observed four-hour period).

A more detailed problem description is presented in Figure 2 in matrix form. It is also adjusted to the model input format. The considered time interval is divided into eight half-hour periods (this is the usual time used in the slot allocation process).

The first matrix represents a discretized demand pattern, i.e. demand per sector-period. Estimated sector capacity is eighteen flights per period and in this case is constant in time and the same for all sectors.

The second matrix shows the occurrence of bottlenecks, i.e. sector-periods where demand exceeds capacity, together with the overload values. It can be seen that such situations occur from the 2nd to the 7th period.

On the other hand, the third matrix shows sector-periods with extra capacity. Naturally the existence of such periods after overloaded periods in all sectors with bottlenecks is a necessary condition for a feasible solution.

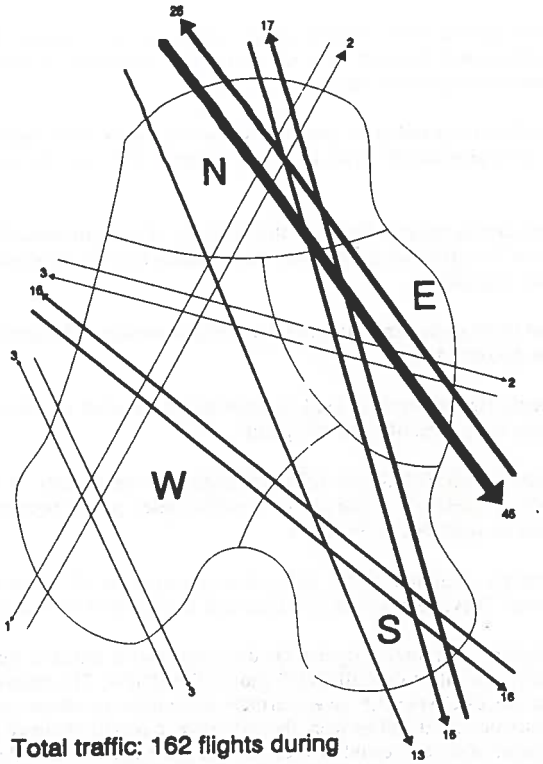


Figure 1 - Example: Traffic Scheme

Sector	Period 1	2	3	4	5	6	7	8
N	10	22	18	21	19	21	18	16
E	9	16	16	18	23	20	21	18
S	11	18	15	9	17	12	11	11
W	5	12	9	6	14	10	9	7

Demand

Sector	Period 1	2	3	4	5	6	7	8
N		4	0	3	1	3	0	
E				0	5	2	3	
S		0						
W								

Overload

Sector	Period 1	2	3	4	5	6	7	8
N	-8		0				0	-2
E	-9	-2	-2	0				-2
S	-7	0	-3	-9	-1	-6	-7	-7
W	-13	-6	-9	-12	-4	-8	-9	-11

Extra capacity

Figure 2 - Example: Model Input Data

In both matrices, sector-periods where demand equals capacity are also shown. These situations are ideal from the demand management point of view. Of course they should be understood as planned on the strategic level, i.e. when schedules are designed.

It should be mentioned that demand in the period after the considered time interval decreases further and in the busiest sector is approximately equal to half of capacity. This was the condition determining the considered interval.

The last two matrices should roughly illustrate the difficulty of the problem. Of course the space/time combinatorial nature of the problem prevents one from finding a trivial solution treating the matrix row by row, i.e. each sector separately.

After the definition of input some explanation of the concrete models and algorithm versions used in the experiment should be presented.

Seven different heuristic algorithms were used to solve the five models of the above- described sample problem. Basically they can be classified in two groups.

The first two algorithms LEXHEUR and LEXLIN minimize the number of delayed flights using a lexicographic approach. LEXHEUR is one of the heuristics based on the bottlenecks, while LEXLIN is based on LP relaxation as described in Section 5.

The other five algorithms minimize global delay costs incurred by all aircraft appearing during the considered time interval. They are based on LP relaxation as explained in Section 6.

The two algorithms dealing with total delay, i.e. the delay cost that is linear in time and all flights having the same unit delay cost, are called GLOBLIN1-T and GLOBLIN2-T. The difference between the two is in the way they treat the case when LP gives partially noninteger solutions, i.e. each uses a different heuristic as explained earlier. As it will be seen, the difference in results obtained using the two heuristics is very small so that either of the two could be used. On the other hand, if the LP solution is integer there is no need for any heuristic at all. These conclusions apply to the following algorithms as well, so only one solution will exist in the next cases.

Algorithm GLOBLIN-CT solves the model in which delay cost is linear in time but flights i have different unit delay costs c_i (case (6) in Section 3). In this experiment it is assumed that coefficients c_i are a function of aircraft size. Aircraft size is represented by the number of seats. The values of coefficients for different aircraft size is given in Table 1. This model can be understood as a minimization of global passenger delay rather then global aircraft delay.

Algorithm GLOBLIN-T2 applies to the case when all flights i have the same delay cost that is quadratic in time, i.e. $c_i(t) = t^2$ (case (7), Section 3). This can be one of the ways to penalize longer delays.

The last algorithm GLOBLIN-CT2 solves the model in which a flight has quadratic delay costs with different coefficients corresponding to its aircraft size i.e. $c_i(t) = c_i\, t^2$.

A comparison of the solutions offered by the above listed algorithms is provided using several levels.

The first and the most general one is shown in Table 2. It presents the number of flights handled on time, its complement, i.e. the number of delayed flights and the total delay.

A more detailed illustration of the incurred delay is shown in Figure 3 where the distributions of delay are presented. This way of presenting does not differ between aircraft.

The next way of comparing the solutions offered by the different models takes into account which particular flights are delayed by each of the solutions. This can be of interest for the final user of the system, the airline or its passengers. The presentation is shown in Figure 4. Mere inspection indicates that one deals with some solutions which are rather similar or some others which differ from each other. This

Figure also includes an illustration of aircraft size which is of interest in understanding some of the solutions. A detail of the same presentation is given in Table 3 for a sequence of forty flights.

Aircraft capacity (seats)	Coefficient
0 - 15	0.1
16 - 50	0.4
51 - 100	0.8
101 - 140	1.2
141 - 180	1.6
181 - 260	2.3
261 - 320	3.0
321 - 450	4.0

Table 1 - Aircraft Classification: Coefficients

		No. of flights with no delay	No. of delayed flights	Total delay [half-hour periods]
LEX	LEXHEUR	150	12	57
	LEXLIN	150	12	42
GLOB	GLOBLIN1-T	143	19	36
	GLOBLIN2-T	141	21	35
	GLOBLIN-CT	145	17	43
	GLOBLIN-T2	124	38	38
	GLOBLIN-CT2	130	32	44

Table 2 - Example: Model Output

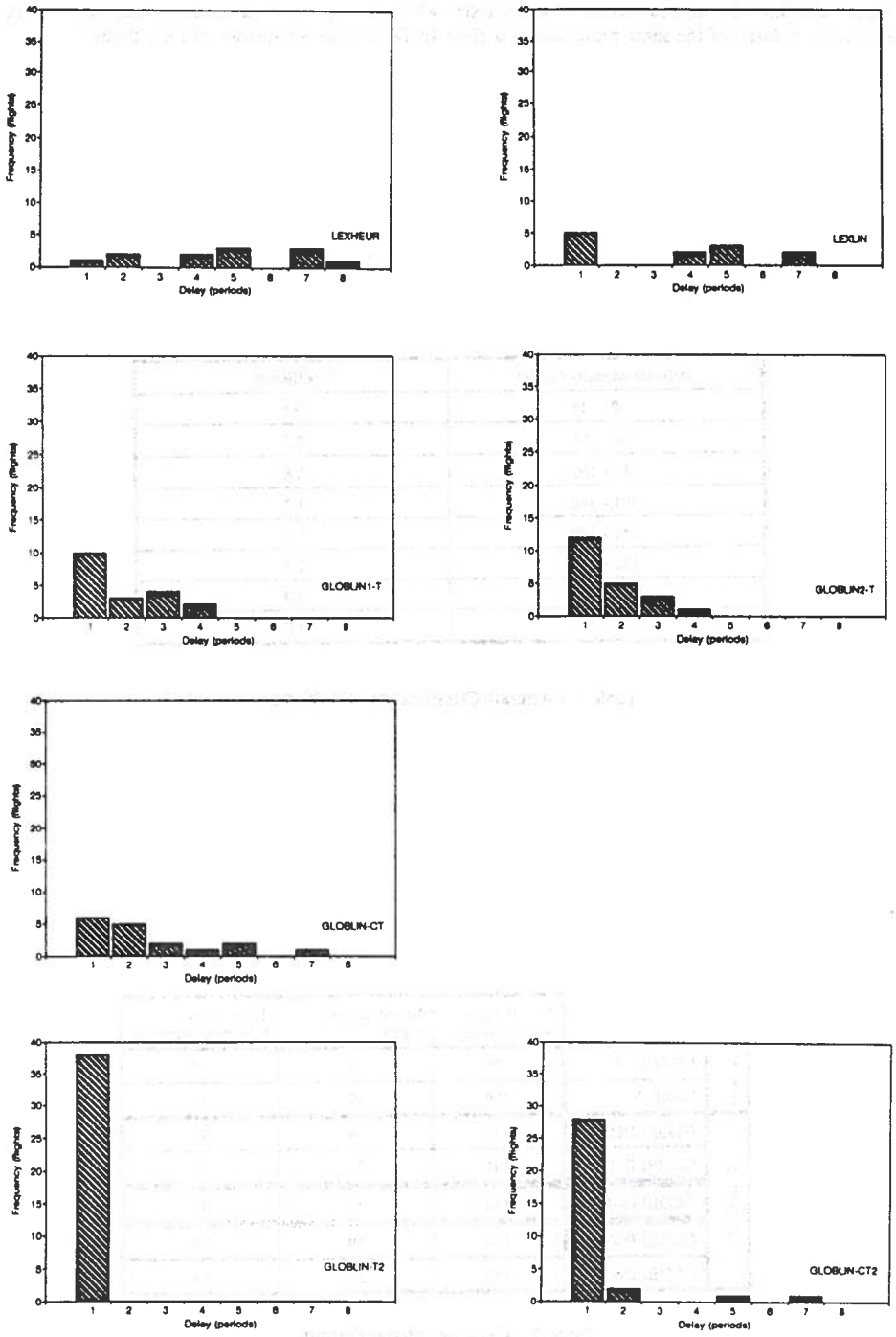


Figure 3 - Example: Model Output - Delay Distribution

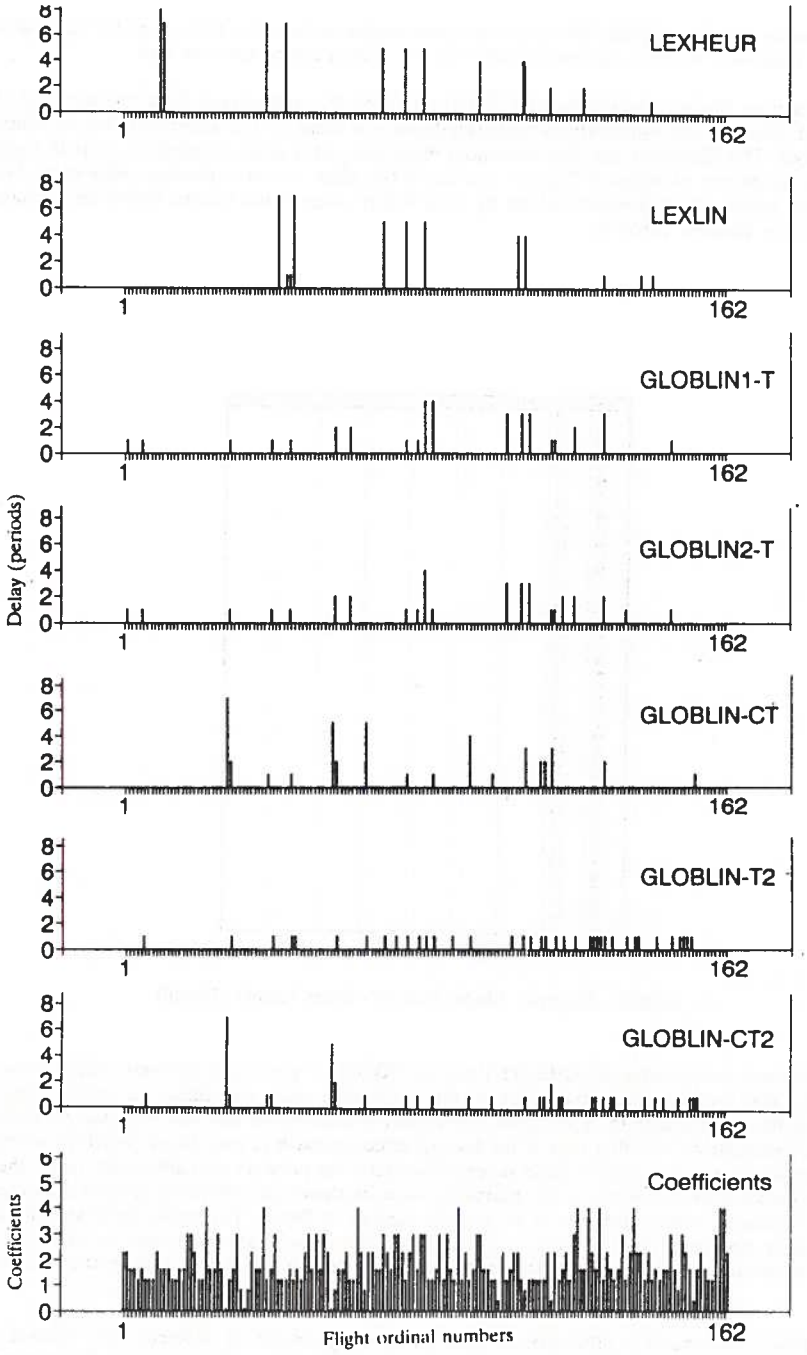


Figure 4 - Example: Model Output - Delay Pattern

Table 2 reveals that the solutions offered for the lexicographic model differ from those for global models. However, differences between the results inside the two groups can be noted as well.

Application of the first two algorithms, LEXHEUR and LEXLIN, produces solutions where only 12 flights are delayed. However the two solutions differ significantly in terms of total incurred delay, 57 compared to 42 periods. This difference can also be noticed when comparing delay distribution plots in Figure 3. Inspecting the pattern of delays in Figure 4 one can notice some similarity between these cases; besides the fact that greater delays are incurred but by fewer flights, some of the delayed flights are the same in both cases (see detail in Table 3).

Fl.No	Cost	LEXHEUR	LEXLIN	GLO1-T	GLOB-T	GLO-CT	GLO-T2	GLO-CT2
40	1.6			1	1		1	1
41	3.0							
42	1.2		7					
43	1.2							
44	1.2	7	1					
45	1.6		1	1	1	1	1	
46	1.2		7				1	
47	1.6							
48	1.2							
49	2.3							
50	3.0							
51	1.6							
52	3.0							
53	1.6							
54	3.0							
55	2.3							
56	0.1					5		5
57	0.8			2	2	2	1	2
58	1.6							
59	1.6							
60	2.3							
61	1.2			2	2			
62	1.2							
63	4.0							
64	2.3							
65	0.8					5	1	1
66	2.3							
67	2.3							
68	1.6							
69	1.6							
70	3.0	5	5				1	
71	2.3							
72	1.6							
73	3.0						1	
74	3.0							
75	2.3							
76	1.2	5	5	1	1	1	1	1
77	2.3							
78	3.0							
79	1.6			1	1		1	1
80	3.0							

Table 3 - Example: Model Output - Delay Pattern (Detail)

The application of algorithms GLOBLIN1-T and GLOBLIN2-T gives quite different results from those offered by the lexicographic model, but at the same time they are similar to each other. Both characteristics are of course to be expected. The former because of the fact that the basic objectives are different, particularly if the first step of the lexicographic approach is considered (which is: as many of flights should leave on time (in the planned period) whatever the price (in later delays) is). Later, the great similarity between the two results is also normal because the algorithms differ only in the non-integer part of the LP solution, which happens to be a small number of flights. The above remarks can only be confirmed by observing Figures 3 and 4. The number of delayed flights is higher but the total delay smaller. It can also be seen that the pattern of penalized flights is not the same as in the two previous cases.

GLOBLIN-CT introduces a difference in treatment among aircraft of different size. Table 2 shows similarity in the number of delayed flights with previous global-type model, but somewhat greater total delay. This delay, as meant by the definition of the objective function, is incurred by different, smaller aircraft, and the solution is quite different as can be seen in Figure 4. The details can also be seen in Table 3.

The solution by GLOBLIN-T2 which penalizes harder (quadratic in time) longer delays, in this case gives equally distributed delay (one period, see Figure 3 and 4) between all delayed flights. Total delay is moderate but the number of delayed flights is significantly greater than in any other case. On the other hand, none of the flights incurs a delay longer than half an hour. This policy is quite opposite from the one imposed by the first two global models.

The last of the applied algorithms GLOBLIN-CT2 balances aircraft size and long delay penalty, and offers again quite a different solution. All of the delayed flights have a short delay except some of the flights performed by the smallest aircraft which are severely penalized, see Figure 4 and Table 3.

Comparing solutions obtained by different models from the "customer's" point of view, one can see that different policies produce quite different effects on the customer's population. This is true even when the customers are unlabeled as presented in this example. However, they all have flight numbers attached to them and no airline would like to be number 42 or 70 in this example if lexicographic policies are applied.

There are two other aspects to be discussed here.

One is how the supply side behaves in the experiment or how well utilized it is. The results of different policies are presented in Figure 5 (number of flights assigned to each sector-period) in such a way as to illustrate how the system is loaded during the considered 4-hour interval. It can be seen that the number of fully loaded sector-periods varies between 11 and 13, and it should be noted that the pattern of "fully utilized" elements (shaded) is not quite the same in different solutions. More careful inspection of the matrices will show that load pattern generally differs between the presented cases (by one or two flights per sector-period). This difference is not so important from the strict capacity aspect because it is known that the acceptable sector load or capacity cannot be easily determined and represented by a "crisp" number (eighteen flights per half hour in this particular case). A more general comment should be underlined here. It is the importance of the interconnection of flights (customers) using (consuming) the ATC system (complex resource). As an illustration of this fact one can notice that in spite of the fact that there is no overload in the first period (see Figure 2) there are differences in sector loads in that period in solutions obtained using different models (see Figure 5). The temporal solution allows only positive delays, but the fact that a flight does not necessarily hit a bottleneck as soon as it enters the considered airspace but somewhere further during its intended path produces this quasi back effect. This of course can be easily explained by the space/time combinatorial nature which is the essence of this problem.

The other aspect of the model application and applicability is computing time. All experiments were conducted using a 486/33 PC and a commercially available LP package. For a problem size similar to the presented one (four sectors, eight time periods, 150-200 flights) the following computing times were typical: lexicographic based heuristics (LEXHEUR) 1 second, lexicographic LP-based heuristics (LEXLIN) 10 seconds, and algorithms for global LP-based models 10 to 30 minutes. It should be mentioned that in these cases there were less than 10 flights with non-integer solutions in the initial step (LP relaxation). The later presented LP with about 1000 variables and about 200 constraints. This poses the question of the applicability of LP-based models for larger sized problems. In the described experiment the authors did not have the opportunity to use more sophisticated computing facilities, but it is reasonable to assume that such equipment (both hardware and software) would be used if some FMP decision support of similar optimization nature if applied in real life. This belief is supported by the fact that there are reports (Vranas, Bertsimas and Odoni, 1992) showing that similar types of problems one order of magnitude larger in size can be solved in reasonable time (taking into account that this problem is FMP rather then ATC and it does not require a strictly real-time solution).

8. DISCUSSION AND CONCLUSIONS

The development of the models herein described and the experience gained by experimenting with them, helped to better understand the nature of the FMP problem and possible solutions. This is particularly true for the core of the problem: the decision who to delay and for how long, which, in the authors' knowledge, has attracted much less attention in European research than research devoted to the two-way communication between flight requesting services and the decision maker authorities.

LEXHEUR

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	18	18	18	<9+6
E	8	15	14	17	18	18	18	18	<9+7
S	11	18	14	9	17	12	11	11	<9+0
W	5	12	9	6	13	10	9	6	<9+2

LEXLIN

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	18	17	18	<9+5
E	9	16	13	18	18	18	18	18	<9+4
S	11	17	13	10	17	12	11	11	<9+0
W	5	11	10	6	14	10	9	6	<9+1

GLO1-T

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	17	17	18	<9+9
E	7	17	16	18	18	18	18	18	<9+8
S	9	18	15	10	16	11	11	12	<9+2
W	5	11	10	5	14	11	9	7	<9+0

GLO2-T

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	18	17	18	<9+9
E	7	17	16	18	18	18	17	18	<9+9
S	9	18	15	10	16	12	10	13	<9+1
W	5	11	10	5	14	11	9	7	<9+0

GLO-CT

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	18	18	18	<9+8
E	9	14	14	18	18	17	18	18	<9+9
S	11	17	14	10	16	13	11	12	<9+0
W	5	11	10	5	14	11	9	7	<9+0

GLO-T2

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	18	18	18	<9+9
E	8	16	15	18	18	18	18	18	<9+9
S	10	17	14	10	15	13	9	13	<9+3
W	5	11	10	4	16	10	8	8	<9+0

GLO-CT2

Sector	Period 1	2	3	4	5	6	7	8	9
N	10	18	18	18	18	18	18	18	<9+8
E	8	14	15	18	18	18	18	18	<9+9
S	10	18	14	9	16	14	8	11	<9+4
W	5	12	9	5	14	11	9	7	<9+0

Figure 5 - Example: Model Output - System Load

It should be underlined that the proposed models have a nice feature in that they can handle any type of delay cost function using basically the LP relaxation algorithm.

To bring this type of model closer to real life, further steps should be undertaken. Some of them concern policy matters while others are of a technical nature.

On the policy side it has been shown that there are differences between results obtained using models based on the assumptions of different ways of expressing delay costs. The choice of the delay cost measurements, as well as delay distribution between flights, i.e. customers, is up to the airlines involved in the process and their associations. Their conclusions should be transformed into policies to be applied in practical decision making in the FMP area.

One can imagine that airlines would come to the conclusion that some type of "fair" or at least controlled delay distribution should be applied. They might want to mark in advance their flights with priority among those considered for delay. This can be based on their importance due to the transfers, rotation policy, importance of the O-D pair, etc. The proposed types of models can easily force any flight into the solution. This even helps in the sense that the size of the combinatorial problem decreases. The other possibility would be to develop some type of a memory of delays incurred by each airline. This memory could be used to balance the delay between airlines over a given period, e.g. a week, which is the scheduling cycle. Namely, towards the end of the cycle the ones that had experienced more delays would have priority, i.e. their flights should be forced without delay in the solutions.

On the technical side, to make this type of model applicable, i.e. possibly useful and used for FMP decision support, significant efforts should be made to develop appropriate preprocessing and particularly postprocessing tools. The latter should examine different suitable ways of presenting the results, whether to be approved by a human operator, or to be analyzed if automatic decision making in FMP is applied.

If the algorithms are made fast enough from the user's point of view (running in minutes in the authors' opinion) the static approach could be transformed into a dynamic one by applying the model periodically during the day, e.g. every hour using updated information about both demand and supply sides.

Finally, it should be mentioned that besides the temporal solution applied here to the FMP, in the case of airspace application a spatial solution should be looked in. The research direction to be explored is in the rerouting area.

8. REFERENCES

Bianco L., Bielli M., Cini M., Grippo L. and Ricciardelli S. (1981), "New Approaches to Air Traffic Control and Management", Transport Research and Social and Economic Progress, ed. S.Yerrel, London, Vol.4, pp.2373-2384.

Bielli M., Callichio G., Nicoletti B. and Ricciardelli S. (1982), "The Air Traffic Flow Control Problem as an Application of Network Theory", Computers and Operations Research, Vol.9, No.4, pp.256-278.

Bailey M.G. and Smith J.U.M. (1983), "A Computer Aid for Dynamic Air-Route Restructuring", paper presented at the 6th EURO Conference, Vienna, July 19-23, 1983.

Odoni A.R. (1987), "The Flow Management Problem in Air Traffic Control", Flow Control of Congested Networks, NATO ASI Series, Vol.38, Springer Verlag, Heidelberg, pp.269-288.

Vranas P., Bertsimas D.J. and Odoni A.R. (1992), "The Multi-Airport Ground-Holding Problem in Air Traffic Control", working paper, Alfred P. Sloan School of Management, MIT, February 1992.

Babic O., Teodorovic D., Hohlacov Dj., Krcmar-Nozic E., Dolzan T. and Tosic V. (1992), Flow Control Algorithms For En Route Air Traffic Control - Initially Proposed Heuristic Algorithms, Division of Airports and Air Traffic Safety, Faculty of Transport and Traffic Engineering, University of Belgrade, Belgrade.

Tosic V. and Babic O. (1992), "Air Traffic Flow Management Modeling, Part I - Problem Statement, Discussion and Definition", Proceedings SYM-OP-IS '92 (Symposium on Operations Research), Belgrade, Yugoslavia, pp. 219-222.

Babic O., Krcmar-Nozic E., Teodorovic D. and Hohlacov Dj. (1992), "Air Traffic Flow Management Modeling, Part II - Problem Formulation and Proposed Solution", Proceedings SYM-OP-IS '92 (Symposium on Operations Research), Belgrade, Yugoslavia, pp. 223-226.

Babic O., Tosic V., Hohlacov Dj. and Dolzan T. (1992), "Air Traffic Flow Management Modeling, Part III - Problem Solution, Numerical Examples and Conclusions", Proceedings SYM-OP-IS '92 (Symposium on Operations Research), Belgrade, Yugoslavia, pp. 227-230.

Bojovic J. and Despotovic S. (1992), "Analysis of Peak Week Traffic in the Belgrade Area Control Center", graduation thesis, Division of Airports and Air Traffic Safety, Faculty of Transport and Traffic Engineering, University of Belgrade.

Tosic V. (1993), "Air Route Flow Management - Problems and Research Efforts", paper at the invited session "International Air Traffic Control", IFORS 93, XIII World Conference on Operations Research, Lisbon, Portugal.

Tosic V., Babic O., Cangalovic M. and Hohlacov Dj. (1993), "Air Traffic Flow Management Modeling, Problem Formulation and LP-based Solution", Proceedings SYM-OP-IS '93 (Symposium on Operations Research), Belgrade, Yugoslavia, pp. 335-338.

Bijedic A. (1994), "Experiments with some Flow Management Problem Models", graduation thesis, Division of Airports and Air Traffic Safety, Faculty of Transport and Traffic Engineering, University of Belgrade.

A dynamic assignment model for Congested Radial Transit Networks

Slim SAÏDI [1,2] *Michel GENDREAU* [2] *François SOUMIS* [1,2,3]

(1) École Polytechnique de Montréal
(2) Centre de Recherche sur les Transports
(3) Groupe d'Étude et de Recherche en Analyse des Décisions
Montréal, Québec, Canada

The urban stretching from city centers to suburbs has created a growing imbalance between business centers, job locations and residential areas. This phenomenon has led to significant increase in both the number and the length of journeys between these locations. In many cities, this increase in transportation demand has not been matched by a comparable increase in supply, resulting in a severe misfunction of urban transit systems. This can be observed especially in high population density cities where slower speeds, irregular frequencies and customer dissatisfaction are common. This situation has added to the difficulty of planning in a generally capacitated network, where any unexpected change may lead to major changes in the whole network operation.

A key element in urban transit planning is the ability to predict how passengers will behave for a given level of supply (i.e. for a given network configuration with fixed frequencies and operating rules). This problem is generally referred to as the passenger assignment problem. Given the

complexities of the problem, it is not possible at this time to derive a general model for it.

So far, transit assignment which corresponds to a dynamic process has been approached from a static point of view: only means values are considered. The reason beyond this is the multitude of parameters and their inter-relation in the general case which has lead to an untractable problem. This limitation permits however good practical solutions that have even been used in applied softwares. Nevertheless, it no longer holds in the case of the capacitated networks, where any unexpected change may lead to major perturbations in the whole network operations.

Due to the complexity of the problem at hand, it becomes illusory, at this stage of knowledge, to pretend the universality of any solution and the particularities of the networks and the context are therefore to be exploited.
In our case we consider a typical congested radial bus network which does not allow free bus exchanges: *Tunis'*.

In this work, we propose a novel approach to model capacity in congested systems. The approach attempts to account for the dynamics of user behaviour when faced with heavily loaded buses by introducing different categories of users and moving the waiting passengers from a category to the next one. Simulations are performed on a small example to validate the choices that were made.

After the presentation of the problem, we introduce a new set of definitions of the capacities while dealing with the problem of an isolated bus stop; we

then describe the global model before giving some results of numerical tests that can justify our modelling choices and finally conclude with some hints about some possible extensions to the study.

AN EQUILIBRIUM ASSIGNMENT MODEL FOR INTERCITY TRANSIT NETWORKS

Agostino Nuzzolo

Department of Civil Engineering

University of Rome "Tor Vergata"

Via della Ricerca Scientifica

00133 Roma

Tel +39+6+72594588 Fax +39+6+72594586

Francesco Russo

Department of System Engineering and Mathematics

University of Reggio Calabria

Via Emilio Cuzzocrea 48

89126 Reggio Calabria

Tel. +39+965+875232 Fax 39+965+875227

ABSTRACT

The majority of models developed in literature for transit assignment use the hyperpath concept and are based on the hypothesis of a level of adaptive user choice. Such a hypothesis results incongruous with systems characterized by a high degree of punctuality and medium-low frequency, like the intercity system. For these systems the problem of departure run choice arises, because this can imply periods of delay or anticipation for the users compared to the Origin or Destination target time. In this case then the real timetable must be taken into account.

In this work the problem of intercity and extraurban transit assignment will be dealt with through a network approach which explicitly considers the timetables and with a random utility model of the choice involving departure time and path. The spatio-temporal network used is of a diachronic type and it therefore allows the dynamic assignment of demand.

In the first part, the assignment approach formulated by the authors is described, reporting the user's departure time and path choice model and the diachronic network model. In the second part, the problem of the demand-supply interaction is dealt with under hypotheses of utility and cost function in order to assess the existence and uniqueness of the equilibrium state. In the final part some applications of the assignment model are described.

This work has been supported by the National Research Council of Italy, grant 93.01882. PF74.

1. INTRODUCTION

From the assignment model point of view, systems of collective transport can be classified on the basis of some characteristics concerning service: the principal ones are the frequency of the offered service and the punctuality involved in the arrival of vehicles compared to the established timetable.

In general a high frequency and a low punctuality, typical in urban areas, leads to a random arrival of the user at stops with a large margin of adaptivity in the choice of path. The models of supply representation, path choice and assignment which are more suited to this type of service and which have been developed in the literature originated in the concept of hyperpath (Nguyen and Pallottino, 1986, 1988; Cascetta and Nuzzolo, 1988; Spiess and Florian, 1989; Russo 1991).

For services with medium-low frequency and high punctuality, typical of national, regional and in general extraurban systems, the hypothesis of adaptive choice proves incongruous and is replaced with the hypothesis of a completely consumptive choice.
Because of the medium-low frequency of the service, which may entail with regard to the user considerable anticipation or delay times relative to departure and arrival target times, specific run choice assumes great importance. Furthermore, if capacity constraints occur, further anticipations or delays connected with them must be allowed for.

The problem of intercity transit assignment taking the timetable into account, has been dealt with in various ways in the literature, through the enumeration of the feasible paths.
From a bibliographical analysis it emerges that various works dealing with some specific problems concerning assignment intercity models have been published. On the other hand there are few examples involving the complete treatment of assignment models. One may recall the works developed for regional areas by Abkowitz (1981) and by Hendrikson and Plank (1984), as well as the procedure developed by the air transport companies in the 1980s (Etschmaier and Mathaisel, 1985) and the specific works by Sangiorgio and Bason (1990), LAST (1991), Nuzzolo and Russo (1993).
At the same time some fundamental steps forward have been achieved by researchers in the sector of urban private transport, where studies have evolved in the direction of treating variable time explicitly. Network assignment models have been widely developed which explicitly establish an intra-period evolution with reference to arrival and departure target times. A general framework of the state of the art of such models was reconstructed by Alfa (1986), while more recent developments have focused on verifying the existence of the solution with deterministic (Drissi, 1990; Janson, 1991) or stochastic user behaviour (Ben Akiva et alii, 1986), or on the evolutions of the system with further day to day dynamics (Cascetta, 1989; Cascetta and Cantarella, 1993).
In this work the problem will be dealt with through a network approach which explicitly considers the timetables and with a random utility model of choice involving departure time and path. The spatio-temporal network used is of a diachronic type and therefore allows the dynamic assignment of the demand too.

In the first part, the assignment model formulated by the authors (1994) is described, reporting the user's departure time and path choice model and the diachronic network model. In the second part, the problem of the demand-supply interaction is dealt with various hypotheses of utility and cost function. In the final part some applications of the assignment model are described.

2. THE UTILIZED APPROACH

2.1 The departure time and path choice model

The existing departure time and path choice models can be classified in different ways: aggregated and disaggregated, according to whether the users are considered in aggregated or disaggregated ways; behavioural and descriptive, according to whether we do or we do not make explicit hypotheses concerning the user's behaviour.
The hypotheses in this approach imposed on the user's behaviour are:
A) With regard to the high level of punctuality of the service the path choice is only of a consumptive nature: the choice concerns contemporaneously path (boarding, alighting and interchange stops) and departure time.
B) In order to choose between the numerous alternatives the user considers as utility components: 1) the access time from origin to first stop; 2) the waiting time at stops; 3) the on board time; 4) the time and inconvenience

for any eventual interchange; 5) the egress time from the last stop to the destination; 6) the fare; 7) the comfort related to the degree of crowding; 8) the disutility for the anticipation or for the delay compared to the origin or destination target time.
C) The user is a rational decision-maker who considers all the paths which connect the relation O-D, and chooses the alternative which involves the perceived minimum utility; this utility is a random variable and therefore the chosen alternative cannot be predicted with certainty but only the probability can be calculated, for example with a random utility model. The random utility model maintains its validity in spite of the punctuality of the trips and the user's hypothesised knowledge of the timetables, for the following reasons:

1) different users can evaluate in a different way the same attributes (for example if the user is more or less aware of access and egress mode or of the interchange time);
2) some attributes evaluated by the user may not be included in those defined in the preceding point B (vehicle characteristics, etc);
3) some attributes can be evaluated by the analyst in an approximated way (interchange, comfort, access and egress times, etc).

Different types of behavioural models exist, the most commonly employed being part of the group of "random utility" models. The differences between various models in this group derive from the hypothesis concerning the distribution of the random term involved in the user's utility. However the models employed in practice mainly refer to two distributions: Weibull-Gumble (WG) and Normal Multivariate (MVN).

The other classification concerning existing models is relative to the path choice: between modes or within modes.
The following belong to the first group:
- Abkowitz (1981) includes 12 departure time choice alternatives each using car and bus, and specific linear and non linear terms for anticipation and delay.
- Hendrikson and Plank (1984) include up to 28 alternatives obtained with 4 modes and 7 different departure times and allow for non linear terms of anticipation and delay.
Various models which generally refer to the car mode belong to the second group:
- Ben Akiva, De Palma and Kanaraglou (1986) present 3 possibilities. Leaving early to avoid the traffic, arriving early, leaving and arriving late, taking an alternative longer path; the authors specify linear terms but introduce a discontinuity by means of an indifference time strip on arrival.
- Alfa (1986) explicitly introduces the variable DTT (destination target time), and verifies the existence of a correlation between the flexibility of arrival at work and the uniformity of the demand in the peak-hour; he considered linear terms without discontinuity.
- Mannering (1988) hypothesized a distribution of path changes of Poissonian type, including the non-change option; he considered flexibility concerning the time at which one starts work.
- Biggiero, Cascetta and Nuzzolo (1992) consider the WG distribution and the width of the indifference strip is estimated compared to the DTT, thereby using a discontinuous function.

Calibrated ad hoc models for home to work and home to school travel will be reported below. They represent departure time models with a path choice between competitive transit modes.
Two classes of models were specified and calibrated (Nuzzolo and Russo 1994): multinomial logit and probit, with the joined utility function of the alternative "j" of the type:

$$U_j^i = \sum_k \beta_k X_{kj}^i + \tau_j \qquad j \in I_i$$

where X_{kj}^i is the k-th attribute for alternative "j" and user "i" τ_j is the random term for alternative j, and I_i is the choice-set for "i", the considered attributes are reported in Tab. 1.

The data utilized were obtained through surveys performed in a region of southern Italy. The destination survey was based on a random sample, with more than a hundred useful interviews with workers and students (University). In the questionnaire there were three different sets of questions; the first concerned the user's socio-economic characteristics, the second, the time and cost of the choice path, and the last, the time and cost of the alternative path by mode and run.

The alternatives considered are a maximum of four for each mode, two that give destination or origin delay time and two that give the anticipation time. The alternatives are enclosed in the choice set if the referred time (origin or destination) is inside the strip of an hour and a half around the target time of the user.
The attributes examined ca be divided into two main groups: a) user specific attributes, obtained by surveys; b) trip level of service attributes for each mode, obtained by official time-table. The trip purposes considered are home-work and home-school. The modes available are: train and bus.
In the calibration of the logit model, the specification was developed by focusing attention of the various components of the travel time and of the anticipation-delay time.

We used various functional forms for the anticipation/delay time attribute: linear, non linear with and without indifference strip. Various specifications on the part of the utility function concerning travel time were tried for each functional form of the anticipation and delay time; furthermore the travel cost was considered also in relation to the user's income group. The best model obtained is reported in Tab.1.
The use of more complicated functional forms than the linear ones of the anticipation and delay time does not offer great advantages concerning the models' performances. For this reason the subsequent calibrations, involving the Probit model, were referred to the linear form.
Various models were calibrated in connection with the typical potentiality of the Probit to introduce different variances among the alternatives, and non zero covariances. From the comparison between logit and probit models, it emerges that the hypothesis of non zero covariance does not always give better results.

2.2 The diachronic network model

The network model used (Fig. 1) represents each run of each line with the explicitation of the variable time; this network is called the diachronic network. The diachronic graph consists of two subgraphs: one relates to run representation and the other is dependent on the demand structure.

For every stop j there is a temporal axis on which for every run r, a node n_j^r which is relative to the stopping time of run r, is positioned. This node is connected by means of a waiting-boarding link to node n_{jb}^r, which is relative to the departure time of run r. It is also connected by means of an alighting link from node n_{ja}^r, relative to the arrival time of run r. The link (n_{ja}^r, n_{jb}^r) represents the standing time at stop j of run r; the link (n_{jb}^r, n_{ka}^r) represents the connection between stop j and stop k by means of run r.

The second subgraph is composed of various subgraphs, at least one per zone centroid, in relation to the structure of the matrix O-D, which is supposedly known for the purpose and the destination target time (DTT) or origin target time (OTT).
In general the distribution of the DTT or OTT can be discretisized with reference to a certain number of time intervals, where it may be accepted that the demand distribution is concentrated at the final point of the interval. A temporal axis is then considered for every zone and a temporal centroid, positioned in the axis on the final point of every interval, will correspond to each time interval.
Every temporal centroid can therefore be the origin and the destination of user trips, and every trip can be with the DTT and/or OTT.

Given a matrix O/D of dimensions $\left[n_O * (i_{OTT} + 1), n_D * (i_{DTT} + 1)\right]$, where n_O is the number of spatial origins, n_D the number of destinations, i_{OTT} and i_{DTT} the number of interval into which the time bands referring to origin and destination for the target departure and arrival times are divided; in each zone it is possible to associate as many temporal centroids concerning origin n_{OTT} and destination n_{DTT} as there are specified time intervals with an additional one for the trips without target.

In order to connect the temporal centroid axis and the spatial centroids with the run subgraph, the access and egress links need to be introduced. Such links have one extremity on the centroid temporal axis or on the spatial centroid and the other extremity on the temporal axis of the stop , each of which guarantees connection with an accessible run in the zone. The first extremity of the generic access link (n_o^r) represents the time at which it is necessary to leave the point of origin in order to reach stop j (second extremity) in time to take the choice run r (n_j^r).

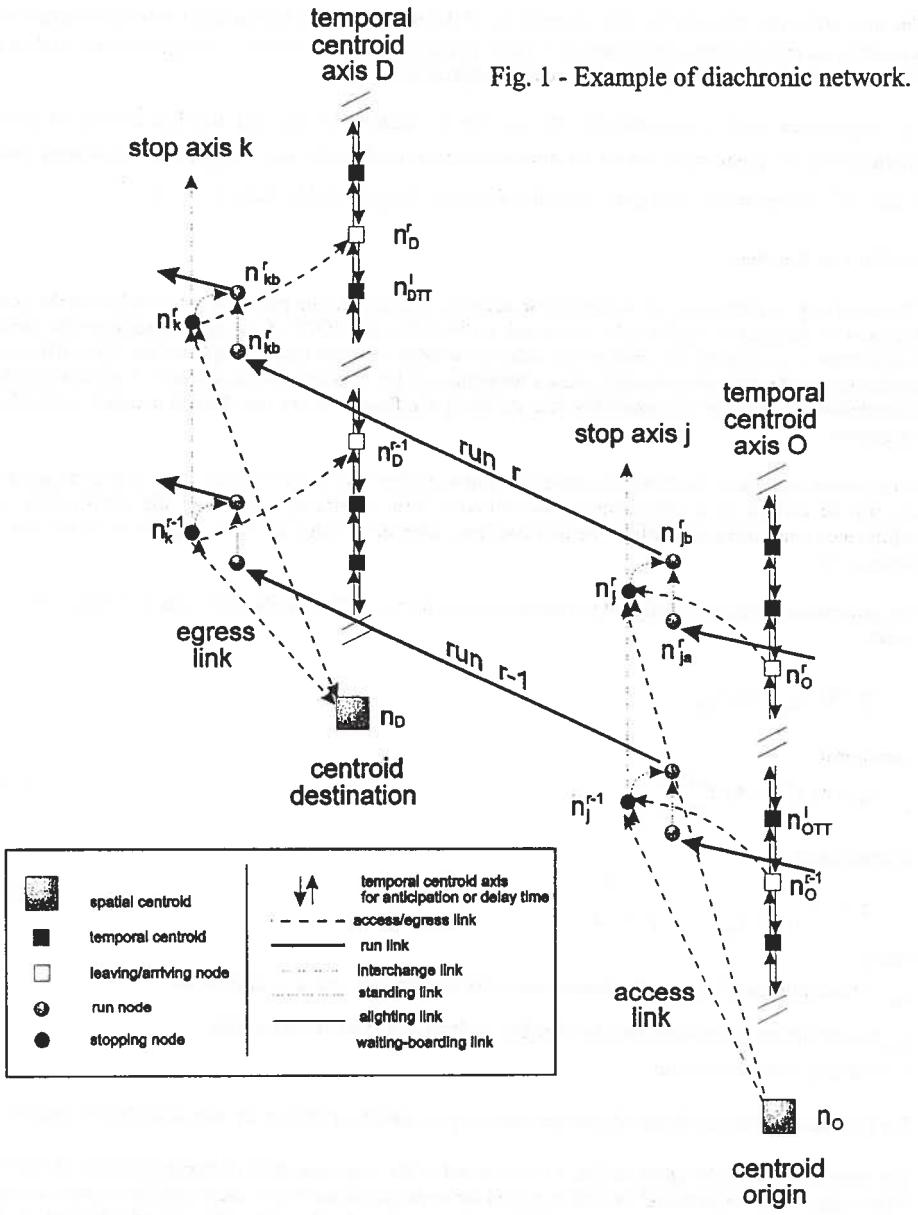


Fig. 1 - Example of diachronic network.

The time difference between the first extremity n_o^r of the access link and the temporal centroid of origin n_{OTT}^t represents exactly the time of anticipation or delay connected to the utilization of the generic run r taken into account. A similar consideration is also true for the egress link.

Fig. 2 represents an OTT-type user trip. The user has two paths at his disposal: the first, leaving the point of origin at time n_o^r along run r, entails the disutility of delay produced by link $\left(n_{OTT}^t, n_o^r\right)$; the second, leaving at time n_o^{r-1} along run r-1, entails the disutility of anticipation produced by link $\left(n_{OTT}^t, n_o^{r-1}\right)$.

2.3 The cost functions

The qualifying characteristic of the diachronic network derives from the possibility of introducing the penalty functions of anticipation and/or delay compared to the DTT and/or OTT of the user, directly on the network. Furthermore it is possible to consider the different weights of delay and anticipation for the different user categories and the different purposes, using a temporal axis for each purpose or category. A second qualifying characteristic is given by the possibility of considering the flow on every run, thereby treating explicitly the congestion.

In the literature penalty functions are expressed both in a linear and in a non linear form; and in addition they can still be divided in a continuous or discontinuous form, according to whether the introduction of an indifference time strip within which the user does not suffer any penalty, has been allowed for or not (see also section 2.1).

The general exemplificative forms of the three functions for the DTT case (the OTT case is similar) are:
- linear

$$t_l = \beta_1 t_{ear} + \beta_2 t_{lat} \tag{2.1}$$

- non linear

$$t_n = \beta_3 t_{ear}^{\beta 4} + \beta_5 t_{lat}^{\beta 6} \tag{2.2}$$

-discontinuous

$$t = \begin{cases} 0 & t < \Delta \\ t_l \text{ or } t_n & t > \Delta \end{cases} \tag{2.3}$$

where:

t_{ear} = time difference between the destination target time and the arrival at destination

t_{lat} = time difference between the time of arrival at destination and the target time

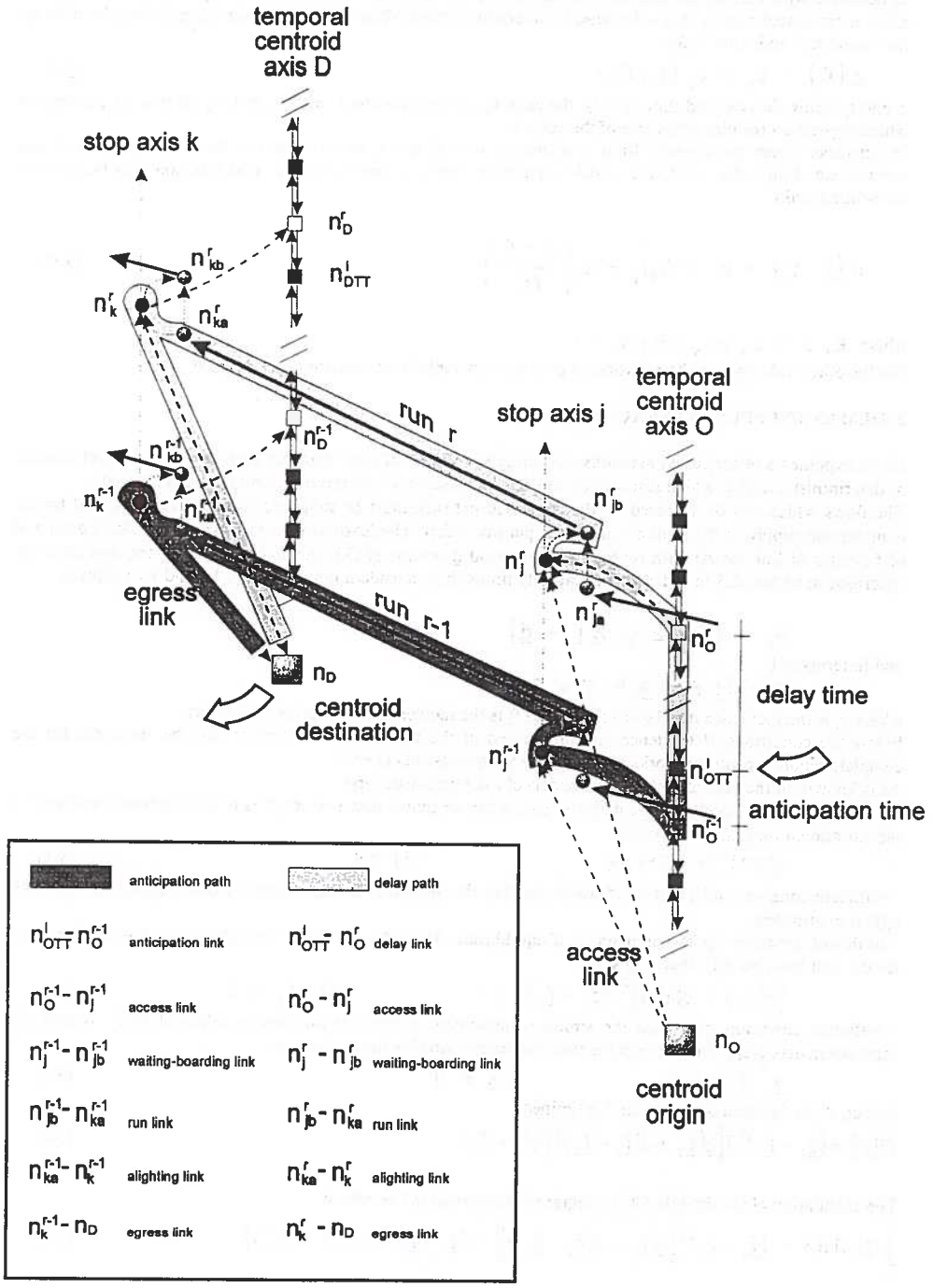
Δ = width of indifference strip.

The β values are obtained from a departure time and path choice model like the ones described in section 2.1.

The representation of the graph in Fig. 1 is connected to the implementation of linear functions of type (2.1). Where algorithms "on network" are still used and the implementation of non linear type functions is desired, it is necessary to introduce more complex forms of the centroid subgraph, with the introduction of direct connecting links between each temporal centroid and the departures of all runs included in the selected interval.
It must be noted that the additivity of link costs, and therefore the use of linear penalty functions are the conditions for the use of the traditional assignment "on network" algorithms.

The treatment of cost functions which are variable with flows is not covered by this paper. Nevertheless, it is appropriate to examine the specifications which may be obtained in relation to "waiting", "standing", and "run" links (Fig. 1), since they are relevant to the treatment of congestion.

Fig. 2 - Alternative paths for declared origin target time.



Congestion produces two main effects: with an increase in crowding user comfort decreases and, once capacity is reached, the "run " link may become inaccessible to new users.
In networks where it may be assumed that the service still remains available to "waiting" users, the congestion effect is translated merely into a decrease in on-board comfort. Thus a separable cost function may be used for the "standing" and "run" links:

$$c_i(f_i) = k_1 + k_2 (f_i / Q_r)^{k_3} \qquad (2.4)$$

c_i and f_i being the cost and flow, and Q_r the capacity of run r on link i, with $k_1 \geq 0$, $k_2 \geq 0$ e $k_3 \geq 1$ parameters which depend on the characteristics of the system.
In networks where the capacity limit is activated, with fj and f_i respectively the flows on the waiting and incident standing links, and Q the vehicle capacity on run r, a "non-separable" cost function may be used for the waiting links:

$$c_j(f_i, f_j) = k_4 + k_5 f_j + k_6 \left(\frac{f_i + f_j}{Q_r} \right)^{k_7} \qquad (2.5)$$

where $k_4 \geq 0$, k_5 e $k_6 > 0$ e $k_7 \geq 1$.
For the other link types in the network, in general "separable" cost functions may be used.

3. DEMAND/SUPPLY INTERACTION

In the hypotheses of congested networks with other hypotheses of rigid demand, path choice models of random or deterministic utility, which respectively created a stochastic or deterministic problem, can be used.
The flows which can be achieved on the congested network must be subjected to conditions imposed by the demand and supply, if the limit of capacity is put into effect. The feasibility set of $\underline{F}$ (vector of path flows) and of $\underline{f}$ (vector of link flows) with regard to the demand $\underline{d}$ (vector of O/D demand) and using the cost function described in section 2.3 in order to take capacity limits in to consideration, with high k_3 and k_7 values is:

$$S_F = \{\underline{F}:\underline{F} \geq \underline{0};\ \underline{B}\,\underline{F} = \underline{d}\}$$

and in terms of $\underline{f}$:

$$S = \{\underline{f}:\underline{f} = \underline{A}\,\underline{F},\ \underline{F} \in S_F\}$$

where $\underline{A}$ is the incidence matrix links-path and $\underline{B}$ is the incidence matrix pairs O/D-paths.
Below the conditions of existence and uniqueness of the equilibrium solution $\underline{f}^*$ will be recorded; for the complete report consult the works by Sheffi(1985) and Cascetta (1990).
As is known, in the area of path choice models of a deterministic type:
- if the vector of link costs is indicated with $\underline{c}(\underline{t})$, it can be shown that a vector $\underline{f}^*$ is of equilibrium if and only if the variational inequality is proven:

$$\underline{c}(\underline{f}^*)^T(\underline{f} - \underline{f}^*) \geq 0 \qquad \forall \underline{f} \in S \qquad (3.1)$$

- sufficient condition such that an $\underline{f}^*$ which satisfies the inequality exists, is that the cost function of each link $c_i(f)$ is continuous;
- sufficient condition for the uniqueness of equilibrium flows $\underline{f}^*$ in S, is the strongly increasing monotonicity for the cost function $\underline{c}(\underline{f})$, that is to say:

$$[\underline{c}(\underline{f}_1) - \underline{c}(\underline{f}_2)]^T (\underline{f}_1 - \underline{f}_2) > 0 \qquad \forall \underline{f}_1, \underline{f}_2 \in S \qquad (3.2)$$

- sufficient condition, such that the strong monotonicity of the cost function is achieved (3.1), is that the Jacobian matrix $\underline{J}(c[\underline{f}])$ of the cost function vector $\underline{c}(\underline{f})$ satisfies the expression:

$$\underline{x}^T \underline{J}\,\underline{x} > 0 \qquad \forall \underline{x} \neq \underline{0} \qquad (3.3)$$

In fact, if the function of the scalar δ is defined:

$$G(\delta) = (\underline{f}_1 - \underline{f}_2)^T \underline{J}\left[\underline{c}(\underline{f}_2 + \delta(\underline{f}_1 - \underline{f}_2))\right](\underline{f}_1 - \underline{f}_2) \qquad (3.4)$$

The calculation of the definite $G(\delta)$ integral with extremes 0,1 results in:

$$\int_0^1 G(\delta)\,d\delta = (\underline{f}_1 - \underline{f}_2)^T \left[\underline{c}(\underline{f}_2 + \delta(\underline{f}_1 - \underline{f}_2))\right]_0^1 = (\underline{f}_1 - \underline{f}_2)^T (\underline{c}(\underline{f}_1) - \underline{c}(\underline{f}_2)) \qquad (3.5)$$

If vectors $\underline{f}_1$ and $\underline{f}_2$ belong to S because of the convexity of such a set, the vector $\underline{f}_2 + \delta(\underline{f}_1 + \underline{f}_2)$ for $0 \leq \delta \leq 1$ will belong to this.
In the positivity hypothesis (3.3) on the whole of S the function $G(\delta)$ will therefore be greater than zero. For δ values belonging to the internal [0,1], the integral (3.4) will also be positive and the sufficient condition for the flow uniqueness will follow on immediately from this (3.2).
Therefore generally speaking, the condition can be presented as follows:

$$\underline{x}^T \underline{J}\left[\underline{c}(\underline{f})\right]\underline{x} > 0 \qquad \forall \underline{x} \neq \underline{0} \quad \forall \underline{f} \in S$$

As is known in the equilibrium stochastic assignment:
- the condition for the existence of the solution is still defined by the continuity of cost functions;
- the condition that $\underline{J}[\underline{c}(\underline{f})]$ is positive definite is sufficient for $\underline{f}^*$ to be unique.

Schematically it is possible to identify two approaches in order to determine vector $\underline{f}^*$ in relation to the use of separable or non cost functions.

3.1 Transit networks with separable cost functions

In these networks for the "waiting-boarding","standing" and "run" links only (2.4) type functions are put in to effect, and therefore all the cost functions are of a separable type.

The Jacobian matrix of the cost functions is in this case a diagonal matrix, with positive elements, producing a positive definite Jacobian matrix, which therefore verifies equation (3.3) for every $\underline{f}$. Thus in this case the equilibrium vector is unique. In the hypotheses where the Jacobian is a symmetrical matrix , in the deterministic area, it is shown that $\underline{f}^*$ can be calculated, thereby solving the minimum problem:

$$\underline{f}^* = \arg \min_{\underline{f} \in S} \oint_0^{\underline{f}} \underline{c}(\underline{v}) d\underline{v} \tag{3.6}$$

and given that cost functions in particular in this case are separable the (3.6) becomes:

$$\underline{f}^* = \arg \min_{\underline{f} \in S} \sum_i \int^{f_i} c_i\ (v_i) dv_i \tag{3.7}$$

In the stochastic area, with the symmetrical Jacobian matrix, $\underline{f}^*$ can still be defined as an optimum problem solution, in different ways(Powell and Sheffi 1982, Fisk 1980).
It must be noted that the main limitation of this procedure is that the users of run links where $\underline{f}_i < Q_i$, are equally penalized, irrespective of the boarding order and therefore the ensuing path flows cannot be the true ones. In order to calculate the equilibrium flows, the algorithms used in road transport networks can be used, as that of Franke and Wolfe in the deterministic case or the algorithm MSA in the stochastic case.

3.2 Transit networks with functions of non separable cost

In the networks type (2.5) functions are used for the waiting-boarding links and type (2.4) functions for the standing and run links.
The sufficient condition for existence, which has already been mentioned, is verified both for the deterministic and the stochastic equilibrium, type (2.4) and type (2.5) being continuous functions.
In the deterministic case the fulfilment of the uniqueness condition (3.3) depends on the type of cost functions used; below it will be seen how it is possible to identify sufficient conditions for which, the solution $\underline{f}^*$ if it exists in S, is unique.

In the case of type (2.5) functions, if the solution exists and if it is unique, the diagonalization algorithms cannot be directly used, because the symmetric condition which allows $\underline{f}^*$ to be calculated, solving the (3.6) are not verified.
However it is possible to use one of the algorithms for asymmetric problems with continuous cost functions (Smith 1983). In the case of stochastic asymmetric equilibrium it emerges that sufficient conditions for uniqueness have not been formulated. However a vector $\underline{f}^*$ can be calculated by means of algorithms which are analogous to those of diagonalisation (Sheffy 1985).
In order to define sufficient conditions of uniqueness in the case of type (2.5) functions with non separable costs it is possible in general to establish that the network be composed of "n" links with "a" waiting links.
Without taking away the generality it is possible to order the links vector, and anologously the function $\underline{c}(\underline{f})$, putting at first begin positions (n-2a) the links with cost separable functions and then the others [waiting and

standing links] in the subsequent positions, in such a way that if position q (after the n-2a) is taken up by a waiting link, position q-1 is occuped by a standing link incident with q.
In this case the Jacobian matrix will have a block framework of the type:

$$\underline{J} = \begin{pmatrix} \underline{D} & \underline{O} \\ \underline{O} & \underline{M} \end{pmatrix} \tag{3.8}$$

with $\underline{D}$ diagonal to positive elements and therefore positive definite and with $\underline{M}$ of (2a, 2a) dimensions, operating on an analogous partition on vector $\underline{x}$ resulting in $\underline{x}^T = \left(\underline{x}_D^T, \underline{x}_M^T\right)$ of dimensions $\underline{x}_D[n-2a,\ 1]$ and $\underline{x}_M[2a,\ 1]$; with this partition it is possible to carry out the product $\underline{x}^T \underline{J} \underline{x}$ and show under which conditions it may be positive, that is to say:

$$\underline{x}^T \underline{J} \underline{x} = \underline{x}_D^T \underline{D} \underline{x}_D + \underline{x}_M^T \underline{M} \underline{x}_M > 0 \qquad \forall\, \underline{x} \in R^n,\ \forall \underline{x}_M \in R^{2a},\ \forall \underline{x}_D \in R^{n-2a} \tag{3.9}$$

sufficient conditions for (3.9) to be valid are:

$$\underline{x}_D^T \underline{D} \underline{x}_D > 0 \qquad \forall \underline{x}_D \in R^{n-2a} \tag{3.10}$$

$$\underline{x}_M^T \underline{M} \underline{x}_M > 0 \qquad \forall \underline{x}_M \in R^{2a} \tag{3.11}$$

Eq. (3.10) is immediately verified with $\underline{D}$ as a positive definite matrix. Eq. (3.11) can be rewritten in relation to the particular block composition of $\underline{M}$:

$$\underline{M} = \begin{bmatrix} \underline{M}_1 & & \underline{O} \\ & \cdots & \\ & \underline{M}_q & \\ & & \cdots \\ \underline{O} & & \underline{M}_a \end{bmatrix}$$

where the $\underline{M}_q$ generic block, which is relative to the $\underline{m}_q$ waiting link, has the shape:

$$\underline{M}_q = \begin{bmatrix} r & o \\ s & t \end{bmatrix}$$

where:

$$r = \frac{\partial c_{q-1}}{\partial f_{q-1}} = k_2\ k_3\ \frac{1}{Q_{q-1}} \left(\frac{f_{q-1}}{Q_{q-1}} \right)^{k_3 - 1}$$

$$t = \frac{\partial c_q}{\partial f_q} = k_5 + k_6\, k_7 \frac{1}{Q_{q-1}} \left(\frac{f_{q-1} + f_q}{Q_{q-1}} \right)^{k_7 - 1}$$

$$s = \frac{\partial c_q}{\partial f_{q-1}} = k_6\, k_7 \frac{1}{Q_{q-1}} \left(\frac{f_{q-1} + f_q}{Q_{q-1}} \right)^{k_7 - 1} \tag{3.12}$$

Therefore eq. (3.11) can be rewritten as:

$$\sum_q \underline{x}_{Mq}^T \underline{M}_q\ \underline{x}_{Mq} > 0 \qquad \forall \underline{x}_{Mq} \in R^2 \tag{3.13}$$

a sufficient condition for (3.13) results in:

$$\underline{x}_{Mq}^T \underline{M}_q\ \underline{x}_{Mq} > 0 \qquad \forall \underline{x}_{Mq} \in \underline{x},\ e\ \forall \underline{M}_q \in \underline{M}$$

The so-called x_b and x_s the diffrences between the flows on the incidental boarding waiting and standing links belonging to two generic flow vectors $\underline{f}_{1q}$ and $\underline{f}_{2q}$ which constitute the generic $\underline{M}_q$, result in:

$$x_{Mq}^T A_q x_{Mq} = \left(x_s, x_b\right) \begin{bmatrix} r & o \\ s & t \end{bmatrix} \begin{pmatrix} x_s \\ x_b \end{pmatrix} = x_s^2 r + x_b x_s s + x_b^2 t \tag{3.14}$$

Tab1 - Logit models for work and school.

continuous linear attributes	*home-work*	*home-school*
On board time	- 2.42	- 1.89
Early time	- 3.96	- 5.52
Early time	- 4.12	- 6.80
Cost		- 0.78
Cost/income	- 0.08	-0.72
Access and egress modal	- 7.39	
RHO^2	0.39	0.40

Tab. 2 . Network characteristics: extraurban transit

	Reggio Calabria		Potenza	
	Nodes	Links	Nodes	Links
Run's graph	1092	2062	1824	2440
Spatial centroids	12	764	70	1164
Temporal centroids	312	3216	912	4520
Anticipation/delay nodes	872		1484	
total	2288	5942	4290	8124

Tab. 3 . Network characteristics: italian air transport

	Nodes	Links
Run's graph	1708	2113
Spatial centroids	22	854
Temporal centroids	282	3574
Antioipation/delay nodes	854	
total	3112	6541

The sign of eq. (3.14) depends on that of r, s and t and on the specification of the utilized functions. In the Appendix I some results of the (3.14) are described.

5. SOME APPLICATION OF THE MODEL

The model described in the previous sections was used for various applications in real cases, above all to test the applicability of the method using standard personal computers.

After an initial series of applications carried out on various test networks, the first application on a real network with an All-or-Nothing assignment, was conducted on part of the collective transport network in the province of Reggio Calabria (population approx. 300,000), with 10 traffic zones, 14 bus lines for a total of 72 bus runs and 37 train runs, considering two trip purposes: home-work and home-school; the day was divided in 25 homogeneous intervals. The characteristics of the graph are reported in Tab 2.
A second application was carried out on the national air transport network, considering 22 first and second level airports, with 146 connections, for a total of 426 flights, distribuited daily in 11 time intervals. The characteristics of the diachronic network are reported in Tab.3.
An AoN assignment model is used, considering 11 trip O/D matrices obtained from official O/D statistics from airport to airport.
A third experiment is currently in progress, which concerns an equilibrium assignment to the extraurban collective transport network in the province of Potenza (population 360.000), with 35 bus lines for an overall 188 runs and 62 train runs, 70 zone centroids and 25 temporal centroids, and three purposes: home-work, home-school and home-other purposes. The graph was developed in the version which makes provision for "standing" links and the doubling of the on-board node in waiting and alighting as in Fig. 1. The characteristics of the graph are reported in Tab. 2.
It is important to note that the various application were carried out using a PC with DOS; in the last application a complete utilisation of the RAM (16M) available was possible.

CONCLUSION

This paper presented a class of assignment models for collective transport with medium and low frequency service, such as the intercity one. The models consider explicitly the origin and destination target time and use a spatio-temporal network known as a diachronic network which allows contemporaneous modelling of the choice of departure time and path, and also permits treatment of congested networks with, "on-network" algorithms. Attention was also focused on the equilibrium problem in relation to the type of cost function used, also taking into consideration the case of "non separable" functions.
The final part reports some applications of the assignment model on extraurban and airline networks, with promising results for its development.

REFERENCES

Abkowitz M. D. (1981) An analysis of the commuter departure time decision. Transportation, 283-297.

Alfa A. S. (1986) A review of models for the temporal distribution of peak traffic demand. Transportation Res. 20B, 491-499.

BenAkiva M. De Palma A. Kanaraglou P. (1986) Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates. Transportation Science 20 .

Biggiero L., Cascetta E., Nuzzolo A. (1992) Analysis and modelling of commuters departure time and route choices in urban networks. Sixth World Conference on Trasportation Research, Lyon 1992.

Cascetta E., Nuzzolo A. (1988) Uno schema comportamentale per la modellizzazione delle scelte di percorso nelle reti di trasporto pubblico urbano. Int. Rep., Dipartimento di trasporti, Università di Napoli .

Cascetta E., Nuzzolo A., Rostirolla P. (1989) Optimal railway pricing: Result of a model for the Italian case, 17th PTRC Summer Meeting, 1989.

Cascetta E., (1989) A stochastic process approach to the analysis of temporal dynamics in trasportation networks. Trasportation Res. 23B pp. 1-17.

Cascetta E., Cantarella G.E., (1993) Modelling dynamics in trasportation networks, Trasportation Science.

Cascetta E., (1990) Metodi quantitativi per la pianificazione dei sistemi di trasporto, Padova: Cedam

Daganzo C. F. (1979) Multinomial probit: the theory and its application to demand forecasting,. Accademic Press, New York.

Drissi O., (1991) A variation inequality formulation of the dynamic traffic assignment problem Universitè de Montreal - Centre de recherche sur les transports- Pubblication n° 735, 1990.
Etschmaier M.M., Mathaisel D.F.X. (1985) Airline Scheduling: An Overview. Transportation Science 19, 127-138.
Hendrikson C., Plank E. (1984) The flexibility of departure times for work trips. Transportation Res. 18A, 25-36. L23
Fisk C., (1980) Some developments in equilibrium traffic assignment. transportation Research vol. 14B, 243-255
Janson B.N.,(1991) Dynamic traffic assignment for urban road networks. Transportation Res. 25 B, 1991, pp. 143-161
LAST (1991)Assegnazione a reti di trasporto collettivo extraurbano, Int. Rep., Facoltà Ingegneria, Università di Reggio Calabria.
Mannering (1989) Poisson analysis of commuter flexibility in changing routes and departure times. Transportation Research 23B, 53-60.
Nguyen S., Pallottino S. (1986) Assegnamento dei passeggeri ad un sistema di linee urbane: determinazione degli ipercammini minimi Ricerca Operativa 38.
Nguyen S., Pallottino S. (1988) Equilibration Traffic Assignment for Large Scale
Powell W.B., Sheffi Y. (1982) The convergence of equilibrium algorithms with predetermined step sires. Transportation Science Vol. 16, 45-55.
Sangiorgio, Bason (1990). La pianificazione della rete dei trasporti pubblici extraurbani nell'area metropolitana torinese. Convegno AIRO, Napoli 1990.
Transit Networks, Eur. J. of Oper. Res. 37, 176-186.
Nuzzolo A., Russo F. (1993) Un modello di rete diacronica per l'assegnazione dinamica al trasporto collettivo extraurbano. Ricerca Operativa 67.
Nuzzolo A., Russo F. (1994) Departure time and path choice models for intercity transit assignment, IATBR-94 Valle Nevado, Santiago, Cile.
Russo F. (1991) Un modello di assegnazione alle reti di trasporto collettivo. Aspetti teorici ed applicazione a casi reali. Seminario SIDT, Reggio Calabria 1991.
Sheffi Y. (1985) Urban Transportation Networks. Englewood Cliffs Prentice-Hall.
Smith M.J. -(1983) An algorithm for solving asymmetric equilibrium problems with a continuous cost-flow function. Transportation Res. 17B, 1983, pp. 365-371.
Spiess H. Florian M. (1989) Optimal strategies: A new assignment model for transit networks. Transportation Res. 23B, 83-102.
Tech. A.V. - Tpl A.V. (1993) Rielaborazione ed aggiornamento delle previsioni di traffico A.V. Milano Napoli.

APPENDIX I

Reference should be made to the relation (3.14) obtained in section 3:

$$x_{Mq}^{T} \Delta_q x_{Mq} = (x_a, x_b) \begin{bmatrix} r & o \\ s & t \end{bmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix} = x_a^2 r + x_b x_a s + x_b^2 t \qquad (1.A)$$

The sign of (1.A) depends on that of r, t, s, and on the discriminant of (1.A), seen as an equation as in x_b o in x_a.

In the case in question, with r and t always being positive, to obtain the positivity of (1.A) for any pair (x_a, x_a) the discriminant must be negative:

$$s^2 - 4rt < 0$$

or

$$s^2 < 4rt \qquad (2.A)$$

From the expressions of r,s and t (3.12) positive it results for k_S :

$$s < t \qquad \forall \underline{f}$$

Therefore we obtain:

$$4rs < 4rt$$

A sufficient condition for (2.A) is therefore:

$$s^2 < 4rs$$

or

$s < 4r$

Inserting the complete expressions (3.12) and simplifying the common terms, the sufficient condition required may finally be expressed as:

$$k_6 k_7 \left(\frac{f_s + f_b}{Q_s}\right)^{k_7-1} < 4\, k_2\, k_3 \left(\frac{f_s}{Q_s}\right)^{k_3-1} \qquad (3.A)$$

In terms of application, the effect to be obtained by using (2.4) and (2.5) is that of a greater disutility for users on waiting links in the whole domain of function $\underline{c}(\underline{f})$. It is then possible to hypothesize that k_3 e k_7 have a value between 3,4 and 5 and the value of the ratio k_6 / k_2 is 2. Equation (3.A) may be re-written as:

$$\left(\frac{f_s + f_b}{Q}\right)^{k_7-1} < 4\, \frac{k_2}{k_6}\, \frac{k_3}{k_7} \left(\frac{f_s}{Q}\right)^{k_3-1}$$

$$\left(\frac{f_s + f_b}{Q}\right)^{k_7-1} < 2\, \frac{k_3}{k_7} \left(\frac{f_s}{Q}\right)^{k_3-1} \qquad (4.A)$$

For $k_3 = k_7 = k$, (4.A) becomes:

$$\left(\frac{f_s + f_b}{f_s}\right)^{k} < 2$$

$$\frac{f_b}{f_s} < 2^{1/k} - 1$$

and supposing $v(k) = 2^{1/k} - 1$, one finally obtains:

$$f_b < v(k) f_s$$

where v(k) assumes the following values with variations in k:

k	1	2	3	4	5
v	1	0.41	0.26	0.19	0.15

For $k_7 - 1 = k_3 = k$, and supposing $f_s / Q = X$, saturation level, (4.A) becomes:

$$\left(\frac{f_s + f_a}{Q}\right)^{k} \left(\frac{Q}{f_s}\right)^{k-1} < 2\, \frac{k}{k+1}$$

$$\left(\frac{f_s + f_a}{f_s}\right)^{k} \frac{f_s}{Q} < 2\, \frac{k}{k+1}$$

$$\left(1 + \frac{f_a}{f_s}\right)^{k} X < 2\, \frac{k}{k+1}$$

$$\frac{f_a}{f_s} < \left(2 \frac{k}{k+1}\, \frac{1}{X}\right)^{1/k} - 1$$

and lastly

$$f_a < v(k, X) f_s$$

where v assumes the following values for k=5, with variations in X:

X	0.1	0.2	0.4	0.8	1
v	0.75	0.58	0.33	0.16	0.1

for k=4 we obtain:

X	0.1	0.2	0.4	0.8	1
v	1	0.68	0.41	0.19	0.1

for k=3 we obtain:

X	0.1	0.2	0.4	0.8	1
v	1.47	0.96	0.55	0.23	0.14

for k=2 we obtain:

X	0.05	0.1	0.2	0.4	0.8	1.0
v	4.10	2.65	1.58	0.82	0.29	0.15

Spatial Externalities and Networks

Jeroen C.J.M. van den Bergh
Peter Nijkamp

Department of Spatial Economics
Faculty of Economics, Vrije Universiteit
De Boelelaan 1105, 1081 HV Amsterdam
E-mail: jvdbergh@econ.vu.nl

Abstract

The spatial dimension of negatieve externalities is discussed in the context of infrastructure networks and transport. An overview is given of network characteristics and related model types, which focus on the relationship between networks, economics and externalities. Among these, multiregional and spatial price equilibrium models receive attention as tools for analysing the linkage between economy, transport and infrastructure. Next, the modelling of externalities in networks is examined by considering a number of possible formulations. The paper concludes by noting critical policy issues in the context of network analysis that should receive attention when applying models such as discussed here.

1. Introduction

Most analyses of externalities generated by economic activities adopt an a-spatial perspective. As in reality every interaction among economic actors is situated in a spatial dimension and thus involves physical transport, it is evident that such analyses always incorporate a high degree of abstraction. Apart from this general deficiency, also other reasons may be mentioned which make it desirable to devote more explicit attention to the spatial dimension of externalities. First, as both immobile economic activities and environmental processes are geographically *localized*, their interaction - on a physical level or in terms of welfare enjoyment - is also localized. By leaving out spatial disaggregation in an investigation of such phenomena, crucial information may therefore be lost. For instance, one may be unable to indicate the extent of local externalities, the substition between externalities over space, or the spatial dispersion of such effects.

In addition to the above localization element, the *interactions* between activities separated in space generates patterns of flows of people and commodities. The transport associated with this takes place over infrastructural networks, where externalities generated by mobile sources emerge on individual links. Depending on both the type of externality and the ramification of the network, the externalities may affect other actors on the link concerned, on a subset of links in the network, in the entire network, or even outside the particular network concerned. Considering such different categories of actors may be relevant from two analytical perspectives. First, one is capable of investigating questions related to different categories of actors, depending on the question whether the externalities incurred are internal or external to the particular system at hand. For some actors environmental - or in a broader sense: physical - effects are partly external and partly internal, notably accidents and congestion, and to a lesser extent pollution and noise. For other actors however, either the internal or external segment dominates, e.g. noise annoyance caused by airplanes to households. In addition, one may consider for each category of actors the effect of the interaction between individual objectives, internal and external effects, policy measures, and the different categories of actors mentioned above. In such a context the interdependence between various effects, such as congestion and environmental emissions, can be given more content.

In general, spatial analyses which focus on immobile sources of externalities deal with non-point pollution or with cross-boundary pollution in multi-region or multi-country settings. With regard to mobile sources it is interesting to note that space is often not explicitly dealt with, but instead transportation is regarded as just another activity that contributes its share in overall externality generation. In an appropriate spatial, disaggregate approach, it seems plausible to separate between externalities from infrastructure and actual movement. Combining the mobile and immobile sources of externalities, e.g., in a network approach, allows one to deal with questions related to (optimal) spatial distribution of activities and interactions, from both a short-term and a long-term perspective. A common characteristic of the non-point source, crossboundary and network externality problems - distinguishing them from non-spatial externality problems - is the fact that control or regulation of the externality is hampered by problems related to information gathering, monitoring and enforcement of regulations. In contrast, the point-source externality problem is much more easy to deal with from a policy angle, but at the same time less applicable to real-world situations.

Focusing on network externality issues leads to various research issues. Interesting questions fall into two categories: (i) what is the systems behaviour in view of the rationality of each actor, and the constraints and externalities he is confronted with? and, (ii) what effects and optimality conditions relate to the use and implementation of policy instruments to the network system? The answer to these questions may be relevant to a variety of fields, including urban, regional, environmental, transport and international economics.

2. Network Models

2.1. Introduction

Network models arise from engineering and operations research approaches which often include a micro orientation (see, e.g., Boyce, 1980; Friesz and Harker, 1985; Harker, 1987; Batten and Westin, 1990). These models are generally very detailed and especially useful for single product analysis. Their usually narrow economic scope is compensated by the appealling characteristic of dealing more accurately with freight transport. For instance, such models may reflect the real transportation infrastructure consisting of a complex network with ports and cargo terminals. Furthermore, such models are useful in dealing with multi-modal issues.

On the whole, it seems that network modelling has incorporated more elements from microeconomic theory than the other way around (see Friesz and Harker, 1985). Florian and Los (1982) provide a formal general representation of extensions of variable transportation cost and multiproduct Spatial Price Equilibrium (SPE) models with network models. Harker (1987) includes analytical, computational and applied results of such an integration. He also discusses the possibility of integrating general economic equilibrium models with transportation network equilibrium models. Harker describes the purpose of his work as '...formally treating the generation, distribution, modal split and assignment of freight movements in a consistent, general equilibrium framework' (Harker, 1987, p. 2). We will first devote now more attention to elements of networks.

2.2. Network Elements

Networks consist of arcs (links) and nodes. A path is a specific sequential combination of

arcs from an origin of supply to a destination of demand. Some nodes represent regions, others transportation infrastructure elements such as ports, terminals, etc. Other important notions often used in network models are 'equilibrium flows', 'origin', 'destination', 'path', 'carrier', 'shipper' (see Friesz and Harker, 1985; or Harker, 1987). Here we will in particular focus on actors in networks.

A shipper[1] is a decision-making entity desiring a particular commodity at a particular destination[2]. A carrier is a decision-making entity that actually executes the transport of freight. Therefore, in simple terms shippers are the demanders of transportation and carriers the suppliers of it. In transportation network analysis an important research area seems to be the simultaneous treatment of shippers and carriers decisions (see Harker, 1987). Up till now most models using submodels for both actors have formulated them sequentially, implying that also computing equilibrium solutions takes place sequentially. In that case they are actually not necessarily equilibrium solutions. Only if both are included in a general equilibrium model that is solvable, then a real equilibrium approach is accomplished. Furthermore, it should be noted that not only a Computable General Equilibrium (CGE) approach integrates the shippers' and carriers' behaviour, but that it also analyzes this behaviour in the same way as that of other actors in the economy. Problems usually occur when non-linear cost curves and imperfect markets for transportation are introduced.

With regard to the network characteristics of combined shipper-carrier models it is usual to employ an aggregate shipper perceived network which is interlinked at certain nodes (transhipment, interline and other key nodes) with more detailed carrier-specific networks.

In relation to shippers, carriers and networks also the two Wardrop's principles (Wardrop, 1952) should be mentioned here. These state conditions for user optimum and system optimum network flows, respectively. Shippers may be modelled by a first Wardropian principle that can be stated as the minimization of delivered commodity prices (see Harker, 1987). A difference between spatial equilibrium models is that they adopt the

---

[1] Notice that the notion of a shipper only applies to freight transport, and not to assenger traffic (by private car), since freight flow patterns are determined by both transportation system users (shippers) and suppliers/owners (carriers) (see Friesz and Harker, 1985, p. 162).

[2] In network models the shipper operates in an environment where demand and supplies at various nodes are fixed. This may be extended to endogenous demand and supply, such as in spatial price equilibrium and general equilibrium models. In both latter cases trip generation becomes endogenous.

second Wardropian system optimum principle (using programming model formulations), while general equilibrium models (using decentralized optimization formulations) start from the user optimum. However, under certain conditions the decentralized and system optimization formulations show a correspondence. In a such a case planning-programming and market-equilibrium formulations may sometimes be substituted for one another[3]. However, one of the conditions is that there are no externalities generated within the system, by one actor onto another. This means that this correspondence looses significance for situations that we are particularly interested here. In that case the decentralized approach shows - or approximates - the actual behaviour, while the planning approach reflects the optimal system's outcome. Matching the two can be done by adding optimal instruments in the decentralized approach. However, for theoretical and implementational reasons one may turn to suboptimal instruments (see Verhoef and Van den Bergh, 1994).

3. Networks and Related Model Types

3.1. Multiregional Models and Network Structure

A multiregional structure can be regarded as a first approximation to a network. The data of such a multiregional structure - for instance, those indicating distances between two points or regions - may be based on a approximate minimum transportation or minimum distance route, as is usual in, for instance, SPE models[4]. Of course, variable transportation costs do not allow for a rigorous treatment of such a minimization process, until all values of variables are known, but still it is not impossible to have some second-best approximation, which is a requirement for multiregional analyses. It is of course also relevant to know which concept of a region is employed. First, a region may be a sort of open and miniature macro-system, containing a variety of production sectors and households. This is the usual meaning. This concept is especially relevant when one is interested in performing transport analysis based on interregional trade between diverse regions. The diversity may result from differences in economic structure, but also in environmental characteristics.

---

[3] A user-optimized equilibrium is reached when no user may lower his transportation cost through unilateral action. A system-optimized equilibrium is reached when the marginal total costs of transportation alternatives are equal (see Friesz and Harker, 1985, p. 162).

[4] Among the various extensions of the 'standard Samuelson/Takayama-Judge SPE models' presented in the very interesting article by Florian and Los (1982) one takes the perspective of a network for dealing with non-constant unit transport costs in a SPE framework.

Another regional concept includes only one actor, be it a production sector or a household. Approaching it differently, one can say that given the set of activities one disaggregates the total space into regions, such that each region contains precisely one activity. This allows one to take into account the materials flows and related transportation service use between sectors mutually as well as between sectors and households. Such a multiregional approach can simultaneously be regarded as a transportation network. This is necessarily a special type of network, including only direct links between each region, based on the above mentioned a priori least cost or least distance routing data. By direct links is meant that it is not possible that transport flows go from node A (sector in A) to household C (in node C) via node B (whatever is on that spot), unless of course the materials are processed into another form at B, in which case one would not speak of a transport flow from A to C but of two separate, independently decided transport flows from A to B and from B to C.

The main disadvantage of this network structure is that congestion in the associated interregional linked transportation network is a limited concept. The same limitation takes place with respect to density factors in the use of transportation. Finally, optimal routes are independent on activity and demand levels in each region, since in a multiregional framework it is assumed that each interregional link represents somehow the optimal connection, i.e. the least transportation cost connection. Of course, the high spatial aggregation level cannot satisfactorily deal with these issues.

To prevent confusion between network and multiregional models, it is emphasized that although a general network-externality model will be introduced in a later section, it is certainly true that many useful lessons may be drawn from multi-regional models. The approach in conventional multiregional models is one of trade between (stucturally) similar systems. As explained above this is traditionally done by way of input-output modelling and econometric modelling (see for example Nijkamp, 1986). We will now proceed with SPE models.

3.2. Spatial Price Equilibrium (SPE) Models

A second type of multiregional approach is the SPE model.[5] They usually adopt a partial

[5] Roson and Westin (1990) distinguish between input-output, transport assignment and gravity models. Transport assignment models probably include a wider class than SPE models, most of which however are less general and therefore of little interest for our discussion. Of more interest may be the combination of gravity and SPE models, for which they however do not give any reference.

view, i.e. consider a single good. However, in principle generalizations to many goods are possible. Besides, the approach shows a clear view on spatially separated markets and transport from an equilibrium point of view. Many steps have already been taken to fill the gaps between various alternative approaches, such as variable transportation costs, network models, dynamic models, and multi-commodity models. The two most important contributions in that respect are the very extensive treatment by Takayama and Judge (1971) and the conceptually new approach by Florian and Los (1982)[6] [7].

The basic SPE formulation aims to solve an interregional trade or spatial price equilibrium, where demand prices equal the sum of supply costs and transportation costs. Samuelson (1952), starting from this equilibrium concept, was the first to derive a formal representation of the SPE problem, as a linear programming model. The problem can be cast as finding, for given transportation costs, domestic demand and supply curves at regional or spatial markets, leading to a spatial competitive equilibrium. The latter consists of a set of prices in all spatial markets, the demands and supplies at each location, and all exports and imports[8]. Samuelson shows that some kind of consumers surplus concept is relevant in trying to convert the problem into a maximum or minimum formulation. He defines a 'net pay-off' (NSP) function as the sum of the social pay-off's in each region (spatial market) minus the transport cost.

A general multiregional one-product formulation of an SPE model is the following:

$$\textit{minimize} \quad \sum_{o} \int_{0}^{x_o} S_o(z)dz - \sum_{d} \int_{0}^{y_d} D_d(z)dz + TC. \tag{1}$$

where S_o are the inverse supply functions and D_d the inverse demand functions, and TC the total transportation cost. These shipments have to satisfy non-negativity restrictions. Supplies x_o at origins o are equal to the sum of all shipments out of o:
and demands y_d at destinations d are equal to the sum of all shipments arriving in region d:
The total transport cost (TC) is a function of all link-specific shipments of goods from

---

[6] The latter seems to be not extensively applied though.

[7] It is also interesting to note that at the other end of the spectrum of spatial equilbrium models there are the more restricted, partial transportation planning problems. Classical programming formulations of such problems are given by Hitchcock, 1941; Koopmans, 1949; Koopmans and Reiter, 1951; Beckman et al., 1956; Many articles refer also to original results by Hartwick in two unpublished papers (1970 and 1971).

[8] The difference with the simpler 'classical' transportation problem is that there demands and supplies at each place are given (fixed).

$$x_o = \sum_d t_o^d, \tag{2}$$

$$y_d = \sum_o t_o^d. \tag{3}$$

sources o to destinations d (t_o^d). Florian and Los (1982) note that the constraints in (2) and (3) may be interpreted as conservation of flows on a bipartite network, for which reason it may be regarded as a generalization of the classical transportation problem.

A different look at the basic SPE optimization formulation is proposed by Florian and Los (1982) by substituting (2) and (3) in (1), so that a non-linear programming problem in non-negative link-specific shipments variables t_o^d is obtained:

$$\textit{minimize} \quad \sum_o \int_0^{\sum_d t_o^d} S_o(z)dz - \sum_d \int_0^{\sum_o t_o^d} D_d(z)dz + TC. \tag{4}$$

According to Florian and Los this is 'a relatively easy non-linear programming problem', for which various solution methods are available.

Now it is of course interesting to investigate the composition of total transportation costs TC. Variation of these is anyway the most that can be done while still remaining in the SPE framework and the class of solvable mathematical constrained optimization problems. Our special interest is to allow for variable cost structures based on assumptions with respect to market structure (number of modes or firms), density effects on links, network characteristics and transportation of multiple commodities[9].

First, assume that the cost function belonging to an interregional link is to be associated with a single unique transportation firm, implying as many firms as links[10]. With one additional assumption, namely that c_o^d is the fixed cost per unit of the good to traverse link o-d, one obtains the most common cost structure, as represented in (5).

An important extension for our survey is the one in which the cost of shipping one unit of product over a certain link is a non-linear function $c_o^d(t_o^d)$ of the total amount of

[9] Notice that we are here, necessarily in the SPE framework, only considering non-constant transportation costs with respect to congestion and some inverse given supply function of transportation services. It is important to realize that this is not so extensive as transportation production in the CGE framework. However, both lead to endogenous transportation prices, stressing different aspects.

[10] They add that the price of transportation services should at least vary in proportion with the costs of supplying them. Though this property is not imposed by the structure of the model, it is interesting to note the similarity to the CGE framework, where this typically is enforced.

$$TC = \sum_o \sum_d c_o^d * t_o^d. \tag{5}$$

transport on that link[11], in which case (5) is replaced by (6):

$$TC = \sum_o \sum_d \int_0^{t_o^d} c_o^d(z)dz. \tag{6}$$

Two other market assumptions are of interest. One represents the other extreme case of a single firm serving all links. In such a case the transportation price per unit of product and distance may be related to the total volume of transportation services. Note that this requires that data is available on the amount of transportation services required to ship a unit of product on a certain link, for each link. The total transportation cost now depends on an inverse transportation supply curve ST, as shown in (7):

$$TC = \int_0^{\sum_o \sum_d t_o^d} ST(z)dz. \tag{7}$$

Note that the transportation cost on link o-d is formulated as $\tau_o^d * ST\left[\sum_o \sum_d \tau_o^d * t_o^d\right]$.

The second market situation is between the two extremes, namely with a small number of firms (m) each controlling a disjoint subset of the routes (V_i), and where each one has an inverse demand function related to the routes it serves[12]:
Further extensions of the latter formulation can be done in a network approach.

---

[11] As Florian and Los suggest, congestion effects imply a convex shape while density effects lead to a concave shape of transport costs as a function of transported volume.

[12] Note that the treatment of this market situation does not require any specific assumptions on oligopolistic price behaviour, since the firms do not compete with each other on the same interregional links. This implies that a next extension would include oligopolistic competion. See Harker (1987) for further discussion on this issue in the context of linked network-SPE models. These issues - interregional link competition and oligopolistic assumptions - are probably more consistently dealt with in a general equilibrium context (although the second one causes many difficulties there as well). Finally, this issue should then be dealt with in a multi-product situation typical for CGE, as discussed hereafter in the context of SPE.

$$TC = \sum_{i=1}^{m} \int_{0}^{\sum_{(a,d)\in V_i} \tau_0^d = t_a^d} ST_i(z)dz. \tag{9}$$

4. General Externality-Network Models[13]

First, the situation with fixed transportation costs will now be considered, where a distinction can be made between transportation and activity analysis models. The first class consists of a (discrete) spatial allocation or distribution model for a single commodity shown in (10), that is based on sectoral revenue, total production costs and transport costs (a variation on a net revenue maximization model in Takayama and Judge, 1971, Chapter 3). Here the problem is to find x_{ij} which are commodity shipments from point or region i to j. The economic environment is fixed, i.e. the quantities available or produced at i (supply), x_i, and the quantities demanded at j, y_j, are all given. Furthermore, all prices are fixed: the sales price of a product in region j coming from region i, p_{ji}; the cost of producing goods in i for j, c_{ij}; the cost of transporting one unit of the commodity from region i to j, t_{ij}, where each region can be both a demander and supplier. Various formulations of this model can be given to include environmental objectives. Presently, the model in (10) includes an environmental tax per unit of commodity transport, e_{ij}, reflecting the environmental cost of transporting one unit of commodity from i to j. This can be interpreted as a fixed markup cost, since both transport and environmental costs include already the impact of the distance and other relevant characteristics of the connection between i and j.

$$\begin{aligned} \underset{x_{ij}\geq 0}{maximize} \quad & \sum_i \sum_j (p_{ji} - c_{ij} - t_{ij} - e_{ij}) x_{ij} \\ s.t. \quad & \sum_i x_{ij} \geq y_j \ , \ \forall j, \\ & \sum_j x_{ij} \leq x_i \ , \ \forall i. \end{aligned} \tag{10}$$

It is interesting to look at the dual problem associated with (10) the solution of which gives the optimal price system at demand and supply regions, so that comparative location and price (dis)advantages are made explicit. One condition of the dual is that p_{ji}-q_j+q^i≤t_{ij}+c_{ij}+e_{ij}, where q_j can be interpreted as the zero-profit price at the production unit,

[13] This section is based on Van den Bergh and Rietveld (1992).

and q^i as the subsidy to satisfy the demand quota. This relationship is interesting because it states the exclusion of reallocation of the product by arbitrage. It shows that environmental taxation will lead to higher optimal prices, or positive subsidies, or it will in the presence of demand quota stimulate lower production costs. Of course, endogenous demand prices would give rise to lower demand levels (see next section).

An alternative model is (11) that represents physical regulation rather than environmental taxing.

$$
\begin{aligned}
&\underset{x_{ij}\geq 0}{maximize} \quad \sum_i \sum_j (p_{ji} - c_{ij} - t_{ij}) x_{ij} \\
&s.t. \ \sum_i x_{ij} \geq y_j \ , \ \forall j, \\
&\quad \sum_j x_{ij} \leq x_i \ , \ \forall i, \\
&\quad (a) \ \ \sum_i \sum_j e_{ij} x_{ij} \leq E, \\
&\quad (b) \ \ \sum_i e_{ij} x_{ij} \leq E_j \ , \forall j, \ \ \sum_j e_{ij} x_{ij} \leq E_i \ , \forall i, \\
&\quad (c) \ \ e_{ij} x_{ij} \leq E_{ij}, \ \forall i,j.
\end{aligned}
\tag{11}
$$

Here one or more constraints are added to a general optimal transportation problem - subject to the conditions in (10) - to reflect an upper bound to environmental impacts. These additional constraints may be formulated for the total system level as shown in (11a), but also for less aggregated environmental indicators linked to origins or destinations of spatial flows, as in (11b), or to each single spatial connection, as in (11c). The right hand sides E, E_i, E_j and E_{ij} in (11a-c) denote environmental regulations on various levels of aggregation. The variables e_{ij} in (11) are not necessarily similar to the ones in (10). This is only true when E, E_i, E_j, E_{ij} are measured in monetary units.

Finally, the model in (12) shows an even more drastic difference compared to (10) and (11). The objective is to minimize environmental effects (costs). The constraints added to this objective are the ones in (10) including a constraint stating that net benefits should

$$
\begin{aligned}
&\underset{x_{ij}\geq 0}{\textit{minimize}} \quad \sum_i \sum_j e_{ij} x_{ij} \\
&s.t. \quad \sum_i x_{ij} \geq y_j \ , \ \forall j, \\
&\qquad \sum_j x_{ij} \leq x_i \ , \ \forall i, \\
&\qquad \sum_i \sum_j (p_{ji} - c_{ij} - t_{ij}) x_{ij} \geq r \ .
\end{aligned} \tag{12}
$$

exceed some mimimum return r (i.e., moving the function in the present objective in (10) to the constraints part of the optimization model). This exchange of objectives and constraints can be interpreted as an alteration in priorities from economic to environmental goals. Of course, the objective function can also be based on the less aggregate conditions (b) and (c) in (11), in which case a multi-objective framework results. Of course, for consistent values of E and r the formulations (11) and (12) generate identical solutions. However, consistency can only be established until one problem is solved. Generally, differences may result because either economic contraints (r) or environmental constraints (E) are set too prudently or loosely.

A second type of model, shown in (13) and (14), includes both production and allocation. In this case, technology and resources are to be included as well. Demands are also here assumed to be given. An important addition to the foregoing problem is the existence of spatial flows of primary products (or intermediate products, since a unisectoral approach is dealt with here) serving as inputs of production units (one type of input is considered here). A case of given locations in this context means thus that production capital or other inputs are non-mobile. The model gives two types of information: (i) spatial flows of inputs and final goods; and (ii) levels of production of final goods and use of inputs at each location. The dual again gives the optimal prices.

$$
\begin{aligned}
&\underset{x_{ij}, x_i, z_{ij} \geq 0}{\textit{maximize}} \quad \sum_i \sum_j (p_{ji}^{x} - t_{ij}^{x}) x_{ij} \\
&\qquad - \sum_i \sum_j (p_{ji}^{z} + t_{ij}^{z}) z_{ij} \ - \ \sum_i c_i x_i \\
&\qquad - \sum_i \sum_j (x_{ij} e_{ij}^{x} + z_{ij} e_{ij}^{z}) \\
&s.t. \quad (10), \ \forall i.
\end{aligned} \tag{13}
$$

The first line of the objective function in (13) represents revenues from selling final

goods net of transport costs from i to j. The second line represents cost of production, composed of purchase and transport costs of inputs and operations costs at the production unit. The final line in (13) is the environmental cost, which include the revenue of taxes levied on transport of inputs to production units and final goods to consumers.

$$\begin{aligned} & x_i \geq \sum_j x_{ij}, \\ & s_i + \sum_j z_{ji} \geq a_i^z x_i + \sum_j z_{ij}, \\ & k_i \geq a_i^k x_i, \\ & \sum_j x_{ji} \geq y_i. \end{aligned} \tag{14}$$

The four requirements in (14) are to be interpreted as follows: the first condition states that for each region the amount of final goods produced is sufficient to allow for the transport flows of final goods to all regions; the second condition assures that the regional availability of inputs s_i and the transport inflow of inputs from other regions together are sufficient to satisfy regional production requirements and regional transport of inputs to all regions (for the production of one unit of output in i one needs a_i^z units of input); the third condition states that the immobile production capital and other resources are sufficient for regional production; the last constraint reflects the satisfaction of given regional demands for the final good. The foregoing statements on the environmental policy regulations in relation to model constraints apply here as well. No matter how environmental policy is included, the present spatial activity analysis allows for comparison of environmental policies directed at different transport flows: input or intermediate products versus final goods; or passenger versus freight, i.e. labour (as a special type of resource input) versus final goods. Finally, only the case of constant transport costs is considered here. It may however be the case that transportation costs are flow dependent. But in essence, this can be considered a special type of network problem. In contrast to the foregoing models, network equilibrium planning problems are associated with different routes between two given spatial centers. In such cases it is from an environmental policy point of view interesting to stimulate route choices that give rise to a minimum of environmental impact. At this point, however, it is more interesting to consider the relationship between transport and economic considerations, so that these network models will be eschewed (the interested reader is referred to Batten and Boyce, 1986).

In the long run spatial economic problems of optimum locations are relevant, which in a sense can be regarded as an expansion to the transport problems of cost minimization or revenue maximization. While in the transport problems all locations are fixed, including markets, resource deposits and firms, the location problem assumes that only markets and resource deposits are fixed. This makes both problems very similar. Finally, relaxation of all fixed factors gives rise to a complex integrated choice problem of locations, production levels and interactions, for given markets and resource deposits (see Batten and Boyce, 1986). This approach is however still very much producer-oriented since consumer demand is included as fixed. Only a general equilibrium approach can relax both fixed producer and consumer choice variables (see Van den Bergh, 1993).

5. The Economic and Policy Context of Network Externalities

Networks can be considered as vehicles through wich regions or nations can control part of the interregional or international competition. Especially in the context of international and interregional coordination and harmonisation, attention should be devoted to the socio-economic benefits which are estimated on the basis of a network which deals with a multimodality of transport systems, monopolistic and oligopolistic spatial market structures, and negative externalities. In the context of Europe, for instance, till recently the interest was mainly focused on single mode solutions, in spite of the fact that since long it is clear that interconnected networks offer a higher value added. In that context it is important to devote attention to the following strategic factors (see Nijkamp, 1993):

- complementarity between different nodes in order to benefit from synergy between networks;
- competition between nodes in order to operate under the most cost-efficiënt conditions (e.g., common carriage);

In order to estimate the effects of specific network compositions, one has to pay attention to both use and management characteristics. For instance, the issue of whether public control, monopolistic control, an oligopolistic market structure, or a atomistic market structure applies to a transport network is relevant. Also it is important to consider the use of demand characteristics related to number of actors, monopsonistic elements, and even (culturally determined) habits and freedom of choice. These organization aspects of transport markets, both at the supply and demand sides, have received far less attention than physical aspects of transport planning over networks, related to complexity of the network and regulation.

Approaches in this respect may borrow elements from various theories, related to transactions of firms, simultaneous development of production and warehousing, synchronous evolution of demand locations and shipment networks, hierarchical networks of local areas, regions, nations and global economies, integral logistics, and the existence of barriers. In the context of externalities and networks this means that many issues need to be investigated before final conclusions regarding network-oriented policies can be drawn.

References

Batten, D.F. and L. Westin, 1990. Modelling commodity flows on trade networks: Retrospect and prospect. In: M. Chatterji and R. Kuenne, *New frontiers in regional science*, Macmillan, London, pp. 135-156.

Batten, D.F., and D.E. Boyce, 1986. Spatial interaction, transportation, and interregional commodity flow models. In: P. Nijkamp., *Handbook of Regional and Urban economics*, Vol. 1. North-Holland, Amsterdam.

Bergh, J.C.J.M. van den, 1993. General equilibrium analysis of freight transport Policy: A multiregion-network approach. Tinbergen Institute Discussion Paper, Regional and Environmental Economics, TI 93-190.

Bergh, J.C.J.M. van den, and P. Rietveld, 1992. Economic effects of environmental policy measures applied to transport and infrastructure. Presented at the Conference of the European Association of Environmental Economists, Krakow, juni 1992.

Boyce, D.E., 1980. A framework for constructing network equilibrium models of urban location. *Transportation Science* 14: 77-96.

Florian, M. and M. Los, 1982. A new look at static spatial price equilibrium models. *Regional Science and Urban Economics* 12: 579-597.

Friesz, T.L., and P.T. Harker, 1985. Freight network equilibrium: A review of the state of the art. In: A.F. Daughety (ed.). *Analytical Studies in Transportation Economics*. Cambridge University Press.

Harker, P.T., 1987. *Predicting intercity freight flows*. VNU Science Press, Utrecht, The Netherlands.

Nijkamp., P., 1986. *Handbook of regional and urban economics*, Vol. 1. North-Holland, Amsterdam.

Nijkamp, P., 1993. The European Network Economy: Opportunities and Impediments. Research Memorandum 1993-63. Faculty of Economics, Free University, Amsterdam.

Samuelson, P.A., 1952. Spatial price equilibrium and linear programming. *American Economic Review* 42: 283-303.

Takayama, T. and G. Judge, 1971. *Spatial and Temporal Price and Allocation Models.* North-Holland, Amsterdam.

Verhoef, E.T., and J.C.J.M. Van den Bergh, 1994. Transport and environmental sustainability: A adapted SPE approach for modelling interactions between transport, infrastructure, economy and environment.

Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers* 1, 325-378.

A NEW EXACT ALGORITHM FOR THE VEHICLE SCHEDULING PROBLEM WITH TIME WINDOW AND OTHER CONSTRAINTS

L. Bianco,[1] V.Maniezzo,[2] A. Mingozzi,[2] S. Ricciardelli[1]

(1) *Department of Electrical Engineering, University "Tor Vergata", Rome. Italy*

(2) *Department of Mathematics, University of Bologna, Bologna, Italy*

ABSTRACT

We consider a vehicle scheduling problem arising in some freight transport systems where a vehicle fleet of different classes, split among different depots, must be used to perform a set of trips over a time period horizon minimizing a given cost function. The set of trips must be partitioned in a number of disjoint sequences (called duties) and each duty must be assigned to a vehicle satisfying
the time windows of each trip, vehicle-trip objection constraints and the working time window of the vehicle. Moreover, a vehicle can perform only one duty and must return to the starting depot. This problem is an extension of simpler problems like the *Multiple Depot Vehicle Scheduling Problem* and the *Crew Scheduling Problem* that are known to be NP *hard*. In this work the problem is formulated as Set Partitioning Problem with side constraints (**SPP**) where each column is a feasible vehicle duty. We describe a procedure for computing a lower bound to the optimal cost by finding an heuristic solution of the dual of the linear relaxation of the **SPP** formulation, without generating the entire **SPP** matrix. The dual solution obtained and a valid upper bound to the optimal cost are used to reduce the set of feasible duties that contain the optimal solution.An exact branch and bound algorithm for finding the optimal **SPP** solution without generating the set of feasible duties is described. The computational results show that the proposed method can be used to solve large size problems.

A MULTIPLIER ADJUSTEMENT METHOD FOR THE MULTI-DEPOT VEHICLE SCHEDULING PROBLEM

Marta Mesquita
Instituto Superior de Agronomia,Tapada da Ajuda,1300 Lisboa, Portugal

J. Paixão
DEIO, Faculdade de Ciências de Lisboa, Bloco C2-Campo Grande,1700 Lisboa, Portugal

Abstract

This paper is concerned with the multi-depot vehicle scheduling problem (MDVSP), which consists of grouping time table trips into vehicle blocks and, at the same time, assign those vehicle blocks to depots in order to minimize the number of vehicles and the operated costs associated with the schedule.
In this paper, we extend a previous work where we describe a lagrangean relaxation, based on the multi-depot problem without requiring that each vehicle returns to the source depot.
Now, a lagrangean dual ascendent procedure that modifies the subset of multipliers associated with each vehicle block, at a time, is suggest by some properties of a restricted lagrangean relaxation.
Simultaneously, two alternative approaches for obtaining upper bounds are described.
Computational experience with real life data from Carris-Lisboa and random test problems, generated as described in literature, shows the success in handling problems with up to 400 trips and 6 depots.

Keywords: vehicle scheduling problem, lagrangean relaxation, adjustement method.

1.Introduction

The vehicle scheduling problem consists of optimally linking timetabled trips into vehicle blocks, so that, trips in the same vehicle block form a feasible schedule for a vehicle. When vehicles are located at different depots, one has, at the same time, to assign those blocks to the depots leading to the multi-depot vehicle scheduling problem (MDVSP)

Bertossi et al.(1987), shown that the problem is NP-Hard and proposed an heuristic algorithm, based on a lagrangean relaxation. In that paper, experimental results are presented for problems with up to 50 trips and 3 depots. Later, Mesquita & Paixão (1990) presented another heuristic procedure based upon a lagrangean relaxation. There, computational experience is reported for randomly generated test problems with sizes up to 250 trips - 2 depots, 200 trips - 3 depots. Also, Dell'Amico el al. (1990) suggested an heuristic algorithm based on the shortest path approach, which always guarantees the use of the minimum number of vehicles. Randomly generated test problems with up to 1000 trips and 10 depots were considered in the mentioned work. Other heuristic algorithms can be found in Bodin et al. (1983), Ceder & Stern (1981), El-Azm (1985), Smith &Wren (1981).

The first exact method for the MDVSP was presented by Carpaneto et al. (1989). There, different lower bound techniques are combined to obtain an additive lower bounding procedure which is embedded in a branch and bound algorithm. Computational experience is given for problems with up to 70 trips and 3 depots.

Alternatively, Ribeiro & Soumis (1991) formulated the MDVSP as a set partitioning problem with side constraints and proposed an exact algorithm based on the continuous relaxation solved by a column generation scheme. The authors report the solving of randomly generated test problems with up to 300 trips and 6 depots.

Recently, Forbes et al. (1994) presented an exact algorithm based on a linear programming relaxation of a multi-commodity network flow type formulation. They consider two basic sets of problems. The first set, generated

accordingly to the procedure described in Carpaneto et al. (1989), includes problems with up to 200 trips and 3 depots, while the second set were extracted from real life data where 72% of the trips may only be operated by one depot, 25% by two depots and 3% by all depots. For such type problems they consider instances with up to 600 trips and 3 depots.

In this paper, we extend our previous work presented in Mesquita & Paixão (1990), where we describe a lagrangean relaxation for the MDVSP. The corresponding relaxed problem can be decomposed into a semi-assignment problem and a multi-depot problem without requiring that each vehicle returns to the source depot. Both problems can be solved by polinomial algorithms. Then, a normal subgradient optimization technique was used for dealing with the resulting lagrangean dual problem.

Now, we present a multiplier adjustement method which exploits the special stucture of the mathematical formulation that we use for the MDVSP. Such procedure is shown to produce significant improvements on the lower bounds provided by using the subgradient optimization method. Likewise in our previous work, heuristic procedures based upon the lagrangean relaxation are developed in order to obtain feasible solutions and, consequently, computing upper bounds on the optimal value for the MDVSP.

The paper is organized as follows. Next section is devoted to describe the mathematical model that we use for the MDVSP. Then, in the following section, a restricted lagrangean relaxation is presented and several properties are stated leading to a multiplier adjustment technique for dealing with the corresponding lagrangean dual problem. In section 4, we show how that technique can be refined in order to cope with the large number of alternative solutions that occurr in one of the lagrangean relaxation subproblems. The heuristic procedures used for producing feasible solutions to the MDVSP are described in section 5. Section 6 reports computational experience with real life data from Carris-Lisbon and randomly generated test problems showing that the method can be sucessful in handling instances with up to 400 trips and 6 depots. Finally, in the last section, some conclusions are drawn.

2. The Mathematical Model

Let us consider a set of timetabled trips $T_1,T_2,...,T_n$ to be operated by vehicles located at k depots denoted by $D_1,D_2,...,D_k$. The vehicles, in number of d_ℓ for $\ell = 1,2,...k$ are supposed to be identical. Each trip T_i is characterized by the corresponding starting and ending times and places. An ordered pair of trips (T_i,T_j) is said to be compatible if the bus released after trip T_i completion can be assigned to trip T_j. One aims to group compatible pairs of trips into vehicle blocks and at the same time assign those vehicle blocks to depots, in order to minimize the costs associated with the schedule.

Let us now formulate the multi-depot problem through a graph theory model. Consider a graph G=(V,A) where the vertex set $V=\{1,...,n+k\}$ is partitioned into $I=\{1,...,n\}$, corresponding to trips $T_1,T_2,...,T_n$, and $D=\{n+1,...,n+k\}$, refering to depots $D_1,D_2,...,D_k$. The arc set A contains arcs (i,j) corresponding to compatible pairs (T_i,T_j) and all arcs $(i,n+\ell)$, $(n+\ell,i)$, for $\ell=1,...,k$ and $i=1,...,n$. For each arc $(i,j)\in A$ a cost c_{ij} is defined.

The objective is to find the minimum cost set of elementary circuits such that:

- each vertex $i\in I$ is covered by exactly one circuit;
- each circuit contains exactly one vertex from set D;
- the number of circuits covering vertex $n+\ell$ never exceeds d_ℓ ($\ell=1,2,...,k$).

Given the decision variables :

$$x_{i,j}\quad (i,j=1,\ldots,n) = \begin{cases} 1, & \text{if trip i is directly connected with trip j} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{i,n+\ell} \quad (i=1,\ldots,n;\ \ell=1,\ldots,k) = \begin{cases} 1, & \text{if after trip i the bus immediately returns to depot } D_\ell \\ 0, & \text{otherwise} \end{cases}$$

$$x_{n+\ell,j} \quad (\ell=1,\ldots,k;\ j=1,\ldots,n) = \begin{cases} 1, & \text{if depot } D_\ell \text{ directly supplies a bus for trip j} \\ 0, & \text{otherwise} \end{cases}$$

$$y_{i,\ell} \quad (i=1,\ldots,n;\ \ell=1,\ldots,k) = \begin{cases} 1, & \text{if trip i is performed by a vehicle housed at depot } D_\ell \\ 0, & \text{otherwise} \end{cases}$$

the graph theory model can be formulated as:

$$(P) \quad \min \quad \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\, x_{ij} + \sum_{\ell=1}^{k}\left[\sum_{i=1}^{n} c_{i,n+\ell}\, x_{i,n+\ell} + \sum_{j=1}^{n} c_{n+\ell,j}\, x_{n+\ell,j}\right]$$

$$\text{s.t.} \quad \sum_{j=1}^{n+k} x_{ij} = 1 \qquad i=1,2\ldots,n \qquad (1)$$

$$\sum_{i=1}^{n+k} x_{ij} = 1 \qquad j=1,2\ldots,n \qquad (2)$$

$$\sum_{j=1}^{n} x_{n+\ell,j} \le d_\ell \qquad \ell=1,2\ldots,k \qquad (3)$$

$$x_{n+\ell,j} - y_{j\ell} \le 0 \qquad j=1,2\ldots,n;\ \ell=1,2\ldots,k \qquad (4)$$

$$x_{i,n+\ell} - y_{i\ell} \le 0 \qquad i=1,2\ldots,n;\ \ell=1,2\ldots,k \qquad (5)$$

$$y_{i\ell} + x_{ij} - y_{j\ell} \le 1 \qquad i,j=1,2\ldots,n;\ \ell=1,2\ldots,k \qquad (6)$$

$$\sum_{\ell=1}^{k} y_{i\ell} = 1 \qquad i=1,2\ldots,n \qquad (7)$$

$$x_{ij} \in \{0,1\} \qquad i,j=1,2\ldots,n+k,\ (i,j)\neq(n+\ell,n+\ell');\ \ell,\ell'=1,2,\ldots,k \qquad (8)$$

$$y_{i\ell} \in \{0,1\} \qquad i=1,2\ldots,n;\ \ell=1,2\ldots,k \qquad (9)$$

A full explanation of the above model has been given in Mesquita & Paixão (1990). Let us recall that the constraints (1) and (2) assure that every trip is chained, from and to, another trip or a depot. The vehicle capacity at the depots is considered in the constraints (3). The constraint set S={(4),(5),(6)} is related with the assignment of trips to the different depots. More explicitly, the constraints (4) and (5) impose that if a trip is directly connected to a depot then it has to be assigned to that particular depot. The constraints (6) guarantee that if a trip i is connected to a trip j then both of them must be assigned to the same depot. Finally, the constraints (7) impose that each trip is assigned to one, and only one, depot.

3. Restricted Lagrangean Relaxation

Likewise we presented in our previous work [Mesquita & Paixão (1990)], a lagrangean relaxation can be considered for the mathematical program described above. There, we add the following redundant constraints which will be useful in the relaxed problem :

$$\sum_{i=1}^{n} x_{i,n+\ell} = \sum_{i=1}^{n} x_{n+\ell,i} \qquad \ell=1,2...,k$$

For the sake of simplicity, we introduce nonnegative auxiliary variables z_ℓ $\ell=1,2...,k$ in the constraints related to the depots capacity. Hence, in the mathematical program (P), we replace the constraints (3) by :

$$\sum_{j=1}^{n} x_{n+\ell,j} + z_\ell = d_\ell \qquad \ell=1,2...,k \qquad (3a)$$

$$\sum_{i=1}^{n} x_{i,n+\ell} + z_\ell = d_\ell \qquad \ell=1,2...,k \qquad (3b)$$

Now, considering nonnegative lagrangean multipliers $\mu_{j\ell}$, $\gamma_{i\ell}$, and $\lambda_{ij\ell}$ associated, respectively, with constraints (4), (5) and (6), and dualizing those constraints, one obtains the following lagrangean relaxation:

$$w(\mu,\gamma,\lambda) = -\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{ij\ell} + \min\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} \left(c_{ij} + \sum_{\ell=1}^{k} \lambda_{ij\ell}\right) x_{ij} + \sum_{\ell=1}^{k}\sum_{i=1}^{n} \left(c_{i,n+\ell} + \gamma_{i\ell}\right) x_{i,n+\ell} \right.$$
$$\left. + \sum_{\ell=1}^{k}\sum_{j=1}^{n} \left(c_{n+\ell,j} + \mu_{j\ell}\right) x_{n+\ell,j} + \sum_{i=1}^{n}\sum_{\ell=1}^{k} \left(-\gamma_{i\ell} - \mu_{i\ell} + \sum_{j=1}^{n} (\lambda_{ij\ell} - \lambda_{ji\ell})\right) y_{i\ell} \right\}$$

s.t.

$$\sum_{j=1}^{n+k} x_{ij} = 1 \qquad i=1,2,...,n \qquad (1)$$

$$\sum_{i=1}^{n+k} x_{ij} = 1 \qquad j=1,2,...,n \qquad (2)$$

$$\sum_{j=1}^{n} x_{n+\ell,j} + z_\ell = d_\ell \qquad \ell=1,2,...,k \qquad (3a)$$

$$\sum_{i=1}^{n} x_{i,n+\ell} + z_\ell = d_\ell \qquad \ell=1,2,...,k \qquad (3b)$$

$$\sum_{\ell=1}^{k} y_{i\ell} = 1 \qquad i=1,2,...,n \qquad (7)$$

$$x_{ij} \in \{0,1\} \qquad i,j=1,2,...,n+k, \quad (i,j)\neq(n+\ell,n+\ell'),\ \ell,\ell'=1,2,...,k \qquad (8)$$

$$y_{i\ell} \in \{0,1\} \qquad i = 1,2,...,n;\ \ell = 1,2,...,k \qquad (9)$$

$$z_\ell \geq 0 \qquad \ell = 1,2,...,k \qquad (10)$$

The resulting problem can be split into the following subproblems:

- $(PX_{\mu,\gamma,\lambda})$ a multi-depot problem without requiring that each vehicle returns to the source depot, regarding only variables x_{ij} and z_ℓ.
- $(PY_{\mu,\gamma,\lambda})$ a semi-assignment problem, including only variables $y_{i\ell}$.

The corresponding lagrangean dual problem is

$$\begin{aligned} (\text{DLP}) \qquad & \max \quad w(\mu,\gamma,\lambda) \\ & \text{st.} \quad \mu \geq 0;\ \gamma \geq 0;\ \lambda \geq 0 \end{aligned}$$

where $w(\mu,\gamma,\lambda)$ can be evaluate in polynomial time by solving $(PX_{\mu,\gamma,\lambda})$ and $(PY_{\mu,\gamma,\lambda})$.

$$w(\mu,\gamma,\lambda) = - \sum_{\ell=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij\ell} + \upsilon(PX_{\mu,\gamma,\lambda}) + \upsilon(PY_{\mu,\gamma,\lambda}) \qquad (11)$$

Let us denote by $\bar{X}$ the optimal solution to $(PX_{\mu,\gamma,\lambda})$ with $\mu=\gamma=\lambda=0$. Then, DLP can be restricted in such way that, regardless the modifications on the multipliers, the solution $\bar{X}$ remains optimal for $(PX_{\mu,\gamma,\lambda})$, while the optimal solution for $(PY_{\mu,\gamma,\lambda})$ may change.The restricted lagrangean dual problem is mathematically described by:

$$\begin{aligned} (\text{RDLP}) \qquad & \max \quad w(\mu,\gamma,\lambda) \\ & \mu, \gamma, \lambda \in W \end{aligned}$$

where W is related to $\bar{X}$ and defined as follows:

$$W = \left\{ (\mu,\gamma,\lambda):\ \mu \geq 0,\ \gamma \geq 0,\ \lambda \geq 0 \text{ and } (u,v,\mu,\gamma,\lambda) \in Z \right\} \text{ with}$$

$$Z = \left\{ (u,v,\mu,\gamma,\lambda) : u,v \in \mathbb{R}^{n+k} \text{ is a linear dual feasible solution for } (PX_{\mu,\gamma,\lambda}) \text{ that satisfies} \right.$$

$$(u_i + v_j - (\sum_{\ell=1}^{k} \lambda_{ij\ell} + c_{ij})\ )\ \bar{x}_{ij} = 0 \qquad i,j=1,2,\ldots,n$$

$$(u_i + v_{n+\ell} - (\gamma_{i\ell} + c_{i,n+\ell}))\ \bar{x}_{i,n+\ell} = 0 \qquad i=1,2,\ldots,n;\ \ell=1,2,\ldots,k$$

$$(v_j + u_{n+\ell} - (\mu_{j\ell} + c_{n+\ell,j}))\ \bar{x}_{n+\ell,j} = 0 \qquad j=1,2,\ldots,n;\ \ell=1,2,\ldots,k$$

$$(u_{n+\ell} + v_{n+\ell}\ )\ z_\ell = 0 \qquad \left. \ell=1,2,\ldots,k \right\}$$

That is, changes in μ, γ, λ, which yield changes in the dual variables u and v, are made in order to keep the solution $\bar{X}$ optimal to $(PX_{\mu,\gamma,\lambda})$. Somehow, this follows the approach proposed in Balas and Christofides (1981) for the asymmetric traveling salesman problem.

Some properties can be derived for the RDLP leading to a procedure for selecting the lagrangean multipliers in order to maximize $w(\mu,\gamma,\lambda)$.

proposition 1

The optimal value for RDLP is equal to the optimum for the following problem

$$(P^*) \qquad \max \left\{ -\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{ij\ell} + \sum_{i=1}^{n}(u_i + v_i) + \sum_{\ell=1}^{k} d_\ell\,(u_{n+\ell} + v_{n+\ell}) + \sum_{i=1}^{n} \rho_i \right\}$$

$$\text{s.t.} \quad \rho_i + \gamma_{i\ell} + \mu_{i\ell} - \sum_{j=1}^{n}(\lambda_{ij\ell} - \lambda_{ji\ell}) \le 0 \qquad i=1,\ldots,n;\ \ell=1,\ldots,k$$

$$(u,v,\mu,\gamma,\lambda) \in Z$$

$$\rho_i \in \mathbb{R} \qquad i = 1,\ldots,n$$

proof

Let $(\bar{\mu},\bar{\gamma},\bar{\lambda}) \in W$ such that $w(\bar{\mu},\bar{\gamma},\bar{\lambda}) = \max_{\mu,\gamma,\lambda \in W} w(\mu,\gamma,\lambda)$

Since the integrality property (Geoffrion, 1974) holds we have $w(\mu,\gamma,\lambda) = w_D(\mu,\gamma,\lambda)$, where

$$w_D(\mu,\gamma,\lambda) = -\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{ij\ell} + \max \left\{ \sum_{i=1}^{n}(u_i + v_i) + \sum_{\ell=1}^{k} d_\ell\,(u_{n+\ell} + v_{n+\ell}) + \sum_{i=1}^{n} \rho_i \right\}$$

$$\text{s.t.} \quad u_i + v_j - \sum_{\ell=1}^{k} \lambda_{ij\ell} \le c_{ij} \qquad i,j=1,2,\ldots,n \qquad (12)$$

$$u_i + v_{n+\ell} - \gamma_{i\ell} \le c_{i,n+\ell} \qquad i=1,2,\ldots,n;\ \ell=1,2,\ldots,k \qquad (13)$$

$$v_j + u_{n+\ell} - \mu_{j\ell} \le c_{n+\ell,j} \qquad j=1,2,\ldots,n;\ \ell=1,2,\ldots,k \qquad (14)$$

$$u_{n+\ell} + v_{n+\ell} \le 0 \qquad \ell=1,2,\ldots,k \qquad (15)$$

$$\rho_i + \gamma_{i\ell} + \mu_{i\ell} - \sum_{j=1}^{n}(\lambda_{ij\ell} - \lambda_{ji\ell}) \le 0 \qquad i=1,2,\ldots,n;\ \ell=1,2,\ldots,k \qquad (16)$$

$$u_i, v_i \in \mathbb{R} \qquad i=1,2,\ldots,n+k \qquad (17)$$

$$\rho_i \in \mathbb{R} \qquad i = 1,2,\ldots,n \qquad (18)$$

$w_D(\bar{\mu},\bar{\gamma},\bar{\lambda})$ attains the optimal value for a set of dual variables $(\bar{u}, \bar{v})$, associate to $(PX_{\bar{\mu}}, \bar{\gamma}, \bar{\lambda})$, and $\bar{\rho}$, associate to $(PY_{\bar{\mu}}, \bar{\gamma}, \bar{\lambda})$, such that:

(i) $\bar{X}$, $(\bar{u}, \bar{v})$ satisfy complementary slackness conditions, so, $(\bar{u}, \bar{v},\bar{\mu},\bar{\gamma},\bar{\lambda}) \in Z$;

(ii) $\bar{\rho}$ satisfy constraints (16);

(iii) $w(\bar{\mu},\bar{\gamma},\bar{\lambda}) = -\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} \bar{\lambda}_{ij\ell} + \sum_{i=1}^{n}(\bar{u}_i + \bar{v}_i) + \sum_{\ell=1}^{k} d_\ell\,(\bar{u}_{n+\ell} + \bar{v}_{n+\ell}) + \sum_{i=1}^{n} \bar{\rho}_i$

So, $w(\bar{\mu},\bar{\gamma},\bar{\lambda}) \le \upsilon(P^*)$.

Now, let us assume that $(\bar{u}, \bar{v}, \bar{\mu}, \bar{\gamma}, \bar{\lambda}), \bar{\rho}$ is the optimal solution for problem (P*). Then

(i) $(\bar{u}, \bar{v}, \bar{\mu}, \bar{\gamma}, \bar{\lambda}) \in Z \Rightarrow (\bar{u}, \bar{v})$ is a linear dual feasible solution to $(PX_{\bar{\mu}, \bar{\gamma}, \bar{\lambda}})$, so constraints (12), (13), (14) and (15) are satisfied;

(ii) $\bar{\rho} \in \mathbb{R}$, satisfy constraints (16);

$$\text{As, } w_D(\bar{\mu},\bar{\gamma},\bar{\lambda}) = -\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{\lambda}_{ij\ell} + \sum_{i=1}^{n}(\bar{u}_i + \bar{v}_i) + \sum_{\ell=1}^{k} d_\ell\,(\bar{u}_{n+\ell} + \bar{v}_{n+\ell}) + \sum_{i=1}^{n}\bar{\rho}_i$$

$$\text{We get } \upsilon(P^*) \le \max_{\mu,\gamma,\lambda \in W} w_D(\mu,\gamma,\lambda) = \max_{\mu,\gamma,\lambda \in W} w(\mu,\gamma,\lambda)$$

For any $(u,v,\mu,\gamma,\lambda) \in Z$ and , such that, $\rho_i + \gamma_{i\ell} + \mu_{i\ell} - \sum_{j=1}^{n}(\lambda_{ij\ell} - \lambda_{ji\ell}) \le 0$ $i=1,\ldots,n$; $\ell=1,\ldots,k$,

$$\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{ij\ell} + \sum_{i=1}^{n}(u_i + v_i) + \sum_{\ell=1}^{k} d_\ell\,(u_{n+\ell} + v_{n+\ell}) + \sum_{i=1}^{n}\rho_i$$

is a valid lower bound on v(P), where v(P) is the optimal value of P.

Before stating the next property, let us remind that, in general, $\bar{X}$ violates the constraints related to the departure and arriving depots; otherwise, it is optimal for P. That is, at least one of the vehicle blocks corresponding to $\bar{X}$ has the following aspect :

$$k_{d_i} \rightarrow i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow i_{m-1} \rightarrow i_m \rightarrow k_{a_i}$$

where k_{d_i} and k_{a_i} correspond to different depots. Suppose that all the multipliers associated with the arcs of the above vehicle block are equal to zero. In such condition we can say that:

proposition 2

Let $\mu_{i_1,k_{d_i}}, \lambda_{i_1,i_2,k_{d_i}}, \ldots, \lambda_{i_{m-1},i_m,k_{d_i}}, \gamma_{i_m,k_{a_i}} \in W$ be nonnegative lagrangean multipliers associated with the corresponding arcs of the above vehicle block, and verifying the following inequalities :

$$\mu_{i_1,k_{d_i}} \le \lambda_{i_1,i_2,k_{d_i}} \le \ldots \le \lambda_{i_{m-1},i_m,k_{d_i}} \le \gamma_{i_m,k_{a_i}}$$

Then, $w(\mu,\gamma,\lambda) + \mu_{i_1,k_{d_i}}$ is a stronger lower bound on the optimal value for P.

proof

Let us denote by $(\tilde{\mu},\tilde{\gamma},\tilde{\lambda})$ the multiplier set after introducing the new penalties $\mu_{i_1,k_{d_i}}, \lambda_{i_1,i_2,k_{d_i}}, \ldots, \lambda_{i_{m-1},i_m,k_{d_i}}, \gamma_{i_m,k_{a_i}}$ and by $w(\tilde{\mu},\tilde{\gamma},\tilde{\lambda})$ the correspondent lower bound. Let $(\tilde{u}, \tilde{v})$ stand for the optimal solution to the linear dual of $(PX_{\tilde{\mu},\tilde{\gamma},\tilde{\lambda}})$. Note that, (μ,γ,λ) stand for the multiplier set before introducing the penalties $\mu_{i_1,k_{d_i}}, \lambda_{i_1,i_2,k_{d_i}}, \ldots, \lambda_{i_{m-1},i_m,k_{d_i}}, \gamma_{i_m,k_{a_i}}$, while $w(\mu,\gamma,\lambda)$ is the correspondent lower bound.

Given, $(\tilde{u}, \tilde{v},\tilde{\mu},\tilde{\gamma},\tilde{\lambda})$, by prop 1 the problem we have to solve is

$$-\sum_{\ell=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\lambda}_{ij\ell}+\sum_{i=1}^{n}(\tilde{u}_i+\tilde{v}_i)+\sum_{\ell=1}^{k}d_\ell\,(\tilde{u}_{n+\ell}+\tilde{v}_{n+\ell})+\max\sum_{i=1}^{n}\rho_i$$

$$\text{s.t.}\qquad \rho_i+\tilde{\gamma}_{i\ell}+\tilde{\mu}_{i\ell}-\sum_{j=1}^{n}(\tilde{\lambda}_{ij\ell}-\tilde{\lambda}_{ji\ell})\le 0 \qquad i=1,\ldots,n$$

$$\rho_i\in\mathbb{R} \qquad i=1,\ldots,n$$

Note, that, lagrangean multipliers assigned to arcs that doesn't belong to the above vehicle block remain unchanged, so the only ρ_i we have to compute are those related to trips involved in the vehicle block presented above.

So, $\rho_{i_m} = -\gamma_{i_m,k_{a_i}}$ since $\lambda_{i_{m-1},i_m,k_{d_i}} \le \gamma_{i_m,k_{a_i}}$ and

$$\begin{cases}\rho_{i_m}\le-\gamma_{i_m,k_{a_i}}\\ \rho_{i_m}\le-\lambda_{i_{m-1},i_m,k_{d_i}}\\ \rho_{i_m}\le 0,\quad k_\ell\neq k_{d_i}\text{ and }k_\ell\neq k_{a_i}\end{cases}$$

$\rho_{i_{m-1}} = 0,\ldots, \rho_{i_2} = 0$ since $\lambda_{i_{m-2},i_{m-1},k_{d_i}} \le \lambda_{i_{m-1},i_m,k_{d_i}}$ and

$$\begin{cases}\rho_{i_{m-1}}\le\lambda_{i_{m-1},i_m,k_{d_i}}-\lambda_{i_{m-2},i_{m-1},k_{d_i}}\\ \rho_{i_{m-1}}\le 0,\quad k_\ell\neq k_{d_i}.\end{cases}$$

$\rho_{i_1} = 0$ since $\mu_{i_1,k_{d_i}} \le \lambda_{i_1,i_2,k_{d_i}}$ and

$$\begin{cases}\rho_{i_1}\le\lambda_{i_1,i_2,k_{d_i}}-\mu_{i_1,k_{d_i}}\\ \rho_{i_1}\le 0,\quad k_\ell\neq k_{d_i}.\end{cases}$$

Recall that solution $\bar{X}$ remains optimal for $(PX_{\tilde{\mu},\tilde{\gamma},\tilde{\lambda}})$, so $\bar{X}$, $(\tilde{u},\tilde{v})$, are primal dual solutions. Then,

$$\upsilon(PX_{\tilde{\mu},\tilde{\gamma},\tilde{\lambda}}) = \sum_{i=1}^{n}(\tilde{u}_i+\tilde{v}_i)+\sum_{\ell=1}^{k}d_\ell\,(\tilde{u}_{n+\ell}+\tilde{v}_{n+\ell}) = \upsilon(PX_{\mu,\gamma,\lambda})+\mu_{i_1,k_{d_i}}+\lambda_{i_1,i_2,k_{d_i}}+\ldots+\lambda_{i_{m-1},i_m,k_{d_i}}+\gamma_{i_m,k_{a_i}}$$

$$\upsilon(PY_{\tilde{\mu},\tilde{\gamma},\tilde{\lambda}}) = \upsilon(PY_{\mu,\gamma,\lambda,}) - \gamma_{i_m,k_{a_i}}$$

and using (11) we have $w(\tilde{\mu},\tilde{\gamma},\tilde{\lambda}) = w(\mu,\gamma,\lambda)+\mu_{i_1,k_{d_i}}$ ♦

The propositions 1 and 2 suggest a lagrangean dual ascendent procedure that operates, at the same time, on all the multipliers related to a particular vehicle block. From that, one generates monotonically non-decreasing values for the lagrangean function $w(\mu,\gamma,\lambda)$.

The process of maximizing $w(\mu,\gamma,\lambda)$, while $\bar{X}$ remains optimal to $(PX_{\mu,\gamma,\lambda})$, tends to create new alternative optimal solutions for this sub-problem. Let us denote by $\hat{X}$ one of such alternative solutions and by $\hat{Y}$ the corresponding solution in $(PY_{\mu,\gamma,\lambda})$.

proposition 3

If $\hat{X}$ and $\hat{Y}$ satisfy at equality all the relaxed constraints for which the lagrangean multipliers are positive then $(\hat{X}, \hat{Y})$ is an optimal solution to (P).

proof

The solutions $\hat{X}$, $(\hat{u}, \hat{v})$ and $\hat{Y}$, $\hat{\rho}$ are the corresponding optimal primal-dual solutions to $(PX_{\mu,\gamma,\lambda})$ and $(PY_{\mu,\gamma,\lambda})$, respectively. So, complementary slackness conditions concerning constraints (1), (2), (3) and (7) are satisfied.

If, in adition, $(\hat{X}, \hat{Y})$ satisfy with equality all the relaxed constraints for which the lagrangean multipliers are positive then the remaining complementary slackness conditions hold and the solution is optimal.

♦

4.Refinement Procedures

As said before, the process of maximizing $w(\mu, \gamma, \lambda)$ generates alternative optimal solutions to $(PX_{\mu,\gamma,\lambda})$. Such solutions yield null lagrangean multipliers, and no increasing in the lagrangean function is obtained. In order to avoid this obstacle and to improve the lower bound, some refinement procedures are proposed in this section.

For the sake of simplicity, and without loss of generality, let us consider again the vehicle block represented by the following chain :

$$k_{d_i} \rightarrow i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow i_{m-1} \rightarrow i_m \rightarrow k_{a_i}$$

The next proposition shows how to penalize the arcs involved in this vehicle block where one considers an alternative arrival depot denoted by k_{a_j}.

proposition 4

Suppose that all alternative solutions to $\bar{X}$, where the arc (i_m,k_{a_i}) is absent, have an arc (i_m,k_{a_j}) with $k_{a_j} > n$ and $k_{a_j} \neq k_{d_i}$. Then new lagrangean multipliers $\mu_{i_1,k_{d_i}}$, $\lambda_{i_1,i_2,k_{d_i}}$, ... , $\lambda_{i_{m-1},i_m,k_{d_i}}$, $\gamma_{i_m,k_{a_i}} \in W$, and $\gamma_{i_m,k_{a_j}} \geq 0$, satisfying $\mu_{i_1,k_{d_i}} \leq \lambda_{i_1,i_2,k_{d_i}} \leq \ldots \leq \lambda_{i_{m-1},i_m,k_{d_i}} \leq \gamma_{i_m,k_{a_i}} = \gamma_{i_m,k_{a_j}}$, can be associated to the corresponding arcs, producing an increase on the lower bound given by $\mu_{i_1,k_{d_i}}$.

proof

Since , $k_{a_j} \neq k_{a_i}$ we have:

$$\begin{cases} \rho_{i_m} \leq -\lambda_{i_{m-1},i_m,k_{d_i}} \\ \rho_{i_m} \leq -\gamma_{i_m,k_{a_i}} \\ \rho_{i_m} \leq -\gamma_{i_m,k_{a_j}} \\ \rho_{i_m} \leq 0, \quad k_\varrho \neq k_{d_i}, k_\varrho \neq k_{a_i} \text{ and } k_\varrho \neq k_{a_j} \end{cases}$$

As $\lambda_{i_{m-1},i_m,k_{d_i}} \leq \gamma_{i_m,k_{a_i}}$ and $\gamma_{i_m,k_{a_i}} = \gamma_{i_m,k_{a_j}}$ it results $\rho_{i_m} = -\gamma_{i_m,k_{a_i}}$. The remaining ρ_j $(j \neq i_m)$ are computed as in prop.2 proof's. Then, using property 2 we obtain an increment on the lower bound given by $\mu_{i_1,k_{d_i}}$. ♦

Concerning a vehicle block with alternative departure depots, we can state the following :

proposition 5

Suppose that all alternative solutions to $\bar{X}$, where the arc (k_{d_i},i_1) is absent, have an arc (k_{d_j},i_1) such that $k_{d_j} > n$, $k_{d_j} \neq k_{a_i}$ and $\gamma_{i_m,k_{d_j}}=0$. Then, new lagrangean multipliers $\mu_{i_1,k_{d_i}} \leq \lambda_{i_1,i_2,k_{d_i}} \leq \ldots \leq \lambda_{i_{m-1},i_m,k_{d_i}} \leq \gamma_{i_m,k_{a_i}}$ and $\mu_{i_1,k_{d_j}} \leq \lambda_{i_1,i_2,k_{d_j}} \leq \ldots \leq \lambda_{i_{m-1},i_m,k_{d_j}}$, with $\mu_{i_1,k_{d_i}} = \mu_{i_1,k_{d_j}}$, and $\mu_{i_1,k_{d_i}}$, $\lambda_{i_1,i_2,k_{d_i}} + \lambda_{i_1,i_2,k_{d_j}}$,..., $\lambda_{i_{m-1},i_m,k_{d_i}} + \lambda_{i_{m-1},i_m,k_{d_j}}$, $\gamma_{i_m,k_{a_i}} \in W$, can be associated to the corresponding arcs, producing a stronger lower bound given by $w(\mu,\gamma,\lambda) + \mu_{i_1,k_{d_i}}$.

proof

Since , $k_{d_j} \neq k_{a_i}$, $\gamma_{i_m,k_{d_j}}=0$ and the lagrangean multipliers satisfy the inequalities stated above, we have that:

$$\begin{cases} \rho_{i_m} \leq -\lambda_{i_{m-1},i_m,k_{d_i}} \\ \rho_{i_m} \leq -\gamma_{i_m,k_{a_i}} \\ \rho_{i_m} \leq 0 \quad k_\ell \neq k_{d_i} \text{ and } k_\ell \neq k_{a_i} \end{cases} \qquad \text{gives } \rho_{i_m} = 0.$$

$$\begin{cases} \rho_{i_{m-1}} \leq \lambda_{i_{m-1},i_m,k_{d_i}} - \lambda_{i_{m-2},i_{m-1},k_{d_i}} \\ \rho_{i_{m-1}} \leq \lambda_{i_{m-1},i_m,k_{d_j}} - \lambda_{i_{m-2},i_{m-1},k_{d_j}} \\ \rho_{i_{m-1}} \leq 0 \quad k_\ell \neq k_{d_i} \text{ and } k_\ell \neq k_{d_j} \end{cases} \qquad \text{gives } \rho_{i_{m-1}} = 0.$$

$$\begin{cases} \rho_{i_1} \leq -\mu_{i_1,k_{d_i}} + \lambda_{i_1,i_2,k_{d_i}} \\ \rho_{i_1} \leq -\mu_{i_1,k_{d_j}} + \lambda_{i_1,i_2,k_{d_j}} \\ \rho_{i_1} \leq 0 \quad k_\ell \neq k_{d_i} \text{ and } k_\ell \neq k_{d_j} \end{cases} \qquad \text{gives } \rho_{i_1} = 0.$$

Since the remainder ρ_j are computed as in proposition. 2 proof, the lower bound is given by $w(\mu,\gamma,\lambda) + \mu_{i_1,k_{d_i}}$. ♦

A stronger result relatively to the previous proposition can be obtained by considering the following new set of additional constraints in the mathematical model (P) :

$$y_{j\ell} + x_{ij} - y_{i\ell} \leq 1 \qquad i,j=1,2...,n;\ \ell=1,2...,k \quad (6')$$

This constraint set is redundant in (P), but will be useful for dealing with the alternative solutions to $\bar{X}$. Denoting by $\varepsilon_{ij\ell}$ the corresponding lagrangean multipliers we can derive the following :

proposition 6

Suppose all alternative solutions to $\bar{X}$, where the arc (k_{d_i},i_1) is absent, have an arc (k_{d_j},i_1) with $k_{d_j} > n$, $k_{d_j} \neq k_{d_i}$ and $\gamma_{i_m,k_{d_j}}=0$. Then lagrangean multipliers $\mu_{i_1,k_{d_j}} \geq 0$, $\mu_{i_1,k_{d_i}}$, $\varepsilon_{i_1,i_2,k_{a_i}}$, ... , $\varepsilon_{i_{m-1},i_m,k_{a_i}}$, $\gamma_{i_m,k_{a_i}} \in W$, satisfying $\mu_{i_1,k_{d_j}} = \mu_{i_1,k_{d_i}} \geq \varepsilon_{i_1,i_2,k_{a_i}} \geq \ldots \geq \varepsilon_{i_{m-1},i_m,k_{a_i}} \geq \gamma_{i_m,k_{a_i}}$, can be associate with the corresponding arcs, producing an increase in the lower bound given by $w(\mu,\gamma,\lambda) + \gamma_{i_m,k_{a_i}}$.

The above proposition makes the case of alternative departure depots become similar to the case of arrival depots expressed by proposition 4. The proof for proposition 6 closely follows the one presented for proposition 4.

5.Upper Bounds

In this section, we describe two alternative approaches for obtaining upper bounds on the optimal value for P:

- a greedy type algorithm;
- a schedule first - cluster second procedure, where a multi-depot problem without requiring that each vehicle returns to the source depot is solved, in the first phase.

In order to suggest the greedy type algorithm, let us consider two vehicle blocks V_i and V_j.

$$k_{d_i} \rightarrow i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow i_{m-1} \rightarrow i_m \rightarrow k_{a_i}$$
$$k_{d_j} \rightarrow j_1 \rightarrow j_2 \rightarrow \ldots \rightarrow j_{t-1} \rightarrow j_t \rightarrow k_{a_j}$$

One wants fo merge the vehicle blocks V_i and V_j so that two new vehicle blocks $V_{i'}$ and $V_{j'}$ are obtained each one assigned to exactly one depot. But, like in lagrangean dual solution, one has to guarantee the non-violation of each depot capacity. The easiest way to do this is to assign the new vehicle blocks $V_{i'}$ and $V_{j'}$, always to the starting depot of V_i and V_j, respectly.

$$k_{d_i} \rightarrow i_1 \rightarrow i_2 \rightarrow \ldots \rightarrow j_{t-1} \rightarrow j_t \rightarrow k_{d_i}$$
$$k_{d_j} \rightarrow j_1 \rightarrow j_2 \rightarrow \ldots \rightarrow i_{m-1} \rightarrow i_m \rightarrow k_{d_j}$$

The basic idea of the algorithm is to merge the vehicle block V_i with the vehicle block V_j that gives the best value for the objective function of P.

Note that, instead of assigning vehicle blocks $V_{i'}$ and $V_{j'}$ always to the starting depot of V_i and V_j we may assign this new vehicle blocks always to the ending depot of V_i and V_j.
Under this condition we can obtain a feasible solution to problem (P) that may be different than the solution obtained before. If so, we keep the solution corresponding to the smaller upper bound.

For more details see Mesquita and Paixão, (1990) .

The alternative heuristic approach we propose for determining upper-bounds is a two phases-algorithm: schedule first - cluster second. The idea is to consider, in the first phase, the vehicle blocks given by the lagrangean dual solution, without the corresponding depots. Then each vehicle block is assigned to one of the depots in order to minimize the total costs and to satisfie the capacity constraints. This second phase is a classical transportation problem where the sources correspond to depots and the sinks to the vehicle blocks. For more details see Branco et al. (1992).

6.Computational Results

Two types of test problems have been considered:

Type 1 - Real life data from CARRIS (Lisbon) as is described in Branco et al. (1992).
This type of test problems includes two classes of instances:
Class A: minimizing total number of vehicles + dead-heading costs
Class B: minimizing total number of vehicles + dead-heading costs + iddle time costs.
All instances have 4 depots and no limit for the number of vehicles available at any depot was imposed by the company. In order to ensure the use of the minimum number of vehicles, a penalty of 5000 is added to costs associate with links to and from the depots.

For such problems, subgradient optimization was unable to improve the lower bounds.
TABLE 1 and TABLE 2 show the performance of the method proposed in §3 and §4.

Type 2 Randomly generated test problems generated as in Carpaneto et al.(1989) and Ribeiro & Soumis (1991). For each value of n and k, 5 instances have been considered.

TABLE 3 shows the average behavior of the algorithm for such type problems. For such type problems, the upper bound is given by the greedy type algorithm, since it has produced better results than the two-phases algorithm.

The method described in the previous sections was coded in PASCAL and implemented on a personal computer SHINE 486 DX, 66 MHz.

Concerning to the multi-depot problem without requiring that each vehicle returns to the source depot, it is solved by a generalized version of the quasi-assignment algorithm Branco, (1989) that takes into account each depot capacity Mesquita & Paixão, (1990).

The heuristic procedures that turns the lagrangean relaxation solution in a feasible solution for problem P are performed whenever an alternate solution to $\bar{X}$ is obtained.

When the optimal solution is not reached, the lower and upper bounds give a value for the aproximation error to the optimum : % gap = (upper bound - lower bound) / (lower bound) × 100 %.

Each TABLE shows the % compatible pairs of trips (i,j), i,j=1,...,n, the % gap concerning $\mu=\gamma=\lambda=0$ (% first gap), the % gap after aplying the multiplier adjustement method, proposed in §3 and §4 (% final gap) and the computing time for the upper bound and the lower bound.

TABLE 1 - Problems Type 1 - Class A

n	k	% first gap	% final gap	avtime (sec) lb	avtime (sec) ub(greedy)	% first gap	% final gap	avtime (sec) ub(two-phas)
192	4	0.0629	0.0457	1.42	214.92	0.0931	0.0711	2.07
192	4	0.0458	0.0303	1.21	248.37	0.0738	0.0614	1.74
192	4	0.0317	0.0205	1.21	246.09	0.0732	0.0588	2.26
192	4	0.0381	0.0274	1.05	229.16	0.0632	0.0533	1.74
255	4	0.0495	0.0392	1.38	587.40	0.0627	0.0515	3.14
255	4	0.0417	0.0340	1.10	534.66	0.0635	0.0529	3.04
255	4	0.0419	0.0340	1.32	556.54	0.0643	0.0537	2.96
255	4	0.0431	0.0354	1.21	616.53	0.0671	0.0566	2.82
264	4	0.0365	0.0211	2.85	363.80	0.0229	0.0117	2.40
264	4	0.0299	0.0167	2.64	341.77	0.0213	0.0130	1.88
352	4	0.0458	0.0223	6.92	1154.49	0.0481	0.0234	4.48
352	4	0.0442	0.0326	5.99	1266.29	0.0518	0.0377	4.44

TABLE 2 - Problems Type 1 - Class B

n	k	% first gap	% final gap	avtime (sec) lb	ub(greedy)	% first gap	% final gap	avtime (sec) ub(two-phas)
192	4	0.0533	0.0394	1.49	264.68	0.1061	0.0851	2.40
192	4	0.0608	0.0386	2.14	261.68	0.0854	0.0701	2.34
192	4	0.0624	0.0390	1.92	261.03	0.0854	0.0707	2.03
192	4	0.0624	0.0390	1.92	263.98	0.0854	0.0707	2.28
255	4	0.0385	0.0346	1.81	688.11	0.0721	0.0625	3.99
255	4	0.0433	0.0382	1.32	642.94	0.0689	0.0609	3.76
255	4	0.0420	0.0374	1.38	619.68	0.0693	0.0614	3.22
255	4	0.0420	0.0374	1.32	616.98	0.0693	0.0614	3.33
264	4	0.0351	0.0171	3.74	365.45	0.0229	0.0091	2.81
264	4	0.0285	0.0105	4.06	371.43	0.0213	0.0081	2.00
352	4	0.0445	0.0283	7.52	1269.49	0.0503	0.0328	3.92

TABLE 3 - Randomly generated test problems

n	k	% first gap	% final gap	avtime (sec) lb	ub(greedy)
100	2	0.0778	0.0370	0.29	19.79
100	4	0.2174	0.1228	0.45	33.33
100	6	0.3244	0.1870	0.60	36.82
200	2	0.0861	0.0511	1.36	337.36
200	4	0.2793	0.1924	2.03	453.01
200	6	0.3306	0.2383	2.39	513.10
300	2	0.1010	0.0640	3.59	1300.72
300	4	0.2494	0.1967	6.74	2095.23
300	6	0.3680	0.2867	7.74	2473.78
400	2	0.1105	0.0951	10.12	4410.63
400	4	0.3231	0.2568	17.51	7239.41
400	6	0.3567	0.2902	21.68	8670.16

In TABLE 3, considering random generated test problems, we didn't show the % gap refering the two phases algorithm since it has always produced worst upper bounds then the greedy- type heuristic.

7.Remarks and Conclusions

In this paper we adressed the multi-depot vehicle scheduling problem. We considered a mathematical program where integer variables and extra constraints are added to a quasi-assignment formulation so as to guarantee that each vehicle returns to the source depot.

A separable lagrangean relaxation problem was obtained by dualizing some of the extra constraints. Then, we restricted the corresponding lagrangean multipliers in order to keep the optimality of the initial quasi-assignment solution. Some properties of the restricted lagrangean relaxation led to a dual ascendent method for computing such multipliers.

Eventual unfeasibilities of the lagrangean relaxation solution can be easily removed by using a greedy type heuristic or a two-phase type algorithm.

When considering real life data from Carris-Lisboa, an usual subgradient optimization method was unable to improve the % first gap. The computational experience reported in the section 6 shows that, now, we are able to improve on the lower bounds for those real life instances (TABLES 1 and 2).

Like for real life data, but concerning the randomly generated test problems we have shown in Mesquita & Paixão, (1994) that the multiplier adjustement method, described in this paper, dominates subgradient optimization since it gives better lower bounds in a shorter time.

In conclusion, the method proposed gives very tight bounds which could be a good starting point for applying a branch and bound technique.

References

Balas, E; Christofides, N; (1981), " A Restricted Lagrangean Approach to the Travelling Salesman Problem ", *Mathematical Programming* 21, 19-46.

Bertossi, A; Carraresi, P; Gallo, G; (1987), " On some Matching Problems arising in Vehicle Scheduling Models ", *Networks*, vol 17, 271-281.

Bodin, L.; Golden, A., Assad; Ball, M.; (1983), " Routing and Scheduling of Vehicles and Crews: The State of the Art ", *Computers and Operations Research* 10, 63-211.

Branco, I. (1989). Algoritmos para Modelos Matemáticos de Quasi-Afectação e Extensões. Tese de Doutoramento, Faculdade de Ciências - Universidade de Lisboa.

Branco, I., M.; Mesquita, M.; Paixão, J.; (1992), " Vehicle Scheduling Problems: Pratical Experience at Large Road Transport Companies ", DEIO - CEAUL, University of Lisbon, *Internal Report*.

Carpaneto, G; Dell'Amico, M; Fischetti, M; Toth, P; (1989), " A Branch and Bound Algorithm for the Multiple Depot Vehicle Scheduling Problem ", *Networks*, 19, 531-548.

Ceder, A.; Stern, I., H.; (1981), " Deficit Function Bus Scheduling with Deadheading Trip Insertions for Fleet Size Reduction ", *Transportation Science* 15, 338-363.

Dell'Amico, M; Fischetti, M; Toth, P; (1990), " Heuristic Algorithms for the Multiple Depot Vehicle Scheduling Problem ", DEIS, University of Bologna, *Technical Report* OR/90/3.

El-Azm, A.; (1985), " The Minimum Fleet Size Problem and Its Applications to Bus Scheduling ", *Computer Scheduling of Public Transport 2* (J.-M. Rousseau, ed.), North-Holland, Amsterdam, 493-512.

Forbes, M. A.; Holt, J. N; Watts, A.M.;(1994), " An Exact Algorithm for the Multiple Depot Bus Scheduling", *European Journal of Operational Research*, 72, 9, 115-124.

Geoffrion, A. M.; (1974), " Lagrangean Relaxation and it uses in Integer Programming", *Mathematical Programming Study*, 2, 82-114.

Mesquita, M; Paixão, J; (1990), " Multiple depot Vehicle Scheduling Problem: A new heuristic based on Quasi-assignment Algorithms ", Computer-Aided Transit Scheduling. *Lecture Notes in Economics and Mathematical Systems* 386. Springer-Verlag, 167-180.

Mesquita, M.; Paixão, J.; (1994), " A Multiplier Adjustement Method for the Multi-Depot Vehicle Scheduling Problem: A Comparison with Subgradient Optimization Techniques ", DEIO - CEAUL, University of Lisbon, *Internal Report*.

Ribeiro, C.C.; Soumis, F.; (1991), " A column Generation Approach to the Multiple-Depot Vehicle Scheduling Problem ", Research Report, GERARD, Montréal.

Smith, B.; Wren, A.; (1981), " Vampires and Tasc: Two Sucessfully Applied Bus Scheduling Programs ", *Computer Scheduling of Public Transport: Urban Passenger Vehicle and Crew Scheduling* (A. Wren, ed.), North-Holland, 1981, Amsterdam, 97-124.

An Optimization Algorithm for the Vehicle Routing Problem with Time Windows based on Lagrangean Relaxation

Niklas Kohl Oli B.G. Madsen

IMSOR
The Institute of Mathematical Statistics and Operations Research
The Technical University of Denmark
DK-2800 Lyngby
Denmark

E-mail: nk@imsor.dth.dk
Revised: May 20, 1994
Submitted to Operations Research

Our paper presents a new optimization method for the Vehicle Routing Problem with Time Windows (VRPTW). The VRPTW is a generalization of the Vehicle Routing Problem, where the service of a customer must start within a given time interval - a so-called time window. Our method is based on a Lagrangean relaxation of the constraint set requiring that each customer must be serviced. The master problem consists of finding the optimal Lagrangean multipliers and the subproblem is a Shortest Path Problem with Time Windows and Capacity Constraints. The optimal multipliers are found using a method exploiting the benefits of subgradient methods as well as a bundle method. The method has been implemented and tested on a series of well known data sets of size up to 100 customers. Our algorithm turns out to be faster than other algorithms considered in the literature, and we have succeeded in solving several previously unsolved problems.

Introduction:

Routing and scheduling problems are major elements of many logistic systems, and a large amount of research has been devoted to find good solutions to these complex combinatorial optimization problems. One of the major research topics has been the Vehicle Routing Problem (VRP) which involves finding a set of routes, starting and ending in a depot, which together cover a set of customers. Each customer has a given demand, and no vehicle can serve more customers than its capacity enables it to. The objective can be to minimize the total distance travelled or the number of vehicles used or a combination of these objectives.

In this paper we consider the Vehicle Routing Problem with Time Windows (VRPTW) which is a generalization of the VRP. A solution to the VRPTW must ensure that the service of any customer starts within a given time interval, a so called time window. The VRPTW can be considered as a special case of Resource Constrained Routing which also contains problems such as the multiple Travelling Salesman Problem with Time Windows, the Dial-a-Ride Problem and Crew Scheduling Problems (see Desrosiers et al. (1993)). The results presented in this article can be generalized to these problems.

The most successful approaches for the VRPTW are based on constrained shortest path relaxations. Desrochers, Desrosiers and Solomon (1992) have used a Dantzig-Wolfe decomposition scheme and Halse (1992) has used a decomposition based on Variable Splitting (also known as Lagrangean decomposition). In both cases the resulting subproblem is a shortest path problem with time and capacity constraints. Even though this problem is $\mathcal{NP}$-complete a pseudo polynomial algorithm exists, and most instances are solvable.

We have developed an algorithm exploiting Lagrangean relaxation of the constraint set requiring that all customers must be serviced. The master problem consists of finding the optimal Lagrangean multipliers. The subproblem is a time and capacity constrained shortest path problem for each vehicle. This approach creates relatively easy instances of the subproblem compared to the Dantzig-Wolfe decomposition and is more simple than the Variable Splitting approach, since the number of Lagrangean multipliers is much smaller.

However, it is often a major problem to find the optimal Lagrangean multipliers, and the commonly used subgradient method gives slow convergence. Therefore we have developed a more sophisticated master iteration method exploiting a number of results in non-differentiable optimization theory. On a number of benchmark problems by Solomon (1987) we obtain optimal solutions faster than reported by Desrochers, Desrosiers and Solomon (1992) and Halse (1992). Furthermore we have solved a number of previously unsolved problems.

In section 1 we present a mathematical model of the VRPTW. Section 2 gives a brief review of existing methods. We focus especially on the methods of Desrochers, Desrosiers and Solomon (1992) and Halse (1992), since these methods are quite similar to our method. Our method is considered in section 3, and we present some results comparing the lower bounds obtained with those of Desrochers, Desrosiers and Solomon (1992) and Halse (1992). Section 4 describes how the subproblem, a constrained shortest path problem, is solved and we provide information on how to estimate the difficulty of a particular instance of the problem. In section 5 the master problem, a non-differentiable concave maximization problem, is considered. Section 6 gives a brief description of the branching strategy implemented and section 7 reports on the computational experiments on a test data set. The last section contains our conclusions.

1 A mathematical model of the VRPTW

The VRPTW is given by a set of vehicles $\mathcal{V}$, a set of customers $\mathcal{C}$ and a directed network connecting the customers and a depot. $|\mathcal{C}| = n$ and the customers are denoted $1, 2, ..., n$. The depot is represented by two nodes denoted 0 and $n+1$. The arcs of the network corresponds to feasible connections between the nodes. No arc terminates in node 0 and no arc originate in node $n+1$. All routes starts in 0 and ends in $n+1$. The set of all nodes is denoted $\mathcal{N}$. A cost c_{ij} and a travel time t_{ij} is associated with each arc (i,j), $i \neq j$ of the network. t_{ij} may include a service time at customer i. $t_{0,n+1} = 0$ and $-c_{0,n+1}$ represents the cost of using a vehicle. A non-existing arc (i,j) can be modelled using a large cost c_{ij}, so apart from node 0 and $n+1$ we can assume that the network is fully connected. Each vehicle has a given capacity q_k, $k \in \mathcal{V}$ and each customer a demand d_i, $i \in \mathcal{C}$.

At each customer, the start of the service must be within a given time interval, a so called time window, $[a_i, b_i]$, $i \in \mathcal{C}$. Also vehicles must leave the depot within the time window $[a_0, b_0]$ and return in the time window $[a_{n+1}, b_{n+1}]$. A vehicle is permitted to arrive before the beginning of the time window, and wait until service is possible, but it is not permitted to arrive after the end of the time window. With out loss of generality we can assume that $a_0 = b_0 = 0$. We will further assume that all data, ie. q_k, d_i, c_{ij}, t_{ij}, a_i and b_i are non-negative integers. For technical reasons t_{ij} must be strictly positive.

The model contains two types of decision variables. x_{ijk} (defined $\forall i, j \in \mathcal{N}$, $\forall k \in \mathcal{V}$, $i \neq j$, $i \neq n+1$, $j \neq 0$) is 1 if vehicle k drives from node i to node j, and 0 otherwise. s_{ik} (defined $\forall i \in \mathcal{N}$, $\forall k \in \mathcal{V}$) denotes the time a vehicle k, $k \in \mathcal{V}$ starts service at customer i, $i \in \mathcal{C}$. If vehicle k does not service customer i, s_{ik} does not mean anything. $s_{0k} = 0, \forall k$ and $s_{n+1,k}$ denotes the arrival time of vehicle k to the depot. The objective is to design a set of minimal cost routes, one for each vehicle, so that all customers are serviced. The routes must be feasible with respect to the capacity of the vehicles and the time windows of the customers serviced.

The VRPTW can mathematically be stated as:

$$\textit{minimize} \quad \sum_{k\in\mathcal{V}}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} c_{ij}x_{ijk}, \qquad \textit{subject to}$$

$$\sum_{k\in\mathcal{V}}\sum_{j\in\mathcal{N}} x_{ijk} = 1 \qquad \forall i \in \mathcal{C} \qquad (1)$$

$$\sum_{i\in\mathcal{C}} d_i \sum_{j\in\mathcal{N}} x_{ijk} \leq q_k \qquad \forall k \in \mathcal{V} \qquad (2)$$

$$\sum_{j\in\mathcal{N}} x_{0jk} = 1 \qquad \forall k \in \mathcal{V} \qquad (3)$$

$$\sum_{i\in\mathcal{N}} x_{ihk} - \sum_{j\in\mathcal{N}} x_{hjk} = 0 \qquad \forall h \in \mathcal{C}, \forall k \in \mathcal{V} \qquad (4)$$

$$\sum_{i\in\mathcal{N}} x_{i,n+1,k} = 1 \qquad \forall k \in \mathcal{V} \qquad (5)$$

$$s_{ik} + t_{ij} - K(1 - x_{ijk}) \leq s_{jk} \qquad \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall k \in \mathcal{V} \qquad (6)$$

$$a_i \leq s_{ik} \leq b_i \qquad \forall i \in \mathcal{N}, \forall k \in \mathcal{V} \qquad (7)$$

$$x_{ijk} \in \{0, 1\} \qquad \forall i \in \mathcal{N}, \forall j \in \mathcal{N}, \forall k \in \mathcal{V} \qquad (8)$$

The objective function states that costs should be minimized. (1) states that each customer must be assigned to exactly one vehicle. We will denote this set of constraints the assignment constraints. (2) states that no vehicle must service more customers than its capacity enables it to. (3) - (5) are the flow constraints requiring that each vehicle k leaves node 0 once, leaves node i, $i \in \mathcal{C}$ if and only if it enters the node and that the vehicle returns to node $n+1$. However since the arc $(0, n+1)$ is included in the network the "empty" tour is permitted. (6) is stating that vehicle k can not arrive at j before $s_{ik} + t_{ij}$ if it drives from i to j. K is a large number. (7) secures that all time windows are respected and (8) are the integrality constraints.

We have not assumed that the triangle inequality, i.e. $c_{ij} \leq c_{ih} + c_{hj}$ and $t_{ij} \leq t_{ih} + t_{hj}$, $\forall h, i, j \in \mathcal{N}$, is satisfied. It is possible to add any scalar to all arc costs c_{ij} with out changing the optimal solution, so any instance of VRPTW can be transformed to an instance satisfying the triangle inequality on cost. This is not true for the time triangle inequality, but it seems that most applications will satisfy this property.

The model permits a fixed number of vehicles or an upper limit on the number of vehicles given by $|\mathcal{V}|$. If this upper limit is sufficiently high, it means that the number of vehicles is free. A free number of vehicles is modelled by setting $c_{0,n+1} = 0$. If the number of vehicles is fixed at $|\mathcal{V}|$ one can set $c_{0,n+1}$ to a large value ensuring that this arc will not be used in any optimal solution. Alternatively the arc $(0, n+1)$ can be removed from the network. We will consider problems with a fixed as well as a free number of vehicles. It is often assumed that all vehicles are identical, ie. $q = q_{k1} = q_{k2}$, $\forall k1, k2 \in \mathcal{V}$. We will make this assumption also, but the results can easily be generalized to the more general case.

This problem includes several well known problems such as the Travelling Salesman Problem, the Vehicle Routing Problem and the Bin Packing Problem. Finding an optimal solution to the VRPTW is a $\mathcal{NP}$-hard problem, and even finding a feasible solution when $|\mathcal{V}| < |\mathcal{C}|$ is $\mathcal{NP}$-hard, since this includes the problem of packing demand on $|\mathcal{V}|$ vehicles.

2 Review of existing optimization methods

Kolen, Rinnooy Kan and Trienekens (1987) have presented the first optimization method for the VRPTW. The method calculates lower bounds using dynamic programming and state space relaxation. Branching decisions are taken on route-customer allocations. The largest problem solved contained 15 customers.

Fisher, Jörnsten and Madsen (1992) describes an algorithm based on the K-tree relaxation of the VRPTW. Capacity constraints are handled by introducing a constraint requiring that some set $\mathcal{S}$, $\mathcal{S} \subset \mathcal{C}$, of customers must be visited by at least $k(\mathcal{S})$ vehicles. This constraint is Lagrangean relaxed and the problem resulting is still a K-tree problem with modified arc costs. Time windows are treated similarly. A constraint, requiring that not all arcs in a time violating path can be used, is generated and Lagrangean relaxed. The method has solved a few problems with up to 100 customers from the so called Solomon data set (Solomon 1987).

At present, the most successful algorithms for the VRPTW are based on shortest path decompositions of the problem. The fundamental observation is, that only constraint (1) is binding the vehicles together. The problem consisting of (2), (3), (4), (5), (6), (7) and (8) is a Shortest Path Problem with Time Windows and Capacity Constraints (SPPTWCC) for each vehicle. This problem is NP-hard in the strong sense (Dror 1993), but very efficient algorithms Dynamic Programming algorithms for a slightly relaxed problem exist. Furthermore most of the integrality gap between the LP- and IP-solution of VRPTW is often explored in the SPPTWCC. Two decompositions has been investigated computationally: Dantzig-Wolfe decomposition and Variable Splitting.

Desrochers, Desrosiers and Solomon (1992), (see also Desrosiers et al. (1993)), have applied column generation to a number of the VRPTW with free number of vehicles. As noted in Desrosiers et al. (1993) this column generation is in fact equivalent to Dantzig-Wolfe decomposition, so we will denote their approach Dantzig-Wolfe decomposition even though this term is not used in Desrochers, Desrosiers and Solomon (1992). The Dantzig-Wolfe decomposition exploits the fact that only constraint set (1) is linking the vehicles together. The method starts out with a solution consisting of $|\mathcal{C}|$ paths, each visiting one customer. The master problem consists of finding a minimum cost set of paths, among all generated paths, that together ensures that all customers are visited. The number of times a path is used is not necessarily an integer, but can be any number in the interval $[0, 1]$. Hence, the master problem is the LP-relaxation of a Set Partitioning problem. For computational reasons one solves a Set Covering problem instead. This does not change the final result, if the triangle inequality on time and cost is satisfied.

The simplex multipliers of the optimal solution are used to modify costs in the subproblems, which are shortest path problems with time windows and capacity constraints. The subproblems are identical, so only one problem is solved. If a path with negative marginal cost is found, it is included in the set of paths in the master problem. Since one might find several negative marginal cost paths, each iteration can generate several paths.

The method terminates when no negative marginal cost paths can be generated in the subproblem. At this point all paths used in the optimal solution to the master problem will have marginal cost 0 and all other paths will have a non-negative marginal cost. If the master problem turns out to have an integer solution, it is optimal. Otherwise a lower bound on the optimal objective value of VRPTW has been found and some branch-and-bound strategy must be employed to find the optimal solution. The method has solved problems with up to 100 customers from the Solomon data set, including many problems other methods have not been able to solve.

Halse (1992) describes three different Variable Splitting methods for the VRPTW. In Variable Splitting the variables in some of the constraints are renamed. A new type of constraint, coupling the original and

the new variables, is introduced and Lagrangean relaxed. This decomposes the problem into two or more independent problems. In the VRPTW this means that $\sum_j x_{ijk}$ is replaced by y_{ik} in some constraints. The constraint $\sum_j x_{ijk} = y_{ik}, \forall i \in \mathcal{C}, \forall k \in \mathcal{V}$ is introduced and Lagrangean relaxed. Now the problem separates into two problems, one in the x_{ijk} and s_{ik} variables and one in the y_{ik} variables. The Variable Splitting method is some times called Lagrangean decomposition, and has been described by Jörnsten, Näsberg and Smeds (1985) and Guignard and Kim (1987).

For the VRPTW there are three natural ways of separating the problem. The model VS_1 consists of a Shortest Path Problem with Time Windows (SPPTW), i.e. constraints (3), (4), (5), (6), (7) and (8), and a Generalized Assignment Problem (GAP), i.e. constraints (1) and (2). VS_2 consists of a Shortest Path Problem with Time Windows and Capacity Constraints (SPPTWCC), all constraints except (1), and a so called Semi Assignment Problem (SAP) consisting of constraints (1) only. SAP is basically a GAP without capacity constraints. In VS_3 constraints (2) are duplicated and modelled into both problems. We get a SPPTWCC and a GAP. Halse (1992) has implemented the VS_2 method and solved problems from the Solomon data set with up to 100 customers. Further more he has solved a 105 customer problem composed of data from two of the Solomon data set problems. Olsen (1988) has implemented the VS_1 method and solved problems with up to 16 customers. The VS_3 model has not been implemented.

3 A new approach based on Lagrangean Relaxation

In this section we present a new solution approach for the VRPTW based on Lagrangean Relaxation. The problem decomposes as in the Dantzig-Wolfe decomposition, but the way the dual variables / Lagrangean multipliers are optimized is new in this context. Further more we will compare the quality of the lower bounds obtained using this decomposition with the lower bounds obtained by the Dantzig-Wolfe decomposition and the VS_2 Variable Splitting method. Our method uses a Lagrangean Relaxation of the assignment constraint (1). The method is closely related to the Dantzig-Wolfe decomposition of Desrochers, Desrosiers and Solomon (1992) and the VS_2 Variable Splitting method of Halse (1992) since the subproblem - a SPPTWCC - is the same for all methods.

Using Lagrangean relaxation is not a new idea in the context of time constrained routing problems. Desrosiers, Sauvé and Soumis (1988) has used Lagrangean relaxation of the assignment constraint to estimate the minimum fleet size in the multiple Travelling Salesman Problem with Time Windows (m-TSPTW). m-TSPTW is just a VRPTW without constraints (2). However, in their recent review of time constrained routing problems (Desrosiers et al. (1993)) the focus is on Dantzig-Wolfe decomposition based methods and it is concluded that Dantzig-Wolfe decomposition has produced better computational results than other approaches. This is because the information obtained in the subproblem is exploited best in the Set Covering master problem, since all generated paths can be exploited in this master problem. They also claim that the Dantzig-Wolfe decomposition provides better information to design the branch-and-bound method. Our work shows how this imformation can be used in a Lagrangean relaxation scheme.

If (1) is Lagrangean relaxed the objective function can be written as:

$$minimize \sum_{k\in\mathcal{V}}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}(c_{ij}-\lambda_j)x_{ijk} + \sum_{j\in\mathcal{C}}\lambda_j, \ subject\ to\ (2),\ (3),\ (4),\ (5),\ (6),\ (7)\ and\ (8)$$

λ_j is the multiplier associated with the constraint requiring that customer j must be serviced. For any given value of the multiplier vector $\underline{\lambda}$, the resulting model decomposes into one subproblem for each vehicle. However, the subproblems are identical since their constraint sets are identical (recall that the vehicles are identical) and the arc costs are independent of the vehicle. The subproblem is a shortest path problem, ie. constraints (3) - (5), from the depot (node 0) back to the depot (node $n+1$) with the additional constraints (2), (6), (7) and (8). (2) are the capacity constraints and (6) and (7) are the time window constraints and (8) is the integrality constraint. The arc costs have been modified with the multipliers, so we can no longer assume that they are neither integers nor non-negative. Compared to the usual shortest path problem, which is easily solved, the SPPTWCC is difficult for two reasons. First because of the additional constraints (2), (6), (7) and (8). Second because the problem may contain negative modified cost cycles, ie. a path starting

and ending in the same node with a total negative modified cost. Even without (2), (6), (7) and (8), the shortest path problem with negative cycles is $\mathcal{NP}$-hard, if we require that no node must be visited more than once, since any instance of TSP can be transformed to this problem in polynomial time. In section 4 we will discuss solution methods for the SPPTWCC.

The Lagrangean multipliers connected to the assignment constraints can be interpreted as a price given for visiting the customer associated with the constraint. Each vehicle chooses its own path, minimizing the cost of the path modified by the multipliers of the customers serviced. If the number of vehicles is free, any vehicle has the option of using the $(0)-(n+1)$ path, ie. no route, at no cost.

If the SPPTWCC has a unique optimum, all vehicles will choose the same path, since vehicles are identical. Unless the problem contains only one vehicle, this can not be a feasible solution to VRPTWCC since customers on the shortest path are visited $|\mathcal{V}|$ times and the other customers are not visited at all.

For any multiplier vector $\underline{\lambda}$ the objective value of the optimal solution to the relaxed problem, denoted $f(\underline{\lambda})$, gives a dual lower bound on the optimal objective value for the VRPTW (Geoffrion 1974). If z denotes the modified cost of the shortest path, i.e. $z = \sum_i \sum_j (c_{ij} - \lambda_j) x_{ijk}$ for the optimal solution to the relaxed problem for any vehicle k, then $f(\underline{\lambda}) = |\mathcal{V}|z + \sum_i \lambda_i$.

The master problem consists of finding the $\underline{\lambda}$ that yields the best lower bound, ie. solving the model $maximize_{\underline{\lambda}}\, f(\underline{\lambda})$. The function f is concave and non-differentiable (Zowe 1985). We will denote the optimal solution $\underline{\lambda^*}$ and the corresponding modified cost of the shortest path z^*. If f is not bounded, the original problem is clearly infeasible. The opposite is, however, not true. A subgradient for $-f(\underline{\lambda})$ is given by $\underline{g} = (g_1, g_2, ..., g_n)$, where $g_i = 1 - \sum_j \sum_k x_{ijk} = 1 - |\mathcal{V}| \sum_j x_{ijk}$, since $\sum_j x_{ijk}$ counts how many times node i is serviced on the shortest path used by all $|\mathcal{V}|$ vehicles. Section 5 will discuss solution methods for the master problem.

If the SPPTWCC has a unique solution, $\underline{g}$ is also a gradient. However, the following proposition shows that the SPPTWCC generally has several optimal solutions when the multipliers are optimal. This means that the optimum is located at a kink, ie. a point where f is not differentiable.

Proposition 1.

Assume f is bounded. If $\underline{\lambda} = \underline{\lambda^*}$ then for any node i, $i \in \mathcal{C}$, there exists a path of minimal cost passing through i.

Proof.

Let $\underline{\lambda} = \underline{\lambda^*}$. Let z_i denote the cost of the shortest path passing node i, $i \in \mathcal{C}$, ie. $z^* = min_{i \in \mathcal{C}}(z_i)$. Assume that node i is not on any shortest path, ie. $z_i - \epsilon = z^*$, $\epsilon > 0$. Now increase λ_i with ϵ, ie. $\lambda_i := \lambda_i + \epsilon$. This decreases z_i down to z^*, but does not change z^*. Since z^* is unchanged the first term of the relaxed objective function is unchanged, but the second term is increased by ϵ. This contradicts that $\underline{\lambda}$ is optimal, and therefore shows that i must be on a shortest path if $\underline{\lambda} = \underline{\lambda^*}$. End of proof.

If there exist a set of minimal cost paths which together cover each customer exactly once, a feasible and optimal solution has been found. Whether this is the case, can be determined by solving a set partitioning problem with columns corresponding to the minimal cost paths and rows corresponding to customers. If the number of vehicles is fixed or bounded this must be included as a constraint. If a feasible solution exists, it is feasible in VRPTW and optimal. If not, we have only obtained a lower bound on the optimal objective value, and it is necessary to use some branch and bound strategy to find the optimal solution. Section 6 discusses branching strategies.

The model does permit a free number of vehicles, but it turns out that the problem is easier to solve if the number of vehicles is fixed, ie. if the arc $(0, n+1)$ is omitted. Therefore we will present a result relating the two problems to each other.

Consider the fixed number of vehicles problem. We will introduce the function $f_m(\underline{\lambda})$ which measures the lower bound obtained using multipliers $\underline{\lambda}$ and exactly m vehicles. $f_m(\underline{\lambda})$ is maximized by $\underline{\lambda_m^*}$ and z_m^* is the cost of any corresponding shortest path. It is easily seen that $f_m(\underline{\lambda_m^*}) = mz_m^* + \sum_i \lambda_{m(i)}^*$, where $\lambda_{m(i)}$ is the i'th element of $\underline{\lambda_m}$. We are now ready to present the following proposition.

Proposition 2.

$[z_m^* \geq 0] \Rightarrow [f_p(\underline{\lambda}_p^*) \geq f_m(\underline{\lambda}_m^*) + (p-m)z_m^*, \ \forall p > m]$.

Proof.

$f_p(\underline{\lambda}_p^*) \geq f_p(\underline{\lambda}_m^*) = pz_m^* + \sum_i \lambda_{m(i)}^* = (p-m)z_m^* + mz_m^* + \sum_i \lambda_{m(i)}^*$.

Proposition 2 can be useful in several ways. In this work we are using it to prove that an optimal solution to the VRPTW with a fixed number of vehicles is also optimal for the corresponding problem with a free number of vehicles. For the problems studied, we have initially calculated a lower bound, denoted m on the number of vehicles needed to serve all the customers. A simple lower bound is calculated by finding the smallest integer not smaller than total demand divided by the capacity of the vehicles. A possibly tighter bound can be found by solving the corresponding bin-packing problem (Fisher 1989), but this has not been necessary in this work. Then we have solved the problem with exactly m vehicles. In all cases z_m^* was positive and using the bounds in proposition 2, we were able to conclude that $|\mathcal{V}| > m$ could not be optimal. This is not a surprising result, since it often will be optimal to use the smallest possible number of vehicles.

Now we consider the quality of the lower bound obtained using our method with the lower bounds obtained using Dantzig-Wolfe decomposition (Desrochers, Desrosiers and Solomon 1992) and the VS_2 Variable Splitting method (Halse 1992). It turns out that we obtain the same lower bound as the Dantzig-Wolfe decomposition and the Variable Splitting decompositions. This bound is generally very good, because a large proportion of the gap between the LP- and IP-solution of the VRPTW is explored in the subproblem, which does not posses the integrality property, but is solved to the IP-optimum. The drawback is - of course - that SPPTWCC is a difficult $\mathcal{NP}$-hard problem.

First consider the Dantzig-Wolfe decomposition. Since the decompositions are identical the bounds obtained are identical. This has been noted by Desrosiers et al. (1993). In the free number of vehicles case it is clear that there exists a dual solution $\underline{\lambda}$ maximizing $f(\underline{\lambda})$ for which $z^* = 0$. In that case $f(\underline{\lambda}) = \sum_i \lambda_i$. This is exactly the objective value of the LP-dual, since all the right hand side values are 1. The fixed number of vehicles case is only slightly more complicated. In the Dantzig-Wolfe decomposition a fixed number of vehicles appears as an additional constraint in the master problem. The simplex multiplier on this constraint, which can be interpreted as the marginal cost of an additional vehicle, is used to modify the cost of the arc $(0, n+1)$ in the SPPTWCC subproblem. At the optimum this marginal cost will be equal to the modified cost of the shortest paths z^*. Therefore the optimal value of the LP-dual is $|\mathcal{V}|z^* + \sum_i \lambda_i$ which is also the value of $f(\underline{\lambda})$.

Now consider the VS_2 method. The two subproblems are SAP and SPPTWCC. The SAP posses the integrality property, i.e. there exists an optimal solution to the LP-relaxation of SAP which is integral. Therefore the lower bound obtained by Variable Splitting is, in this case, not better than the lower bound obtained by an ordinary Lagrangean relaxation of constraint set (1) (Guignard and Kim 1987).

4 Solution method for the subproblem

The subproblem is given by

$$minimize \sum_{k\in\mathcal{V}} \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} (c_{ij} - \lambda_j) x_{ijk} + \sum_{j\in C} \lambda_j, \ subject\ to\ (2),\ (3),\ (4),\ (5),\ (6),\ (7)\ and\ (8)$$

Compared to the usual shortest path problem with non-negative arc costs, which can be solved in $O(n^2)$ time (Nemhauser and Wolsey 1988), this problem is complicated for two reasons. The existence of side constraints, ie. (2), (6) and (7), is the first reason. The possible existence of negative costs cycles is the other reason (note that (6) rules out any path containing a cycle). Methods to overcome the first difficulty are well known, but not much is known on how to deal with negative costs cycles.

The time window constraints and capacity constraints are modelled into the network by making one

copy of each node in the original network for each permitted combination of time and (used) capacity. To keep the network finite, we must require that time and capacity take values only from a discrete set, such as the natural numbers $\mathbb{N}$. The number of nodes in the new network is given by $\sum_{i \in \mathcal{N}}(b_i - a_i + 1)(q + 1)$. For the test problems studied in section 7 this approximately means between 500,000 and 17,000,000 nodes in the new network. The networks must not necessarily be created explicit, but can be treated implicitly by lists of labels. However, the number of nodes gives an upper bound on the running time of the algorithm.

The shortest path in the new expanded network does not contain cycles since time is increasing on every arc. However, a node (in the original network) might still be visited more than once. Fortunately the cycles will end sooner or later, because time or capacity runs out. We have implemented the dynamic programming (DP) algorithm for the SPPTW by Desrochers (1988). The algorithm has been modified to handle capacity constraints as well as time windows.

Houck et al. (1980) has suggested a method to eliminate 2-cycles, ie. cycles of the form $... - (i) - (j) - (i) - ...$ but the method can not easily be generalized to cycles of more nodes. Therefore we are forced to solve a relaxed version of SPPTWCC permitting cycles of 3 nodes or more. We have implemented this 2-cycle elimination. Several strategies to decrease the running time of the algorithm has been proposed. Desrochers, Desrosiers and Solomon (1992) suggest a method to decrease the width of the time windows and also proposes to use a pulling algorithm instead of the more straight forward reaching algorithm usually used in DP- algorithms for shortest path problems. We have not implemented any of these suggestions, since we believe that they will only lead to minor improvements.

The solution time for the SPPTWCC is extremely dependent on the concrete instance. It is well known, that wide time windows and/or large capacity generally creates more difficult subproblems, because the number of nodes in the expanded network grows with the width of the time windows and capacity, and because wide time windows creates the possibility of cycles. For this reason the VRPTW generally becomes easier when the time windows become narrower.

Our work reveals that the running time is also very dependent on the arc cost, ie. on $\underline{\lambda}$. Low arc costs, ie. high multipliers, makes the problem much more difficult since very long paths come into consideration. Figure 1 illustrates this point. For 4 different problems from the Solomon data set, see section 7, we have fixed the multipliers at $\underline{\lambda_0} = (\lambda_0, \lambda_0, ..., \lambda_0)$ and found the running time of the SPPTWCC-algorithm on a HP-9000-735 computer. The arc $(0, n+1)$ was not included, so any path must visit at least one customer. This is a very primitive experiment, since the multipliers generally vary from customer to customer, but it illustrates our general point - that large multipliers makes the SPPTWCC much more difficult.

The problem $c101.100$ has very narrow time windows, $c102.100$ and $c103.100$ have wider time windows, and in $c104.100$ the time windows are hardly constraining. For $c101.100$ the running time increases by a factor 5 when λ_0 goes from 0 to 300. For the $c104.100$ the running time is increased by a factor 106. This is because the problems with very wide time windows are especially sensitive to large multipliers, since the number of time feasible paths (possibly including cycles) is virtually infinite.

We can of course not control the value of $\underline{\lambda^*}$, but we do have some control over the way $\underline{\lambda}$ is modified in the master problem. It seems vise to start out with relatively low multipliers, ie. a low λ_0 value, and gradually increase them using small steps. This way we go through a series of relatively easy instances of SPPTWCC before the optimum is reached.

In the Dantzig-Wolfe decomposition method the multipliers are not directly controlled, but are determined by the simplex multipliers in the Set Covering master problem. If the "good" paths have not been generated yet, the simplex multipliers will tend to be to high, thus creating difficult subproblems. We believe this is the main reason for the success of our method compared to the Dantzig-Wolfe decomposition.

5 Solution method for the master problem

The master problem is a concave non-differentiable maximization problem. We have implemented a method exploiting the advantages of both traditional subgradient optimization, see eg. Shor (1985), and a bundle method, see eg. Lemaréchal, Strodiot and Bihain (1981). For a general introduction to non-differentiable

optimization, see eg. Zowe (1985). In non-differentiable optimization it is common to study a minimization problem. Therefore we formulate the master problems as

$$minimize_{\underline{\lambda}} \; h(\underline{\lambda}) = -f(\underline{\lambda})$$

Both subgradient methods and bundle methods use the current iterate $\underline{\lambda}$, a step direction $\underline{d}$, $\underline{d} \in \mathbb{R}^n$, and a step size t, $t \in \mathbb{R}$, to calculate the new iterate $\underline{\lambda}$. If the suffix denotes the iteration number, the master iteration number p consists in finding $\underline{d_p}$ and t_p and setting

$$\underline{\lambda_{p+1}} := \underline{\lambda_p} + t_p \underline{d_p}$$

In subgradient optimization one uses a current subgradient as step direction. Note that the solution of SPPTWCC automatically provides us with a subgradient. The step size must be chosen a priori, ie. without a line search in the direction d_p. Several rules for choosing t_p exist. A major drawback of the subgradient optimization method is, that it only gives guaranteed convergence to optimum if t_p is a divergent series, or if the optimal objective value is known. In the first case guaranteed convergence even requires an infinite number of iterations. In practical computation this means that one must often settle with a solution which is not guaranteed optimal. Subgradient optimization does not provide a stop criterion and no primal solution is obtained.

In a bundle method a convex combination of previously obtained subgradients (the bundle of subgradients) is used to calculate $\underline{d_p}$. A line search is performed and a step is taken if $\underline{d_p}$ is a direction of "sufficient" descent. Otherwise a so called null step is taken. The new obtained subgradient is included in the bundle, and a new step direction is calculated. The algorithm terminates at optimum after a finite number of iterations (Lemaréchal, Strodiot and Bihain 1981). Calculating $\underline{d_p}$ involves solving a quadratic programming problem (QP). The number of variables in the QP is equal to the number of subgradients one stores. Ideally one should store all subgradients previously obtained, but our experience shows that one can settle with far less. For the test data considered in section 7, we have used a bundle size of at most 50.

The advantages of subgradient optimization are, that it is easy to implement and usually gives rapid improvement in the first iterations. Later convergence slows down, and for some difficult problems convergence in a reasonable time is not possible. The advantages of the bundle method are, that it secures reasonably fast convergence to the proven optimum, when one is relatively close to optimum. Far from optimum the performance is relatively poor because the line search often requires several iterations of little value. Therefore it is natural to consider a strategy initially using a number of subgradient iterations and then finding the optimum using the bundle method. We have implemented a very simple version of this idea, which turns out to produce very competitive running times. In Phase I, we ran a fixed number of subgradient iterations using step size $t_p = \alpha\beta^p$, where α and β are parameters. This series is not divergent, but since we anyway are not going to make an infinite number of iterations this does not matter. The initial multipliers λ_0 was chosen slightly larger than 0. Depending on problem size we used the values stated in table 1.

The parameters are in no way optimized, but the relationship between the number of iterations and β secures that the last step size is reasonably small. The objective is to get close to the optimum in Phase I. Since one can not easily estimate the distance to optimum, we have chosen parameters which will give a fair chance of getting close to optimum. As section 7 shows, this sometimes means that Phase I is carried on to long, and other times means that it is ended prematurely.

In Phase II we execute the bundle algorithm from the best solution obtained during the subgradient optimization. In Phase III the bundle algorithm is executed from the optimum found in Phase II. From a lower bound viewpoint Phase III is not necessary, but it serves to generate all shortest paths that are needed to prove that $\underline{0}$ is included in the subdifferential. Actually Phase III gives exactly the same information as the final solution of the master problem in the Dantzig-Wolfe decomposition method, since the set P, see section 3, of all minimum cost paths are generated. Whether the solution is feasible to the VRPTW or not, can be determined by inspecting the result of the last solved QP. In some cases Phase III is not needed since Phase II provides us with a feasible and optimal solution.

Compared to the Dantzig-Wolfe decomposition master problem (LP relaxation of Set Covering), our

master problem in Phase II and III (QP) is more difficult, and each iteration is more time consuming. This is, however, not very important since most of the time anyway is spent solving the subproblem.

The results in section 7 indicates that much can be done to improve the performance of the master problem method. We are currently working on several ideas combining subgradient optimization, bundle methods as well as linear programming.

6 Branching strategy

Often a large proportion of the integrality gap is explored in the subproblem, ie. SPPTWCC. In many cases it is not even necessary to apply a branch and bound method, when the optimal Lagrangean multipliers have been found. When branching is necessary, we have chosen to branch on the time windows. If we choose to branch on the time window of customer i we separate $[a_i, b_i]$ into two time windows $[a_i^1, b_i^1]$ and $[a_i^2, b_i^2]$ by setting $a_i^1 = a_i$, $b_i^2 = b_i$ and $b_i^1 + 1 = a_i^2$. A simple option is to choose b_i^1 as the largest integer not larger than $(a_i + b_i)/2$. More sophisticated strategies are suggested in Gélinas et al. (1992). This creates two subproblems. If the primal solution to the parent node of the branch and bound tree contained one or more paths visiting customer i in as well the first part of the time window as the second, both subproblems are restrictions of the parent problem.

To select the time window to branch on we determine which customers are more than once on a path used in the optimal primal solution. The first one detected is selected as branching customer. As initial value for the Lagrangean multipliers if any branching node, we used the optimal multipliers in from the parent node. It turns out that this simple branching strategy is sufficient to solve the problems considered in section 7.

7 Numerical results

As test data we have used the so called Solomon data set (Solomon 1987), which Fisher, Jörnsten and Madsen (1992), Desrochers, Desrosiers and Solomon (1992) and Halse (1992) have tried to solve as well. We have only considered the 27 data sets with clustered customers, because these data sets are known to require very little branching. Since our work has been focused on the calculation of lower bounds and does not contain any new results on branching strategy, we find that these problems provides the best basis for evaluation of our results.

Of the 27 problems Desrochers, Desrosiers and Solomon (1992) succeeded in solving 21. Fisher, Jörnsten and Madsen (1992) solved 12. Halse (1992) solved 20, and together optimal solutions for 23 problems were known, when we started this work. Since Fisher, Jörnsten and Madsen (1992) does not report any running times we will not consider their results for further comparison. As table 2 shows, we have now found the optimal solution to all 27 problems. 25 of the 27 problems can be solved without branching. For the two remaining problems the branch and bound strategy of section 6 was executed. The optimal solution was found within a few branching nodes, see table 3.

One should note that the instances considered by Desrochers, Desrosiers and Solomon (1992), Halse (1992) and us are slightly different. Halse considers a problem with a fixed number of vehicles, calculates t_{ij} and c_{ij} with one decimal and uses ordinary rounding. The number vehicles is fixed at the number of clusters, which is also the optimal number of vehicles for all 27 instances. We have considered the problem with free number of vehicles, calculated t_{ij} and c_{ij} with one decimal and used truncation. Desrochers, Desrosiers and Solomon (1992) also solved the problem with free number of vehicles, calculates c_{ij} with one decimal and truncation, but calculates t_{ij} with no decimals and truncation (Desrosiers and Gelinas 1993). This reduces the worst case number of nodes by a factor 10, since t_{ij} takes value from a 10 times smaller set. Potentially the worst case running time is reduced by a factor 100. However, we doubt that the actual reduction in running time is very large.

The solutions obtained are in all cases identical. Note however that Desrochers, Desrosiers and

Solomon (1992) contains one tiny misprint. Halse obtains a slightly different objective value due to the rounding of c_{ij}. However, in c109.25 Desrochers, Desrosiers and Solomon (1992) must branch because the optimal solution can not be obtained in the initial node. Therefore the running time on this instance can not be compared with the results of Halse and our results.

Desrochers, Desrosiers and Solomon (1992) used a SUN SPARK 1 computer and coded their algorithm in Fortran. liti used a HP 9000-835 computer and coded the algorithm in Pascal. Our algorithms has been coded in the C programming language and executed on a HP 9000-735 computer. The HP 9000-835 is about 10% faster than the SUN SPARK 1 (SPEC Release 1.0 1991). We have been running some of the problems on a HP 9000-835 as well, and the results indicate, that the the HP 9000-735 is about 8 times faster. Table 2 shows that our running times in general are more that 8 respectively 9 times faster than the results obtained by Halse (1992) respectively Desrochers, Desrosiers and Solomon (1992). The running times of Halse (1992) are best for 7 of the most easy problems. We do not think this shows very much since I/O-times and other program overheads might dominate when the running time is only a few seconds. The running time of Desrochers, Desrosiers and Solomon (1992) is only best for one instance. However the results, compared to our results, are better for large problems than for small problems. We think this is due to inefficiencies in our master iteration method.

Table 4 shows how the cpu-time in the initial node is distributed between the three phases. In the easy cases up to about 80% of the time is spent in Phase I. This indicates that these problems might have been solved faster, if a fewer number of iterations in Phase I was chosen. For the c101.100 problem we tried a method consisting of 100 subgradient iterations (instead of 500 iterations) and the bundle method afterwards. This reduced the cpu-time from 51.44 sec. to 28.33 sec. A similar improvement is probably possible for the other easy problems. On the other hand, Phase I seems to be ended prematurely for the most difficult problems. This suggests that a more flexible stop criterion in Phase I would improve the results.

8 Conclusions

We have presented a new optimization method for the VRPTW based on a Lagrangean relaxation of the assignment constraint. The method is closely related to previously suggested successful algorithms, which decomposes the problem into a series of Shortest Path Problems with Time Windows and Capacity Constraints.

Theoretically our method produces the same lower bounds on the objective value as the methods of Desrochers, Desrosiers and Solomon (1992) and Halse (1992), but our method creates relatively easy instances of the subproblem, and therefore we have been able to solve a number of previously unsolved problems, and have obtained better running times even though many possible improvement of our algorithm exists.

References

[1] Desrochers, M. (1988), A Generalized Permanent Labelling Algorithm for the Shortest Path Problem with Time Windows, *INFOR 26, no. 3*, pp. 191-211.

[2] Desrochers, M., J. Desrosiers and M. Solomon (1992), A New Optimization Algorithm for the Vehicle Routing Problem with Time Windows, *Operations Research 40*, pp. 342-354.

[3] Desrosiers, J., M. Sauvé and F. Soumis (1988), Lagrangean Relaxation Methods for Solving the Minimum Fleet Size Multiple Travelling Salesman Problem with Time Windows, *Management Science 34, no.8*, pp. 1005-1022.

[4] Desrosiers, J., Y. Dumas, M.M. Solomon and F. Soumis (1993), *Time Constrained Routing and Scheduling*, to appear on North-Holland.

[5] Desrosiers, J. and E. Gelinas (1993), *Private Communication.*

[6] Dror, M (1993), *Note on the Complexity of the Shortest Path Models for Column Generation in VRPTW*, Working Paper, Decision Sciences, College of Business, The University of Arizona.

[7] Fisher, M.L. (1989), *Optimal Solution of Vehicle Routing Problems using Minimum K-Trees*. Research Report 89-12-13, Department of Decision Sciences, The Wharton School, University of Pennsylvania.

[8] Fisher, M.L., K.O. Jörnsten and O.B.G. Madsen (1992), *Vehicle Routing with Time Windows*, Research Report 4C/1991, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[9] Gélinas, M, M. Desrochers, J. Desrosiers and M.M. Solomon, *Vehicle Routing with Backhauling*, Research Report G-92-42, GERAD, Montreal.

[10] Geoffrion A.M. (1974), Lagrangean Relaxation and Its Uses in Linear Programming, *Mathematical Programming Study 2*, pp. 82-114.

[11] Guignard, M. and S. Kim (1987), Lagrangean Decomposition: A Model Yielding Stronger Lagrangean Bounds, *Mathematical Programming 39*, pp. 215-228.

[12] Houck Jr., D.J, J.-C. Picard, M. Queyranne and R.R. Vemuganti (1980), The Travelling Salesman Problem as a Constrained Shortest Path Problem: Theory and Computational Experience, *Opsearch 17*, pp. 93-109.

[13] Halse, K. (1992), *Modeling and Solving Complex Vehicle Routing Problems*, Ph.D. Dissertation no. 60, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[14] Jörnsten, K., M. Näsberg and P. Smeds (1985), *Variable Splitting - A new approach to some Mathematical Programming Models*, Report LITH-MAT-85-04, Department of Mathematics, Linköping Institute of Technology, Sweden.

[15] Kolen, A.W.J., A.H.G. Rinnooy Kan and H.W.J.M. Trienekens (1987), Vehicle Routing with Time Windows, *Operations Research 35*, pp. 256-273.

[16] Lemaréchal, C., J.J. Strodiot and A. Bihain (1981), On a Bundle Algorithm for Nonsmooth Optimization, *Nonlinear Programming 4*, ed.: Mangasarian, O.L. et. al., Academic Press.

[17] Nemhauser, G.L. and R.A. Wolsey (1988), *Integer and Combinatorial Optimization*, Wiley - Interscience.

[18] Olsen, B. (1988), *Routing with Time Constraints: Exact Solutions*, in Danish, M.Sc. Thesis, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, DK-2800 Lyngby, Denmark.

[19] Shor, N.Z. (1985), *Minimization Methods for Non-Differentiable Functions*, Springer Verlag, Originally published in Russian in 1979.

[20] Solomon, M (1987), Algorithms for the Vehicle Routing and Scheduling Problem with Time Window Constraints, *Operations Research 35*, 254-265.

[21] *SPEC Release 1.0 (1991) Summary of Results.*

[22] Zowe, J. (1985), Nondifferentiable Optimization, *Computational Mathematical Programming*, ed.: Schittkowski, K., NATO ASI Series, Vol. F15, Springer Verlag.

Customers	Iterations	α	β	λ_0
25	100	5.00	0.9500	30
50	200	5.00	0.9750	30
100	500	5.00	0.9875	30

Table 1: Parameter values in Phase I.

Problem	Optimal solution	Vehicles	Branch nodes	Time	Time_{VS}	Time_{DW}	Rel_{VS}	Rel_{DW}
c101.25	191.3	3	1	0.82	6.3	18.6	1.0	2.5
c101.50	362.4	5	1	6.60	46.6	67.1	0.9	1.1
c101.100	827.3	10	1	51.44	616.0	434.5	1.5	0.9
c102.25	190.3	3	1	2.11	35.0	79.7	2.1	4.2
c102.50	361.4	5	1	12.96	289.4	330.3	2.8	2.8
c102.100	827.3	10	1	151.95		1990.8		1.5
c103.25	190.3	3	1	5.95	47.4	134.7	1.0	2.5
c103.50	361.4	5	1	23.87		896.0		4.2
c103.100	826.3	10	1	241.49				
c104.25	186.9	3	1	9.51	151.9	223.9	2.0	2.6
c104.50	358.0	5	10	890.71				
c104.100	822.9	10	1	2270.58				
c105.25	191.3	3	1	0.96	3.6	25.6	0.5	3.0
c105.50	362.4	5	1	7.86	73.1	99.1	1.2	1.4
c105.100	827.3	10	1	60.55	968.1		2.0	
c106.25	191.3	3	1	0.83	3.5	20.7	0.5	2.8
c106.50	362.4	5	1	7.47	46.4	91.3	0.8	1.4
c106.100	827.3	10	1	79.27	4241.6	724.8	6.7	1.0
c107.25	191.3	3	1	0.84	3.9	31.7	0.6	4.2
c107.50	362.4	5	1	7.58	56.4	170.6	0.9	2.5
c107.100	827.3	10	1	63.36	2914.4	1010.4	5.7	1.7
c108.25	191.3	3	1	1.42	6.9	43.1	0.6	3.4
c108.50	362.4	5	1	12.43	158.2	245.6	1.6	2.2
c108.100	827.3	10	1	100.29		1613.6		1.8
c109.25	191.3	3	1	1.91	18.2	585.4	1.2	
c109.50	362.4	5	1	17.52	235.8		1.7	
c109.100	827.3	10	3	367.85				

Table 2: Results on the clustered data from the Solomon data set (Solomon 1987). Times are given in seconds. Time_{DW} denotes the time reported by Desrochers, Desrosiers and Solomon (1992). Time_{VS} denotes the time reported by Halse (1992). Rel_{VS} and Rel_{DW} denotes the relative performance of the methods, ie. $\text{Rel}_{VS} = \text{Time}_{VS}/(8\text{ Time})$ and $\text{Rel}_{DW} = \text{Time}_{DW}/(9\text{ Time})$.

Problem	Nodes	Depth tree	Time in init. node	Time after init. node	Total time
c104.50	10	5	234.82	655.89	890.71
c109.100	3	2	233.09	134.76	367.85

Table 3: The problems solved by Branch-and-Bound. Times in seconds.

	Phase I		Phase II		Phase III		Total	
Problem	Iter.	Time	Iter.	Time	Iter.	Time	Iter.	Time
c101.25	100	0.59	21	0.19	3	0.04	124	0.82
c101.50	200	4.48	58	1.91	5	0.21	263	6.60
c101.100	500	41.13	79	9.08	10	1.23	589	51.44
c102.25	100	1.54	22	0.49	3	0.08	125	2.11
c102.50	200	10.10	31	2.42	5	0.44	236	12.96
c102.100	500	111.70	123	37.14	10	3.11	633	151.95
c103.25	100	2.93	74	2.88	3	0.14	177	5.95
c103.50	200	18.63	34	4.52	5	0.73	239	23.87
c103.100	500	216.82	42	24.67			542	241.49
c104.25	100	4.68	69	4.54	4	0.29	173	9.51
c104.50	200	58.39	321	144.00	66	32.43	487	234.82
c104.100	500	502.77	1103	1767.81			1603	2270.58
c105.25	100	0.64	21	0.27	3	0.05	124	0.96
c105.50	200	5.37	58	2.24	5	0.25	263	7.86
c105.100	500	48.55	79	10.53	10	1.47	589	60.55
c106.25	100	0.58	21	0.21	3	0.04	124	0.83
c106.50	200	5.14	58	2.10	5	0.23	263	7.47
c106.100	500	63.90	79	13.56	10	1.81	589	79.27
c107.25	100	0.60	21	0.19	3	0.05	124	0.84
c107.50	200	5.16	58	2.20	5	0.22	163	7.58
c107.100	500	48.98	96	12.91	10	1.47	606	63.36
c108.25	100	0.92	31	0.44	3	0.06	134	1.42
c108.50	200	8.55	58	3.56	5	0.32	263	12.43
c108.100	500	82.55	69	15.40	10	2.34	579	100.29
c109.25	100	1.45	20	0.39	3	0.07	123	1.91
c109.50	200	12.67	49	4.37	5	0.48	254	17.52
c109.100	500	131.43	268	93.75	23	7.91	791	233.09

Table 4: Number of iterations and time in seconds in the three phases in the initial Branch-and-Bound node.

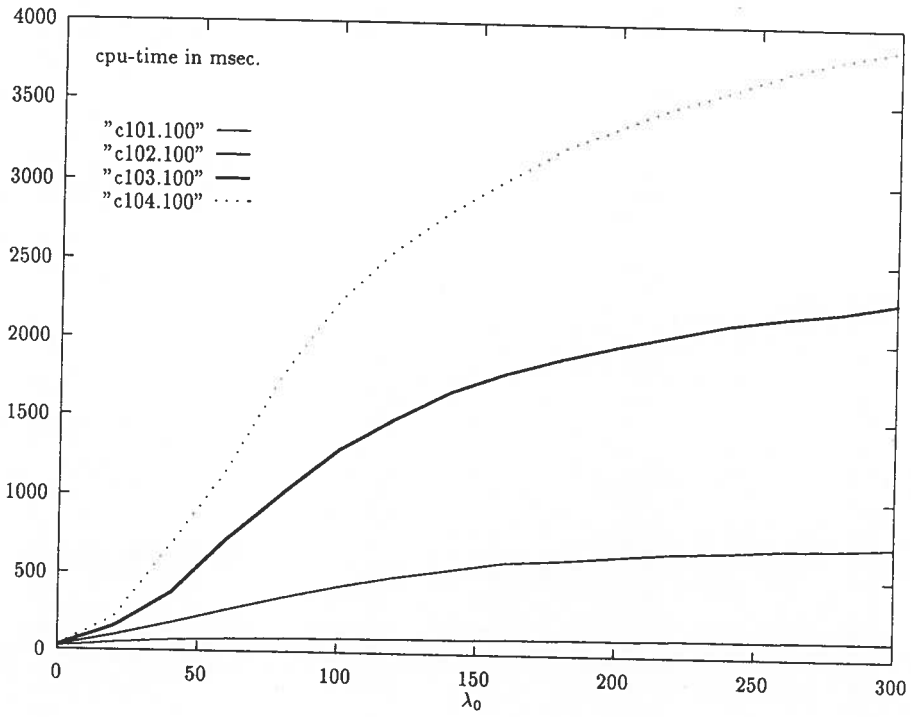


Figure 1: cpu-time in msec. for the SPPTWCC (with 2-cycle elimination) as function of λ_0 (HP 9000-735)

Lower Bounds for the VRP via TSSP+1 Decomposition

Charles Noon, Management Science Program, The University of Tennessee, USA

In the capacity-constrained vehicle routing problem (VRP), V vehicles are based at a single depot and must be used to make deliveries of a single product to N customers. Each customer must be supplied with a particular amount of product, while each vehicle has a specific limit on the total amount of product it can supply. The problem is to determine a set of capacity-feasible vehicle routes with minimum total travel cost so that each customer is serviced by exactly one vehicle.

In this paper, we present a decomposition strategy for multi-vehicle routing problems and test its effectiveness for generating lower bounds for the capacity-constrained VRP. The decomposition is based on a Lagragian relaxation of the set of constraints which require each vehicle to be visited by exactly one vehicle. The relaxed problem can be solved as a collection of independent single vehicle problems. Each single vehicle problem is modelled as a Travelling Salesman Subset-Tour Problem with an additional constraint (TSSP+1). The TSSP+1 is a variant of the classic TSP whereby the vehicle need not visit all cities (customers) and the route must be feasible with respect to the constraint on vehicle capacity. For any given set of multipliers corresponding to the relaxed constraints, we can produce a lower bound on the VRP by optimally solving V TSSP+1 problems.

The decomposition can be related to an alternative model for the VRP based on a set partitioning formulation. In the set partitioning formulation for the VRP, each column corresponds to a feasible vehicle route and each row corresponds to a customer. The LP relaxation of the set partitioning model is known to provide a tight lower bound for the VRP, but it is hampered by the fact that the number of columns grows exponentially as the number of customers increases. Except for very small VRP problems, it is impossible to provide a full statement of the set partitioning representation. We analytically show, however, that for the case of identical vehicles and the proper choice of multipliers, the TSSP+1 decomposition is capable of producing a lower bound that is superior to the bound produced by the linear programming relaxation of the set partitioning formulation for the VRP.

The analytical results establish a complementary relationship between the TSSP+1 decomposition and the set partitioning formulation for the VRP. The TSSP+1 decomposition can be used to identify entering columns within a columns generation approach for the LP relaxation of the set partitioning representation. The optimal LP dual variables of the set partitioning representation provide very good multipliers for the decomposition. As a consequence, the two models can be used in an iterative fashion to produce a tight lower bound for the VRP. In this work, we utilize a branch-and-cut algorithm for solving the TSSP+1 problems in conjunction with an LP solver for the set partitioning representation. For some problems, we achieve bounds for the VRP that are probably equivalent to the LP relaxation of the full set partitioning statement.

QUEUING OPTIMIZATION OF A SIGNALIZED INTERSECTION

BY

AVI CEDER* AND ISRAEL RESHETNIK**

* Chief Scientist, Ministry of Transport Israel and professor at the Civil Engineery Department.
Technion-Israel Institute of Technology, Haifa, Israel 32000.
(Tel. : +972-4-293054, +972-2-319444, FAX : +972-4-225716)
** Chief Engineer, Haifa Traffic Control Center
Haifa Municipality, P.O.B. 4811, Haifa, Israel 31047.
(Tel. : +972-4-356419, +972-4-726548, FAX : +972-4-965825)

QUEUING OPTIMIZATION OF A SIGNALIZED INTERSECTION

by

Avi Ceder and Israel Reshetnik

Abstract

Previous research regarding optimization of isolated intersections was carried out in the past as well as in recent years, yet, optimization techniques implemented on the signal controller's level are very primitive and unsatisfying. Three major obstacles prevent the controller from adapting programs to varying traffic conditions :

1. The necessity to deal with several constraints, such as intergreen times, minimum green times, phase sequence, green wave constraints etc.
2. Complex non-polynomial optimization problems, especially for saturated and oversaturated conditions
3. The difficulty to collect the data required to control an intersection at oversaturated conditions.

In the following article, a self-tuning system for controlling signalized intersections in undersaturated and oversaturated conditions is depicted. The system was successfully tested in one of Haifa's most congested intersections. The system tailors some well-known control techniques together with some specially developed methods and algorithms to overcome the difficulties.

The philosophy of system is based on a very well known actuated traffic control which was proved to be very efficient at unsaturated intersections and is very simple to implement. A method of constructing a basic program that meets the required constraints and can be modified by means of traffic detectors is discussed briefly in the following section. The article formulates several optimization problems that can be solved in polynomial time like Simplex algorithms (virtually in some fractions of a second) for oversaturated as well as undersaturated conditions. A method for collecting the required data at reasonable cost is presented at the end of the article.

Optimization at Stopping Points

The creation of a signalization program for an intersection generally requires consideration of constraints such as: phase sequence, minimal green times for every movement, inter-green times, internal green-waves (i.e. for adjacent intersections which are operated by one traffic

controller, etc.). The construction of an optimization problem which is intended to ensure optimal green for all directions and also to maintain all the planning constraints is a difficult and highly complex process. A possible method which permits control of a signalized intersection is control by means of a stopping point.

According to the above mentioned method a basic program, which satisfies all the constraints, is inserted into the controller. Points are established, in the basic program, at which the program stops and waits for a certain condition to be fulfilled. This condition can be a lack of demand or the presence of a queue at a particular detector, the arrival of a releasing pulse from an adjacent intersection or from a control center, or a manual release provided by a policeman. As a result of the controller's pause at the stopping point the green light is prolonged for the movements "stuck" at this point and the waiting times at all the stopping points create the overall picture of the cycle.

The basic program is constructed, not in real time, manually or by means of algorithms (a type of Gantt algorithm). The program is constructed to establish activities (green phases) in the timetable, subject to the constraints of minimum green and inter-green times and other constraints. The stopping points described above are established in the basic program. These points are generally established as close as possible to the conclusion of the phase which will become prolonged at this point. At the stopping point a transition can occur to another basic program as a result of lack of demand at a particular detector. An example of a signalization program for an intersection is described in Appendix A.

Constructing a signalization program in this way permits optimization only of the duration of waiting time at the stopping point without consideration for the constraints expressed when in outline program. The number of decision variables is less than 10 in this case and only at a few intersections with a complex geometry does the number of stopping points exceed 6.

Problems of this order of magnitude can be resolved in real time for the time ranges required by traffic light control systems when the optimization problem is not linear. The applicability of various optimization problems which are suitable for operating a signalized intersection will be examined in the next section.

Conversion Matrix of Waiting Times to Green Times

The various optimization problems are mostly expressed in respect of the green times which are allotted to the different movements at an intersection. In order to express the various

mathematical problems in respect of the waiting times at the stopping points it is necessary to find a way of presenting the green times by means of the waiting times at the stopping point.

Definition: A conversion matrix $\mathbf{W_{m*n}}$ is a matrix in which the number of rows **m** is equal to the number of movements on which the optimization is done and the number of columns **n** is the same as the number of stopping points which can be prolonged. The value of the term $\mathbf{W_{ij}}$ becomes **1** if movement **i** is prolonged as a result of waiting at stopping point **j**, otherwise the value of $\mathbf{W_{ij}}$ is **0**.

Let **g** be the green time vector of the different movements and let $\mathbf{g_m}$ be the minimal green times according to the outline program; one can then write :

$$\mathbf{g} = \mathbf{W} \cdot \mathbf{y} + \mathbf{g_m} \tag{1}$$

where **y** is the vector of waiting times at the various stopping points.

By substituting **y** for **g** one can formulate various target functions and constraint functions in accordance with the operating strategy.

Unsaturated Traffic Conditions

The allocation of times at unsaturated intersections does not constitute a special problem and there are a number of target functions which can be optimized relatively rapidly. The various target functions are mostly based on minimization of the delay caused to an isolated vehicle or to all the vehicles at an intersection. The first delay function was developed by Webster[18] in 1958 and since then was improved many times by other researchers (Miller[16], Newell[17], McNeil[13], et al). It is presently customary to use the delay function presented at HCM 1985[11] and which can be minimized by expressing the optimization problem quoted below:

$$\text{Find min } \sum_i d_i \qquad \text{s.t. } \langle \mathbf{1}, \mathbf{g} \rangle = c \tag{2}$$

where the delay $\mathbf{d_i}$ for every movement i, is defined by the formula :

$$d_i = \left\{ \frac{0.38 \cdot c \cdot (1 - g_i / c)^2}{1 - g_i / c \cdot X_i} + 173 \cdot X_i^2 \left[(X_i - 1) + \sqrt{(X_i - 1)^2 + \frac{16 \cdot X_i}{g_i \cdot S_i}} \right] \right\} \cdot Fr \tag{3}$$

and c - the cycle time (in seconds).
X - the degree of the traffic saturation (X=(c·V)/(g·S)).
V - hourly traffic volume (vehicles/hour).
S - saturation flow of the traffic (vehicles/hour).
Fr - correction coefficient due to non-random arrival (in a green or red wave).

Minimization of the target function proceeds quickly over the only constraint when the green times **g** are represented by the waiting times **y**, as described in Equation 1. In this case the optimization problem will appear as follows :

$$\text{Find min } \sum_i d_i(\mathbf{y}) \qquad (4)$$
$$\text{s.t. } \langle \mathbf{1}, \mathbf{W}\cdot\mathbf{y} + \mathbf{g_m} \rangle = c$$

By carrying out a linear search over the range of the logical cycle times one can deduce the duration **c** of the optimal cycle although this is not critical since, at intersections operated by detectors the cycle time can be modified in real time by the intersection detector.

Another method which is simpler and as logical is based on the desire to remove all the movements at the intersection during the course of the cycle. The green time allotted by this method suffices for the discharge of the volume accumulating during the cycle. In this case the green allocation problem becomes a linear programming problem as follows :

$$\text{Find min } c = \sum_i g_i \qquad (5)$$
$$\text{s.t. } 1)\ \frac{V_i \cdot c}{S_i \cdot g_i} \le 1 \ \Rightarrow\ S_i \cdot g_i - V_i \cdot \sum_i g_i \ge 0 \quad \forall i = 1..m$$
$$2)\ \mathbf{g} \ge \mathbf{g_m}$$

As the traffic volumes vary from one cycle to another it is usual to slightly increase the volumes arriving (for example to calculate the 70th or the 80th percentile of the volume in accordance with the Normal or Poisson distribution formula), when the local detectors are supposed to shorten the green times for the actual traffic volumes.

Here also the optimization problem can be simplified by substituting waiting times at the stopping point for the green times (according to Equation 1). The new optimization problem becomes :

$$\text{Find min } \sum_i y_i \qquad (6)$$

$$\text{s.t. } 1)\ S_i \cdot \left\langle \mathbf{W}^i, \mathbf{y} \right\rangle - V_i \cdot (\mathbf{1Wy} + c_m) \geq 0 \qquad \forall\, i = 1..m$$

$$2)\ \mathbf{y} \geq \mathbf{0}$$

This problem is also a linear programming problem and the complexity of the green time allocation problem under unsaturated conditions of is therefore a polynomial problem at the stopping point.

Traffic Control for Congested Conditions

The operation of signalized intersections under conditions of congestion cannot be carried out based on counting the number of vehicles arriving at the intersection. The reasons are as follows :

- The accepted delay models (Webster, HMC) for steady state traffic flow are not quite suitable for describing delay during congestion.
- The use of the method for clearing an intersection is not practicable, since during congestion the intersection cannot be cleared at the end of the cycle.
- The demand of traffic aiming at passing the intersection cannot be assessed since the detector can gauge only vehicles which passed over it on one hand while on the other hand this quantity becomes smaller as congestion increases (because of the reduction in speed).

These limitations have led to the development of methods, based on data from a queue detector, for the purpose of allocating green-times to the traffic lights. Models based on an optimization of queue lengths have been suggested in the past by a number of researchers (Michalopoulos & Stephanopoulos[15], Church & Revelle[6], et al), although these methods required prior knowledge at the time on the mode of development of the queue and were not appropriate to the method of control which adapts itself in real time to traffic volumes.

The first adaptive method was proposed by Gartner[9] in 1983. The method is based on the measurement of the delay by integrating the length of the queue over a certain interval of time when the decision variables are the points in time when the green light has to change. The foregoing objective function is described in general terms as follows :

$$\text{Find min } \sum_i \Phi_i(x_1, x_2 \ldots x_p) \qquad (7)$$

$$\Phi_i(x_1, x_2 \ldots x_p) = \int_0^T \left(Q_0 + A(t) - D(x,t)\right)$$

where :

$x_1, x_2, \ldots, x_p$ - decision points when green light has to change.

Q_o - The queue at the start of the time interval T being examined.

A(t) - The vehicles arrival function.

D(x,t) - The vehicles discharge function.

However these methods were based on programming with integers or on dynamic programming and were unsuitable for optimization which has to be executed in real time. This paper suggests a number of alternative methods for controlling a signalized intersection in a congested condition.

Queuing Optimization Model

The principle of the proposed method is that the optimization has to be done on a restricted number of stopping points and that the objective function has to be calculated in real time in accordance with data arriving once per cycle and not to be based on data for a longer interval of time (apart from slight corrections in accordance with past observations). Since the accumulating queue cannot be discharged in all traffic directions during a single cycle one of the following strategies can be employed :

- The increment to the accumulating queue of vehicles during a single cycle in all the critical movements will be minimal.
- The maximum accumulating queue in all the critical movements will be minimal.

If one does not take into account the number of lanes or the differences in the saturation flow between the various movements, the use of both strategies will produce the same outcome which is a queue of more or less the same length in all the critical directions. Otherwise priority will be given to movements in which the total saturation flow is higher.

The following is the mathematical expression of the optimization problem described above :

$$\text{Find min } \sum_i \Delta Q_i$$

$$\begin{aligned}
\text{s.t.} \quad & 1) \quad Q_i^0 + \Delta Q_i \le Q_i^{max} && \forall\, i = 1 \ldots m \\
& 2) \quad Q_i^0 + \Delta Q_i \ge 0 && \forall\, i = 1 \ldots m \\
& 3) \quad \sum_i y_i + c_{min} \le c_{max} \\
& 4) \quad y_i \ge 0 && \forall\, i = 1 \ldots m
\end{aligned} \qquad (8)$$

where :

ΔQ_i - increment to vehicle queue in movement i at the end of the cycle.
Q^0_i - vehicle queue in movement i at the end of the green light.
y_i - duration of prolongation at the stopping point.
c_{min} - duration of the minimal cycle according to the outline program.
c_{max} - the maximal desired cycle duration.

The queue increment to the movement at the traffic light can be expressed by means of the green times which all the movements get at the traffic light or evenly by prolongation of the stopping points.

$$\Delta Q_i = \alpha_i \cdot \left(\langle \mathbf{1y} \rangle + c_{min} \right) - s_i \cdot \langle \mathbf{W'y} \rangle \qquad (9)$$

where :

α_j - average arrival rate of vehicles at the intersection during the cycle.
s_j - saturation flow of traffic.

The first constraint limits the maximal length of the queue which a particular movement is allowed to accumulate. Without the said bound a trivial solution will be obtained according to which the movement with the largest saturation flow will receive most of the green time (see also the results of research by P.G.Michalopoulos & G.Stephanopoulos). The second constraint is to prevent the allocation of green time to a movement in which the entire queue has been discharged and to avoid wasting time. The third constraint restricts the length of the cycle at the intersection and the last constraint dictates only constructive solutions to the problem.

Notwithstanding that the problem appears to be a linear programming problem, it is not possible to use the objective function described in Equation 7 as it stands because the problem is liable to converge into very short cycles due to the queue increment getting smaller as the measuring time shortens, i.e. the cycle duration. In order to obtain a reliable index the objective function has to be normalized in relation to the length of the cycle at the intersection. The final optimization problem thus acquires the following form :

$$\text{Find min } \frac{\sum_i \Delta Q_i}{\sum_i y_i + c_{min}} \Leftrightarrow \text{Find max } \frac{\sum_i s_i \langle \mathbf{W}^l \mathbf{y} \rangle}{\sum_i y_i + c_{min}}$$

$$\begin{aligned}
\text{s.t.} \quad & 1) \quad Q_i^0 + \Delta Q_i \le Q_i^{max} && \forall\, i = 1 \ldots m \\
& 2) \quad Q_i^0 + \Delta Q_i \ge 0 && \forall\, i = 1 \ldots m \\
& 3) \quad \sum_i y_i + c_{min} \le c_{max} \\
& 4) \quad y_i \ge 0 && \forall\, i = 1 \ldots m
\end{aligned} \tag{10}$$

One can see that the problem of minimizing a queue is commensurate with the problem of maximizing the normalized saturation flow in relation to the length of the cycle.

In spite of the defined optimization problem not being a classic linear programming problem one can easily see that the level sets of the objective function are hyperplanes having the form :

$$\sum_i s_i \langle \mathbf{W}^l \mathbf{y} \rangle = Const. \cdot (\sum_i y_i + c_{min}) \Rightarrow \sum_i (s_i \langle \mathbf{W}^l \mathbf{y} \rangle - Const. \cdot y_i) = Const. \tag{11}$$

The significance of this is that the solution to the problem is obtained in one of the apexes of the linear constraint polyhedron. In the scope of our research a computer program was developed which implements an algorithm to search the polyhedron apexes, similarly to the Simplex algorithm, in which the transit direction from apex to apex is determined by the gradient size in the direction of the apex to which the transit is made.

Assessing Queuing Length

The use of the optimization problem which is based on the lengths of queues requires the gathering of data on these queue lengths. It turns out that the practicable gathering of the data is not an altogether simple problem. Covering the approaches to the intersection over a long distance with a queue detector every few meters will be make the system unreasonably expensive and will not be implemented.

It was decided, in order to cope with the problem of data gathering, to use only two detectors: a detector over which there is a stop-and-go flow (for example the detector closest to the intersection) and a detector over which there is a free flow (the detector farthest from the intersection). In practice the "loaded" detector counts the number of vehicles leaving the system while the "free" detector counts those arriving at the traffic light. The difference

between the readings would yield the length of the queue between the two detectors if the initial length of the queue was known at the instant the traffic light was operated. It is also necessary to provide a self-calibrating controller for the system to correct the errors which accumulate in the system.

For example, in a real-life test case explained in the next section, two possible calibration points were established :

1. If the detector near the intersection identifies a suitable gap between vehicles during the green light it means that the system has emptied and the queue behind the detector is zero.
2. If the distant detector identifies a stationary queue during the red light it means that the length of the queue is equal to the distance of the detector from the intersection.

Under normal operation of the traffic light, it is possible, but not always, to ascertain the above mentioned calibration points during one cycle. At volumes less than the capacity the first calibration point will mostly be revealed. At volumes exceeding the capacity a second calibration point will mostly be apparent. A problem arises in cases where volumes are close to capacity. If, in the course of a number of cycles, no calibration point transpires it becomes necessary to run a special calibration cycle in which one of the calibration points appears in any possible instance.

Let **L** be the measured distance in the number of vehicles between the two detectors (the loaded one and the free one). If, at the beginning of the green light for the movement, the free detector is still free then one has to actuate green for the movement for a time $g \geq L/s$ where **s** is the saturation flow measured at the detector for the preceding cycles. If, during the green light, the first calibration point was not apparent this shows that :

1. The queue between the two detectors is $Q \leq L - N_a^r$ because the queue did not reach the distant detector during the red light. Here N_a^r is the number of vehicles entering the section between the two detectors during the red light.
2. The queue between the two detectors is $Q \geq N_s - N_a$ because the first calibration point was not found. Here N_a is the number of vehicles entering the section between the two detectors during the cycle and N_s is the number of those leaving the section during the cycle.

The upper and lower bounds for the length of the queue in the section between the two detectors have thus been obtained.

In order to ensure the definite detection of the queue, the allocation of green and red times for each of the movements must ensure the following condition :

$$N_s - N_a \geq L - N_a^r \quad \Rightarrow \quad s \cdot g - \alpha \cdot c \geq L - \alpha \cdot (c - g) \quad \Rightarrow \quad g \geq \frac{L}{s - \alpha} \tag{12}$$

One cannot use the previously unknown rate of arrival α in order to determine the location of the detectors although one may assume that in unsaturation conditions the condition $\frac{\alpha \cdot c}{s \cdot g} \leq 1$ i.e. $\alpha \leq s \cdot \left(\frac{g}{c}\right)$ is fulfilled (otherwise the queue would rapidly diverge beyond the range of the free detector). By using this assumption in Equation 12, it is possible to derive the equation :

$$g_i^2 - g_i \cdot c + L_i \cdot c / s_i \geq 0 \qquad \text{for all sampled movements, i=1,...,n} \qquad (13)$$

One can deduce, from an interactive solution of the equations for all the movements being examined at an intersection, the practicable and logical connection between the distances between the detectors and the allocation of the green times in the calibration cycle.

An Implementation Example

Apendix A presents a test case system was built and implemented in order to examine the theoretical and practical problems in the proposed method of control at the Check Post intersection in Haifa.

By solving Equation 13 a calibration cycle time of 140 seconds was selected with the following green times and inter-detector distances :

Movement No.	Green Duration	Inter-Detector Distance	Number of Loop Detectors in the Lane
1	60 seconds	120 meters	3
4	25 seconds	75 meters	3
5	40 seconds	100 meters	2

A plan of the intersection with the location of the detectors is given on Page A-1 of Appendix A.

Pages A-2 in the Appendix describe the phases and the flow diagram according to which the traffic light controller operates. Page A-3 presents the intergreen time table. Pages A-4 and A-5 show an example of some outline programs with the stopping points at which the optimization is done, for the branches which were presented in the flow diagram. Page A-6 shows an example of a possible extended operating program.

At certain points in the flow diagram the optimization program can be read. For test purposes this program run on an external 386 PC computer. The maximal green times for each phase have been directly written to the memory of the traffic light controller via the communications port. The data for volumes and queues were processed by the traffic light controller and transmitted to the external computer. It was further determined that if the intersection does not reach the calibration points within three cycles a self-calibration cycle will be carried. The description of the complete Queue Control System (QCS) is presented in Figure 1.

Running the system during peak hours showed a drastic improvement in the length of the queues in the directions of conjested traffic as opposed to only a slight worsening in the other directions. The overall state of the traffic at the intersection improved considerably and the duration of the rush hour was shortened. Figures 2 and 3 illustrate the results obtained.

REFERENCES

1) R.AKcelik, THE HIGHWAY CAPACITY MANUAL DELAY FORMULA FOR SIGNALIZED INTERSECTIONS, ITE Journal Vol. 58, No 3, pp.23-27, 1988.

2) R.E. Allsop, A DIAGRAM SHOWING PRACTICABLE SETTINGS FOR A FIXED TIME TRAFFIC AND THEIR EFFECT ON DELAY, Transp. Research, Vol.5, no 1, pp.59-65, 1971.

3) K.L. Bang, OPTIMAL CONTROL OF ISOLATED TRAFFIC SIGNALS, Traffic engineering and control, vol.17, pp.288-292, 1976.

4) M.S. Bazaraa and C.M Shetty, NONLINEAR PROGRAMING, Jhon Wiley & Sons, 1979.

5) I. Catling, a TIME DEPENDENT APPROACH TO JUNCTION DELAYS, Traffic Engineering and Control, nov. 1977.

6) R. Church and C. Revelle, MODELING AN OVERSATURATED INTERSECTION, Transp. Research, vol.12, No. 3, pp. 185-189, 1978.

7) G.C.D'ans and D.C. Gazis, OPTIMAL CONTROL OF OVERSATURATED STORE-AND-FORWARD TRANSPORTATION NETWORKS. Transp. Science, vol.10, no.1, pp.1-19, 1976.

8) N.H. Gartner and J.D.C. Little and H.Gabbay, OPTIMIZATION OF TRAFFIC SIGNAL SETTINGS BY MIXED-INTEGER LINEAR PROGRAMING. PART - II : THE NETWORK SYNCHRONIZATION PROBLEM. Transp. Science, vol.9, No.4, pp.344-363, 1975.

9) N.H. Gartner, OPAC : A DEMAND-RESPONSIVE STRATEGY FOR TRAFFIC SIGNAL CONTROL, Transp. Research Record' No.906, pp.75-81, 1983.

10) HIGHWAY CAPACITY MANUAL, Transp. Research Board, Special Report, No. 209, 1985.

11) A.J.Mayne, TRAFFIC SIGNALS FOR CONGESTED TRAFFIC, Proc., Engineering Foundation Conference on Research Directions in Computer Control of Urban Traffic Systems, Calif., Feb. 1979.

12) D.R.McNiel, A SOLUTION TO THE FIXED-CYCLE TRAFFIC LIGHT PROBLEM FOR COMPOUND POISSON ARRIVALS, J. Appi. Prob., No.5, pp.624-635, 1968.

13) P.G.Michalopoulos and G.Stephanopoulos, OVERSATURATED SIGNAL SYSTEMS WITH QUEUE LENGTH CONSTRAINTS - I : SINGLE INTERSECTION, Transp. Research, Vol.11, No.6, pp.413-421, 1977.

14) P.G. Michalopoulos and G.Stephanopoulos, OVERSATURATED SIGNAL SYSTEMS WITH QUEUE LENGTH CONSTRAINTS - II : SYSTEMS OF INTERSECTIONS, Transp. Research, Vol.11, No.6, pp.423-428, 1977.

15) A.J.Miller, SETTINGS FOR FIXED-CYCLE TRAFFIC SIGNALS, Opnl. Res. Quart. Vol.14, pp.373-386, 1963.

16) G.F.Newell, APPROXIMATION METHODS FOR QUEUES WITH APPLICATION TO THE FIXED-CYCLE TRAFFIC LIGHT, SIAM Rev., No.7, pp.223-239, 1965.

17) F.V.Webster, TRAFFIC SIGNAL SETTINGS, Road Research Technical Paper, No.39, Her Majesty's Stationery Office, London, 1958.

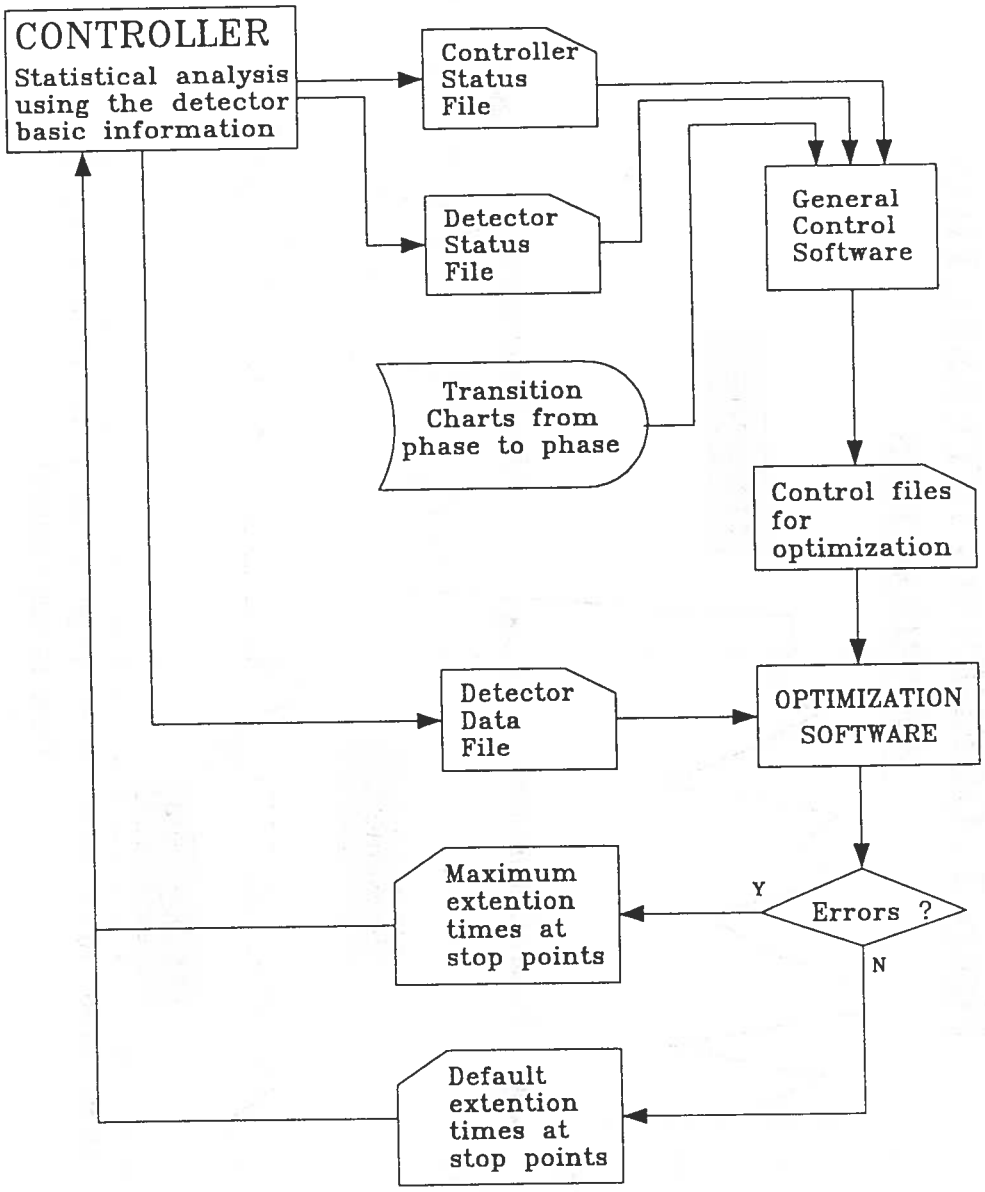


Figure 1 : The main components of the QUEUE CONTROL SYSTEM

Figure 2 : QUEUES ON INTERSECTION APPROACHES

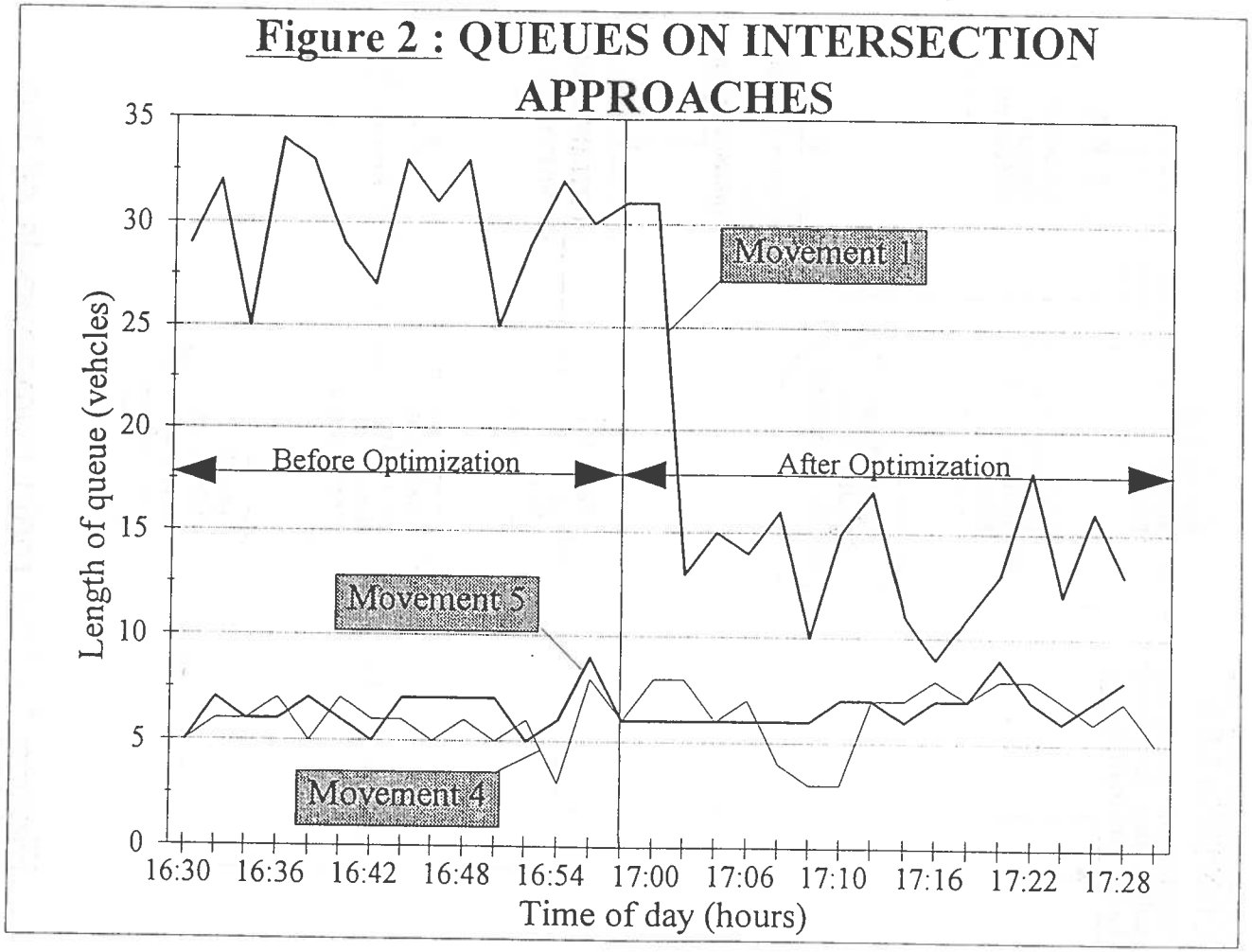
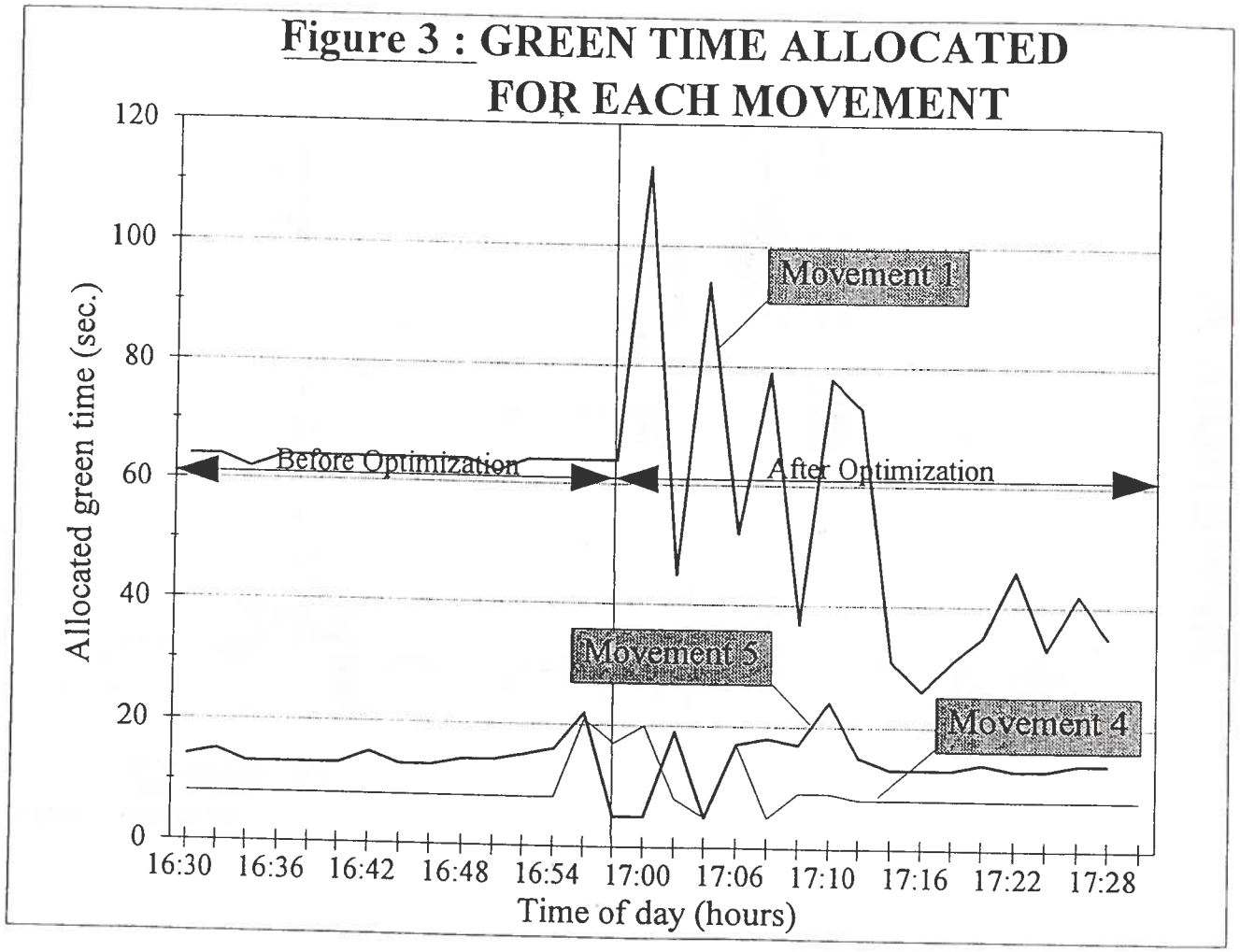
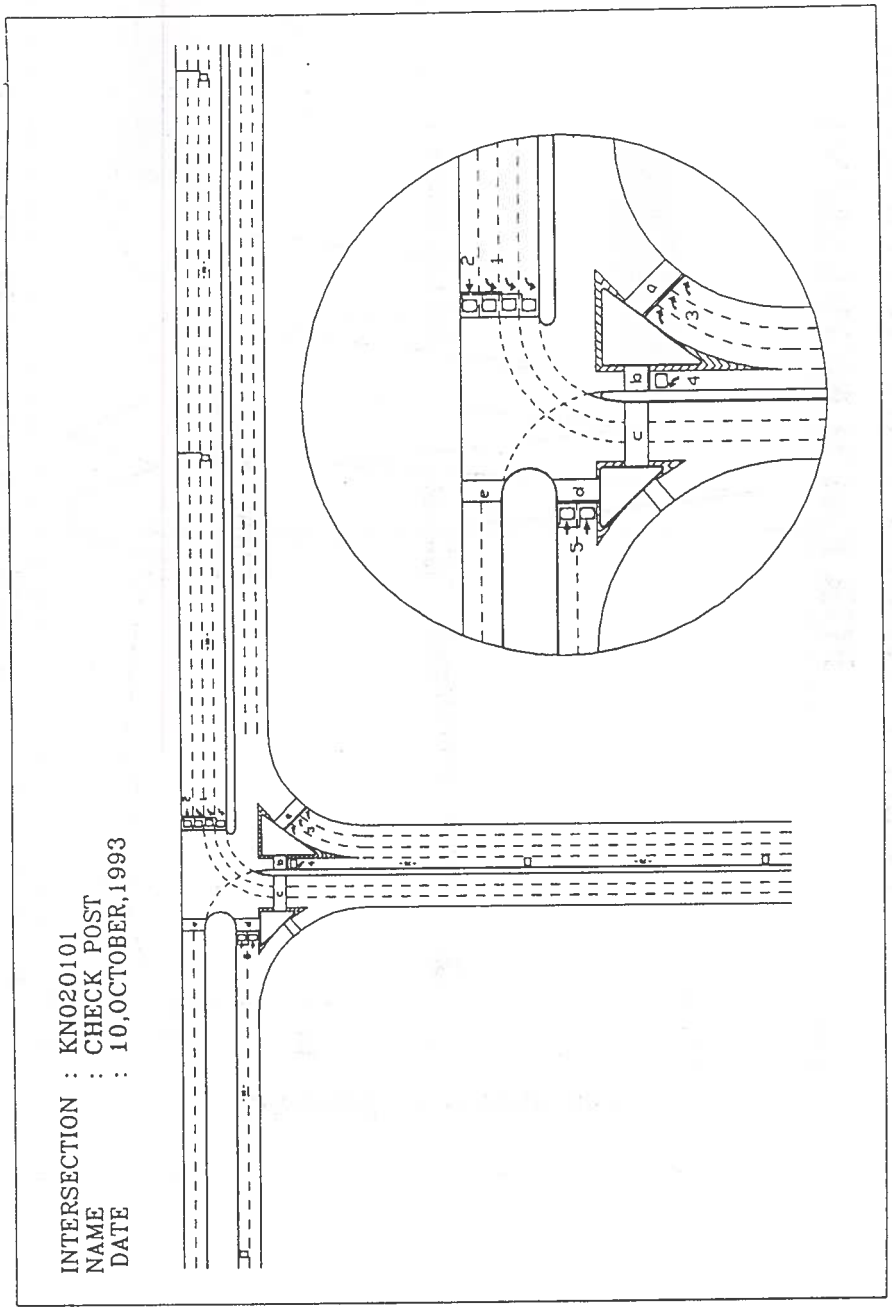
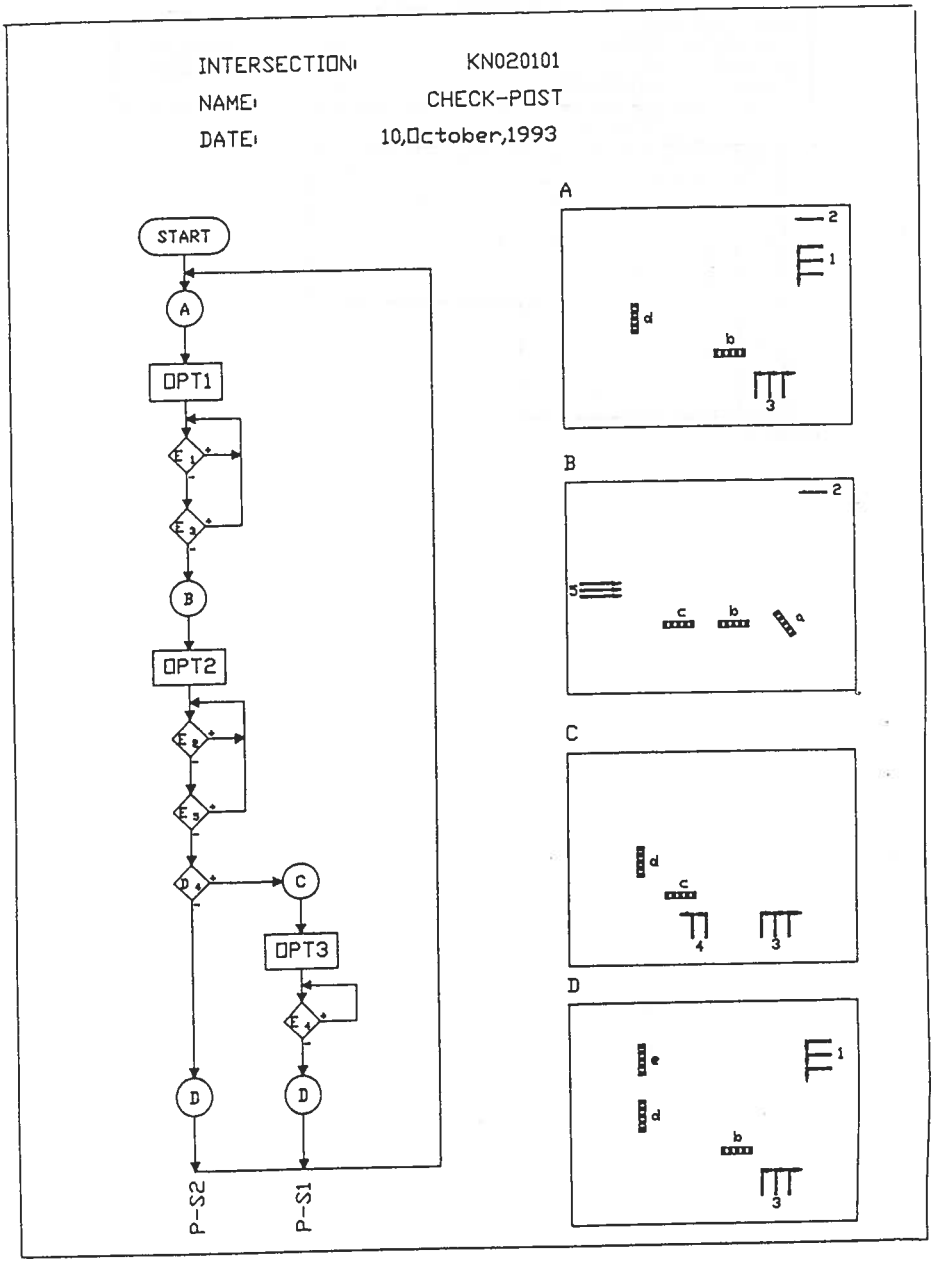


Figure 3 : GREEN TIME ALLOCATED FOR EACH MOVEMENT



APPENDIX A


INTERSECTION : KN020101
NAME : CHECK POST
DATE : 10,OCTOBER,1993


INTERSECTION: KN020101
NAME: CHECK-POST
DATE: 10,October,1993
START
A
OPT1
B
OPT2
C
OPT3
D
D
P-S2
P-S1
A
B
C
D

Name: CHECK POST
Intersection: KN020101 Date : 08/04/93
GBR group : 02 Controller: MTC 3000
Program : INTERGREENS Cycle:

Entering

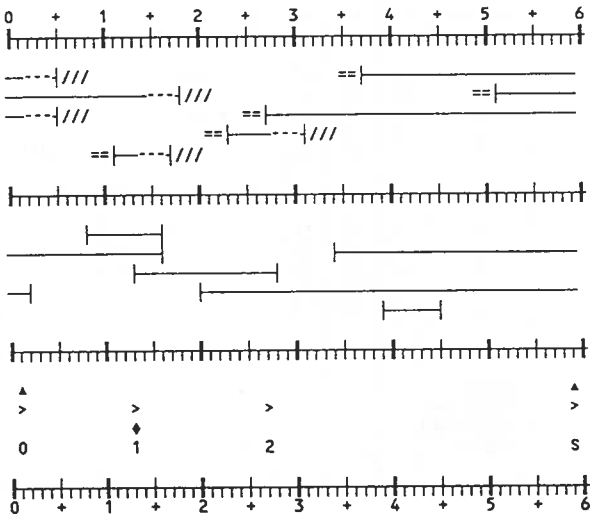
mvm	1	2	3	4	5	a	b	c	d	e
1	--	--	--	6	6	--	--	8	--	--
2	--	--	--	5	--	--	--	--	--	7
3	--	--	--	--	4	3	--	--	--	--
4	5	7	--	--	4	--	3	--	--	8
5	6	--	8	6	--	--	--	--	3	--
a	--	--	11	--	--					
b	--	--	--	7	--					
c	9	--	--	--	--					
d	--	--	--	--	9					
e	--	6	--	6	--					

Clearing

Name: CHECK POST
Intersection: KN020101
GBR group : 02
Program : P-S1
Phases : A-B-C-D

Date : 08/04/93
Controller: MTC 3000
Cycle: 60"

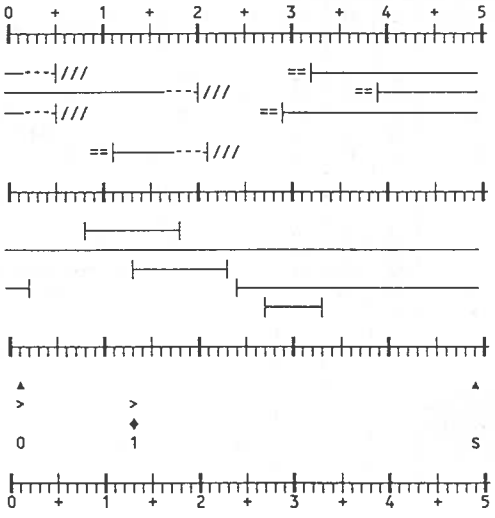
Movement	Ts	Tf	Tt
1	37	5	28
2	51	18	27
3	27	5	38
4	23	31	8
5	11	17	6
a	8	16	8
b	34	16	42
c	13	28	15
d	20	2	42
e	39	45	6
Pulse Extension Detection Number			



Name: CHECK POST
Intersection: KN020101
GBR group : 02
Program : P-S2
Phases : A-B-C

Date : 08/04/93
Controller: MTC 3000
Cycle: 50"

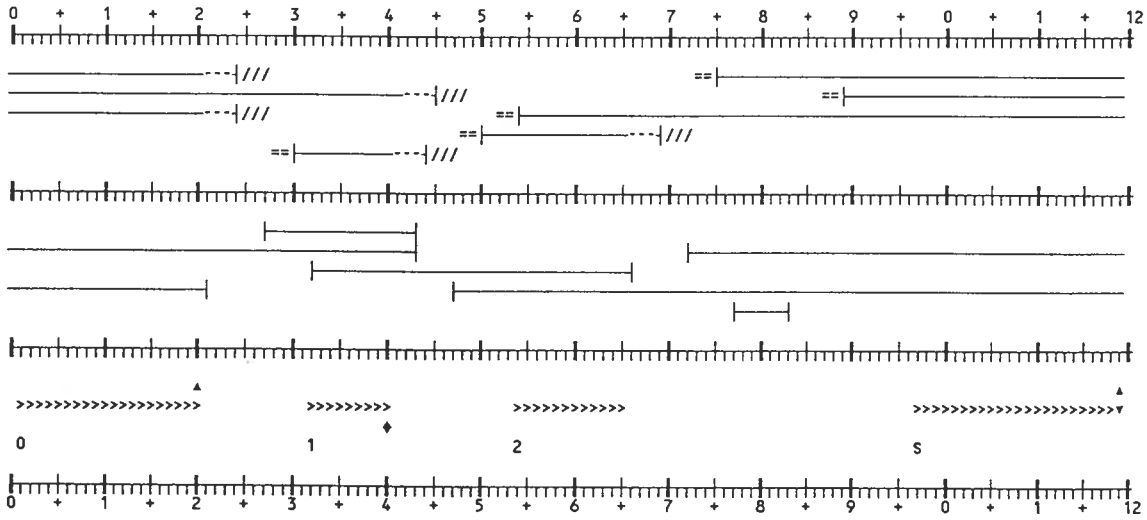
Movement	Ts	Tf	Tt
1	32	5	23
2	39	20	31
3	29	5	26
4	0	0	0
5	11	21	10
a	8	18	10
b	0	50	50
c	13	23	10
d	24	2	28
e	27	33	6
Pulse Extension Detection Number			



Name: CHECK POST
Intersection: KN020101
GBR group : 02
Program : PR-1
Phases : A-B-C-D

Date : 08/04/93
Controller: MTC 3000
Cycle: 120"

Movement	Ts	Tf	Tt
1	75	24	69
2	89	45	76
3	54	24	90
4	50	69	19
5	30	44	14
a	27	43	16
b	72	43	91
c	32	66	34
d	47	21	94
e	77	83	6
Pulse Extension Detection Number			



REALBAND: AN APPROACH FOR REAL TIME COORDINATION OF TRAFFIC FLOWS

P. Dell'Olmo
I.A.S.I. - C.N.R. - Viale Manzoni 30 - 00185 Roma
Pitu B. Mirchandani
S.I.E. Dpt. University of Arizona, Tucson - AZ 85721

Abstract

The paper deals with the coordination of network intersection phasing. Presently, network coordination is done using off-line methods such as MAXBAND/PASSER II for arterials and TRANSYT for networks. On-line approaches like SCOOT adapt these methods by constantly inputting average loads computed from detector data. The method proposed here consideres available real-time information on computing signal timings. It first identifies platoons in the network- the sizes and the speeds. These are identified by fusing and filtering the traffic data obtained, from various sources, in the last few minutes. Then, using an appropriate traffic model, these platoons are propagated through the network, for a given time horizon. The signals are set so that the identified platoons are provided appropriate green times to optimize the performance criterion. It is obvious that two platoons demanding conflicting movements may arrive at an intersection at the same time. In that case, then either one or the other will be giving priority on the green time, or it may be necessary to split one of these platoons, to maximize a given measure of performance. The paper discusses both the models for fast propagation of platoons in the network and the resolution of these conflicts.

Keywords: real time simulation, network coordination, traffic control

1. Introduction

Since the early seventies, several cities in Australia, Europe, and USA have implemented traffic control systems where a network of intersections are centrally controlled by a mainframe or a mini-computer. By and large, most of these systems have (1) magnetic loop detectors near the intersection, and (2) a micro-processor based local controller at each intersection where some traffic control parameters are either input manually or downloaded through communication links composed of telephone lines, twisted pair cable, cable television lines, and the like.

Current implementations, even those which are state of the art, have some drawbacks which are due to the inherent constraints imposed on the system design by available technologies, and now are gradually being eliminated with the rapid advances in detector, communication, and computer technologies, with the impetus being provided by the Intelligent Vehicle-Highway Systems (IVHS) program in the USA, and similar high-technology based programs in Europe and Japan.

These modern technologies, combined with methodological advances in control theory and operations research, have been used to develop a control architecture for real time traffic management, which has recently been proposed in the RHODES research project, a prototype hierarchical control system being developed for the city of Tucson, Arizona, (Head, Mirchandani and Sheppard (1992)). Such a system allows for a modular implementation of many subsystems within the control structure, namely: network loading,

capacity allocation, network coordination, intersection scheduling and intersection dispatching.
The papers deals with the coordination of network intersection phasing within the real time hierarchical traffic control system. Coordination of intersection phasing poses the most challenging sub problem within the control system.

2. Network Coordination

Prototypical off-line approaches to network coordination are TRANSYT, MAXBAND, and PASSER II, the latter two being predominately for arterial coordination. The basic ingredients within these methods are (1) a traffic flow model and (2) an algorithm for optimizing a specified performance criterion (this criterion could be a weighted sum of several performance indices). For example, in TRANSYT, vehicles are "loaded" on to network at given *origins* and are propagated through the network in accordance with a traffic flow model. Traffic controls affect the movement of these vehicles, and numerical optimization (gradient search) is performed to find controls that optimize the specified performance criterion. In MAXBAND and PASSER II, vehicles are loaded on an arterial and traffic signals on that arterial are coordinated to optimize a performance criterion, which often relates to the number of stops. Since these are off-line methods, assumptions on the traffic loads are based on historical average volumes and these are uniformly loaded on to the arterial. This results in an assumption of platoons of uniform size and identical speeds.

One may use TRANSYT in an off-line fashion and compute signal settings every few minutes and download those settings to the field. In a way, this exactly what SCOOT does. However, the current versions of SCOOT have the disadvantage that arterials within the network may no have sufficient platoon progression. Band-aid approaches have been suggested to enhance SCOOT to consider this; however, to our knowledge, it has not been implemented. On the other hand, TRANSYT has been modified to include progression opportunities [Hadi and Wallace, 1992] but has not been implemented for real-time applications. It is not obvious that this approach is amenable for real time applications due to its excessive computational requirements.
Furthermore, TRANSYT, and for that matter SCOOT, do not explicitly consider the current traffic slows (i.e.; platoons and their speeds) but rather takes the current data and assumes a uniform flow of the current volumes.

3. The REALBAND approach

The method being proposed here explicitly considers available real-time information on computing signal timings. It first identifies platoons in the network - the sizes and their speeds. These are identified by the fusing and filtering the traffic data obtained, from various sources, in the last few minutes. Then, using an approximate traffic model, these platoons are propagated through the network, for a given time horizon, say, T. The signals are set so that the identified platoons are provided appropriate green times to optimize the performance criterion. It is obvious that two platoons demanding conflicting movements may arrive at an intersection at the same time. In that case, then either one or the other will be given priority on the green time, or it may be necessary to split one of the platoons, to maximize the given measure of performance. Resolving such conflicts in the heart of the optimization process within of the proposed system.

Consider the time-distance diagram on a single arterial as shown in Figure 1. The goal of arterial progression algorithms, such as MAXBAND and PASSER II, is to set the signal timings such that number of vehicles that can traverse the arterial in either direction without stopping (other similar criteria may be incorporated) is maximized. The figure shows these

bands of green times. Note that following drawbacks: it is assumed (1) that platoons are or equal size, (2) that platoons travel at the same constant speed, and (3) that streets intersecting with this arterial have small traffic volumes and therefore movements in those directions have low priorities (that is the reason why MAXBAND and PASSER II coordinate signals on an arterial rather than on a network).

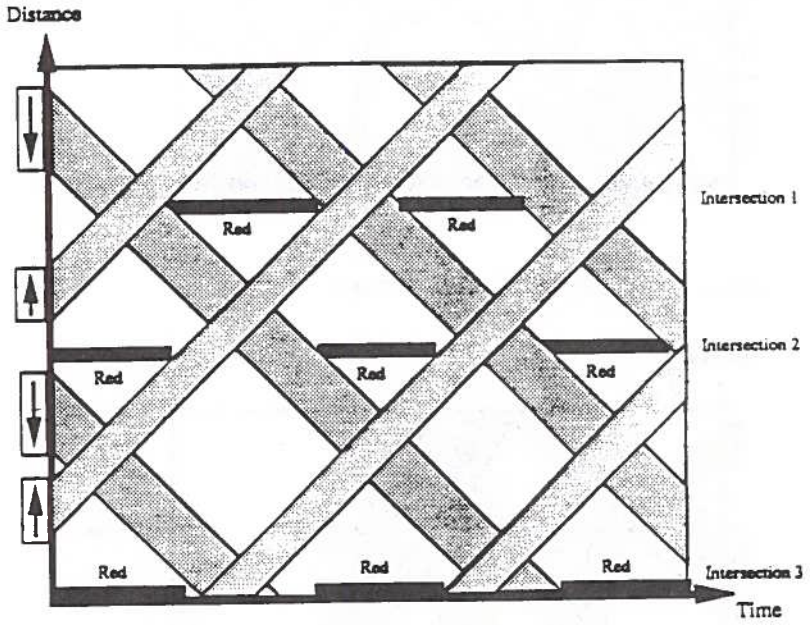


Figure 1: The MAXBAND Concept

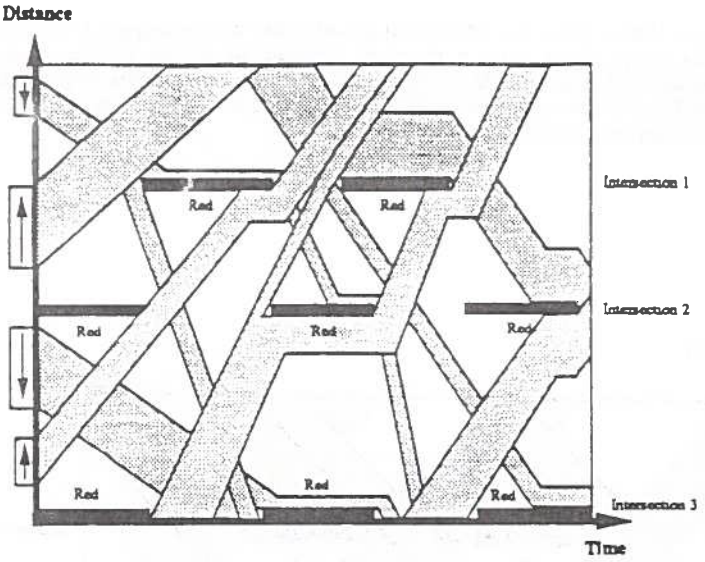


Figure 2: Actual MAXBAND Performance

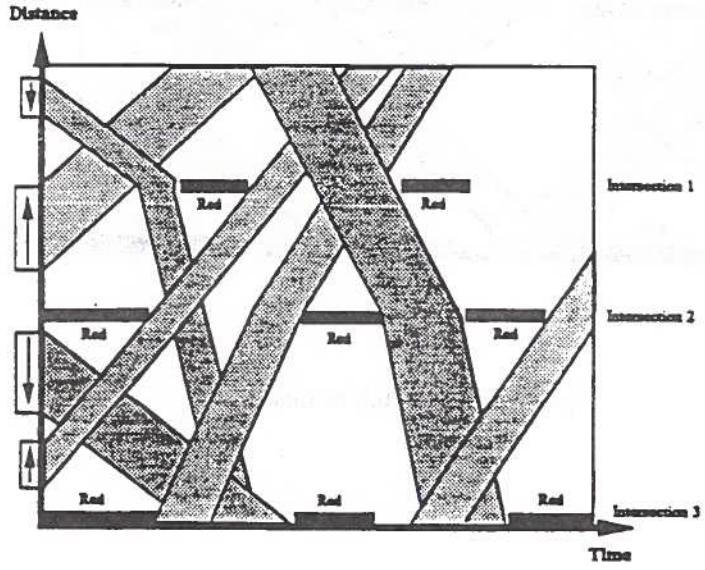


Figure 3: The REALBAND Concept for Single Arterial

Now consider the time-distance diagram as shown in Figure 2. Here we have platoons of different sizes and different speeds. The green times required for these platoons is different than those required for the uniform case shown in Figure 1, and therefore the smooth anticipated progression is disrupted. By slightly adjusting the red times, we now reinstate the green bands for the given platoons with their own sizes and speeds, as shown in Figure 3. In this manner we can also include platoons dispersion and compression but for convenience we haven't shown this in the figure. In fact, in the illustrations to follow, we have not shown turning vehicles which could grow or decrease platoons sizes and have purposefully exaggerated the speed differences in the figures. This is so that proposed concept for network flow control is visualized easier. The platoon flow prediction algorithm to be developed in this effort will take into account these underlined characteristics.

The platoons, if any, on intersecting arterials have still not been considered above. If the intersecting platoons fitted exactly within the red times shown here, then we do not need to resolve green-time demand of conflicting movements. On the other hand, if flows at an intersection produced a concurrent green-time demand for conflicting movements, then the conflict has to be resolved in determining to which movement the green time must be allocated.

Figure 4 now shows platoons on two other perpendicular arterials, at intersections 2 and 3. The approach being proposed here - for brevity we shall referto this as *REALBAND* - makes a forward pass in time. When a conflict arises, a decision node in a tree is formed; the types of decisions at this node are (1) give green time to platoon A, (2) give green time to platoon B,

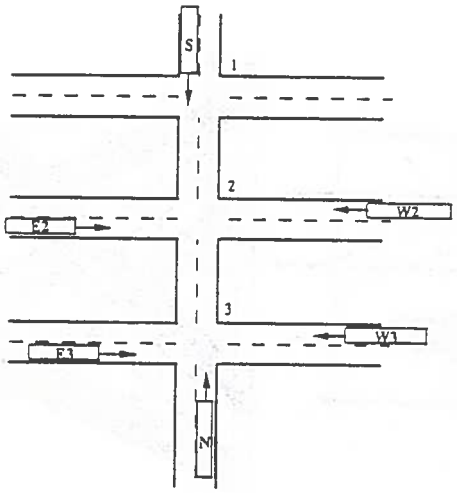


Figure 4: REALBAND Example Network

(3) split platoon A (or platoon B, since only one or the other platoon needs to be split)? Each branch of the tree is propagated over time, keeping track of the total performance up to the decision node plus the performance on the link associated with the potential decision. Note that we use an implicit approximation on the additive nature of the performance measure as we propagate from node to node in the decision tree.

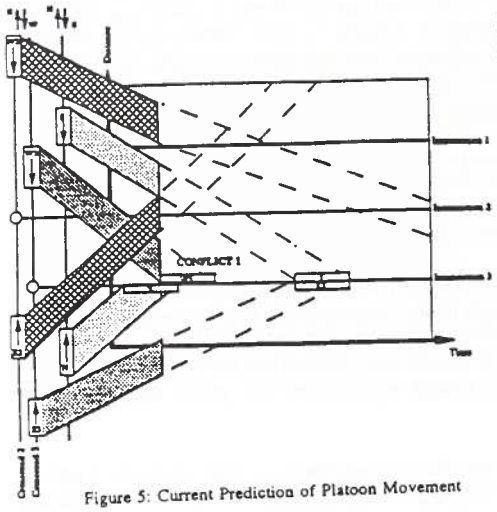


Figure 5: Current Prediction of Platoon Movement

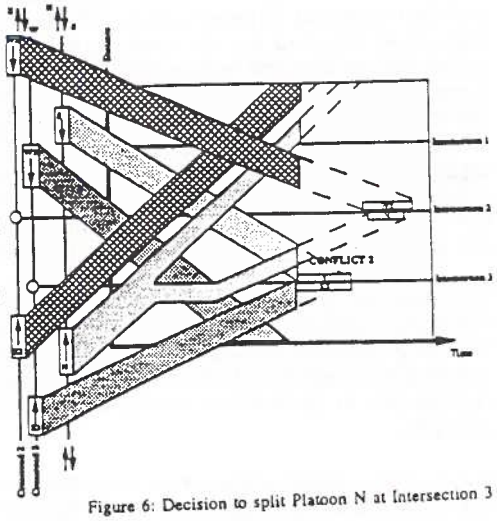


Figure 6: Decision to split Platoon N at Intersection 3

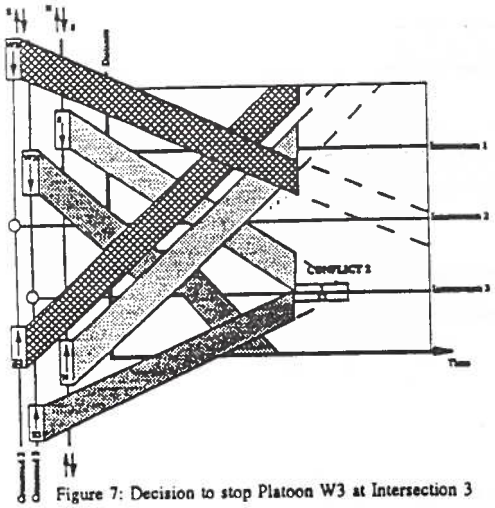


Figure 7: Decision to stop Platoon W3 at Intersection 3

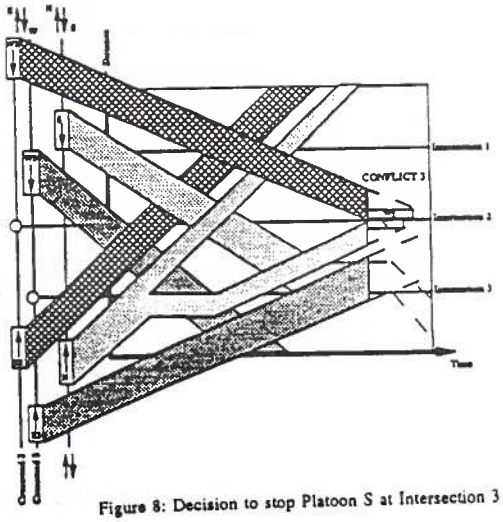


Figure 8: Decision to stop Platoon S at Intersection 3

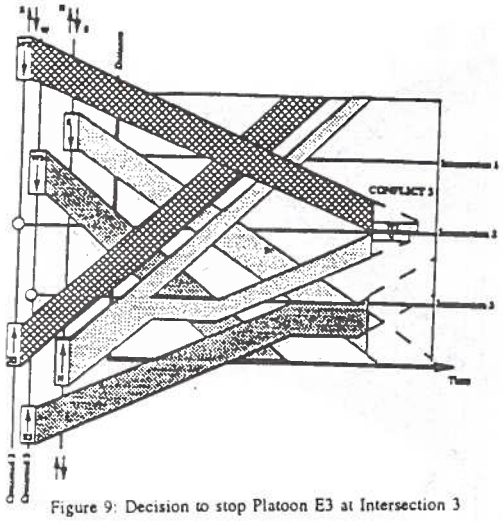


Figure 9: Decision to stop Platoon E3 at Intersection 3

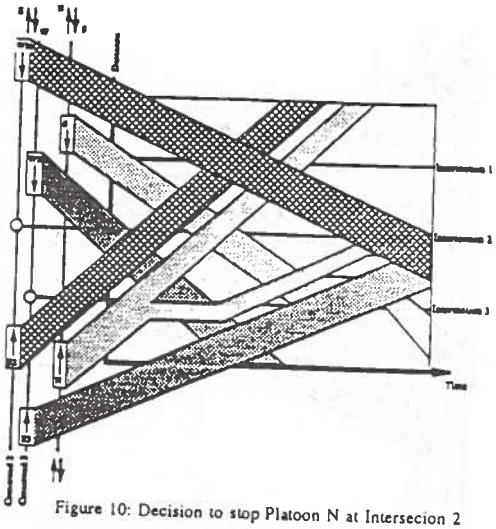


Figure 10: Decision to stop Platoon N at Intersecion 2

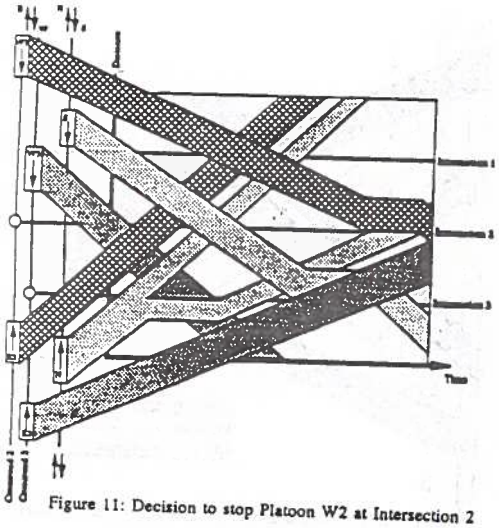


Figure 11: Decision to stop Platoon W2 at Intersection 2

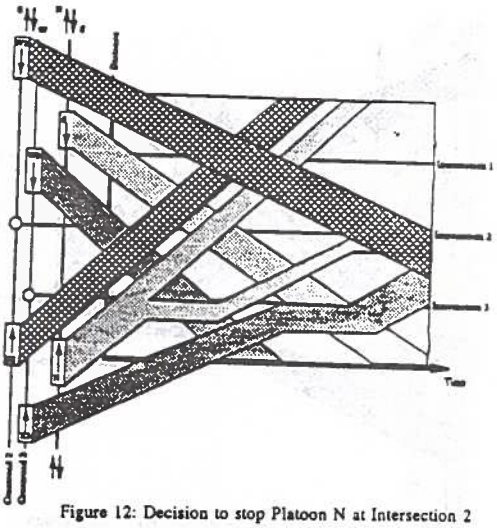


Figure 12: Decision to stop Platoon N at Intersection 2

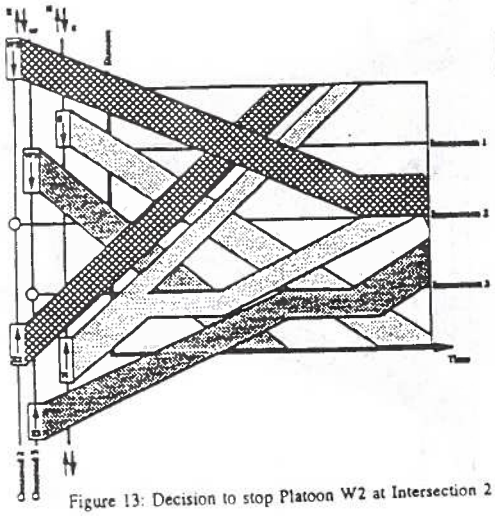


Figure 13: Decision to stop Platoon W2 at Intersection 2

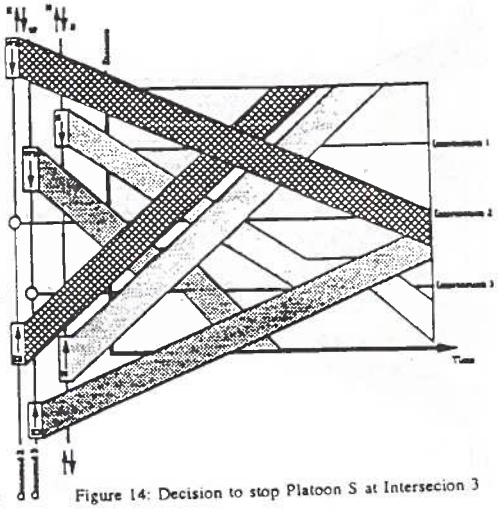


Figure 14: Decision to stop Platoon S at Intersecion 3

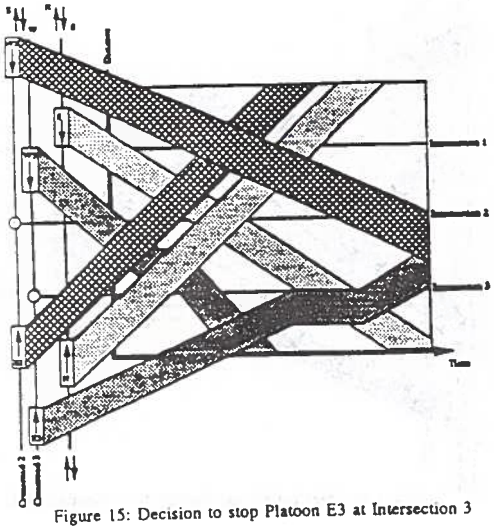


Figure 15: Decision to stop Platoon E3 at Intersection 3

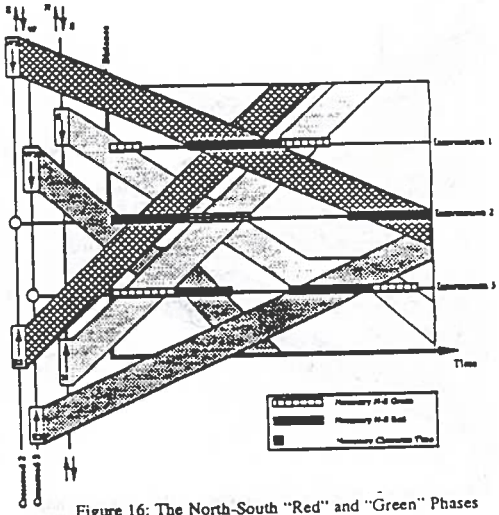


Figure 16: The North-South "Red" and "Green" Phases for the Decision of Figure 14

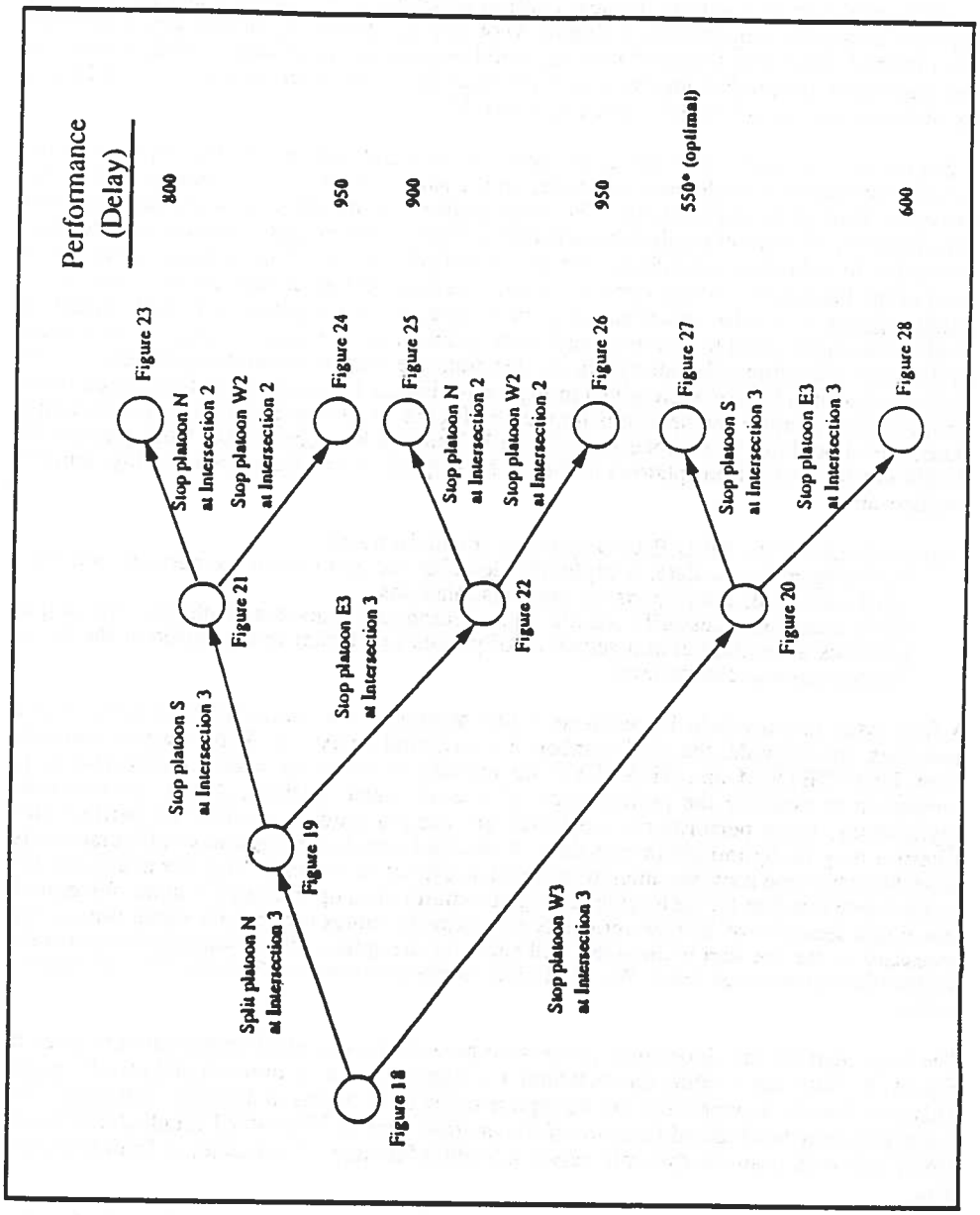


Figure 17: Decision Tree for the Illustrative Problem

Figure 5 is the current prediction of the movement of the platoons shown in Figure 4. The first demand conflict arises between platoon N and W3 at Intersection 3. At this point we either split platoon N (Figure 6) or stop platoon W3 (Figure 7). Considering the resulting predictions shown in Figure 6, the next conflict arises between platoon S and E3. Here the decision is to either stop platoon S (Figure 8) or stop E3 (Figure 8). In this way, a decision tree i formed which keeps track of various candidate decision ad demand conflicts arises. For this illustration, the predictions that arise for various decisions are given in Figures 19-28 and the decision tree that is formed is given in Figure 16.

When the time horizon is reached, associated with each end node will be the total cost of the cumulative decisions made up to the node, on the path from the root of decision tree to the end node (leaf) of the decision tree. Choosing the one with minimum cost provides the least cost trajectory of conflict resolution decisions. A final backward pass provides a phase plan, within the time horizon considered, for the identified platoons. This is passed to the third level of the hierarchical traffic control system (intersection/interchange control logic) as an initial solution for further optimization as more detailed data is gathered at the intersection level. Choosing the root to end-node path with optimum performance (in this case minimum total delay), the optimal decision from the decision tree for this illustrative problem, for the platoons shown in Figure 4, are given in Figure 14. Figure 17 shows the red and green times for the N-S arterial for the decisions implied by Figure 14. It indicates that at Intersection 3 platoon N should not be stopped but platoon W3 should be stopped and, later, platoon E3 should not be stopped but platoon S should be stopped, when the corresponding demand conflicts arise.

Two advantages of the REALBAND approach should be noted:

(1) Using real-time data, it explicitly identifies the platoons in the network and their sizes and speed, and response to these real platoons.

(2) It does not necessarily require a pre-determined sequence of phases. The output provides an initial cut at a sequence of phases for further optimization at the lower intersection/interchange level.

A final issue that needs to be resolved in this method is the computation of performance measures, for example, the total number of stops, total delay, etc. This is where concepts from TRAFNETSIM and TRANSYT are utilized to come up with a *quick-and-dirty* simulation to evaluate the performance of a set of signal settings. Again, for real time applications, these performance measures are needed quickly so that the performance criterion may be optimized in real time. A detailed simulation becomes computationally unwieldy when one uses the simulation model as a *function evaluator* (i.e., for evaluating the performance function for each candidate signal setting) in an optimization routine. We remark that at this second level of hierarchy only approximate values for optimal signal timings are necessary so that we start in the right "ball park" for the optimizations being performed at the intersection/interchange level. We shall refer to this simulator as *Platoon Flow Prediction* model.

The flow chart for the algorithmic process for network flow control optimization is given in Figure 18. Note that to being the recursion it is suggested that an initial signal plan be given. This is so that the network flow control optimization itself begins in the right "ball park". The initial plan may be obtained from an off-line method such as TRANSYT [Hadi and Wallace, 1992] with data from the dynamic network loading function, if available, or from historical data.

In relation to the use of the fast simulation model for function evaluation, the spatial region for the simulation model needs to be bigger than the area of control for network coordination,

so that one can predict the movements of all real-time platoons within the region of control for several minutes in the future. As an illustration, Figure 19 shows the region of control for the network flow control logic and Figure 20 shows the area simulated using the *Platoon Flow Fast Prediction* simulator.

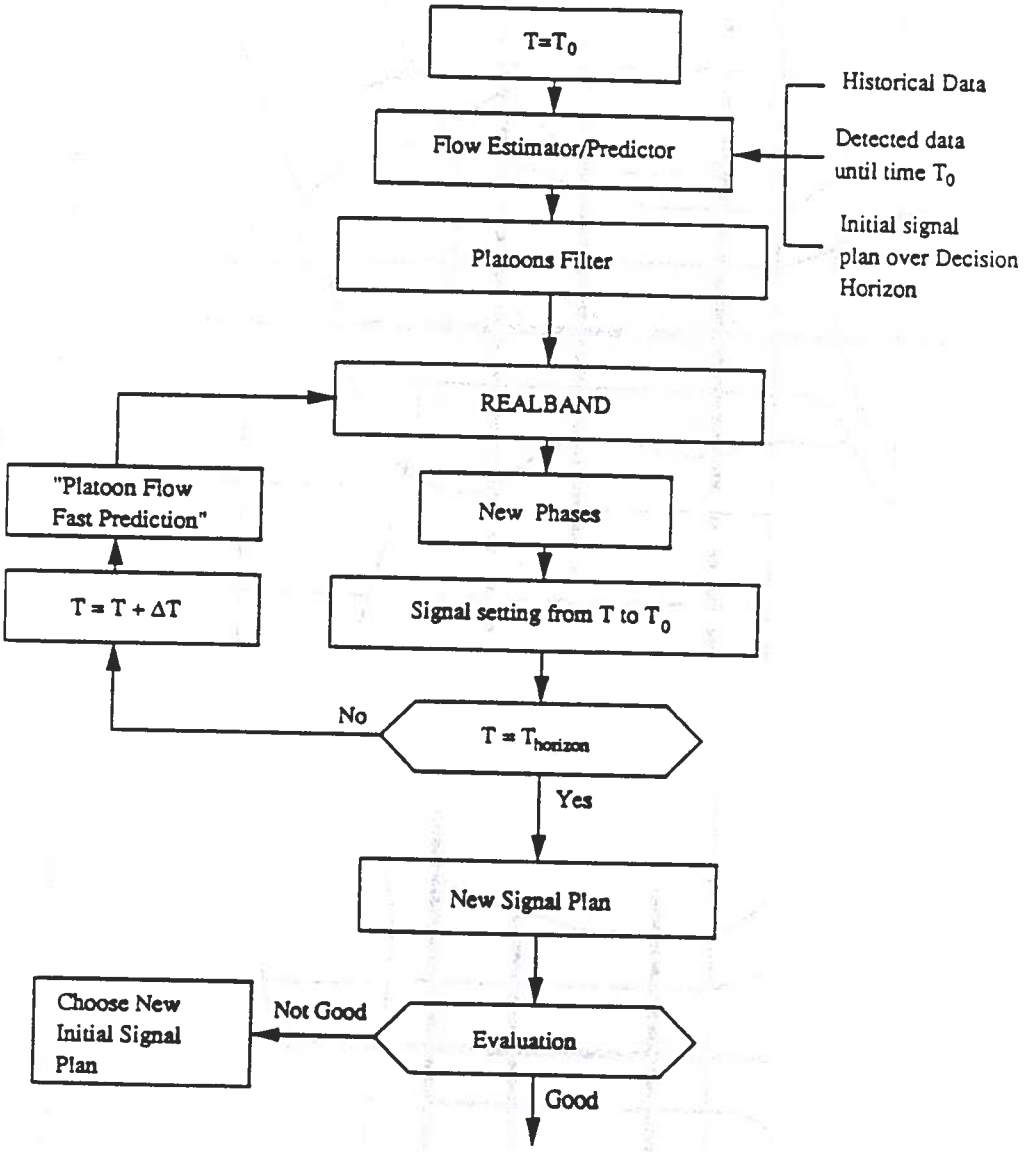


Figure 18: Flow Chart for Network Flow Control Optimization

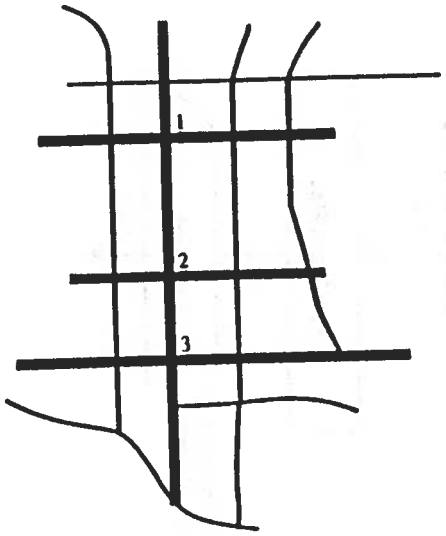


Figure 19: Region for Network Flow Control

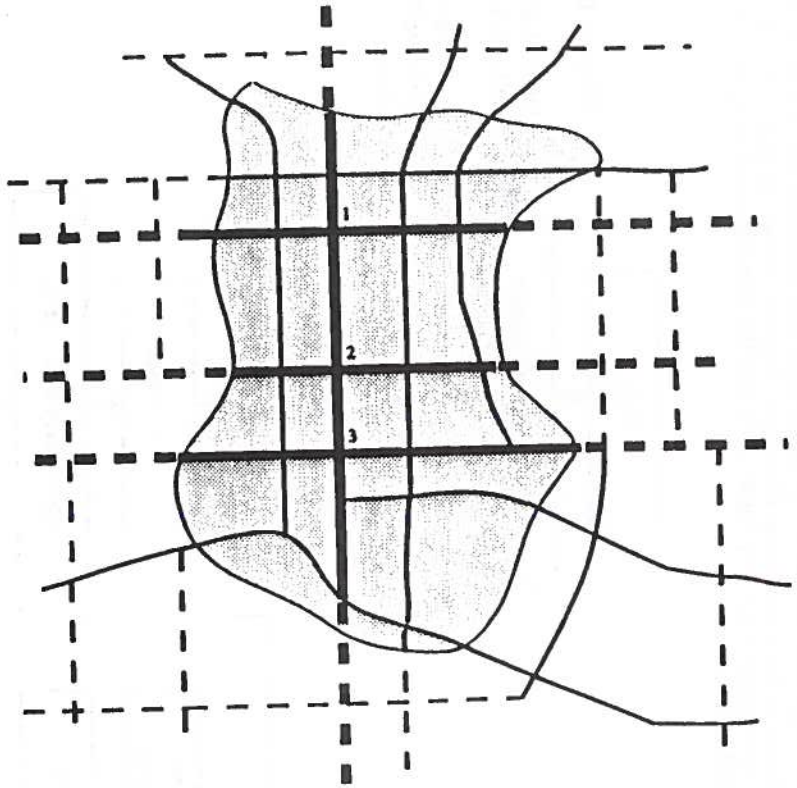


Figure 20: Region for Simulation

REFERENCES

Chang, E.C-P., B.G. Marsden, and R. Derr (1987), "PASSER II-84 Microcomputer Environment System - Practical Signal-Timing Tool", Journal of Transportation Engineering, Vol. 113, 1987, 625-641.

Gartner, N.H. (1985), "Demand-Responsive Traffic Signal Control Research", Transportation Research 19A, 369-373.

Head, K.L., P.B. Mirchandani and D. Sheppard (1992), "A Hierarchical Framework for Real-Time Traffic Control", Transportation Research Record.

Head, K.L. and P.B. Mirchandani (1993), "A Hierarchical Architecture for Real-Time Traffic Control", Proceedings of the IFAC Conference, Sidney (1993).

Hunt, P.B., D.I. Robertson, R.D. Bretherton, and R.I. Winton (1981), SCOOT - A Traffic Responsive Method of Coordinating Signals, Report No. LR 1014, Transport and Road Research Laboratory, Crowthorne, Berkshire, England.

Khoudour, L., J-B. Lesort, J-L. Farges (1991), "PRODYN - Three Years of Trials in the ZELT Experimental Zone", Recherche - Transports - Securite, English Issue, Special Traffic Management, 89-98.

Little, J.D.C., M.D. Kelson, and N.H. Gartner (1981), "MAXBAND; A Program for Setting Signals on Arteries and Triangular Networks", Transportation Research Record 795, 40-46.

Mauro, V. and D. Di Taranto (1990), "UTOPIA", Proceedings of the 6th IFAC/IFIP/IFORS Symposium on Control Computers and Communication in Transportation, Paris, France.

McShane, W.R. and R. Roess (1990), Traffic Engineering, Prentice-Hall, New York, NY.

Mirchandani, P.B. and H. Soroush (1987), "Generalized Traffic Equilibrium with Probabilistic Travel Times and Perceptions", Transportation Science, 21, 133-152.

Papageorgiou, M. (1983), Applications of Automatic Control Concepts to Traffic Flow Modelling and Control, Springer-Verlag, New York.

Rathi, A. K. and A.J. Santiago (1989), "The New NETSIM: TRAF-NETSIM Version 2.00 Simulation Program", presented at the Transportation Research Board's 68th Annual Meeting, Washington, D.C., January 1989.

Sims, A.G. (1988), "SCATS Computer Aided Traffic Engineering System (SCATES)", ARR 160, Australian Road Research Board, in L.Y.K. Luk, Ed., Proceedings - Urban Traffic Control Workshop, 14th ARRB Conference, Canberra, August 1988, 17-33.

Wallace, C.E., K. Courage, D.R. Reaves, G.W. Schoene, and G.W. Euler (1981), TRANSYT-7F User's Manual, Federal Highway Administration.

Wardrop, J. G., (1952), "Some Theoretical Aspects of Road Traffic Research," Proc. Inst. Civil Engr., Part II, 1, 325-378.

A METHOD FOR INTEGRATED DEMAND RESPONSIVE SIGNAL CONTROL

Bernhard Friedrich

Fachgebiet Verkehrsplanung und Verkehrswesen
Technische Universität München

EXTENDED ABSTRACT

The paper describes a method for on line signal control that is improved, implemented and tested in the context of the CEC-ATT research project MUNICH_COMFORT at an aterial road section in Munich.

The increasing demand of private transportation with its negative impacts on environmental quality leads city councils to support a policy of reducing car traffic by giving priority to public transport systems. Beside these well known requirements for public transport priority additional, often conflicting, objectives have to be considered by the control method. Thus the requests of general public and local authorities focus on the minimization of vehicle emissions and other negative impacts of car traffic.

These new requirements for signal control cannot be solved in a sufficiant way by traditional methods like vehicle actuation and priority pre-emption since the explicit

formulation of an objective function that describes the balance of the different goals is not possible.

The request for an impact sensitive method suggests that demand responsive methods which are based on impact modelling and consider an explicit objective function would fulfill the mentioned requirements. However a review of demand responsive control methods of the second and third generation showed that the major part does not take into account in particular the above mentioned goals. The investigation does not aim to be complete since some interesting approaches, particularly recent ones, are not published providing enough information. Summing up it may be said that most of the approaches are characterised by a number of shortcomings :

-The requests for high computation speed restricts the necessary accuracy of some traffic models in a way that impacts cannot be realistically calculated (e.g. components of vehicle emissions).

- The investigated methods do not allow for an active priority of public transport vehicles.

- The information on network flows via on-line estimation is not used.

In the context of the real requirements of the City of Munich (which may be applicable to other European cities too) and under consideration of the applicability of the method to the local situation (complex intersection geometries which make a consideration of more than two stages necessary) a new control method has been developed which fulfills the mentioned requirements.

Building on the traditional way of vehicle actuation and priority pre-emption the new method overcomes the mentioned shortcomings by integrating an advanced demand responsive component. Thus the new method bases on a master signal plan that can be customized to each local situation and that allows for favouring buses and for reacting to microscopic events. This basis is completed by an higher level component responding to medium term changes in demand and balance control according to the goals formulated

in the ojective function. The kernel of the demand responsive component are medium term forecasts (for a rolling horizon of about 15 min.) of the traffic flows in the network and a microscopic traffic model which allows for the consideration of traffic actuation, the consideration of single public transport vehicles and for the calculation of different peformance criteria. Special features of the traffic model are the high discretisation of time (1/10 sec) and space (1/10 m), a modelling of network components in a way that every road and intersection geometry can be represented and the consideration of single driver-vehicle units that follow precise car following rules. Designing the model components with respect to recent research studies a representation of traffic flow close to reality could be achieved.

The control algorithm of the medium term level varies, based on the results of the impact analysis (which is the result of the on line runs of the microscopic traffic model and the corresponding impact model), the control variables and searches a new combination by optimizing the objective function. The function can be formulated with regard on the criteria delay and stops of public transport vehicles, delay and stops of car traffic, queue lengths and emissions. The variables are parametrized in order to allow for a flexible weighting.

Based on this knowledge decisions are derived in a heuristic way to modify the control variables stage scheme, stage sequence, split and offset. These variables are assigned to a multi level control strategy which shows up their mechanism with regard to time and space. The optimization problem formulated in the objective function is a non-convex integer problem which is solved by heuristic algorithms.

The following algorithms seems to be suited for the on line solution of the problem and are actually tested :

- A SCOOT-like algorithm that modifies split and offset in small modifications and choose the stage scheme and the cycle time according to the saturation criterion.

- An algorithm based on the steepest descent for the optimization of split and offset by previously determined stage scheme and cycle time according to the saturation criterion.
- An genetic algorithm for the optimization of split, stage scheme, stage sequence and offset by predetermined and cycle time

In order to close the control loop the rolling horizon technique is applied.

A discussion shows which requirements concerning data acquisition, transmission and control-hardware for the implementation of the method within the Munich test site have to be fulfilled with respect to the local situation.

First results on the applicability of the method in the field test in Munich and its potential benefits will be presented in the paper.

A Decentralized Traffic Control System Based on Logic Programming

Giovanni Rinaldi
IASI - CNR, Roma, Italy

Klaus Truemper
University of Texas at Dallas, USA

We describe a new system for controlling the traffic lights in a urban network. The basic component of the system is a control unit that governs the traffic lights at a single intersection.

The behavior of each control unit is governed by a logic program. The possible actions that the control unit may take, depending on various traffic conditions, are expressed in a quasi-English language that is close to first order logic. The decision of switching from the current light configuration to another, then becomes a minimum cost satisfiability problem. Although such logic problems generally are difficult, the instances with about 300 clauses and 300 variables that typically occur in a control unit are solved in less than 1 second on a low cost microcomputer. That solution time is made possible by the Logic Programming System *Leibniz*, which automatically produces fast solution algorithms for given logic formulations. *Leibniz* also produces a guaranteed upper bound on computation time for each algorithm it produces. That bound is well below 1 second for the traffic control problem, thus guaranteeing reliable and timely decision making by each control unit.

The decision process of each control unit is independent from that of the other control units, and is based on the traffic conditions of the streets that converge to the intersection. Thus, all control units are at the same hierarchical level, and there are no units that supervise the activities of others. Communication is allowed only between control units located at neighboring intersections, mainly to provide information on incoming traffic.

The hardware required to execute the algorithm of a control unit can be any low cost commercial microcomputer. The entire system is fully decentralized and requires a very simple low speed communication network.

The advantages of the system are low cost of installation and maintenance, robustness of decisions made, as well as extraordinary flexibility and modularity.

We also describe a graphics-based tool for the design and evaluation of the control units and their logic formulations. That tool is an interactive graphic system that simulates several intersections, executes the control units of these intersections, and switches traffic lights according to the decisions by the control units.

Simplex-based Tabu Search for the Multicommodity Capacitated Fixed Charge Network Design Problem

Teodor Gabriel Crainic[1,2], Michel Gendreau[1,3]
Judith M. Farvolden[4], and Patrick Soriano[1]

(1) Centre de recherche sur les transports
Université de Montréal
(2) Département des sciences administratives
Université du Québec à Montréal
(3) Département d'informatique et recherche opérationnelle
Université de Montréal
(4) Algorithmics Inc., Toronto, Canada

May, 1994

1 Introduction

The fixed charge network design problem is a well–known problem, of both practical and theoretical significance. We address, in particular, the fixed charge capacitated multicommodity network design problem with linear costs and no additional side constraints (CMND). Special cases of this instance of the problem are service network design, telecommunication network design, power system design, etc.

Interesting results have been derived for the uncapacitated design problem [13, 14, 1]. Less effort has been directed towards the capacitated problem, which is more difficult to solve; [12], for example, derives a lower bound for a simple instance of the problem. For capacitated problems of any interesting size, only heuristics specially tailored have proved of any help. The simplest of these are drop and add procedures based on reduced cost calculations which determine the marginal value of including or excluding an arc from the network [15, 11]. The more sophisticated determine the marginal value of creating a new path in the network [2, 10, 6]. These local search heuristics find a local optimum and generally require significant computation times.

Tabu search is a metaheuristic introduced by Glover [8, 9], designed to escape the trap of local optimality. In essence, it is a local search scheme that explores iteratively the space of potential solutions by moving from one of these solutions to another at each step. To each solution is associated a *neighbourhood* consisting of

all solutions that differ only slightly from it. At each iteration, the neighbourhood of the current solution is examined and the candidate yielding the best reward (the objective value or other measure) within it is selected to become the new current solution. In order to avoid cycling, a recent history of the search is maintained in constantly updated *tabu lists*; this information is used to prevent the search from returning to solutions visited in the most recent iterations. As the search proceeds, one can record the best solutions encountered; the attributes of these solutions may then be used to guide the exploration of the solution space.

Tabu search procedures for mixed integer network design problems have already yielded promising results: first, for uncapacitated multicommodity location–allocation problems with balancing requirements [3] and, more recently, for the CMND [4]. In both cases, the search is conducted by exploring the space of the design (integer) variables. This makes for rather cumbersome procedures that require heavy computations to evaluate neighbours and select the best move. For the problem of interest here, this translates into a requirement to solve a capacitated multicommodity minimum cost network flow problem (MCNF) for each neighbour evaluation. Since this is too costly, sampling and approximation schemes are used. Although this effort represents one of the first successful algorithms for the CMND, we believe that a more elegant and efficient tabu search-based procedure is possible. The goal of this paper is to present and analyze such a procedure, in which the tabu search framework is combined with simplex pivoting rules to produce a search that explores the space of the continuous flow variables instead.

2 Formulations

The general form of the arc–based formulation of the network design problem that we aim to solve is:

$$\text{Minimize } z(x,y) = \sum_{a \in D} f_a y_a + \sum_{p \in P} \sum_{a \in A} c_a^p x_a^p \tag{1}$$

$$\text{Subject to } \sum_{j \in N} x_{ji}^p - \sum_{j \in N} x_{ij}^p = d_i^p \quad \forall i \in N, p \in P \tag{2}$$

$$\sum_{p \in P} x_a^p \leq u_a y_a \quad \forall a \in D \tag{3}$$

$$\sum_{p \in P} x_a^p \leq u_a \quad \forall a \in A \backslash D \tag{4}$$

$$x_a^p \geq 0 \quad \forall a \in A, p \in P \tag{5}$$

$$y_a = \{0,1\} \quad \forall a \in D, \tag{6}$$

where:

$$d_i^p = \begin{cases} -w^p & \text{if } i = r^p, \\ w^p & \text{if } i = s^p, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

and $\mathcal{G} = (N, A)$ is the graph describing the network, N the set of nodes, $A = \{a = (i,j) \,|\, i \in N, j \in N\}$ the set of arcs, and $D \subseteq A$ the set of design arcs. P represents

the set of commodities to be distributed. A commodity $p \in P$ is defined to be the flow between an origin–destination pair (r^p, s^p). The following notation is required to describe the supply of and demand for the set of commodities:

$\mathbf{r} = (r^p)$: origin node of commodity $p \in P$,
$\mathbf{s} = (s^p)$: destination node of commodity $p \in P$,
$\mathbf{d} = (d_i^p)$: demand pattern of commodity $p \in P$ at node $i \in N$,
$\mathbf{w} = (w^p)$: size of commodity $p \in P$.

The decision variables are:

$\mathbf{x} = (x_a^p)$: amount of flow of commodity $p \in P$ on arc $a \in A$,
$\mathbf{y} = (y_a)$: network design variables,

while the parameters of the model are:

$\mathbf{c} = (c_a^p)$: unit cost of flow of commodity $p \in P$ on arc $a \in A$,
$\mathbf{f} = (f_a)$: fixed cost of introducing capacity on arc $a \in D$,
$\mathbf{u} = (u_a)$: capacity of arc $a \in A$.

We note that, for any set of values of the design variables $\bar{y}$, the arc–based formulation of the CMND (equations 1 to 7) becomes a MCNF problem. The objective of such a formulation is to distribute a number of items between their points of origin and destination, over a network in which some or all of the arcs have limited capacity. Multicommodity network flow problems are linear programs. However, those of interest are usually too large to be solved in reasonable time by standard linear programming solvers, and specialized solution techniques which exploit the structure of the constraint matrix have to be developed. Farvolden, Powell and Lustig [5] derive an efficient primal partitioning solution for the path–based formulation of the MCNF. We use this approach to design our simplex framework.

The path-based formulation of the CMND is:

$$\text{Minimize } z(h,y) = \sum_{a \in D} f_a y_a + \sum_{p \in P} \sum_{l \in L^p} k_l^p h_l^p \tag{8}$$

$$\text{Subject to } \sum_{l \in L^p} h_l^p = w^p \quad \forall p \in P \tag{9}$$

$$\sum_{p \in P} \sum_{l \in L^p} h_l^p \delta_a^p \leq u_a y_a \quad \forall a \in D \tag{10}$$

$$\sum_{p \in P} \sum_{l \in L^p} h_l^p \delta_a^p \leq u_a \quad \forall a \in A \backslash D \tag{11}$$

$$h_l^p \geq 0 \quad \forall p \in P, l \in L^p \tag{12}$$

where:

L^p = the set of paths for commodity $p \in P$,
h_l^p = the flow of commodity $p \in P$ on path $l \in L^p$,

$$\delta_{al}^p = \begin{cases} 1 & \text{if arc } a \text{ is in the } l^{th} \text{ path for commodity } p, \\ 0 & \text{otherwise,} \end{cases}$$

$k_l^p = \sum_{a \in A} c_a^p \delta_{al}^p$, the cost of path $l \in L^p$ for commodity $p \in P$,

$L_a^p =$ the set of paths using design arc $a \in D$, $L_a^p \subset L^p$.

As before, for any given $\bar{y}$, the path-based formulation of the CMND becomes a MNCF (in path formulation). If we denote by $H(\bar{y})$ the set of feasible path flows for a given $\bar{y}$, an optimal solution to this MNCF (assuming a non-empty $H(\bar{y})$) can always be found at an extreme point of $H(\bar{y})$. Denote by $\bar{H}(\bar{y})$ the set of such extreme points. Consider now $H(\bar{y} = 1)$, the set of feasible flows when all design arcs are open, that we will simply denote by H in the following, and its associated set of extreme points $\bar{H}$. H is defined by the set of constraints

$$\sum_{l \in L^p} h_l^p = w^p \quad \forall p \in P \tag{13}$$

$$\sum_{p \in P} \sum_{l \in L^p} h_l^p \delta_a^p \leq u_a \quad \forall a \in A \tag{14}$$

$$h_l^p \geq 0 \quad \forall p \in P, l \in L^p \tag{15}$$

and it is easly verified that $\bar{H} = \bigcup_{\bar{y} \in Y} \bar{H}(\bar{y})$, where Y is the set of possible design variable vectors.

For any given flow pattern $\bar{h} \in \bar{H}$, a minimal cost design relative to the flows, denoted $y(\bar{h})$, can be obtained by setting

$$y_a(\bar{h}) = \begin{cases} 0 & \text{if } \sum_{p \in P} \bar{h}_a^p = 0 \ , \\ 1 & \text{otherwise.} \end{cases} \tag{16}$$

An optimal solution of the CMND can thus be found by complete or implicit enumeration of the elements of $\bar{H}$, but such a procedure is likely to be extremely time-consuming, even for moderately-sized instances. Alternatively, the previous property may be used to define local search heuristics for the CMND: the adjacency relationships in $\bar{H}$ define a natural neighbourhood structure, while the pivoting rules of the simplex provide an efficient way to determine the neighbours of any given solution.

3 The Tabu Search Framework

In typical real-life network design problems, the set of possible paths is too large to allow its complete enumeration. The solution adopted in [5] to overcome this difficulty in the MCNF context is to resort to a column generation scheme: the method starts with a small number of paths for each origin-destination pair, and new paths are added as the algorithm proceeds. The new paths are obtained by solving shortest path problems on a graph where arc costs correspond to reduced costs relative to the current basis.

Adapting this idea within our tabu search framework represents an interesting challenge, since several of the standard assumptions upon which column generation schemes are based are no longer verified. In particular, the monotonic decrease of the

objective function value is no longer ensured, and both variable and fixed costs characterize arcs. On the other hand, the precise definition and generation of the tabu search neighbourhoods in a column generation context poses an additional challenge, linked particularly to the description and identification of path variables. The implementation of a path dictionary and of the corresponding access scheme is therefore an important algorithmic efficiency factor.

The approach developed here consists in exploring the space of the continuous commodity-flow variables by using pivoting as the basic tabu search move. At each iteration, all non-basic variables present in the master problem are evaluated using the exact costs and the best non-tabu move is implemented. Note that the objective function of the CMND provides us with a natural aspiration criterion.

The algorithm resorts to column generation only when a local optimum has been reached relative to the paths already generated that are not currently tabu and this local optimum has not been improved upon over a certain prespecified number of iterations. At that point, the method generates new "shortest paths" for each commodity. However, the reduced costs used in the column generation subproblems are not exact costs but correspond to a continuous relaxation of CMND with the following cost function: $\bar{c}_a^p = c_a^p + f_a/u_a \quad \forall a \in A, p \in P$. This modification, by accounting for the fixed design costs, should promote the generation of potentially attractive paths with respect to the original problem.

The tabu mechanism used by the algorithm consist in preventing exiting path variables from re-entering the basis for a certain number of iterations by associating to them a *tabu tag* drawn randomly from $[t_{min}, t_{max}]$, where both t_{min} and t_{max} are input parameters. Finally, the method stops when a predetermined number of column generation phases have been performed without improving the best solution found.

4 Preliminary Computational Results

In the following table we report the preliminary results obtained on a set of 18 randomly generated test problems. The identificators listed under the *Problem* column provide the main characteristics of each instance. These are respectively the number of vertices, links and commodities defining the network and two letter indicators. The fisrt specifies which costs are predominant, the variable flow costs (V) or the fixed design costs (F), while the second indicates if the capacity restrictions are tight (T) or loose (L).

The value of the best solution found by the Simplex Based Tabu Search algorithm (SBTS) is listed under *Z_Best*, the iteration at which it was found under *B_Iter*, the total number of iterations performed by the algorithm under *T_Iter*, and the amount of cpu time corresponding to each iteration count under *B_Time* and *T_Time* respectively. In order to evaluate the performance of the approach we have also included in the table the values of the best upper and lower bounds (under *UBound* and *LBound* respectively) found by the bounding procedure developed by Gendron and Crainic [7] and the time required to compute them. This procedure combines a resource-decomposition heuristic to determine upper bounds and both Lagrangean relaxation and subgradient optimization for generating lower bounds. All times reported are cpu times on a SUN SPARC10 workstation and are given in seconds.

Table 1: Preliminary Computational Results

	Bounding Procedure		Simplex Based Tabu		
Problem	LBound (Time)	UBound (Time)	Z_best	B_Iter (B_Time)	T_Iter (T_Time)
25,100,10,F,L	12,545 (5.86)	27,347 (1.31)	24,541	6 (0.08)	9 (0.12)
25,100,10,F,T	42,067 (5.40)	69,379 (1.92)	54,426	69 (1.15)	99 (1.70)
25,100,10,V,L	14,577 (4.37)	16,086 (1.80)	14,925	8 (0.09)	9 (0.12)
25,100,30,F,L	31,784 (19.40)	59,337 (5.75)	42,225	29 (0.50)	138 (3.00)
25,100,30,F,T	80,119 (18.16)	107,123 (4.14)	87,983	136 (4.00)	312 (15.00)
25,100,30,V,T	356,462 (18.10)	427,228 (3.87)	383,826	266 (17.18)	373 (36.00)
100,400,10,F,L	19,093 (24.45)	74,956 (7.89)	28,940	8 (0.21)	11 (0.32)
100,400,10,F,T	46,908 (22.95)	150,732 (9.29)	85,334	28 (0.87)	29 (0.97)
100,400,10,V,L	27,426 (15.60)	40,834 (10.17)	28,443	27 (0.81)	34 (1.07)
100,400,30,F,L	42,236 (91.42)	178,564 (31.16)	75,889	5 (0.35)	8 (0.67)
100,400,30,F,T	112,653 (71.97)	285,236 (23.12)	158,756	205 (11.69)	411 (25.88)
100,400,30,V,T	367,491 (64.28)	412,416 (27.26)	386,450	408 (52.38)	494 (85)
20,230,40,V,L	421,726 (50.01)	501,400 (19.98)	431,179	111 (3.08)	208 (7.91)
20,230,40,V,T	368,516 (50.38)	403,574 (17.96)	372,723	79 (2.65)	242 (8.93)
20,230,40,F,T	633,040 (51.32)	X (14.62)	649,365	40 (1.22)	233 (7.44)
20,230,200,V,L	89,237 (303.35)	190,558 (107.37)	194,384	4 (0.34)	168 (6.09)
20,230,200,F,L	128,466 (305.83)	232,287 (95.50)	306,100	2 (0.32)	169 (6.02)
20,230,200,V,T	94,236 (298.75)	144,530 (92.63)	172,824	2 (0.33)	14 (0.86)
	393 (34.04)				

It should be noted that the results we are reporting here were obtained with a rather basic implementation of the approach in which no intensification or diversification mechanism were incorporated. In addition, the setting of the parameters used by the algorithm is still crude and has yet to be refined. Nevertheless, these numbers are already quite interesting. Indeed, for most problems the present version of SBTS found significantly better solutions than the ones provided by the upper bounding procedure, and generally this was accomplished while expending less cpu time. In only three instances, the last ones in the table, was the algorithm unable to improve on the upper bound. These particular instances have a much larger number of commodities than the others, however no special attention was given to this aspect in the present experiments and they were solved without changing the parametter settings. We feel that a more adequate setting should enable the SBTS to perform as well on problems of this type as it does on the others.

Further numerical experiments are currently under way to complete the development of this promising approach. In addition, several tabu search refinements will be examined. In particular, we intend to document the use of medium and long term memories in conjunction with intensification and diversification processes in the near future.

References

[1] Balakrishnan A., Magnanti T.L. and Wong R.T. (1989), *A Dual-Ascent Procedure for Large-Scale Uncapacitated Network Design*, **Operations Research** 37(5), 716–740.

[2] Crainic T.G. and Rousseau J.-M. (1986), *Multimode, Multicommodity Freight Transportation: A General Modelling and Algorithmic Framework for the Service Network Design Problem*, **Transportation Research B** 20(3), 225–242.

[3] Crainic T.G., Gendreau M., Soriano P. and Toulouse M. (1993), *A Tabu Search Procedure for Multicommodity Location-Allocation with Balancing Requirements*, **Annals of Operations Research** 41, 359–383.

[4] Crainic T.G., Farvolden J.M. and Gendreau M. (1993), *Tabu Search Heuristics for the Capacitated Fixed Charge Network Design Problem*, Working Paper, Centre de recherche sur les transports, Université de Montréal (forthcoming).

[5] Farvolden J.M., Powell W.B. and Lustig I.J. (1992), *A Primal Partitioning Solution for the Arc-Chain Formulation of a Multicommodity Network Flow Problem*, **Operations Research** 41(4), 669–694.

[6] Farvolden J.M. and Powell W.B. (1991), *Subgradient Methods for the Service Network Design Problem*, Working Paper, Department of Industrial Engineering, University of Toronto.

[7] Gendron B. and Crainic T.G. (1994), *Relaxations for Multicommodity Capacitated Network Design Problems*, Working Paper CRT-965, Centre de recherche sur les transports, Université de Montréal.

[8] Glover F. (1989), *Tabu Search – Part I*, **ORSA Journal on Computing** 1(3), 190–206.

[9] Glover F. (1990), *Tabu Search – Part II*, **ORSA Journal on Computing** 2(1), 4–32.

[10] Jarvis J.J. and Mejia de Martinez O. (1977), *A Sensitivity Analysis of Multicommodity Network Flows*, **Transportation Science** 11(4), 299–306.

[11] Koskosidis Y.A., Powell, W.B. and Solomon M.M. (1992), *An Optimization-based Heuristic for Vehicle Routing and Scheduling with Soft Time Windows Constraints*, **Transportation Science** 26, 69–85.

[12] Lamar B.W., Sheffi Y. and Powell W.B. (1990), *A Capacity Improvement Lower Bound for Fixed Charge Network Design Problems*, **Operations Research** 38, 704–710.

[13] Magnanti T.L. and Wong R.T. (1984), *Network Design and Transportation Planning: Models and Algorithms*, **Transportation Science** 18(1), 1–55.

[14] Minoux M. (1989), *Network Synthesis and Optimum Network Design Problems: Models, Solution Methods and Applications*, **Networks** 19, 313–360.

[15] Powell W.B. (1986), *A Local Improvement Heuristic for the Design of Less-than-Truckload Motor Carrier Networks*, **Transportation Science** 20(4), 246–357.

PARALLEL IMPLEMENTATIONS OF BOUNDING PROCEDURES FOR MULTICOMMODITY CAPACITATED NETWORK DESIGN

Bernard Gendron
Teodor Gabriel Crainic
Centre de recherche sur les transports
Université de Montréal
C.P. 6128, succursale A
Montréal, Québec
Canada H3C 3J7

May 1994

In this paper, we present parallel algorithms for solving large-scale multicommodity capacitated network design problems. It is well-known that these problems model various transportation applications [5]. In particular, our aim is to use our parallel algorithms to solve large-scale applications that arise in the motor-carrier industry [1].

In an earlier paper [2], we study a general model for the design of multicommodity capacitated networks. In particular, we identify several relaxations and compare them, both theoretically and empirically. We also show how to efficiently derive upper bounds by using the resource-decomposition principle, as applied to multicommodity network flow problems [4]. In both cases, the problems obtained decompose into several independent subproblems, thus making attractive their resolution in a parallel computing environment. Our objective is to present computational results of implementations of these bounding procedures on coarse-grained parallel architectures, and to show their usefulness in solving large-scale capacitated network design problems.

1 Formulation

Given:

A network $G = (N, A)$, where N is the set of nodes and A is the set of arcs;

A set of commodities P that move through the network;

For each commodity $p \in P$, a set of origin nodes $O(p)$;

For each commodity $p \in P$, a set of destination nodes $D(p)$;

For each commodity $p \in P$ and each origin node $i \in O(p)$, a supply $o_i^p > 0$;

For each commodity $p \in P$ and each destination node $i \in D(p)$, a demand $d_i^p > 0$;

For each arc $(i, j) \in A$, a total capacity $u_{ij} > 0$;

For each commodity $p \in P$ and each arc $(i, j) \in A$, a partial capacity $b_{ij}^p \geq 0$;

For each arc $(i, j) \in A$, a fixed cost $f_{ij} \geq 0$ incurred to include the arc in the design;

For each arc $(i, j) \in A$ and each commodity $p \in P$, a per unit routing cost $c_{ij}^p \geq 0$.

We seek to minimize total cost, while satisfying supply and demand requirements, flow conservation at transshipment nodes, and capacity constraints, under the following hypotheses:

$$\sum_{i \in O(p)} o_i^p = \sum_{i \in D(p)} d_i^p = t_p \quad \forall\, p \in P$$

$$b_{ij}^p \leq \min(u_{ij}, t_p) \quad \forall\, (i, j) \in A,\ p \in P$$

$$u_{ij} \leq \sum_{p \in P} b_{ij}^p \quad \forall\, (i, j) \in A$$

We define the following variables:

$x_{ij}^p \geq 0$ represents the amount of flow of commodity $p \in P$ moving on arc $(i, j) \in A$;

y_{ij} models the design decision for arc $(i, j) \in A$, i.e.

$$y_{ij} = \begin{cases} 0, & if\ x_{ij}^p = 0,\ \forall\, p \in P \\ 1, & otherwise \end{cases}$$

The model is formulated as follows (where $N^+(i) = \{j \in N \,|(i, j) \in A\}$ and $N^-(i) = \{j \in N \,|(j, i) \in A\}$ for each $i \in N$):

$$\min \sum_{p \in P} \sum_{(i,j) \in A} c_{ij}^p x_{ij}^p + \sum_{(i,j) \in A} f_{ij} y_{ij}$$

$$\sum_{j \in N^+(i)} x_{ij}^p - \sum_{j \in N^-(i)} x_{ji}^p = \begin{cases} o_i^p & if\ i \in O(p) \\ -d_i^p & if\ i \in D(p) \\ 0 & otherwise \end{cases} \quad \forall\, i \in N,\ p \in P \tag{1}$$

$$\sum_{p \in P} x_{ij}^p \leq u_{ij} y_{ij} \quad \forall\, (i,j) \in A \tag{2}$$

$$x_{ij}^p \leq b_{ij}^p y_{ij} \quad \forall\, (i,j) \in A,\ p \in P \tag{3}$$

$$x_{ij}^p \geq 0 \quad \forall\, (i,j) \in A,\ p \in P \tag{4}$$

$$0 \leq y_{ij} \leq 1 \quad \forall\, (i,j) \in A \tag{5}$$

$$y_{ij}\ integer \quad \forall\, (i,j) \in A \tag{6}$$

In the remainder, we refer to 1 as the *flow conservation* constraints, to 2 as the *weak linking* constraints, and to 3 as the *strong linking* constraints.

2 Bounding Procedures

2.1 Relaxations and Subproblems

We define two Lagrangean relaxations. The first consists in relaxing the weak and strong linking constraints, while the second consists in relaxing the flow conservation constraints. In the first case, the subproblems reduce to $|P|$ minimum cost network flow problems. In the second case, the subproblems decompose into $|A|$ continuous knapsack problems. Thus, computing lower bounds through these Lagrangean relaxations appears highly attractive in a parallel environment.

The Lagrangean decomposition principle [3] also allows to define interesting relaxations of the model. First, it consists of adding the following copy constraints:

$$x_{ij}^p = w_{ij}^p \quad \forall\, (i,j) \in A,\ p \in P \tag{7}$$

$$y_{ij} = z_{ij} \quad \forall\, (i,j) \in A \tag{8}$$

Both strong and weak linking constraints are then rewritten (and the other constraints, with the exception of flow conservation constraints, are duplicated) in terms of the new variables. A Lagrangean relaxation with respect to the copy constraints then splits the

problem into two subproblems: a *flow subproblem*, containing all constraints, with the exception of linking constraints, and a *linking subproblem*, containing all constraints, with the exception of flow conservation constraints. We note that the flow subproblem has the same structure as the one obtained from the Lagrangean relaxation relative to the linking constraints. Similarly, the linking subproblem has the same structure as the one obtained from the Lagrangean relaxation relative to the flow conservation constraints.

2.2 Upper Bounds

To compute upper bounds, we consider the following multicommodity network flow problem, defined relative to a subset $\bar{A}$ of A:

$$\min \sum_{p \in P} \sum_{(i,j) \in \bar{A}} (c_{ij}^p + f_{ij}/u_{ij}) x_{ij}^p$$

$$\sum_{j \in N^+(i)} x_{ij}^p - \sum_{j \in N^-(i)} x_{ji}^p = \begin{cases} o_i^p & if\ i \in O(p) \\ -d_i^p & if\ i \in D(p) \\ 0 & otherwise \end{cases} \quad \forall\, i \in N,\ p \in P \tag{9}$$

$$\sum_{p \in P} x_{ij}^p \leq u_{ij} \quad \forall\, (i,j) \in \bar{A} \tag{10}$$

$$0 \leq x_{ij}^p \leq b_{ij}^p \quad \forall\, (i,j) \in \bar{A},\ p \in P \tag{11}$$

To obtain feasible solutions to this problem, we use the resource-decomposition principle [4]. The computational effort required by the method involves a projection problem, which decomposes into $|\bar{A}|$ singly-constrained quadratic problems, and a reource-decomposition problem, which decomposes into $|P|$ minimum cost network flow problems. The upper bounding procedure is thus highly parallelizable.

References

[1] Farvolden J.M. and Powell W.B. (1991), *Subgradient Methods for the Service Network Design Problem*, Working Paper 90-06, Department of Industrial Engineering, University of Toronto.

[2] Gendron, B. and Crainic T.G. (1994), *Relaxations for Multicommodity Capacitated Network Design Problems*, Publication 965, Centre de recherche sur les transports, Montréal (also presented at the Netflow93 conference).

[3] Guignard M. and Kim S. (1987), *Lagrangean Decomposition: A Model Yielding Stronger Lagrangean Bounds*, Mathematical Programming, 39, 215-228.

[4] Kennington J.L. and Helgason R.V. (1980), *Algorithms for Network Programming*, Wiley-Interscience.

[5] Magnanti T.L. and Wong R.T. (1984), *Network Design and Transportation Planning: Models and Algorithms*, Transportation Science, 18(1), 1-55.

A MULTICRITERIA ANALYSIS FOR URBAN NETWORK DESIGN AND PARKING LOCATION

Giulio Erberto Cantarella, Antonino Vitetta
Università degli Studi di Reggio Calabria
via E. Cuzzocrea, 48
89128 Reggio Calabria
Italy
Tel. +39 - 965 - 875209/26/28
Fax. +39 - 965 - 875227

ABSTRACT

The Urban Network Design Problem consists in determining the optimal configuration of an urban transport network, which is link capacities (including signal setting) and topological configuration, assuming that the user path choice behaviour is modelled through an equilibrium assignment. Exact procedures to solve the problem is not computationally suitable for a real-size network. In this paper the Urban Network Design is extended to include a multicriteria solution approach for Parking Location and the results of an application of genetic algorithms to solve the problem are presented. The procedure has been applied to a real network, producing a set of non dominated solutions to the problem, generated with the multicriteria analysis. Some comparisons between the different configurations of the network so generated are carried out.

1 INTRODUCTION

The Urban Network Design Problem (UNDP) is an abstract, simplified version of a problem often faced by agencies responsible for street management in the urban area: namely, how to best design those systems in response to transportation demand. The UNDP generally consists in determining the optimal configuration of the elements of the urban network (topological characteristics and capacities of links), with respect to a set of criteria. In the Network Design Problem (NDP) the optimum links direction and links capacity (defined by lanes allocation) are determined; in the UNDP, link capacities also depend on the signal setting. This optimisation problem is particularly difficult to solve with an exact optimal procedure, since an user equilibrium flow has to be calculated at each step of the optimising search process system. Currently the most promising solution methods are heuristic, that is they not guarantee even the local optimality of their final results.

The NDP has been proposed in two different forms: a discrete form dealing with the addition or modification of roadway segments of a street network, and a continuous form dealing with the optimal capacity increases of existing roadway segments. The NDP has been mostly studied through a mono-criterion analysis by several authors. Billheimer and Gray (1973), Foulds and others (1981), Poorzahedy and Turnquist (1982), Chen and others (1991) use an heuristic algorithm with discrete variables; Le Blanch and Boyce (1986), Suwansirikul and others (1987), use continuous variables. The traffic signal setting problem has been studied in three different ways: single junction (Webster 1958, 1966, Gazis 1964, 1965, Allsop 1971, 1972) interacting junctions (Hilier 1966, Gartner and others 1972,1975) and equilibrium network signal setting (Allsop and

Charlesworth 1977; Gartner and others 1980, Cantarella and others 1987, 1991).

In this paper a previously proposed model (Cantarella and others 1994 a) for multicriteria UNDP is extended to include lane allocation and parking spaces: every lane can be used for vehicular movements or it can be used for parking spaces or it can be closed. A heuristic algorithm with modest computational requirements for finding non-dominated approximate solutions to the multicriteria UNDP is proposed. A multi-level optimisation framework is adopted. At the outer level a new configuration of network is evaluated with a genetic procedure and discrete variables are used (link direction and lane allocation); at the inner level continuous variables are used (traffic signal setting parameters and link flows) in an iterative procedure inside the traffic equilibrium assignment. The optimal non-dominated configurations of the network (link directions, lane allocation, signal setting in at junctions) are determined, in respect users' utility, investment funding, impacts on the community (e.g. air pollution) . In sections 2 and 3, a formulation of the UNDP and a solution procedure are presented. In section 4 a numerical example for a large-scale network is discussed. Some conclusions are reported in section 5.

2. THE URBAN NETWORK DESIGN PROBLEM

The UNDP assumes that a transport urban area is represented as a network consisting of a set of points, or nodes, connected by directed links. The representation of the topological characteristic of the network can be made with a link-route matrix **A**. A subset of the network nodes serves as points of trip origins, while another (possibly different) subset of nodes serves as trip destinations; the total demand for trips between each origin-destination (O-D) pair over the planning interval (such as the morning peak period) is inserted in the O-D vector **d**.

Five variables are attached to each link i of the network inserted in vectors **c, g, f, y, b**:

- the cost of travelling link i, c_i, function of flow on link i, lane allocation and traffic signal setting parameters;
- the traffic signal setting parameters g_i, function of flows and topological characteristics and which expresses the traffic signal setting method utilised;
- the link flow f_i, obtained from the equilibrium assignment to the network: **f=APd**, where **P** is the matrix of route choice probabilities (**P=*P*(c,y)**);
- a three entry vector $\mathbf{y_i}$, representing the lane allocation for link i (fig. 1): the first entry is the number of lanes in the direction ij (N_{ij}), the second is the number of lanes in the direction ji (N_{ji}) and the third the number of lanes for parking spaces (N_p). The total number of lanes in the link i (N) is known and the number of closed-lanes is obtained from: $N_c=N-(N_{ij}+N_{ji}+N_p)$. The generic decisional variable $\mathbf{y_i}$ can be associated at a single link or a set of links that must have congruent modifications in the design. The use of one variable for a set of congruent links, add further constraints to the problem but the computational complexity is reduced.
- the cost of construction or modification of link i b_i respect to the present configuration.

The mono-criterion UNDP consists in determining a configuration of network **y*** and corresponding signal setting at the junctions which minimise the measure of total system cost:

$$\begin{cases} \mathbf{y^*} = \arg\min_{\mathbf{y}} \; Z(\mathbf{y}) = \mathbf{c}^T \; \mathbf{f} + \beta \; \mathbf{b}^T(\mathbf{y}) \; \mathbf{u} \\ \text{subject to:} \\ \mathbf{f} = \mathbf{A} \; \mathbf{P}(\mathbf{c},\mathbf{y}) \; \mathbf{d} \\ \mathbf{c} = \mathbf{c}(\mathbf{f},\mathbf{y},\mathbf{g}) \\ \mathbf{g} = \mathbf{g}(\mathbf{f},\mathbf{y}) \end{cases}$$

where β denotes a scaling coefficient which converts units of construction cost into units of travel cost and **u** is the vector containing 1 in all its components.

If there is a budget constraint B on the investment in the network, the problem can be express in a knapsack problem form. In the objective function Z others components of cost can be present. In particular in

urban areas it is important to design the network in relation to the travel cost and in relation to the air pollution.

The UNDP more generally requires the simultaneous optimisation of multiple competing criteria. This problem is usually solved by combining different criteria into a single criterion according to some utility function not well known prior to the optimisation process. In this paper the UNDP is treated using a multicriteria approach and the non-dominated configurations of the element of the urban network (link directions and link capacities, signal setting at the junctions) are determined, in respect to users' utility, to investment funding, to the impacts on the community (e.g. air pollution).

3. SOLUTION APPROACH

A multi-level heuristic algorithm is used (Cantarella and others 1994 a) in order to find non-dominated approximate solutions to the UNDP (fig. 2). As said before at the outer level a new configuration of network is evaluated through a genetic procedure and discrete variables are used as described above. At the inner level the traffic signal setting and link flow assignment are carried out by an iterative method inside the traffic assignment procedure where, the signal setting, the delays on the network and the temporary flows are cyclically computed until two successive flow patterns are closed within a specified tolerance.

The genetic algorithms at the outer level allow efficient research in a very large search space and they are based on the mechanics of natural selection and natural genetics. They iteratively update a population of tentative solutions described by chromosomes; in this problem the chromosomes are the hypothetical decisional vector y of the topological characteristics of the urban network. Every position in the chromosome is called gene and it is relative to a design link on the network. The hypothetical value of a gene is called allelic value. Allelic values may vary on a definite set of values called allelic alphabet. In many application the allelic alphabet belongs to the {0,1} set. In the procedure adopted in this paper an allelic alphabet variable with the gene is used in relation to the possible design configurations of the correspondent link.

In particular three genes (fig. 1) are associated to each design link as previously described. The use of this representation consents the utilisation of small chromosomes to represent the configuration of the network. In this case the algorithm is more efficient than the representation of the chromosome with the alphabet {0,1}. The search is carried out by applying genetic operators reproduction, crossover and mutation.

According to the reproduction operator, the chromosomes of the population are reproduced using a Montecarlo method, according with reproduction probability proportioned to their fitness function values. Intuitively, the fitness function is a measure of profit, utility or assets to be maximised. In single criterion optimisation the fitness function is the value of a transformed objective function. In multicriteria optimisation the notion of objective function is not obvious. The method proposed by Goldberg (1989) and Fonseca and Fleming (1993) assumes the association of a rank to each solution, which it represents the number of solutions that dominated the solution plus one. The rank is used as fitness function value, which it may be the same for different solution. In this paper a different method to value the fitness function is proposed. The fitness function is obtained by transforming the Euclidean distance between each chromosome and the "ideal optimum solution", through a decreasing monotonic function. An "ideal optimum solution" is an ideal solution containing the best criteria value of current solutions. In general for the transformed function a negative exponential one is used. By means of this operation the reproduction probability is calculated by a Logit model.

The crossover operator (fig. 3) is repeated in each iteration the cardinality of genetic population multiplied by the crossover rate times and it takes as input two randomly chosen individuals and then combines them to generate two sons. The combination is performed by choosing, with uniformly distributed probability, two crossing points in the original chromosomes and then exchanging the allelic values between the two crossing points. In the crossover operator it is necessary to position the cut points in correspondence to genes referring to different links (fig. 4).

The mutation operator is repeated in each iteration the cardinality of genetic population multiplied by the mutation rate times ant it causes the change of an allelic value of a randomly chosen gene. This operator guarantees the possibility of exploring the whole search space independently from the specific initial population. The mutation must be executed without changing the total number of lanes in each link. For this reason in the mutation operation the random choice of chromosome, the random choice of a link and a new random feasible link configuration are successively executed (fig. 5).

After each crossover and mutation operation the connection of the network must be checked, by verifying if there is at least one route between each O-D pair. Genetic algorithms are different from traditional

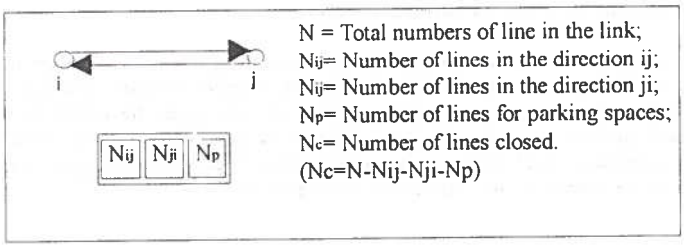


Fig. 1 - Part of chromosome relative to link ij

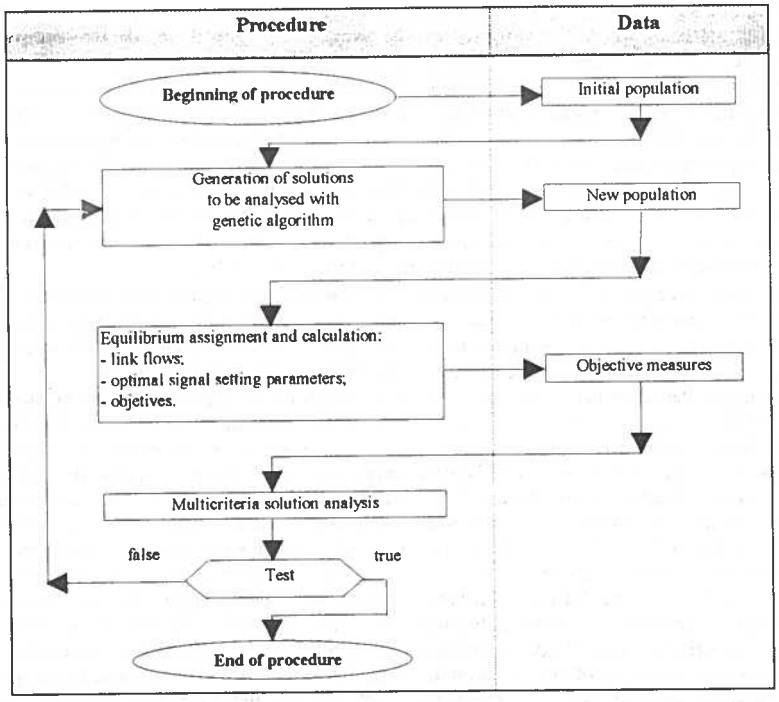


Fig. 2 - The global procedure of Urban Network Design Problem

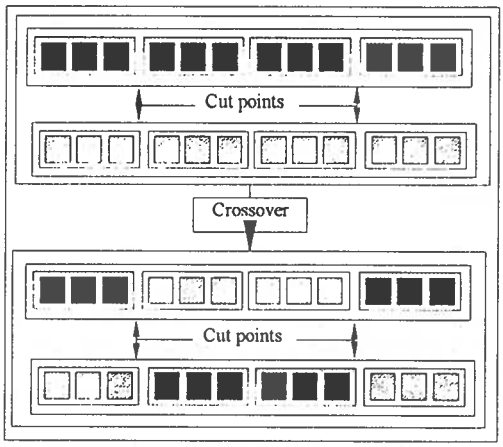


Fig. 3 - Crossover Operator

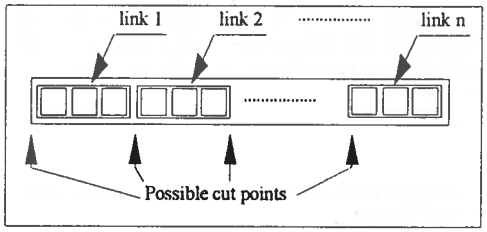


Fig. 4 - Chromosome and cut points

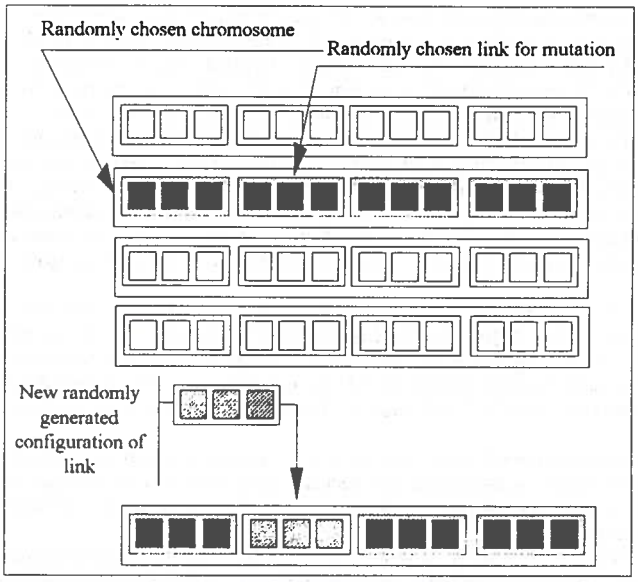


Fig. 5 - Mutation operator

methods because they search from a population of solutions not a single solution, and use probabilistic transition rules not deterministic ones. A genetic procedure can be temporarily interrupted to modify the current population in relation to other constraints not considered or decision goals. In general they are quicker than traditional ones.

4 NUMERIC APPLICATION

The global procedure has been applied to define the optimal configuration of the central private network of Messina (Italy). This area occupies a narrow coastal strip between the sea and the hills. This city has 280.000 inhabitants. The transportation demand has been computed through a four stage model calibrated with measured flows on the network. The supply is defined by a network model (fig. 6). The city is divided into two parts: the central design area and the suburb area. In the central area there are 100 design links and 40 signalised junctions. In the design area 21 congruence (see par. 2) design routes are used: 5 routes with 10 lanes, 2 routes with 5 lanes, 6 routes with 4 lanes, 3 routes with 3 lanes and 4 routes with 2 lanes. The number of possible allocations of n lanes in a link is given by $\sum_{i=0}^{n}(i+1)(i+2)/2$. In total, at maximum, $286^5*84^1*56^2*35^6*20^3*10^4$=7.41E34 different lane allocations are possible. The flows assignment has been performed through a Stochastic User Equilibrium with Probit route choice model. The signal setting is valued, for simplicity, through the Webster method (1958) assuming only isolated junctions.

In the genetic algorithm the following parameters are used (Cantarella and Vitetta 1994 b):

- randomly chosen initial population containing the present configuration;
- population cardinality 100;
- number of iterations 100;
- crossover rate 0.4;
- mutation rate 0.3.

In the multicriteria analysis, relative to the design area, three criteria are used:

- total user time spent on the network;
- global emission of CO which has been computed as a function of equilibrium traffic flows and equilibrium link speeds;
- number of parking spaces to be built to save the present number of parking spaces.

The application of the procedure generates 10 non-dominated solutions. In tab. 1 the values of criteria of non-dominated solutions and of the present state are reported with an increasing number of new parking spaces. In particular the first solution does not request a new parking spaces (park=0) and the values of total time and CO emission is a bit less than the correspondent values of the present state.

The solutions cover almost continually different values of parking spaces to be put in the network. The total time spent on the network and the CO emission in the different solutions increase simultaneously while the number of parking spaces decreases. In fig. 7, in a three-dimensional drawing, the different planes of solution criteria are represented and in fig. 8 the difference of criteria of non-dominated solutions compared to the present configurations of network are reported. It is shown that to reduce the total time spent on the private network and to reduce the CO emission, it is necessary to build new parking spaces and possibly to close present links.

In fig. 9 the ratio between the reduction of total user time spent and the total emission of CO in the design area and the number of parks to build is represented. Solution 1 is like the present state and it is no necessary to build new parking spaces but there is not appreciable reduction of the others criteria. The solutions 3 and 7 present the best reduction of time and CO per unit of parking spaces but solution 3 can be carried out with little investment and solution 7 with great investment (fig. 10). The worst solutions, of this point of view are the 2 and 10.

The solutions present few open links and a lot of the area is closed for vehicular movement. Generally the number of open routes in the design area decrease from solution 1 to solution 10. There is some route present in almost all the non-dominated configurations and there is generally a hierarchization of route in the network with some principal rings.

In tab. 2 the sets of solutions with the maximum percentage deviation between criteria less than the value present in the first column are reported. If the maximum percentage difference between the criteria value of the pairs of solution is less than 10%, it is not possible to identify large sets of similar solution. If the maximum percentage difference increases, it is possible to identify three subsets: similar solutions with few

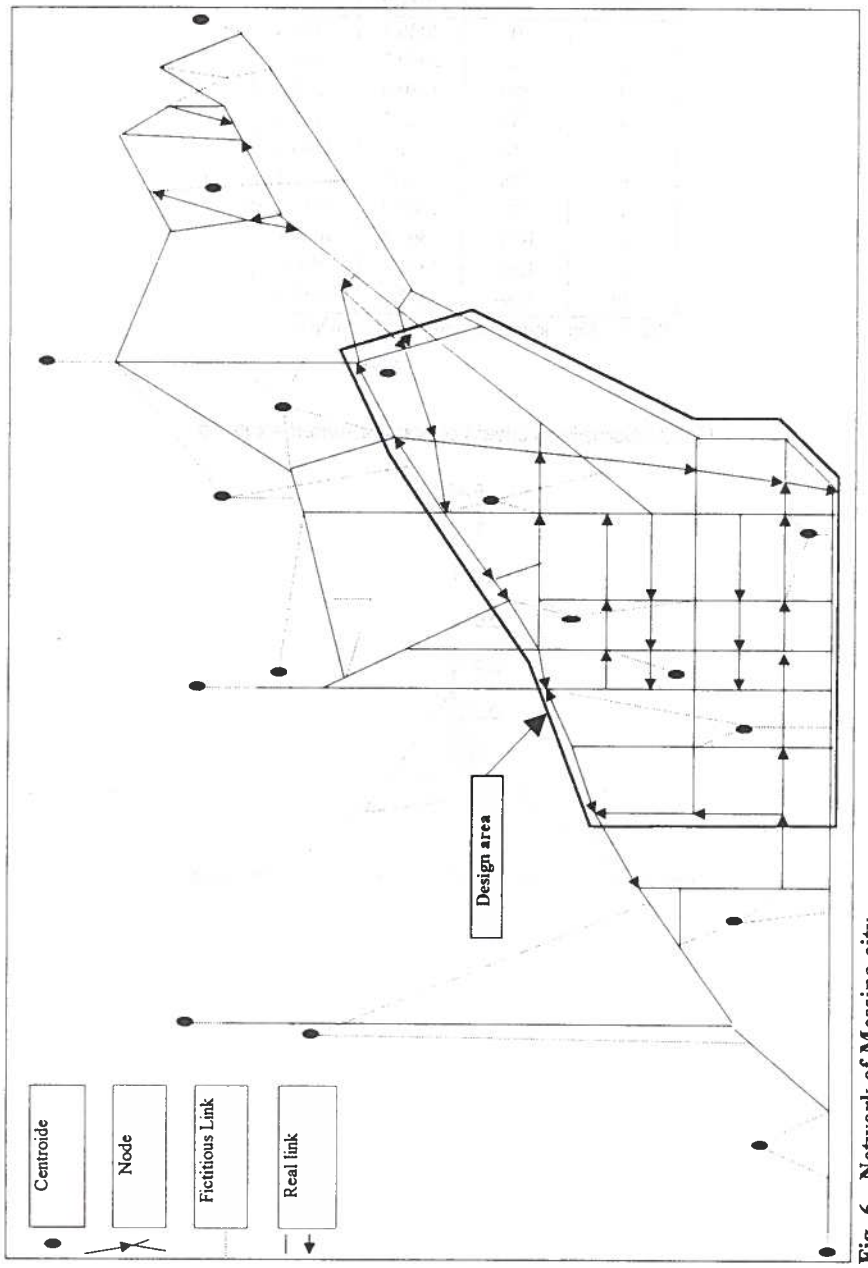


Fig. 6 - Network of Messina city

Solution number	PARKING spaces	TIME h	CO Kg/h
present	0	2047.9	1516.7
1	0	2026.1	1500.8
2	212	2000.7	1444.2
3	280	1780.9	1243.6
4	654	1723.7	1137.9
5	714	1711.9	1206.2
6	738	1717	1127.4
7	740	1561.2	1046.7
8	1058	1501.2	1007.1
9	1266	1477.6	993.3
10	2284	1462.9	1015.9

Tab. 1 - Non-dominate solutions design area

Fig. 7 - Normalised criteria of non-dominated solutions

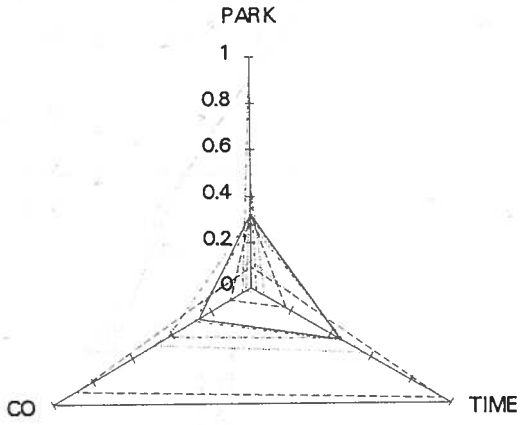


Fig. 8 Difference of criteria concerning non-dominate solutions respect to the present configuration of network

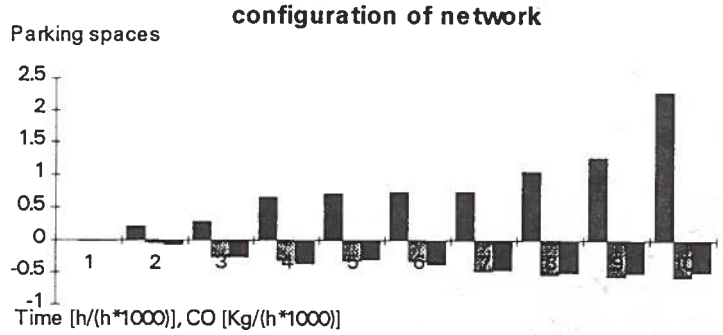
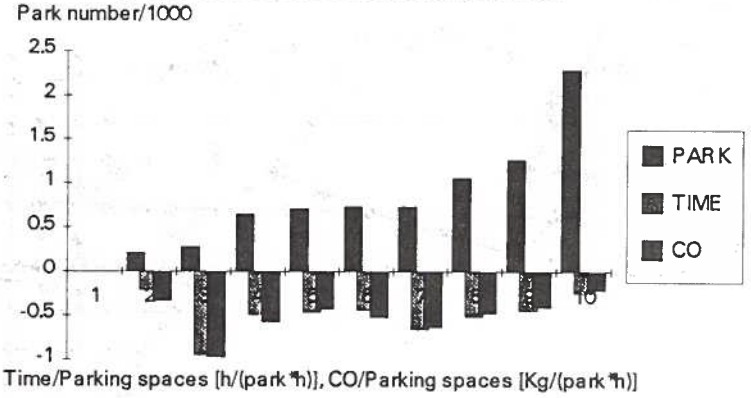


Fig. 9 - Rate between the decrement of Time or CO emission over parking spaces



	Solutions (1-10) and similar sets (a-g)									
Maximum percentage difference	1	2	3	4	5	6	7	8	9	10
< 10%				c	c-d	d-e	e			
< 20%				c	c	c	c	g	g	
< 30%	a	a	a	c	c	c	c	g	g	
< 40%	a	a	a	c	c	c	c-f	f-g	g	
< 50%	a	a	a	c	c	c	c	c	c-g	g
< 60%	a	a	a-b	b-c	c	c	c	c-g	c-g	g
< 70%	a	a	a	a-c	c	c	c	c	c	c
< 80%	a	a	a	a-c	a-c	a-c	a-c	a-c	c	c
< 90%	a	a	a-c	a-c	a-c	a-c	a-c	a-c	a-c	c
<100%	a	a	a	a	a	a	a	a	a	a

Tab. 2 - Similar solutions in relation to the maximum percentage deviation between criteria

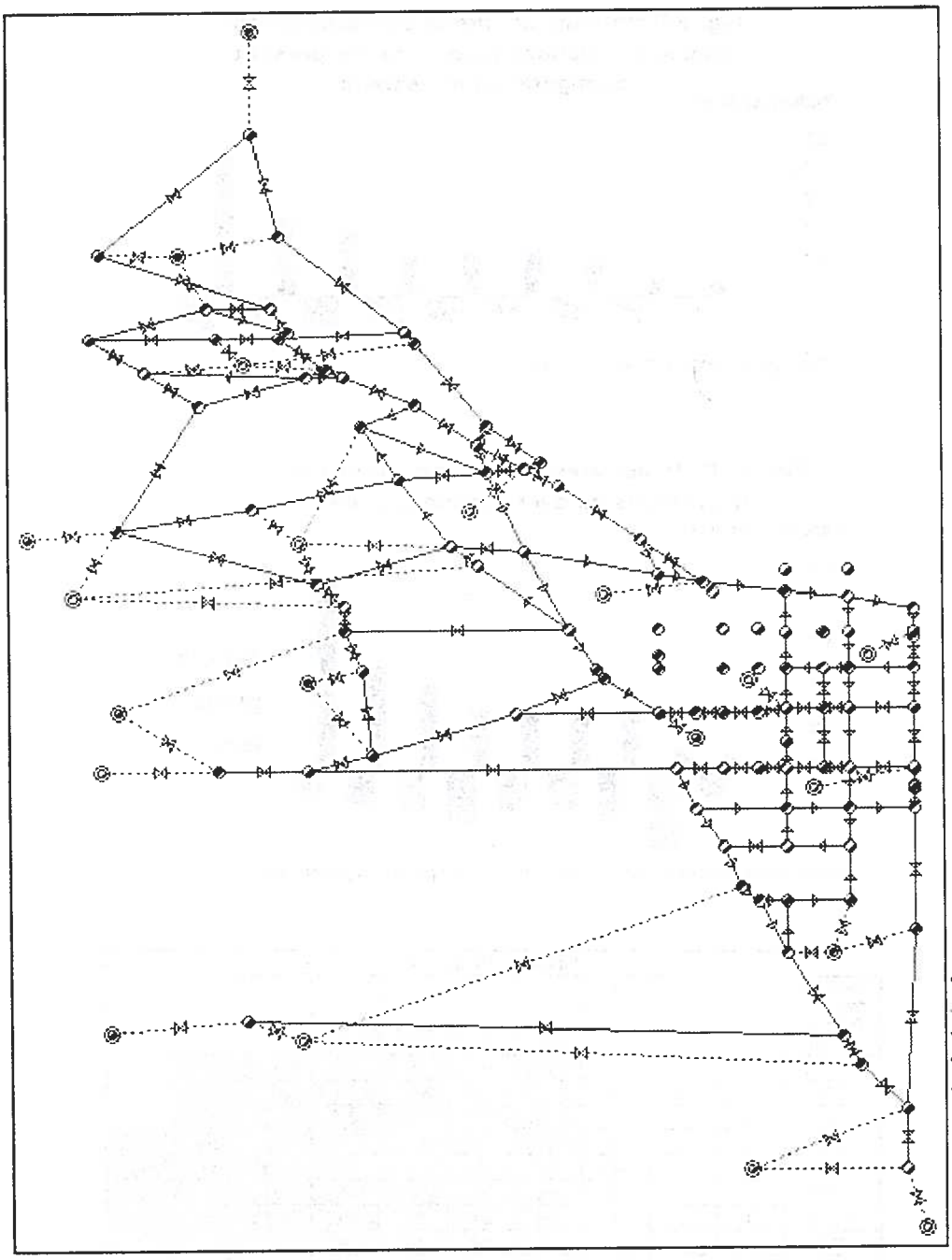


Fig. 10.A - Network of solution 3

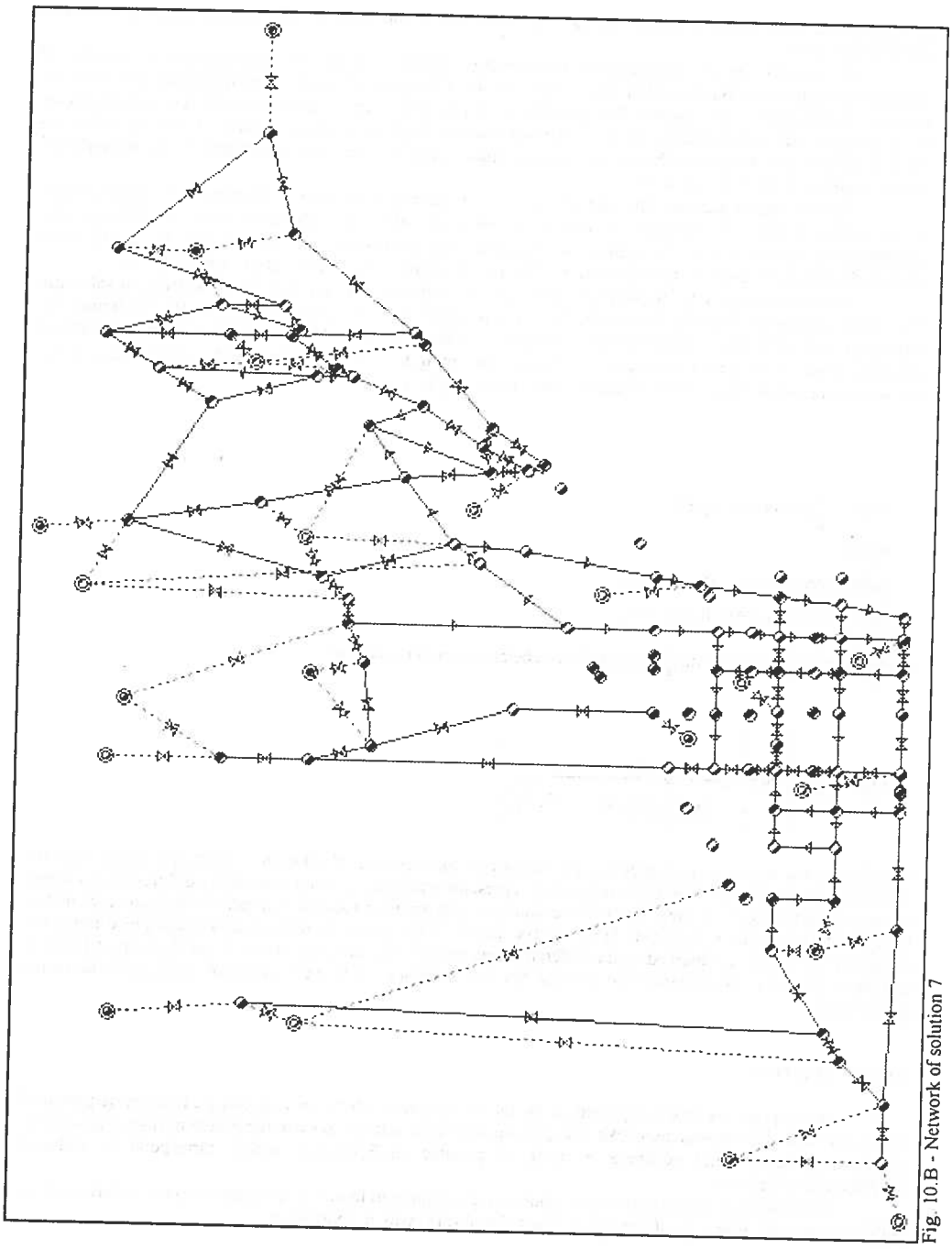


Fig. 10.B - Network of solution 7

parking spaces (1-3), solutions with many parking spaces (8-10) and solutions with an intermediate number of parking spaces (4-7).

An analysis for the classification of normalised criteria and for the classification of distance of different configurations between solutions is done. In tab. 3 the pairs of solutions with distance less than 0.3 (values 1 in the matrix) are reported. It is possible to identify three subsets: solutions with few parking spaces (1-2), solution with many parking spaces (7-10) and solutions with intermediate number of parking spaces (3-6). It is obvious that a high number of new parking spaces permits a low total time spent on the network and global emission of CO and vice versa.

Another cluster analysis has been carried out considering the number of identical lane configurations in the different pairs of solutions. In tab. 4 the pairs of solutions with fewer than 25 different lane configurations (elements 1 in the matrix) are reported. Also in this analysis it is possible to identify three classes of solution but there is more dispersion of data in the matrix than in the matrix reported in tab. 3.

In order to select a set of good solutions it can be useful to carry out a sensitive analysis of solutions. The different criteria adopted, for simplicity, are considered with the same weight. At the beginning the dominance and discordance indexes between couples of solution are used. Given a set S of N_S non-dominated solutions; given $\mathbf{o^i}$ the vector of criteria of solution i belonging to S with n_c components (number of criteria), the dominance index dom_{ij} of two solutions i and j belonging to S is given as:

$$dom_{ij} = \sum_{k=1}^{n_c} dominate(i,j,k)$$

where:

$$\begin{cases} dominate(i,j,k) = 1 & \text{if } o^i{}_k > o^j{}_k \\ dominate(i,j,k) = 0 & \text{if } o^i{}_k \le o^j{}_k \end{cases}$$

and the discordance index dis_{ij} of two solutions belonging to S is given as:

$$dis_{ij} = \max_{k=1..n_c} \left[\max \left(0, \frac{o^j{}_k - o^i{}_k}{\max_{h=1..N_s} \left[\max_{y=1..N_s} \left(o^h{}_k - o^y{}_k \right) \right]} \right) \right]$$

The dominance index dom_{ij} represents the number of total criteria of solution i which are better than the criteria of solution j; the discordance index dis_{ij} represents a measure of maximum utility difference of solution j compared to solution i, in the criteria where solution j is better than solution i. In tab. 5 the dominance matrix is reported and in tab. 6 the discordance matrix. In tab. 7 the global dominance and discordance index are reported (the index i is obtained as the difference between the sum of terms in row i and the sum of terms in column i). Solution 10 appears to be the best but also solutions 1,2,8 and 9 occupied good positions in the global results.

5 CONCLUSIONS

In this paper an heuristic procedure for the multicriteria Urban Network Design Problem is presented and it is suitable for real-size networks. The genetic algorithm, utilised to solve the problem, finds a set of non-dominated solutions and generates a range of possible configurations which correspond to different combinations of criteria.

In the future a more detailed flow model can be adopted to improve the computation of criteria such as pollutant emission. Elasticity of demand (at least modal split) will be considered.

1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0
0	0	1	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	0	0	0	1

TAB. 3 - Pairs of solutions (elements 1) with normalised distance less than 0.3

1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0	0
0	0	0	1	0	1	0	1	0	0
0	0	0	0	0	0	1	1	1	0
0	0	0	0	1	0	1	1	1	1
0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	0	1	0	1

Tab. 4 - Pairs of solution (elements 1) with less than 25 different line configurations

0	1	1	1	1	1	1	1	1	1	0
2	0	1	1	1	1	1	1	1	1	0
2	2	0	1	1	1	1	1	1	1	0
2	2	2	0	2	1	1	1	1	1	0
2	2	2	1	0	2	1	1	1	1	0
2	2	2	2	1	0	1	1	1	1	0
2	2	2	2	2	2	0	1	1	1	0
2	2	2	2	2	2	2	0	1	2	0
2	2	2	2	2	2	2	2	0	2	0
2	2	2	2	2	2	2	1	1	0	0
18	17	16	14	14	14	12	10	9	11	

Tab. 5 - Dominance matrix between solutions

0.0	0.1	0.5	0.7	0.6	0.7	0.9	1.0	1.0	1.0	6.5
0.1	0.0	0.4	0.6	0.5	0.6	0.8	0.9	0.9	1.0	5.8
0.1	0.0	0.0	0.2	0.1	0.2	0.4	0.5	0.5	0.6	2.7
0.3	0.2	0.2	0.0	0.0	0.0	0.3	0.4	0.4	0.5	2.3
0.3	0.2	0.2	0.1	0.0	0.2	0.3	0.4	0.4	0.4	2.6
0.3	0.2	0.2	0.0	0.0	0.0	0.3	0.4	0.4	0.5	2.3
0.3	0.2	0.2	0.0	0.0	0.0	0.0	0.1	0.2	0.2	1.2
0.5	0.4	0.3	0.2	0.2	0.1	0.1	0.0	0.0	0.1	1.9
0.6	0.5	0.4	0.3	0.2	0.2	0.2	0.1	0.0	0.0	2.5
1.0	0.9	0.9	0.7	0.7	0.7	0.7	0.5	0.5	0.0	6.5
3.5	2.8	3.3	2.9	2.3	2.8	4.0	4.3	4.4	4.1	

Tab. 6 - Disconcordance matrix between solutions

Solution	dominance	position	discordance	position
1	-9		**3.0**	**1**
2	-7		**3.0**	**1**
3	-5		-0.6	
4	-1		-0.6	
5	-1		**0.2**	**4**
6	-1		-0.5	
7	3	4	-2.8	
8	7	**2**	-2.4	
9	9	**1**	-1.9	
10	**5**	**3**	**2.4**	**3**

Tab. 7 - Global dominance and discordance index

BIBLIOGRAPHY

R. E. ALLSOP (1971): Delay-Minimizing Setting for Fixed-Time Traffic Signal at Single Road Junction. J. Inst. Maths Applics.

R. E. ALLSOP (1972): Delay at a Fixed Time Traffic Signal. I. Theoretical Analysis. Transp. Science.

R. E. ALLSOP, J. A. CHARLESWORTH (1977): Traffic in a signal controlled road network: An example of different signal timings including different routing. Traffic Eng. Control 18, 262-4

J. W. BILLHEIMER, P. GRAY (1973): Network Design with Fixed and variable Cost Elements. Trans. Sci. 7, 49-74.

G. E. CANTARELLA, A. SFORZA (1987): Methods for Equilibrium Traffic Signal Setting. Flow Control of Congested Networks. Vol. F38.

G. E. CANTARELLA, A. IMPROTA, A. SFORZA (1991): Road Network Signal Setting: Equilibrium Condition. Concise Encyclopedia of Traffic & Transportation Systems. Editor Papageorgiou. Pergamon Press, pag. 366-371.

G. E. CANTARELLA, F. VIOLA, A. VITETTA (1994 a): Un modello per il progetto di una rete di trasporto urbano. Atti convegno SIDT 1994. Roma, Università "La Sapienza".

G. E. CANTARELLA, A. VITETTA (1994 b): Algoritmi genetici per il progetto delle reti di trasporto urbane. Ricerca Operativa N. 69.

M. CHEN, A. S. ALFA (1991): A Network Design Algorithm Using a Stochastic Incremental Traffic Assignment approach. Transportation Science 25.

C. M. FONSECA, P. J. FLEMING (1993): Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. Prooceding of the fifth international conference on Genetic Algorithms. Morgan Publishers, USA.

L. R. FOULDS (1981): A Multi-Commodity Flow Network Design Problem. Transpn. Res. 15 B, 273-283.

N. H. GARTNER (1972). Constraining Relations Among Offsets in Synchronized Networks. Transp. Sci. 67: 88-93.

N. H. GARTNER, J. D. C. LITTLE, H. GABBAY (1975): Optimization of Traffic Signal Settings by Mixed-Integer Linear Programming. Part I: The Network Coordination Problem; Part II. Trans. Sci., 9: 321-363.

N. H. GARTNER, S. B. GERSHWIN, J. D. C. LITTLE, P. ROSS (1980): Pilot Study of Computer-Based Urban Traffic Management. Trans. Res., 14B: 203-217.

D. C. GAZIS (1964): Optimum Control of a System of Oversatured Intersection. Ops. Res.

D. C. GAZIS, R. B. POTTS (1965): The Oversatured Intersection. Proc. of the 2nd Int. Symp. on the Theory of Road Traffic Flow, Organization for Economic Cooperation and Development, Parigi.

D. E. GOLBERG (1989): Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

J. A. HILLIER (1966): Appendix to Glasgow's Experiment in Area Traffic Control. Traffic Eng. Control, 7: 569-571.

L. LE BLANC, D. E. BOYCE (1986): A Bilivel Programming Algorithm for exact Solution of the Network Design Problem with User-Optimal Flows. Transpn. Res. 20 B, 259- 265.

H. POORZAHEDY, M. A. TURNQUIST (1982): Approxsimate Algorithms for the Discrete Network Design Problem. Transpn. Res. 16 B, 45-55.

C. SUWANSIRIKUL, T. L. FRIESZ, L. T. TOBIN (1987): Equilibrium Decomposed Optimization: A Heuristic for the Continuous Equilibrium Network Design Problem. Tansportation Science 21, 254-263.

F. W. WEBSTER (1958): Traffic Signal Settings. Road Res. Technical Paper n. 39.

F. W. WEBSTER, B. M. LOBBE (1966) Traffic Signals. Ministry of Transport, Road Research Technical Paper n. 56, HMSO Londra.

Network Design Problems with Applications to Transportation Industries*

Patrick Jaillet† Gang Yu‡ Mark Song§

September 1993

Abstract

Despite the popularity of the hub-and-spoke networks (in air transportation, but also in ground transportation, satellite communication, telephone networks and other logistic systems), no systematic mathematical analysis has been conducted in order to (i) justify and evaluate the current hub-and-spoke system; (ii) provide an optimal network structure; and (iii) suggest alternative routing policies. Most of the literature has been concerned with finding the location of hubs and designing routing policies for hub-and-spoke systems. In this paper, we present integer linear programming models as an effort to answer the above questions.

EXTENDED ABSTRACT

1 Introduction, Objectives and Contributions

Since the 1978 Airline Deregulation Act, profound changes have taken place in the U.S. air transportation. During the past 14 years, the average real airfare has declined by 15%, the number of air travelers has almost doubled, and the number of flights has increased by one-third (see [6]). The Deregulation Act allowed airlines to choose the routes to serve and the fares to charge. This, together with the intensive competitions among airlines, and between airline and other transportation industries, has greatly affected airline operations, including airfares, service, quality and safety. One of the most important innovation has been the hub-and-spoke system implemented by U.S.

*Submitted to Tristan II

†MSIS Department, The University of Texas at Austin, Austin, TX 78712-1175, and Laboratoire de Mathématiques Appliquées (affiliated CNRS), ENPC, Paris, France.

‡MSIS Department and the Center for Cybernetic Studies, The University of Texas at Austin, Austin, TX 78712-1175.

§MSIS Department, The University of Texas at Austin, Austin, TX 78712-1175.

airline industry. This development is based on the idea that the transportation cost could be reduced by consolidating passengers from different origins to different destinations across the hubs. Essentially, the centralization and the broader scope of operations allow airlines to take advantage of the economy of scale. The successful application of this concept has not only been present in air transportation, but also in ground transportation, satellite communication, telephone networks, and other logistic systems (see [4] for one example).

Despite the popularity of the hub-and-spoke systems, very few studies have appeared in the literature on these systems. In fact, no systematic mathematical analysis we are aware of has been conducted in order to (i) justify and evaluate the current hub-and-spoke system; (ii) provide an optimal network structure; and (iii) suggest alternative routing policies. Most of the existing literature fall into two categories: a) empirical studies in airline economics analyzing the advantages of hub-and-spoke networks (see for example [2, 3, 7]); and b) operations research analyses solely concerned with finding the optimal location of hubs and designing routing policies within a given hub-and-spoke network (see [1, 8, 9]).

In this paper, we study the three previous questions in the unified framework of Network Design. We have formulated mixed integer linear programming models to answer the following questions: 1) What type of network structure should an airline adopt such that it economically serves its market demand; 2) What types of aircraft should an airline purchase/lease in order to optimally accomplish its service; 3) What routing strategies should an airline apply. The models use as input the origin-destination flow demand matrix, the capacity and mileage cost of different types of aircraft, and the fixed cost for purchasing or leasing aircraft. With the above information, our goal is to design a network which minimizes the total transportation and aircraft purchasing/leasing cost. This particular formulation has the quality of being both very compact and general enough (this has been one of the challenges that kept this problem from being tackled in the past). Let us indicate the spirit of our approaches in the next section.

2 Mixed Integer Linear Programming Models

We first describe a very simple model with only one type of aircraft and at most one intermediate stop for each flight, then we present more complex models involving more intermdiate stops and several types of aircrafts.

2.1 A Simple Model

As a simplification, we consider the network design problem from the point of view of one airline company that has to serve a given fixed share of the market. We thus assume that the demand is given and does not depend on the resulting design of the network. This would be the case of a monopoly, for which the demand is captive (e.g., Olympic Airways having to serve all the Greek Islands). The network design problem can thus be stated as follows:

> Given a fixed origin-destination flow demand matrix, the capacity and mileage cost of different types of aircraft, and the fixed cost for purchasing or leasing aircraft, design a network which satisfies the demand and minimizes the total transportation and aircraft purchasing/leasing costs.

Let us define the inputs to the model:

K = number of aircraft types (e.g. Boeing 747's or DC 10's).

n = number of cities the airline serves.

f_{ij} = number of passengers who desire to fly from city i to city j per day by the designated airline.

d_{ij} = air distance from city i to city j.

c_k = cost per mile for aircraft type k.

b_k = capacity for aircraft type k.

F_k = fixed cost for aircraft type k per day. This cost can be interpreted as the aircraft leasing cost or the depreciation cost of the airline's own aircraft.

Assume that the airline offers two types of service: nonstop flights and one-stop flights. This very simple model already includes all the basic ingredients of the design problem, and will be able to show the advantages, if any, of the hub-and-spoke network. Under this assumption, the following decision variables are defined.

Type 1: Variables for selection of routes

x_{ij} = fraction of the flow f_{ij} served by a direct flight from i to j.

x_{ilj} = fraction of the flow f_{ij} served by an indirect flight from i to an intermediate city l and then to j.

Type 2: Variables for choice of aircraft

y_{ij}^k = number of aircraft of type k used on the link from city i to city j.

With the above variables, we have the following mixed integer linear programming model.

$$(NW1) \qquad \min \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{K} d_{ij} c_k y_{ij}^k$$

subject to

$$f_{ij} x_{ij} + \sum_{l=1}^{n} f_{lj} x_{lij} + \sum_{l=1}^{n} f_{il} x_{ijl} \leq \sum_{k=1}^{K} b_k y_{ij}^k \qquad i,j = 1,...,n \tag{1}$$

$$x_{ij} + \sum_{l=1}^{n} x_{ilj} = 1 \qquad i,j = 1,...,n \tag{2}$$

$$x_{ij}, x_{ilj} \geq 0 \qquad i,j,l = 1,...,n \tag{3}$$

$$y_{ij}^k \in Z_+ \qquad i,j = 1,...,n; k = 1,...,K. \tag{4}$$

The objective function minimizes the total transportation cost for the airline. Constraints (1) limit the total capacity for a selected set of aircraft to carry demand for every pair of cities i and j. Constraints (2) imply that all demand is satisfied. Note that the following is not accounted for in this simplified model:

1. Fixed cost for purchasing/leasing aircraft.

2. Limit on total number of available aircraft. Omission of this limit is not an unreasonable approximation since we have assumed the possibility of leasing aircraft at any city by any amount. The objective function also gives an incentive to reduce the number of aircraft.

3. Periodic airline operations. The airline operations should be scheduled in a regular bases. After a certain length of time, crews have to return to their base.

4. Two or more intermediate stopping cities from origin to destination. This issue will be discussed in the next subsection.

However, this model does catch the essence of our intended goal in the following sense: (i) If a hub-spoke system is beneficial, we believe this model will reveal the pattern; (ii) This model will provide decisions both for routing the passengers and for selections of aircraft.

It is in fact quite easy to modify this simple model in order to include the fixed cost of purchasing/leasing aircraft and in order to consider of aircraft connections and periodic operations of airlines. We have developed such an improved model ($NW2$) and will also present computational results for it as well.

2.2 More Complex Models

We now model the case with possibly more than one intermediate stops for each flight. Let P_{ij} be a path from city i to city j and let the number of cities traversed by the path be $|P_{ij}|$. Mathematically,

P_{ij} can be represented by a sequence $v_1, v_2, ..., v_{|P_{ij}|}$ with $v_1 = i, v_{|P_{ij}|} = j$. We use expression $(i,j) \in P_{kl}$ to indicate that the arc (i,j) is in the path P_{kl}, or i and j are two consecutive cities in the path. The extension of model $(NW1)$ is then easy to write. Let $(NW3)$ be the resulting model. Since a path may involve up to n cities, this model may have many variables. However, we believe that the model with at most two intermediate cities is adequate to describe realistic situation, since more intermediate stops for a flight will make the travel unacceptable and may thus lose its constomers.

Model $(NW3)$ with unlimited path length can actually be solved by a transformation to $(NW1)$ with one additional constraint. We now describe the transformation. Let k^* be the most economic aircraft; i.e., the cost per seat c_{k^*}/b_{k^*} is lowest for aircraft type k^*. In model $(NW1)$, we add a constant demand for each pair of cities: $\bar{f}_{ij} = f_{ij} + mb_{k^*}$, where m is a large integer, say $mb_{k^*} \geq \sum_{ij} f_{ij}$. We also add a constraint

$$y_{ij}^{k^*} \geq m \qquad i,j = 1, ..., n. \tag{5}$$

We have the following theorem.

Theorem 1 *Let the solution to $(NW1)$ with the transformed demand $\bar{f}_{ij}$ and the additional constraint (5) be $\bar{x}_{ij}, \bar{x}_{ijl}, \bar{y}_{ij}^k$, then $\hat{y}_{ij}^k = \bar{y}_{ij}^k - m$ is the optimal solution to $(NW3)$ with unlimited path length.*

Using the above theorem, the unlimited path length case can be solved by the following procedure:

Step 1:Solve $(NW1)$ to get solution $\hat{y}_{ij}^k$.

Step 2: Use the longest path in the graph determined by nonzero $\hat{y}_{ij}^k$ to limit the possible path length.

Step 3: Solve $(NW3)$ as linear program with y_{ij}^k fixed at $\hat{y}_{ij}^k$ and the path length limited by what has been found in step 2.

References

[1] Aykin, T. (1990). On a Quadratic Integer Program for the Location of Interacting Hub Facilities. *European Journal of Operations Research*, 46, 409–411.

[2] Bailey, E., D. Graham, and D. Kaplan. (1985). *Deregulating the Airlines.* The MIT Press. Cambridge.

[3] Bauer, P. (1987). Airline Hubs: A Study of Determinant Factors and Effects. *Economics Review - Federal Reserve Bank of Cleveland*, Forth Quarter, 13–19.

[4] Brandon, W. (1987). Opportunities for Small, Low Cost, SHF Satellite Communications Terminal. *IEEE Mil. Conf. Proc.*, 265–270.

[5] Garey, M. and D. Johnson. (1979). *Computers and Intractibility: A Guide to the Theory of NP-completeness.* Freeman. San Francisco.

[6] Hahn, R. and R. Krosner. (1989). The Mismanagement of Air Transport: A Supply-Side Analysis. *The Public Interest*, **95**, 100–111.

[7] Morrison, S. and C. Winston. (1986). *The Economics Effects of Airline Deregulation.* The Brooking Institute. Washington DC.

[8] O'Kelly, M. (1986). The Location of Interacting Hub Facilities. *Transportation Science*, **20**, 92–106.

[9] O'Kelly, M. (1987). A Quadratic Integer Program for the Location of Interacting Hub Facilities. *European Journal of Operations Research*, **32**, 393–404.

The Traveling Salesman Problem with Delivery and Backhauls

by
Shoshana Anily
Faculty of Management
Tel-Aviv University
Israel

and

Gur Mosheiov
School of Business Administration
and the Statistics Department
The Hebrew University
Jerusalem
Israel

Abstract

The problem that we consider here deals with a single vehicle of a given capacity that needs the others need to deliver stock from their location to the warehouse. The objective is to find a shortest feasible tour visiting all customers, emanating from and ending at the warehouse. We introduce an efficient ($O(N^2)$) heuristic for this problem. The heuristic improves the worst-case bound known in the literature from 2.5 to 2. However, its average performance is shown in our numerical study to be slightly worse than that of a previously published ($O(N^3)$) solution method.

Keywords: Traveling Salesman Problem, Vehicle Routing, Heuristic, Worst-case Analysis, Computational Complexity.

1. Introduction:

The problem we study here deals with a single period, single warehouse distribution system with two sets of customers dispersed in the plane. The first set, denoted by D, contains "delivery customers" that require a delivery of goods from the warehouse, whereas the second set, denoted by B, contains "backhaul customers" that need to deliver goods from their location to the warehouse. (Note that the literature distinguishes between "backhaul" customers and "pick-up" customers: a "pick-up" customer may deliver goods to any of the other locations including the warehouse. In this research we restrict ourselves to delivery and backhaul customers only). In order to avoid excess notation we will assume a single commodity problem (however the results can be easily generalized to the multiple commodity case). Each customer i is characterized by its geographic location and its requirement size d_i (b_i) units for delivery (backhaul). The stock is delivered and backhauled by a single vehicle of limited capacity. The distance function between distinct locations is assumed to obey the triangle inequality. The objective is to design a route of minimum total length so that all customers are served: the desired route should emanate from the warehouse, stop at customers for delivery/backhaul purposes and finally terminate at the warehouse without violating the vehicle's capacity restriction along the route. Note that the vehicle is not allowed to visit the warehouse in between its initial and final stops there; also, backhaul stock cannot be used to satisfy delivery requirement. Thus, the resulting tour is a closed tour on the set of customers and the warehouse that visits each location exactly once. We call this problem the Traveling Salesman Problem with Delivery and Backhaul (TSPDB).

The multiple vehicle case which we call the Vehicle Routing Problem with Delivery and Backhaul (VRPDB) and we address in a later paper, has many real life applications. Bodin et al.[1] and Casco, Golden and Wasil [2] describe some applications and review some solution techniques.

The problem is faced, for example, by mailing parcel services as UPS: trucks load parcels at the warehouse and on their way they unload parcels at addressees and collect parcels from senders. All parcels collected must be processed at the warehouse before being delivered to the addressees. Another application described by Casco, Golden and Wasil [2], deals with the grocery industry in the USA: this industry has recognized the cost-cutting potential of servicing backhaul points on predominantly delivery routes. As the authors report, the industry has saved upwards of $160 million a year since 1982 on its distribution costs by allowing vehicles on their delivery routes to collect large volumes of inbound products. Mosheiov [5] reports about a different application where inner-city under privileged children are transported to summer vacations at volunteer families living out of town. The transportation service is made by buses and is confined to a limited number of days when some children start their vacation and others end theirs. In all of these problems, if a heuristic based on cluster-first-route-second is applied then the routing problem related to each of the clusters is a TSPDB type.

There is a vast literature on the Traveling Salesman Problem (TSP) with or without additional constraints, however the problem discussed here has obtained very little attention in spite of its applicability. It is easy to see that if $D=\phi$ or $B=\phi$ then the TSPDB reduces to the classical TSP, proving that the problem is NP-hard (see Garey and Johnson [4]). As such, exact solution methods become impractical when the problem size increases, arising the need for developing heuristic solution methods whose running time is polynomial in the problem size. In this paper we propose a heuristic for the TSPDB and prove its worst-case performance: a heuristic is said to have a worst-case bound of $1+\alpha$ if for any data set the heuristic will produce a solution whose cost does not exceed $100(1+\alpha)\%$ the optimal cost. (For example, the best known heuristic in terms of the worst-case bound for the classical TSP, is Christofides' algorithm (see Christofides [3]), whose

worst-case bound is 1.5). Mosheiov [5] in his paper, considers the TSPDB: he shows that any closed tour on $D \cup B$ contains a certain (initial) point and a certain direction so that starting at that point with a vehicle loaded with the total amount to be delivered and following the tour at that direction, while delivering (loading) stock at delivery (backhaul) customers as required, will not violate the capacity restriction at any point on the tour. Therefore, there exists an arc and a direction on any closed tour on $D \cup B$ so that disconnecting the arc and connecting its two end points to the warehouse will result in a feasible tour. As a consequence any heuristic for the TSP whose worst-case bound is $1+\alpha$ may be used for constructing the initial tour over $D \cup B$ and by searching the arc that can be disconnected and connecting its two end points to the warehouse will generate a feasible tour for the TSPDB whose worst case bound is $2+\alpha$. Since Christofides's heuristic provides the best known worst-case bound (of 1.5) for the TSP, applying it as suggested above will result in a heuristic for the TSPDB whose worst-case bound is 2.5 and complexity of $O(N^4)$. In this paper we propose the 2MST heuristic for the TSPDB which is based on doubling a minimum spanning tree. We prove that the worst-case bound of the 2MST is 2 which is an improvement of Mosheiov's [5] bound. In terms of complexity, 2MST requires $O(N^2)$ operations, which is the size order of the data set. Similarly to the TSP, the average performance of various heuristics may be far away from their worst-case bound: e.g., Christofides algorithm, which became a milestone in the theory of the TSP, is rarely used in practice. In our computer study, one version of a heuristic proposed by Mosheiov [5]- the Cheapest Insertion with Delivery and Backhaul (CIDB) which has a worst case bound of 3, usually outperforms 2MST.

The paper is organized as follows: in Section 2 we introduce the notation and preliminaries. In Section 3 we propose the heuristic for the TSPDB. We conclude with Section 4 that presents an example and a short computer study in which we compare our heuristic to the heuristic proposed in

Mosheiov [5].

2. Notation and Preliminaries:

D (B)= the set of delivery (backhaul) customers.

d_i = the demand size of delivery customer i, $i \epsilon D$.

b_i = the amount to be loaded at backhaul customer i, $i \epsilon B$ (d_i and b_i are assumed to be non-negative integers).

Note that if a location i serves both as a delivery and a backhaul customer, the overall effect on the vehicle's load of serving customer i is the difference between its backhaul size and its delivery size. When such a customer is visited then first the vehicle should be unloaded by its delivery size and then should be loaded by its backhaul size. Thus, we can assume without loss of generality that i is a delivery customer if its demand size is greater or equal to its backhaul size and that i is a backhaul customer otherwise. The delivery (backhaul) size should be set to the absolute value of the difference between its delivery and backhaul sizes. Observe that it is not possible to ignore customers for which the above difference is 0 since the location cannot use its own stock to cover its demand and thus must be visited by the vehicle, even though the visit does not effect the vehicle's load.

$N_D = \Sigma_{i \epsilon D} d_i$ ($N_B = \Sigma_{i \epsilon B} b_i$) total amount to be unloaded (loaded) at delivery (backhaul) customers.

$N = D \cup B$ the total set of customers. We also use N to denote the total number of customers. Note that for each $i \epsilon N$ either $d_i=0$ or $b_i=0$.

We denote the warehouse by 0.

$N^0 = N \cup \{0\}$, the set of customers and the warehouse.

Throughout the paper we assume that the triangle inequality holds, i.e. for any $i,j,k \epsilon N^0$:

$d(i,k) \leq d(i,j)+d(j,k)$.

q = vehicle's capacity.

Since we are not allowing the vehicle to return to the depot in between the beginning and the end of the tour, we require, in order to ensure feasibility, that both N_B and N_D do not exceed the vehicle capacity q.

Let G(N,A) be a graph on the set of nodes N, and the set of arcs $A \subseteq NXN$. Let $l(G)$ denote the total length of the arcs of G. The graph G may be directed or undirected. In an undirected graph $(i,j) \epsilon NXN$ represents an arc connecting i and j. In a directed graph we indicate the arc direction by using the symbol $i \rightarrow j$ to denote an arc connecting i to j. An undirected tree is a connected graph that does not contain any cycles. Each rooted tree is a tree with a distinguished node which is called the root. The depth of a node in a given rooted tree is the number of arcs on the path connecting the root to the node. In a rooted directed tree with all arcs directed away from the root, each node i except the root, is associated with a single node k so that the arc $k \rightarrow i$ is an arc of the tree; node k is said to be the parent of i. All nodes j that are connected to i via an arc $i \rightarrow j$ are called the children of i. The leaves of a tree are all the nodes which do not have any children.

For any given problem, we let V^{OPT} represent its minimum cost. If heuristic H is applied to solve the problem, the cost generated is denoted by V^H.

3. The 2MST Heuristic:

In this section we propose a simple polynomial heuristic for the TSPDB whose worst-case bound is 2. We follow a common procedure where at the first step we solve a relaxation of the problem and in the second step we extend the solution obtained at step 1, into a feasible solution while increasing its cost by a constant factor.

The proposed heuristic, called hereafter 2MST, first constructs a Minimal Spanning Tree (MST) on the set of nodes N^0, which is obviously a relaxation of the TSPDB. In the second step we

take two copies of the MST. We will show below that it is possible to transform the two copies of the MST into a feasible tour without additional cost. A similar heuristic for the TSP is well known, however in the TSPDB the feasibility issue in terms of the vehicle's capacity constraint needs to be considered as well. We first present the algorithm, and then explain its various steps.

Algorithm 3.1 (Heuristic 2MST)

Step 0: let TOUR be an array of length N where TOUR(ℓ) denotes the ℓ-th location served by the vehicle after the vehicle leaves the warehouse. Set $\ell:=1$. Also let I(v) be an indicator variable for $v\epsilon B$ assuming the value 1 if location v was already served and 0 otherwise. Set $I(v):=0$ for all $v\epsilon B$.

Step 1: construct a minimum spanning tree on N^0 by using the Greedy algorithm. Let MST denote the directed tree obtained, rooted at 0, where all arcs are directed towards the leaves. For each $i\epsilon N$ let C(i) represent the set of children of node i, i.e. $C(i)=\{k:i\rightarrow k\epsilon MST\}$. Let also p(i) represent the parent of node i for $i\epsilon N$, i.e. $p(i)=\{k:k\rightarrow i\epsilon MST\}$.

Step 2: scan the nodes of the MST from the leaves to the root as follows: if node i is a leaf then let $e(i)=-d_i$ in case i is a delivery customer, and otherwise let $e(i)=b_i$. For any node j which is not a leaf let $e(j)=b_j-d_j+\Sigma_{k:j\rightarrow k}\, e(k)$. ($e(0)=N_B-N_D$.)

Start at 0 with a vehicle loaded with N_D units and set i=0.

Step 3: Set C'(i)=C(i);

While $C'(i)=\phi$ do begin

if i=0 stop.

if {$i\epsilon B$ and I(i)=0} do begin

load b_i units at i; set TOUR(ℓ):=i; $\ell\leftarrow\ell+1$; I(i):=1; end;

set $i\leftarrow p(i)$; endwhile;

Select node $v \epsilon C'(i)$ so that $e(v) = \min_{k \epsilon C'(i)} e(k)$. Set $C'(i) \leftarrow C'(i) - \{v\}$.

If $v \epsilon D$ do begin

unload d_v units at v; set $TOUR(\ell) := v$; $\ell \leftarrow \ell + 1$; end;

If $\{v \epsilon B$ and $C'(v) = \phi\}$ do begin

load b_v units at v; set $TOUR(\ell) := v$; $\ell \leftarrow \ell + 1$; $I(v) := 1$; end;

$i \leftarrow v$; repeat Step 3.

end of algorithm.

As explained below at the end of the algorithm, array TOUR consists of exactly N different locations. The scanning procedure of the tree as proposed by the algorithm traverses twice through each of its arcs: firstly, from its tail to its head and secondly, from its head to its tail. We note that each node i is visited exactly $1 + |C(i)|$ times by the algorithm where C(i) is the set of children of node i in the MST. Initially, we set $C'(i) = C(i)$. At each visit at i one element of the set C'(i) is deleted. The last visit at i occurs when $C'(i) = \phi$. If i is a delivery customer then it is served at the first visit of the algorithm at i and if it is a backhaul customer it is served only at the last visit of the algorithm at i. Thus each node except 0 is traversed by the algorithm $|C(i)|$ times without performing any work. In the array TOUR we keep by the order of their occurrence only those visits where a loading or unloading is taking place.

In view of the above considerations it is easy to see that the algorithm's complexity is determined by the complexity of generating a minimal spanning tree which is $O(N^2)$.

In the next theorem we prove that starting at the warehouse with a vehicle loaded with N_D units and following the tour $0 \rightarrow TOUR(1) \rightarrow TOUR(2) \ldots TOUR(\ell-1) \rightarrow TOUR(\ell) \ldots TOUR(N) \rightarrow 0$ where at each location the vehicle is unloaded or loaded according to the customer's requirement produces a feasible tour, i.e. the total load of the vehicle never exceeds the vehicle's capacity nor it falls below

0.

Theorem 1:

The tour produced by Algorithm 3.1 (heuristic 2MST) is feasible.

Proof:

We will show that the vehicle's load is feasible in each of the $|C(i)|+1$ epochs the algorithm traverses through node i, $i \epsilon N^0$. Since the vehicle was initially loaded by $N_D \leq q$ units which is the total amount unloaded at the customers, its load can never drop below 0. We will first demonstrate that each time the algorithm traverses through 0 the vehicle's load must be feasible. Observe that e(i) represents the net effect on the vehicle's load if node i and all its successors are served. Thus $\Sigma_{i \epsilon C(0)} e(i) = N_B - N_D$. Suppose that the children of 0 are indexed in non-decreasing order of their e( ) values, i.e. $e(1) \leq e(2) \leq \ldots \leq e(|C(0)|)$. Assume, for the sake of contradiction, that the first time infeasibility occurs at 0 is when the service of the subtree rooted at k, $1 \leq k \leq |C(0)|$ is terminated. Infeasibility means that the load at that visit exceeds q, i.e., $N_D + \Sigma_{i=1}^{k} e(i) > q$. This implies that $\Sigma_{i=1}^{k} e(i) > q - N_D \geq 0$. Therefore, e(k), e(k+1),...,e(|C(0)|) must be strictly positive implying that $N_B - N_D = \Sigma_{i=1}^{|C(0)|} e(i) \geq \Sigma_{i=1}^{k} e(i) > q - N_D$ resulting in $N_B > q$ which contradicts our assumption that the problem is feasible.

Suppose now that infeasibility occurs at a certain node of the MST. It is easy to see that according to the algorithm, for any node i there exists an integer k(i), $1 \leq k(i) \leq |C(i)|+1$, such that the vehicle's load during the sequence of visits at node i is non-increasing up to the k(i)-th visit, and non-decreasing from the (k(i)+1)-st visit up to the last visit at i. (Note that $k(i)=1$ ($k(i)=|C(i)|+1$) means that all the e( ) values of the children of i are non-negative (non-positive).) Therefore the vehicle's load reaches its maximum value in the $|C(i)|+1$ times the algorithm traverses through node i either at the first entrance to i, i.e. on the arc p(i)→i, or at the last exit from i, i.e. on the arc

$i \rightarrow p(i)$. Suppose node i^* is a node satisfying the following two conditions: 1) infeasibility occurs on one of the arcs of the MST connecting node i^* to an adjacent node; 2) i^* is a node with the minimum depth on the MST among all nodes satisfying 1). In view of the above discussion infeasibility occurs either on the arc $i^* \rightarrow p(i^*)$ or on the arc $p(i^*) \rightarrow i^*$ implying that $p(i^*)$, whose depth is smaller by one than the depth of i^*, satisfies condition 1) above. In view of condition 2) i^* must be the root node. However, according to our earlier argument at the beginning of the proof, infeasibility cannot occur at the root node, completing the proof. □

In the next theorem we prove that the worst-case bound of the proposed heuristic is 2 and that this bound is tight, i.e. there exists an instance for which the ratio V^{2MST}/V^{OPT} is arbitrarily close to 2.

Theorem 2:

a) The following inequality holds for any given instance of the problem TSPDB

$$V^{2MST} / V^{OPT} \leq 2.$$

b) The worst-case ratio in a) is tight.

Proof:

a) As mentioned above the heuristic 2MST traverses twice through each of the arcs of MST. Array TOUR with 0 the initial and the final node, which is the solution produced by the heuristic, is a subsequence of the sequence of nodes traversed by the algorithm. According to the triangle inequality $V^{2MST} \leq 2l(\text{MST})$. Furthermore, $V^{OPT} \geq l(\text{MST})$; this is because any traveling salesman tour on N^0 can be transformed into a tree by erasing one of its edges and the length of the shortest spanning tree is at least as short as the resulting tree, yielding the desired bound.

b) In order to show that the worst-case ratio of 2 is tight it is sufficient to present a specific instance on which this ratio is achieved. We use a simple modification of the example used by Papadimitriou

and Steiglitz [6] pp. 415-416 to show that the corresponding heuristic for the TSP has a worst-case ratio of 2. Suppose we are given n delivery customers and n backhaul customers so that $d_i=b_i=1$ for $i=1,\ldots,n$. For simplicity we denote the delivery customers by $d_1,d_2,\ldots,d_n$ and the backhaul customers by $b_1,b_2,\ldots,b_n$. Suppose the customers are located as depicted in Figure 1(a) with all inner points being backhaul customers and all outer points being delivery customers. Number the customers clockwise such that delivery point i and backhaul point i fall on the same polar through the center of the graph. Suppose also that the warehouse's location coincides with customer d_1. Let the vehicle's capacity $q=n+1$. The MST is depicted in Figure 1(a). Without loss of generality assume that in the MST the warehouse has a single child and moreover $e(d_1)=0$, $e(b_1)=1$, $e(d_i)=-1$ for $i\neq 1$, and $e(b_i)=0$ for $i\neq 1$. Heuristic 2MST produces the graph depicted in Figure 1(b) where the optimal solution is depicted in Figure 1(c). As is shown in Papadimitriou and Steiglitz [6], choosing $R=1$, $c=1/n^2$ and n arbitrarily large, the ratio between the lengths of these two graphs can be made arbitrarily close to two. □

It is interesting to observe that Christofides Algorithm (see Christofides [3] or Papadimitriou and Steiglitz [6]) for the TSP does not have a simple analog for the TSPDB. Applying Christofides algorithm on the set of nodes N^0 may result in a graph which can not be transformed into a feasible solution. Consider the following example with $q=2$, two delivery customers d_1, d_2 and two backhaul customers b_1, b_2 with all demand and backhaul sizes equal one. Suppose the MST is such that the warehouse is connected to the two backhaul customers and each of the backhaul customers is connected to one delivery customer. According to Christofides Algorithm the MST should be completed to an eulerian graph connecting the two delivery customers one to the other resulting in the cycle $0\rightarrow b_1\rightarrow d_1\rightarrow d_2\rightarrow b_2\rightarrow 0$. As can be seen neither of the cycle's directions is feasible.

4. Numerical Test:

In this section we compare the 2MST heuristic to the heuristic proposed in Mosheiov [5]. As mentioned in the Introduction, any closed tour on the set N contains a certain arc and a specific direction such that the tour can be made a feasible solution by disconnecting that arc and connecting its two endpoints to the warehouse. Thus, if the original tour is constructed by a TSP heuristic having a worst-case bound of $1+\alpha$ then the generated solution will be within $100(2+\alpha)\%$ of the optimal solution. In this numerical test we use the "cheapest insertion" algorithm for the TSP to construct the initial tour. We call this procedure "Cheapest Insertion with Delivery and Backhaul" or shortly CIDB. The worst-case bound of the "cheapest insertion" heuristic for the TSP is known to be 2, see Rosenkrantz et al. [7], therefore, the worst-case bound of CIDB for the TSPDB is 3. Mosheiov [5] shows that the average relative error of CIDB is bounded by 10% on randomly generated problems with up to 25 customers. He compares the cost of CIDB to the optimal Traveling Salesman Tour on the set of customers and the warehouse.

Four sets of 10 problems were generated and solved by both heuristics. All programs were coded in PASCAL on an AT-386. In all problem sets the total delivery size equals the total backhaul size and both are equal to the vehicle's capacity. Although, this is not required by any of the heuristics tested, it represents the case where the capacity constraint is the tightest possible. The sets differ in the number of customers N and the range [1,U] from which the delivery/backhaul sizes were generated. After fixing the number of customers N, their locations were generated according to a uniform distribution on a square of size 500×500. The warehouse location was set to be the square's center. All customers, except the last one, were scanned one by one to determine their type (delivery or backhaul) and their requirement size: if the up to date absolute difference between the total delivery and backhaul sizes exceeds U then the next customer type was selected to minimize

this difference. Otherwise the current customer was set to a delivery (backhaul) one with probability 0.5. The requirement size of each customer except the last one was uniformly generated on [1,U]. The type and requirement size of the last customer were determined to ensure that $N_B=N_D$.

In all four sets CIDB performed somewhat better than 2MST on average, see Table 1 below. Note that 2MST is of complexity $O(N^2)$, where CIDB has complexity of $O(N^3)$. Indeed, our computer study indicates that the running time of 2MST is substantially shorter in all sets (in set 4 e.g., 2MST requires on average about 20% of the running time of CIDB).

Table 1

Average performance of heuristics 2MST vs. CIDB

Set Number	N	U	V^{2MST}/V^{CIDB}
1	10	10	1.038
2	30	20	1.080
3	50	20	1.072
4	100	50	1.046

In Figures 2(a) and 2(b) we demonstrate the actual tours obtained by each of these two heuristics on a problem from set 4. The tour produced by CIDB is shorter by 1.95% than the one produced by 2MST.

References:

[1] L. Bodin, B. Golden, A. Assad and M. Ball (1983), Routing and Scheduling of Vehicles and Crews-State of the Art, Computers and Operations Research 10, no.2., 63-211.

[2] D.O. Casco, B. L. Golden and E.A. Wasil (1988), Vehicle Routing with Backhauls: Models, Algorithms, and Case Studies, in "Studies in Management Science and Systems/ Vehicle Routing: Methods and Studies" edited by B.L. Golden and A.A. Assad, North Holland, 127-147.

[3] N. Christofides (1976), Worst-Case Analysis of a New Heuristic for the Traveling Salesman Problem, Technical Report, GSIA, Carnegie-Mellon University.

[4] M.R. Garey and D.S. Johnson (1979), Computers and Intractability: A Guide to the Theory of NP Completeness, Freeman, San Francisco.

[5] G. Mosheiov (1991), The Traveling Salesman Problem with Pick-Up and Delivery, Working Paper, School of Business Administration and the Statistics Department, The Hebrew University, Jerusalem, Israel.

[6] C.H. Papadimitriou and K. Steiglitz (1982), Combinatorial Optimization Algorithms and Complexity, Prentice Hall Inc. Englewood Cliffs, New Jersey 07632.

[7] D.J. Rosenkrantz, R.E. Stearns and P.M. Lewis II (1977), An Analysis of Several Heuristics for the Traveling Salesman Problem, SIAM J. Comput. 6, 563-581.

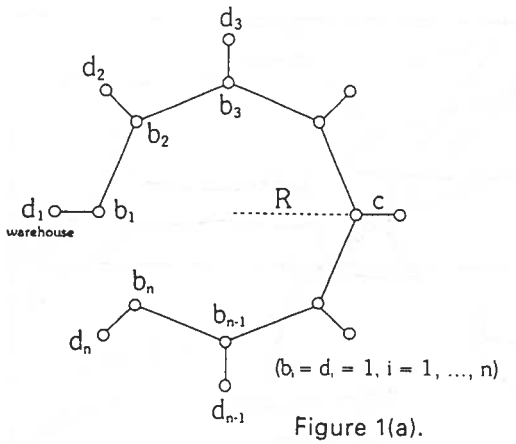


Figure 1(a).
Minimum spanning tree

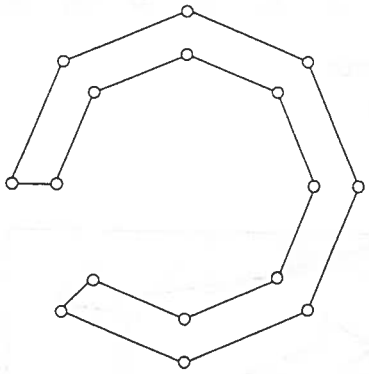


Figure 1(b).
Tour obtained by heuristic 2MST

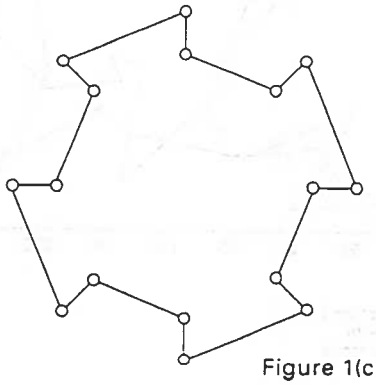


Figure 1(c).
Optimal tour

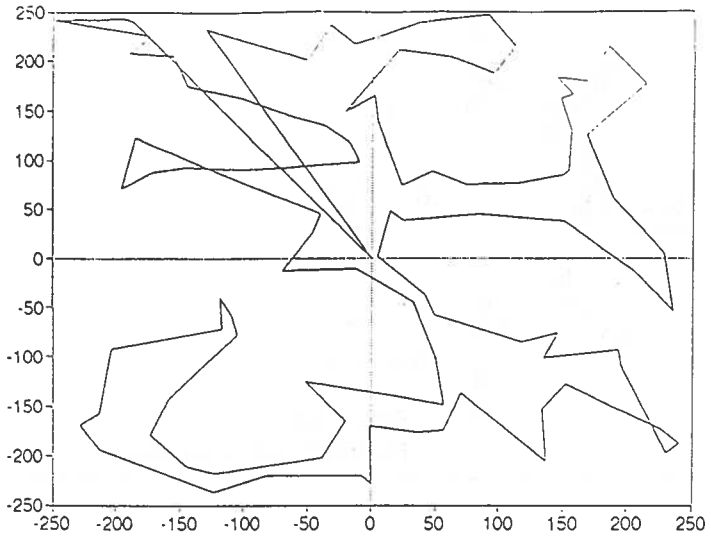


Figure 2(a). Tour obtained by CIDB.
N = 100, U = 50,
Total length = 5108.2.

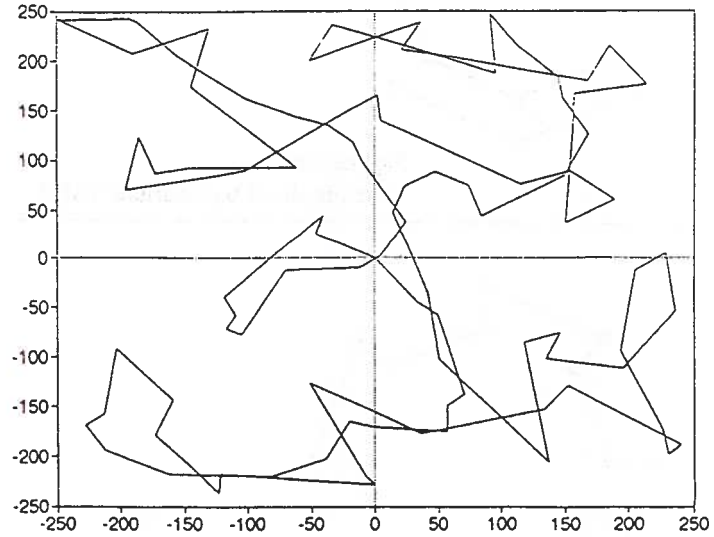


Figure 2(b). Tour obtained by 2MST.
N = 100, U = 50,
Total length = 5276.8.

The Vehicle Routing Problem with Delivery and Backhaul Options

by
Shoshana Anily
The Recanati Graduate School of Business Administration
Tel-Aviv University
Israel

Abstract:

In this paper we consider a version of the Vehicle Routing Problem *(VRP)*: a fleet of identical capacitated vehicles serves a system of one warehouse and N customers of two types dispersed in the plane. We distinguish between "delivery customers" that require a delivery of stock from the warehouse, and "backhaul customers" that need to deliver stock from their location to the warehouse. The objective is to design a set of routes of minimum total length to serve all customers, without violating the capacity restriction of the vehicles all along the routes. The capacity restriction here, in contrast to the *VRP*, complicates the problem significantly as special attention should be given to the order the customers are visited along the routes. The problem is *NP-hard*. We develop a lower bound on the optimal total cost and a heuristic solution which is shown to converge to the optimal solution, under mild probabilistic conditions, as fast as $N^{-0.5}$. The complexity of the heuristic as well as the computation of the lower bound is $O(N^3)$ if all customers have unit demand size and otherwise $O(N^3 \log N)$ independently of the demand sizes.

Key Words: Vehicle Routing Problem *(VRP)*, Heuristic, Worst-case Analysis, Probabilistic Analysis, Asymptotic Analysis, Strongly Polynomial Algorithm.

A Computerized Approach to the New York City School Bus Routing Problem

Jeffrey Braca
Director of Information Systems, Fund for the City of New York
Julien Bramel
Graduate School of Business, Columbia University
Bruce Posner
Vice President, Fund for the City of New York
David Simchi-Levi
Dept. of MS & IE, Northwestern University, and
Dept. of IE & OR, Columbia University.

September 1993

School bus routing and scheduling is an area where operations research has had much success. Many communities around the country and the world have implemented and used computerized routing systems that in most cases has lead to substantial reduction in operational cost. It is an area where the average person can understand (to some degree) the complexities and intricacies of the problem and how suboptimal solutions can cause vast amounts of resources to be "wasted". It is clearly a problem that must be solved in almost every community.

The City of New York has one of the most extensive school systems in the world. A majority of its 1,000,000 students either walk or take public transportation to school: transit buses or subways. About 125,000 students ride school buses leased by the Board of Education. The majority, or about 83,000, of these students are classified as General Education students. These students walk to their neighborhood bus stop in the morning and wait for a bus to bring them to school. In the afternoon, a bus takes them from their school and drops them off back at their bus stop. The rest of these students have special needs. They are classified as Special Education students and are picked up and dropped off directly at their homes.

This distinction makes the routing of Special Ed students fundamentally different from routing General Ed students. Another difference is that in many cases Special Ed students must traverse the entire City to reach one of the few schools which provides for their special needs. General Ed students usually go to schools only a few miles from their homes; almost exclusively to schools within the same borough. Also, Special Ed students, such as wheel-chair bound students, are transported in specially designed vehicles with much smaller carrying capacities.

Currently in New York City, transporting all General Ed students for the academic year requires leasing approximately 1,150 buses. These buses must pick-up and drop-off students in a vast network of 19,000 bus stops and 750 schools. Several companies bid for the contract to transport the students and currently the companies winning contracts design the routes. Independent of the company, the leasing cost to the City of a bus and its driver is approximately $80,000 annually. The total yearly budget for General Education student transportation alone is therefore close to $100 million.

The routing of buses for over 40,000 Special Education students is done in house. A team of inspectors/routers at the Board of Education Office of Pupil Transportation use pins to mark the students homes and schools on large maps. Then, using experience they have acquired in years of work, routers use string to link pins together and form routes. The routes are checked for feasibility by sight or by measurement; while the inspectors have a good idea whether a route can be driven, this is clearly a time consuming process.

The Board of Education is appropriating funds to develop a computerized system, called CATS (Computer Assisted Transportation System), that can design routes for both the General and Special Ed students. The project, which has already received the approval of the New York State Education Department, has three phases.

The first phase consists of installing a geographic information system (GIS). This software, widely available only in the last few years, makes it possible to see complicated maps on a PC monitor. In addition, by "clicking and dragging" an area with a mouse, one

can pull up information about the area outlined. This information can include U.S. Census data such as number of households, median age, income etc.... More importantly in this application, by designating two points, the GIS can calculate exact locations (latitude and longitude coordinates) and also the distance between the two points. When this phase is completed, the router's job will be a lot easier since this will allow him or her to "click" on a location with a mouse and "string" together points on the computer screen. This phase consists of emulating on the computer screen what was previously done with maps, pins and string. At the end of this phase, CATS will provide support for Special Education door-to-door service by replicating on the computer the point-to-point manual system now employed by the inspectors. An example of this is what is called "blocking and stringing". There, a router can block a set of bus stops together on the screen and then let the computer figure out the best route through these stops.

The second phase extends the functionality developed in the first phase to include regular education stop-to-school service. This implies installing a database that keeps track of all relevant student and school information. Data on each student includes: address, bus stop and school. For each school, the data is an address, as well as starting and ending times for all sessions. This makes it possible to keep all relevant data up to date, and provide some of the basic input data needed for routing and scheduling.

The third and final phase of the project consists of installing an optimization module. The aim here is to build software that uses the student and school data and the GIS to generate efficient bus routes and schedules that meet all existing transportation policies. This is the phase we are the most interested in.

In this talk, we present a range of issues involved in developing this optimization module. We discuss many of the intricate details involved in routing in New York City. We focus on routing the General Ed students; the routing of Special Ed is currently being done at the Office of Pupil Transportation using the "blocking and stringing" approach.

We first present a short review of some of the literature on computerized school bus

routing and scheduling. Then we present details of the school bus routing and scheduling problem in Manhattan.

Since any optimization must begin with estimates of distance and travel time, we then discuss these issues as they pertain to the New York City street network.

To assess the magnitude by which an optimization module that routes and schedules buses can reduce total operating costs, we have tested a number of routing algorithms on the Manhattan data. We explain the algorithm we designed to generate routes and demonstrate preliminary results on the Manhattan data.

MANAGING URBAN SNOW REMOVAL AND DISPOSAL OPERATIONS

James F. Campbell[1] and André Langevin[2]

[1]School of Business Administration, University of Missouri-St. Louis, 8001 Natural Bridge Road, St. Louis, MO 63121, USA
[2]GERAD and École Polytechnique de Montréal, 5255, avenue Decelles, Montréal, Québec Canada H3C 3J7

Abstract

Many cities face a winter problem of clearing large quantities of snow from streets and sidewalks. Plowing, along with warmer temperatures, may be sufficient to control moderate snowfalls. However, in urban areas that experience heavy snowfall and persistent low temperatures, snow plowed to the sides of streets and sidewalks accumulates and impedes pedestrian and vehicular traffic. In this case, the snow has to be loaded into trucks and hauled to snow disposal sites. Urban snow removal and disposal operations include a number of challenging strategic and tactical problems that have been little studied by the OR/MS community.

Snow Removal and Disposal Operations

There are four major steps in snow removal and disposal operations: (1) spreading de-icers and abrasives, (2) snow clearing (plowing), (3) snow loading, and (4) snow disposal. Large cities are usually divided into sectors, and operations are conducted simultaneously in each sector by a small fleet of vehicles, including spreader trucks (to spread salt and abrasives),

graders, snowplows, sidewalk snowplows, front-end loaders, snowblowers, service vehicles and trucks for hauling snow. The following description is typical of operations in Montreal.

Montreal receives about ten major snow storms each winter and the average annual snowfall exceeds 200 cm. When snow begins to fall, salt and abrasives are spread first on major arteries and then on secondary streets, as well as on sidewalks. When the snow accumulation reaches 2.5 cm, snow clearing begins using street plows and sidewalk plows. The snow is plowed as close as possible to the side of the road or the sidewalk. Once the precipitation ends, snow loading begins. Street plows and sidewalk plows work together in every sector to create a long row of snow in the middle of the street. The snow is then removed by a snowblower which moves slowly along the row of snow and loads the snow into trucks moving alongside. As soon as a truck is filled, it departs for its assigned snow disposal site and another truck takes its place. Each disposal site has an hourly receiving rate capacity, which depends on the logistics and configuration of unloading facilities at the disposal site, and an annual capacity corresponding to the finite space available for storing snow.

One important development in Montreal is recent legislation that will phase out use of the least expensive disposal method (dumping snow in the St. Lawrence River) by 1996. River disposal sites must be replaced by other sites that allow the melted snow to be processed in waste water treatment plants before release to the environment.

Operations management problems

Urban snow removal and disposal operations involve a host of engineering and managerial problems. Two fundamental strategic problems that provide the foundation for snow removal and disposal operations are to partition the city into snow removal sectors and to locate disposal

sites. The partitioning problem shares some similarity with school and political (re)districting problems that have been addressed by mathematical programming and by decision support systems. The location of snow disposal sites can be viewed as a facility location problem in discrete space, since a set of candidate disposal sites can usually be identified based on infrastructure (e.g., sewers), geography and land use (e.g., vacant parcels). Snow disposal sites may be viewed as seasonally obnoxious facilities, because they generate a large volume of truck traffic and around-the-clock activity during snow disposal operations.

Once the sectors and disposal sites have been defined, the snow disposal assignment problem is to assign each sector to a disposal site. The assignments may have to satisfy an annual capacity (m^3 of snow per year) and an hourly receiving capacity (m^3 of snow per hour) for each site. We have modeled the snow disposal assignment problem as a multi-resource generalized assignment problem. The snow disposal assignment problem and the partitioning problem to define sectors are complicated by interdependencies: the size and shape of the sector should probably depend on the assignment. Separating these decisions simplifies the analysis, but will likely produce a sub-optimal system.

There are also fleet mix, equipment selection, and crew scheduling problems. One problem is to allocate trucks to each sector to haul snow to the assigned disposal sites. We have developed preliminary models for allocating trucks to sectors and have shown how numbers and sizes of trucks depend on the distances from sectors to disposal sites, determined by the assignment of sectors to sites.

Snow removal and disposal also involves a number of routing problems for spreaders, graders, snowplows, snowblowers, and trucks. These problems are complicated by precedence

constraints (e.g. clearing main arteries first) and by time-window constraints (e.g. maximal time interval, parking restrictions, and other traffic issues). Arc routing problems are the most studied of any snow removal and disposal problems. These are important practical examples of the Chinese Postman Problem and more general capacitated arc routing problems. Another operational problem is to route trucks from sectors to disposal sites and back to the sectors. This is not simply a shortest path problem, because of the dynamic nature of traffic and the movement of the snow blower while a truck travels to and from the disposal site.

Decision support system for snow removal and disposal

Although there has been some work on arc routing problems in the context of snow removal, there is almost no analytical research on many of the issues specific to snow disposal, including the design of sectors for snow removal, the assignment of sectors to disposal sites and the allocation of trucks to sectors. Analysis of individual problems would be useful, but research that treats snow removal and disposal from a global perspective would be most beneficial.

Our research is concerned with the development of a snow removal and disposal decision support system (DSS) that addresses several major problems, including assigning sectors to sites and allocating trucks to sectors, in an integrated fashion. Because of the inherent complexity of the problems, an interactive DSS would be most useful to management to allow solutions to be adjusted to incorporate political and difficult to quantify factors. Hence, fast heuristic algorithms that produce good approximate solutions will be preferred over optimal algorithms requiring excessive computer time or power.

Two important uses for the snow removal and disposal DSS are to help find solutions to the fundamental problems of partitioning the city into sectors and evaluating potential disposal

site locations. It would also be useful for real-time response to contingencies (e.g. equipment breakdowns) and sensitivity analyses. Such a DSS will be especially important to redesign snow removal and disposal systems to reduce costs and decrease environmental impacts.

Study on O-D Demand Adjustment Problem: A Bilevel Programming Approach

Michael Florian and Yang Chen*

June 1993

The O-D demand adjustment (or estimation) problem (DAP) is studied extensively in this paper in a bilevel programming approach. The DAP may be formulated as a nonlinear bilevel programming problem in the following way:

Find an estimate demand g such that g solves

$$\min F(g,v) = \frac{1}{2}[\sum_{i\in I}(g_i - \hat{g}_i)^2 + \sum_{a\in A}(v_a - \hat{v}_a)^2] \tag{1}$$

$$s.t. \qquad g_i \geq 0, \quad i \in I, \tag{2}$$

where v solves the equilibrium problem for given g

$$\min f(v,g) = \sum_{a\in A}\int_0^{v_a} c_a(t)dt \tag{3}$$

$$s.t. \quad \sum_{k\in K_i} h_k - g_i = 0, \quad i \in I \tag{4}$$

$$h_k \geq 0, \quad k \in K_i, \quad i \in I \tag{5}$$

$$v_a = \sum_{i\in I}\sum_{k\in K_i} \delta_{ak} h_k, \quad a \in A \tag{6}$$

where $c(v) = (c_a(v_a))$ is link cost function, I, A and K_i are index sets of O-D pairs, links and paths corresponding to O-D pair i, respectively.

*Centre de recherche sur les transports, Department d'informatique et recherche operationalle, Université de Montréal, Montréal, Què. H3C 3J7

This DAP is reformulated as a single level optimization problem, where the marginal function of the equilibrium problem (??)-(??) is involved in the constraints and the variable of the lower level problem is replaced by path proportions. The gradient of the marginal function is obtained in terms of the path proportions and the link cost function. As a direct consequence, the DAP is shown to be a differentiable problem, though the bilevel problem is not differentiable in general. A set of necessary optimality conditions for the DAP is derived based on the theory developed in [?] and [?] for general bilevel programming. The interpretation of these conditions are analyzed for degenerate and nondegenerate cases, which both coincide with the Wardrop's principle in certain sense.

Two solution procedures are proposed to solve the DAP.

An Augmented Lagrangean method is adopted to solve the problem by taking the advantage of the marginal function formulation and the gradient information of it. The resulting penalized problem, which is nonconvex with simple linear constraints, is solved by successive linear approximation techniques by virtue of the fully decomposition feature of the subproblems. It is shown that the sequence of solutions generated by the algorithm converges to points which are equilibrium and satisfy the necessary optimality conditions of the DAP.

Aiming at dealing with large-scale problems, a Gauss-Seidel type method is investigated in order to avoid explicit enumeration of the paths or path proportions of the network. The main feature of this method is to solve two single level optimization problems at each iteration by fixing the upper and lower level variables in turn such that the path information may not be necessarily obtained directly. Instead, the path information may be obtained implicitly from equilibrium link flows by some special features of the software package EMME/2. There is no general

convergence property for this algorithm in theory. However, it is shown that if the algorithm terminates, it will terminate at a point which satisfies the necessary optimality conditions.

These two methods are implemented for test networks and real traffic networks by using macro commands of EMME/2, which enable one to carry out the computations on link space rather than the path space. Computational results are reported and discussed.

References

[1] Chen, Y. and M. Florian, The Nonlinear Bilevel Programming: Formulations, Regularity and Optimality Conditions, Publication CRT-794 (revised version, 1993), Université de Montréal, Montréal, Canada.

[2] Chen, Y., Bilevel Programming Problems: Analysis, algorithms and applications", Ph.D thesis, Université de Montréal, Montréal, Canada (1993)

AN EFFICIENT ALGORITHM FOR A BICRITERION TRAFFIC ASSIGNMENT PROBLEM

Patrice MARCOTTE and Daoli ZHU
Centre de recherche sur les transports
Université de Montréal

May 1994

Abstract

We consider a bicriterion traffic assignment model where the network users have different values of time. We present an infinite-dimensional formulation of this model that lends itself to an extremely simple and efficient algorithm that takes into account the network structure of the problem.

Traffic assignment procedures, with fixed or elastic demand, are frequently based on Wardrop's assumption that, at equilibrium, the cost of any used path is minimal among all used or unused paths. In this context, the term "cost" is used as a synonym for "generalized transportation cost", i.e. a combination of several criteria: actual cost, travel time, waiting time, average speed, etc. In this paper, having in mind the problem of pricing of the network arcs, we consider explicitly two criteria, namely time and money. Following Leurent [5], Dafermos [2] and Dial [3], the preferences of the users are represented by a density function $h(\alpha)$ that yields information pertaining to the "number" of people whose value of time equals α. Our model is similar to that of Dafermos and Leurent. However, our analysis is simpler and yields a solution procedure that is more efficient and obviates path enumeration.

For the sake of simplicity, we will assume that demand is fixed and that the function that characterizes the distribution of time values is common to all origin-destination pairs. These restrictions can easily be relaxed. More specifically, define

$G = (N, A)$:	A network with node set N and arc set A
$FS(i)$:	The forward star associated with node $i \in N$
$BS(i)$:	The backward star associated with node $i \in N$
α:	The value of time (state variable)
k:	A commodity index (origin-destination pair index)
K:	The set of origin-destination couples
d^k:	The demand for commodity (origin-destination) k
$h(\alpha)$:	A probability density function over the values of time. The number of people having values of time in the interval $[\alpha_1, \alpha_2]$ is given by $d^k \int_{\alpha_1}^{\alpha_2} h(\alpha)\, d\alpha$. We have: $\int_0^\infty h(\alpha)\, d\alpha = 1$
$x_a^k(\alpha)$:	The flow density along arc a with respect to the value of time α
$X_a(\alpha)$:	The total flow density of class α along arc a: $X_a(\alpha) = \sum_k x_a^k(\alpha)$
$x(\alpha)$:	The vector of functions $x_a^k(\alpha)$, $a \in A$, $k \in K$
X_a:	The total flow along arc a. One has: $X_a = \sum_{k \in K} \int_0^\infty x_a^k(\alpha)\, d\alpha = \int_0^\infty X_a(\alpha)\, d\alpha$
G_a:	The fixed monetary cost of travel along arc a
$C_a(X_a)$:	The delay (congestion) function associated with arc a.

The feasible set of this problem corresponds to the flow density functions lying in the set Ω, where Ω is defined by the nonnegativity constraints on the commodity flows and the flow conservation equations:

$$\sum_{a \in FS(i)} x_a^k(\alpha) - \sum_{a \in BS(i)} x_a^k(\alpha) = \begin{cases} d^k h(\alpha) & \text{if } \ i = \text{origin of } k \\ -d^k h(\alpha) & \text{if } \ i = \text{destination of } k \\ 0 & \text{otherwise.} \end{cases}$$

In our assignment model, equilibrium is reached when, for every possible value of time α, the flow density $x_a^k(\alpha)$ is assigned to a shortest path with respect to α, given the congestion levels $C_a(X_a)$ on the network. Equilibrium flows can be characterized as feasible flows satisfying the infinite dimensional variational inequality:

$$\sum_{k \in K} \sum_{a \in A} (x_a^k(\alpha) - y_a^k(\alpha))(C_a(X_a) + G_a/\alpha)\, d\alpha \leq 0$$

for all feasible flow patterns $y_a^k(\alpha)$, $a \in A$, $k \in K$.

As the congestion functions are arc-separable, this variational problem can be reformulated as the convex minimization program:

$$\min_{x(\alpha)\in\Omega} \sum_{a\in A} \left\{ \int_0^{X_a} C_a(t)\,dt + G_a \int_0^{\infty} \frac{X_a(\alpha)}{\alpha}\,d\alpha \right\}.$$

This infinite dimensional problem can be solved efficiently using Frank and Wolfe's linear approximation algorithm, without resorting to numerical discretization or path enumeration. The convergence of the method follows from a result established by Auslender and Brodeau [1].

For a current solution vector x and corresponding total flow vector X, the linear subproblem resulting from the linearization of the objective with respect to the flow variables is:

$$\min_{y\in\Omega} \sum_{k\in K} \sum_{a\in A} (C_a(X_a) + G_a/\alpha) y_a^k(\alpha) \qquad \forall \alpha \geq 0.$$

This problem is separable by commodity. For a given commodity k and value of time α, its solution can be found by computing the shortest path corresponding to k, based on the current arc costs $C_a(X_a) + G_a/\alpha$. Actually, all these shortest paths can be computed in one pass by solving a parametric shortest path problem with respect to the parameter α. This can be performed efficiently using a variant of the network simplex method. The solution to this problem will take the form

$$y_a^k(\alpha) = h(\alpha) d^k$$

if arc a lies on a the computed shortest path corresponding to commodity k and value of time parameter α, and $y_a^k(\alpha) = 0$ otherwise. Let $Y_a(\alpha) = \sum_{k\in K} y_a^k(\alpha)\,d\alpha$ and $Y_a = \int_0^{\infty} Y_a(\alpha)\,d\alpha$ be the associated total arc flow. The next iterate is then determined by solving the one-dimensional problem

$$\begin{aligned} \min_{t\in[0,1]} \quad & \sum_{a\in A} \int_0^{X_a + t(Y_a - X_a)} C_a(t)\,dt + G_a \int_0^{\infty} \frac{1}{\alpha} \left[X_a(\alpha) + t(Y_a(\alpha) - X_a(\alpha)) \right] d\alpha \\ & = \phi(t) + (1-t)Q + tR \end{aligned}$$

where $Q = \int_0^{\infty} X_a(\alpha)/\alpha\,d\alpha$ and $R = \int_0^{\infty} Y(\alpha)/\alpha\,d\alpha$.

Now let

$$t^* \in \arg\min_{t\in[0,1]} \phi(t) + t(R - Q)$$

and update

$$X_a^+ = (1 - t^*)X_a + t^* Y_a.$$

It is remarkable that the above update can be realized without explicit reference to the commodity flow variables, exactly as for the standard traffic assignment problem. One must only be careful to keep track of the integrals $\int X_a(\alpha)/\alpha$ and $\int Y_a(\alpha)/\alpha$. Indeed:

$$\int_0^\infty X_a^+(\alpha)/\alpha \, d\alpha = (1 - t^*) \int_0^\infty X_a(\alpha)/\alpha \, d\alpha + t^* \int_0^\infty Y_a(\alpha)/\alpha \, d\alpha.$$

One can obtain a closed form formula for the latter integrals only if the function $h(\alpha)/\alpha$ admits a closed form primitive. If this is not the case, an approximate but closed form expression can be derived from polynomial quadratures for $h(\alpha)$.

It follows from the previous discussion that the above bicriterion traffic assignment problem can be solved in a similar fashion to the standard problem, with a parametric shortest path problem being substituted to a standard shortest path problem. Furthermore one can enhance the basic model in several ways. For instance:

1. If the Lebesgue-Stieltjes integral $\int dH(\alpha)$ is used in place of the Lebesgue integral $\int h(\alpha)$, the model subsumes the multiclass model where only finitely α-values are considered.

2. The model can be generalized to take into account asymmetric congestion effects, if the functions $C_a(X)$ are dependent on the entire total flow vector. In this situation, there would not be an equivalent convex optimization formulation of our problem, and we would use a restriction strategy such as Lawphongpanich and Hearn's [4] restricted simplicial decomposition to solve it.

3. If no two paths associated with a given commodity have identical monetary costs, then it can be proved that the path-flow functions are unique, with the possible exception of a set with measure zero. This is not the case for the standard single-class model, where the arc flows are unique (under suitable assumptions) but the decomposition into path flows is *not*. Moreover, in the multiclass model, path flows can easily be recovered from the sole knowledge of equilibrium arc costs. These uniqueness relationships between path and arc flows hold without any monotonicity or convexity assumptions on the congestion functions C_a.

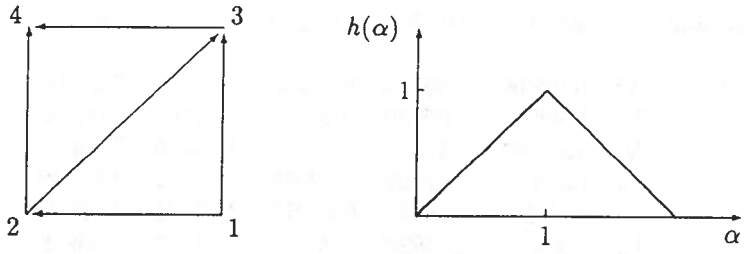


Figure 1: Network topology and VOT density.

arcs:	1–2	1–3	2–3	2–4	3–4
delay function:	$10x$	$50 + x$	$10 + x$	$50 + x$	$10x$
G_a:	100	90	10	100	100

Table 1: Exogenous data.

A seemingly innocuous extension of the multiclass model would be to consider two or more flow-dependent criteria. However we can show that this situation leads to difficult non monotone, asymmetric problems, thus contradicting recent claims made in the paper by Dial [3].

We conclude with a numerical illustration on a simple example inspired from the "Braess paradox" network. The network topology and the value of time function are illustrated in Figure 1, and the exogenous data given in Table 1. The total demand from node 1 to node 4 has been set to 6 units of flow. Table 2 illustrates the convergence of the arc flows to their equilibrium values. Notice that the extremal flow vector y also converges to the equilibrium, albeit at a slower pace. Figure 2 gives the equilibrium flow densities along each arc.

The behavior of the algorithm is quite similar to what could be observed on the single-class Braess paradox problem, for which convergence is geometric, given that all equilibrium arc flows are all positive. This is atypical of the Frank-Wolfe algorithm, whose convergence cannot be expected to be better than sublinear for large-scale problems.

iteration		arc 1–2	arc 1–3	arc 2–3	arc 2–4	arc 3–4
1	X:	6.00000	0.00000	6.00000	0.00000	6.00000
	Y:	0.00000	6.00000	0.00000	0.00000	6.00000
2	X:	1.52284	4.47715	1.52284	0.00000	6.00000
	Y:	5.87631	0.12369	0.00000	5.97631	0.12369
3	X:	3.13266	2.86734	0.95973	2.17294	3.82706
	Y:	1.43172	4.56828	0.00000	1.43172	4.56828
4	X:	1.92045	4.07955	0.27576	1.64469	4.35531
	Y:	5.58180	0.41820	0.00000	5.58180	0.41820
5	X:	2.60862	3.39138	0.22393	2.38469	3.61531
	Y:	3.55351	2.44649	0.33760	3.21591	2.78409
6	X:	3.12049	2.87951	0.28551	2.83499	3.16501
	Y:	0.00000	6.00000	0.00000	0.00000	6.00000
7	X:	2.71402	3.28598	0.24832	2.46570	3.53430
	Y:	2.38555	3.61445	0.89214	1.49341	4.50659
8	X:	2.58165	3.41834	0.50776	2.07389	3.92611
	Y:	4.62834	1.37166	0.00000	4.62834	1.37166
9	X:	2.77794	3.22206	0.45907	2.31887	3.68113
	Y:	2.96092	3.03908	0.00111	2.95981	3.04019
10	X:	2.89472	3.10528	0.16680	2.72792	3.27208
	Y:	0.00000	6.00000	0.00000	0.00000	6.00000
20	X:	2.72173	3.27827	0.35748	2.36425	3.63575
	Y:	3.03223	2.96777	0.15852	2.87371	3.12629
30	X:	2.72430	3.27570	0.34687	2.37743	3.62257
	Y:	2.92814	3.07186	0.23070	2.69743	3.30257
40	X:	2.73527	3.26473	0.34056	2.39471	3.60529
	Y:	2.73657	3.26343	0.33405	2.40252	3.59748
50	X:	2.73507	3.26493	0.33947	2.39560	3.60440
	Y:	2.73186	3.26814	0.34008	2.39178	3.60822
60	X:	2.73476	3.26524	0.33934	2.39543	3.60457
	Y:	2.73516	3.26484	0.33910	2.39606	3.60394
70	X:	2.73478	3.26521	0.33932	2.39546	3.60454
	Y:	2.73479	3.26521	0.33931	2.39548	3.60452
80	X:	2.73478	3.26521	0.33932	2.39546	3.60454
	Y:	2.73478	3.26522	0.33932	2.39545	3.60455
90	X:	2.73478	3.26521	0.33932	2.39546	3.60454
	Y:	2.73478	3.26521	0.33932	2.39546	3.60454

Table 2: Convergence of arc-flow values.

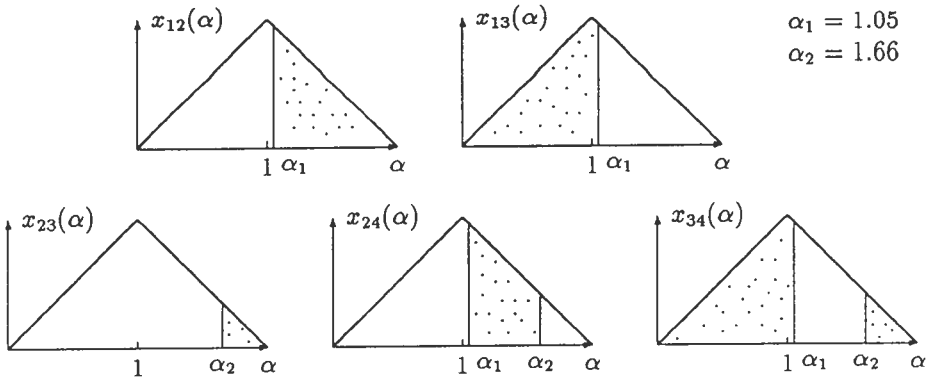


Figure 2: Equilibrium arc-flow density functions.

References

[1] Auslender, A. and Brodeau, F., "Convergence d'un algorithme de Frank et Wolfe appliqué à un problème de contrôle", *RIRO* **2** (1968) 3–12.

[2] Dafermos, S., "A multicriteria route-mode choice traffic equilibrium model", unpublished manuscript, Lefschetz Center for Dynamical Systems, June 1981.

[3] Dial, R. B., "T2: A multicriteria equilibrium traffic assignment model. Theory and algorithms", unpublished manuscript, Volpe National Transortation Systems Center, July 1993.

[4] Lawphongpanich, S. and Hearn, D. W., "Simplicial decomposition of the asymmetric traffic assignment problem", *Transportation Research B* **18B** (1984) 123–133.

[5] Leurent, F., "Cost versus time equilibrium over a network", *EJOR* **71** (1993) 205–221.

Modelling and Algorithmic Improvements in DSS for Urban Traffic Planning

by DOMENICO INAUDI - CSST, Torino, Italy
VITO VELARDI - ELASIS, Salerno, Italy

ABSTRACT

Decision Support Systems for Urban Traffic and Transport planning are widely used and consolidated tools. A DSS is composed of three main components: a set of Mathematical Models, a structured Data Base and a finally a Human Machine Interface. All these components are continuously up dated and chanced according to the main scientific and technological development.

In this paper we will present the improvement carried out, in the ELASIS Traffic Laboratory during the development of the DSS named MTU (Mobility and Traffic in Urban area), with special respect to the modelling features. Two aspects of the modelling are reported: the implementation of new approaches in traffic modelling, and the development of computation algorithms with high levels of computing performances.

With respected to the models three research line will be described:

A1. Demand Model [Velardi, Cascetta, Nuzzolo]. Two different demand models have been specified, calibrated and implemented; the first is simplified model for the estimation of origin–destination matrices internal to the study area, per trip purpose and mode, in the morning peak period; the second is an extended, behavioural model simulating trip chaining with a sequential approach for all time–bands of the day. The *simplified demand model* is a typical *partial share* model. It is made up of sub models for trip frequency, spatial distribution and modal choice; the first is of a descriptive type while the remaining two have a Nested Logit (Ben Akiva e Lerman, 1985) structure and can be sees as behavioural. *The trip chaining model* follows a sequential approach to simulate home–based tours , i.e. sequences of trips starting and ending at home. The disaggregate model system is based on the primary destination approach (Ben Akiva, 1993). The hypothesis at the base of this approach is that every tour has a primary activity exists motivating travel towards a primary destination and that other travel choices (such as to make another trip to a secondary destination) are conditioned to the primary ones.

A2. Assignment with parking facilities [Bifulco, Cascetta, Tartaro, Nuzzolo, Velardi]. A generalisation of the SUE probit model was adapted for simulating path choices in the road + parking assignment model (Bifulco, 1992) (Cascetta,

Nuzzolo, Bifulco, 1991). Unlike the above models, the generic alternative is now made up by a set *(p, k)* where p denotes the parking type (on street, off street, charged, etc.) and its location in the study area, and *k* denotes the path by the origin to the destination, including pedestrian links for access and exit by parking link. A multi-user approach is necessary supposing that different categories of users perceive and trade off differently component costs.

A3 OD estimation bilevel programming [Cascetta, Inaudi]. The models for estimating the Origin/destination demand, so far developed, can be divided into three categories: direct sample estimation, model estimation and estimation from traffic counts. A great importance is generally given to the last category because these methods use data that are often available, are the cheaper ones when survey are needed and are very interesting for on line application of the traffic control. These methods can be applied for the correction or the update of an OD matrix already estimated (for example with models belonging to the first two categories): so they can be used to estimate directly the OD matrix or to correct it. The Bilevel Programming rises from the analysis of the connection between the OD estimation and the assignment procedures: looking for a correct method for estimating the OD demand bring to consider together the two models and consequently to get into a problem of two levels. The Bilevel Programming gives an appropriate theoretical framework to state an optimisation problem comprising a constraint which is itself an optimisation problem.

With respected to the algorithms also three aspects will be discussed and results presented:

B1. Alternative methods to solve the Users Equilibrium Assignment model have been investigated [Inaudi, Nguyen, Velardi]. We first review the impact of embedding the PARTAN (parallel tangent) technique into Frank-Wolfe basic code as a simple way to obtain a good acceleration of the initial code without sacrificing the initial robustness and simplicity. In addition, we investigate the viability of applying a commodity decomposition strategy. This strategy consists in cyclically solving traffic assignment problems with a single origin, the vehicle flow associated with all other origins being fixed at their current levels. We next take a critical look at a disaggregate simplicial decomposition technique recently reported in literature. The interest on this approach stems from the claimed efficiency and accuracy achieved in solutioning large scale traffic assignment

problems. Finally, in the light of the numerical results obtained so far, we design a new hybrid algorithm that blends a second order descent direction approach based on successive quadratic programming, and the simplicial decomposition method.

B2. Shorted path algorithm to incorporate turn penalties [Pallottino, Tartaro, Della Valle]. In order to solve the minimum route problem for transportation networks with turn penalties and prohibitions a new method has been developed. In the standard methods turn penalties are represented through the expansion of nodes and the shortest path tree problem is solved by a classic node-labelling algorithm. It has been also evaluated the link-labelling algorithm. Moreover it has been developed an "hybrid" algorithm which save computational time and efficiency and doesn't require the expanded transportation supply. Performances of this "hybrid procedure" have been compared with the previous two on real case (Salerno private network).

B3. Post optimisation approach to the bilevel programming (as for the model A3) [Inaudi, Nguyen, Pallottino]. the basic idea of the proposal algorithm is: starting from a feasible solution of a DUE with a given OD demand find a new feasible solution related to a new OD demand not too much dissimilar from the previous one. The developed algorithm performs quite well in solving the bilevel programming problems saving more than the 75% of the original computation time.

A PLANNING AND DESIGN MODEL FOR TRANSIT ROUTE NETWORKS WITH COORDINATED OPERATIONS

Mao-Chang Shih
Hani S. Mahmassani

Department of Civil Engineering
The University of Texas at Austin

Abstract

This paper presents a heuristic approach for the design of bus transit networks with coordinated operations. Previous approaches have almost exclusively focused on uncoordinated route operations. This model incorporates the ability to construct the route network around the transit center concept, and relies on a trip assignment procedure that explicitly recognizes coordinated routes. In addition, this model is able to determine the appropriate vehicle size for each bus route; it also incorporates demand responsive service for demand that cannot be served effectively by fixed-route, fixed-schedule service.

These features are responsive to application needs of trip patterns associated with suburban, multi-nucleated urban development. This model is comprised of four major procedures. A route generation procedure (RGP) constructs the transit network around the transit center concept. A network analysis procedure (NETAP) incorporates a trip assignment model for both coordinated and uncoordinated operations and a frequency setting and vehicle sizing procedure. A transit center selection procedure (TCSP) identifies suitable transit centers for route coordination. A network improvement procedure provides improvement modifications for the set of routes generated by RGP. An illustrative application to data generated from the existing transit system in Austin, Texas, is included.

Introduction

The transit network design problem seeks to configure a bus transit network consisting a set of routes and determine associated frequencies. Mathematical formulations of this problem have been concerned primarily with the minimization of a generalized cost measure, usually a combination of user costs and operator costs. User costs consist of access cost, waiting cost, and in-vehicle travel time cost; operator cost is estimated by total vehicle operating miles or time. Constraints to the problem may include, but are not limited to 1) minimum operating frequencies on all or selected routes, 2) a maximum load factor on bus routes, and maximum available resources (fleet size or capital). The complex formulation and the combinatorial nature of the transit network design problem preclude solution by exact optimization models (Newell, 1979; Baaj, 1990). Consequently, heuristic approaches intended to find good solutions but do not guarantee a global optimal solution have been proposed to solve this problem. Past heuristic approaches include those of Lampkin and Saalmans (1967), Rea (1971), Silman, et al. (1974), Dubois, et al. (1979), Hasselstrom (1981), Ceder and Wilson (1986), Van Nes, et al. (1988), Baaj and Mahmassani (1991), and Israeli and Ceder (1991).

To overcome the complexity of transit network design, most approaches partition the problem into two parts, route construction and frequency setting. Usually, a good set of routes is determined in the first stage; the corresponding route frequencies are then obtained. In some studies, acceptable transit links are identified and then aggregated to form routes according to some heuristic functions. In other cases, route skeletons are generated between preselected termini, and then expanded to form routes through some node selection and insertion strategy. As indicated by Baaj and Mahmassani (1991), the shortcomings of the previous heuristic approaches include: 1) failure to consider the inherent multi-objective nature of the problem; 2) limited responsiveness to the transit demand matrix in the route layout; and 3) failure to incorporate service planning guidelines (such as route length and route circuitry) reflecting the professional judgment and practical experience of transit planners. Baaj and Mahmassani developed a model incorporating AI search techniques to overcome the above shortcomings. However, that approach, like most others, is also limited to conventional uncoordinated transit service.

The design of coordinated route transit systems (timed-transfer transit systems) differs from that of uncoordinated systems in both network construction and route frequency setting. The timed-transfer system is based on the concept of transit centers, consisting of major community retail and/or employment centers, that function as effective hubs. These centers are served by local buses that serve a regional collection-distribution function, as well as trunk or main lines that interconnect various centers. Routes at the transit centers are coordinated so as to reduce the transfer waiting time. Therefore, the design of coordinated transit networks includes the following three tasks: 1) selection of transit centers, 2) construction of transit route network, and 3) determination of route service frequencies for both coordinated and uncoordinated routes.

In this paper, a hybrid heuristic approach which has evolved from Baaj and Mahmassani's design methodology is presented to accomplish the above tasks. The next section presents the overall solution approach while sections 3, 4, 5, and 6 address, respectively, the four major procedures of the solution approach, namely, the route generation procedure (RGP), the network analysis procedure (NETAP), the transit center selection procedure (TCSP), and the network improvement procedure (NIP). An illustrative application of the solution approach using the data generated from the Austin, Texas, transit network is also described. The final section presents some concluding remarks.

Solution Approach

For a given set of input information on transit demand and street network connectivity of a projected service area, the design process starts with the generation of a set of routes (using the route generation procedure, RGP) that achieves the minimum desired service levels in terms of system service directness and coverage. The minimum directness and coverage levels are provided by the designer as design parameters. Two network design concepts are offered by the RGP, which configure route networks either with or without the transit center concept. For the set of routes generated by the RGP, the network analysis procedure (NETAP) is then utilized to 1) assign the given transit demand and compute an array of network-level, route-level, and node-level descriptors, 2) determine frequencies and vehicle sizes for all bus routes, and 3) compute a variety of system performance measures. The NETAP follows an iterative procedure, starting with an initial set of frequencies associated with the given routes. In each iteration, a new set of route frequencies is determined and compared to the input frequencies. If the revised frequencies are significantly different from the input values, the NETAP iterates with the revised values serving as input frequencies until convergence is attained. In other words, the iterative process seeks to obtain internal consistency of route frequencies and vehicle sizes.

When the design is desired around the transit center concept, candidate centers are identified with the transit center selection procedure (TCSP) for the given route configuration, using several node-level descriptors computed by the NETAP. Given these centers, the routes are restructured by the RGP to insure good transfer opportunities at the centers, as well as fast and direct service between centers. The new route structure is evaluated by the NETAP. For timed-transfer systems, passenger trips are assigned to the enhanced network according to a trip assignment procedure that explicitly recognizes coordinated operation. Route frequencies are set to the same or multiple integer values for all coordinated routes. The set of transit centers can be obtained using either only one TCSP run or iterating until two consecutive sets of candidate transit centers converge. The network improvement procedure (NIP) considers a set of improvement actions which modify the previously generated bus network so that ineffective bus routes are improved, eliminated, or replaced by demand responsive service. The approach is summarized in Figure 1.

The Route Generation Procedure (RGP)

The RGP has evolved from the route generation algorithm (RGA) developed by Baaj and Mahmassani (1994), which does not incorporate the transit center concept. The RGP structures the route network around the transit center concept, if a set of transit centers is specified for a particular design. Otherwise, the RGP employs the same procedure as in the RGA. Networks generated around the transit center concept are enhanced by providing good transfer opportunities at transit centers and fast and direct service between centers. The RGP is comprised of three main components : initial skeleton formation, skeleton expansion to complete routes, and RGP termination.

The RGP starts by generating initial skeletons. In this step, the designer's knowledge is reflected in the likely minimum number of routes (M) required for the service area. The RGP does not check for termination until M routes are generated so as to avoid unnecessary steps and increase search efficiency. The initial skeletons are constructed in two sub steps. First, the RGP selects M demand node pair seeds for the initial skeletons. In the earlier initial RGA (Baaj and Mahmassani, 1994), the M highest demand node pairs in a sorted demand matrix are selected as node pair seeds. Consequently, the resulting network closely tracks the transit demand matrix, favoring direct (no transfer) service between high demand pairs, thereby contributing to lower overall user cost. To also ensure direct service between transit centers, the RGP first considers the feasible transit center pairs for which the travel time along the connecting shortest path is less than a preset maximum. If the number of feasible transit center pairs (N) is less than the specified minimum number of routes M, the RGP selects the (M - N) highest demand node pairs from the list of sorted demand node pairs, after removing the feasible transit center pairs from the list. These high demand node pairs are then eliminated from the list. After M node pairs are selected, the RGP connects each of these node pairs along either the shortest path or an alternate short path to form M initial skeletons. The alternate short path is the next shortest path in which the nodal composition is substantially different from that of the shortest path.

Through a node selection and insertion strategy, each skeleton is expanded to a complete route. A set of candidate nodes for insertion is defined for each skeleton by requiring each candidate node to satisfy the following six criteria : 1) the resulting route does not form a loop after inserting the node; 2) the node still has a low percentage of its total originating demand satisfied directly after previous insertions in other routes; 3) the

resulting route does not become circuitous ; 4) the ratio of the contributed demand satisfied per insertion cost (in-vehicle travel time) exceeds a preset desired level; 5) the required service frequency on the resulting route does not exceed the maximum operationally implementable value; and 6) the round trip time of the resulting route does not exceed a maximum allowable value. The test of route circuitry is based on a preset circuitry factor. In order to obtain faster and more direct service between transit centers, a lower circuitry factor is used for routes expanded from feasible transit center node pair seeds.

The order of inserting the candidate nodes follows one of the four sorting properties suggested by Baaj and Mahmassani (1994), including 1) direct demand satisfied as a result of inserting the candidate node (MD); 2) direct demand satisfied per increase of the total in-vehicle travel time as a result of the insertion (MDMT); 3) direct demand satisfied per increase in round trip time of the route under expansion as a result of the insertion (MDML); and 4) direct demand satisfied per increase of the sum of the total in-vehicle travel time and the round trip time (MDMC). In order to provide better service quality (in terms of transfer opportunity and accessibility) at transit centers, these centers are inserted with priority because of the expected large number of transferring trips. In other words, if transit centers are contained in the set of candidate insertion nodes, the insertion procedure considers these centers first, and before proceeding to other candidate insertion nodes

The initial M routes are examined to determine whether any route is substantially represented by any other route, by checking whether its nodal composition is a subset of another route's nodal composition. Baaj and Mahmassani suggested that this condition may be relaxed to a check on the ratio of the number of nodes of a given route that are traversed by some other route to the route's total number of nodes. If there is overlap, the RGP eliminates the subset routes and removes from the list of sorted demand node pairs all elements for which the demand is satisfied directly by the set of remaining routes. The RGP reiterates by selecting the highest demand node pair in the remaining list of sorted demand node pairs to generate an additional route. The iterative procedure continues until M 'independent' routes are generated.

The last step of the RGP is to check for termination. The RGP terminates when the resulting set of routes collectively satisfies system service directness and system coverage levels, both user specified. In this study, the system directness level (SDL) is defined as the minimum percentage of total demand satisfied directly without transfers (denoted by *dsdirmin*), and the system coverage level (SCL) is defined as the percentage of demand satisfied within two transfers (denoted by *dsmin*). This is based on the assumption that passengers will not be willing to travel by bus if their trips cannot be completed within two transfers. The termination step first checks the system directness level. If it is not satisfied, all demand pairs satisfied directly by the current set of routes are removed from the list of sorted demand node pairs. The RGP then picks the first (highest) demand node pair in the remaining list to generate an additional route using the same process for forming and expanding a skeleton to a route as described above. New routes are generated one at a time until the system directness level is satisfied. The next step is to check the system coverage level. If it is not satisfied by the current set of routes, all node pairs currently satisfied within two transfers are eliminated from the remaining list. For conventional bus service design, additional routes are generated following the same process until the system coverage level is satisfied. For timed-transfer system design, the demand between each selected node pair is checked to see if it can be satisfied by connecting one of the two nodes to a transit center. If this is the case, the additional route is generated by connecting the center to the closest (least travel time) of the two nodes. The intent is to satisfy the remaining dispersed demand in the service area through the closest feasible center to reduce unnecessary vehicle miles and obtain a more effective route network. It should be noted that varying the minimum system directness level and the minimum system coverage level results in different sets of routes. This enables the designer to address different trade-offs between service directness and system coverage. The flow chart of the RGP is shown in Figure 2.

The Network Analysis Procedure (NETAP)

The NETAP is a bus transit network evaluation tool. For the purpose of network design, it is also used to determine route frequencies and vehicle sizes. Once the RGP generates a set of routes, it is analyzed via the NETAP to determine route frequencies and vehicle sizes (if the variable vehicle sizes are allowed; then the resulting system is evaluated to determine the following eight types of information:

1) Network descriptors consisting of the following three types of data:
   - Node information, including originating flow, terminating flow and transferring flow at each demand node.
   - Link information, including link flows along each route.
   - Route information, including the round trip time for each route and the total number of passengers on each route.
2) Demand: the total number of trips in the system as well as the percentages of demand that are unsatisfied, or satisfied with 0, 1, or 2 transfers.
3) User cost: the total travel time experienced by users in the system, and the respective percentages of in-vehicle travel time, waiting time, transfer time (reflecting a prespecified time penalty considered to be equivalent to a transfer), equivalent in-vehicle-travel time cost, and equivalent waiting cost.
4) Level of service: the service frequency, vehicle size, and load factor associated with each route.

5) Operator cost: the system operation cost and the required fleet size.
6) Fuel consumption: the total system fuel consumption.
7) System utilization: the ratio of the system actual user miles to the total user miles that can be provided by the system.
8) Network shape: classification of network into one of the following four shape categories: radial, spinal, grid, or delta. In addition, the network is classified as a one-nucleus, two-nucleus, three-nucleus, or multiple-nucleus network.

The NETAP employs an iterative procedure that seeks to achieve internal consistency of frequencies and vehicle sizes. This iterative procedure consists of two major components: a trip assignment model, and a frequency setting and vehicle sizing procedure. Figure 3 shows the flow chart of NETAP.

Since route frequencies are required before performing the trip assignment process, an initial set of frequencies is assumed for each route when the NETAP is used for design purposes. The NETAP computes the round trip time for each route. Next, it utilizes the trip assignment model, described below, to assign known demand between origin-destination pairs to specific paths in the transit network; the resulting flow pattern forms the basis for computing desired network descriptors, including a array of performance measures reflecting service quality, and user and operator costs.

The trip assignment model at core of the NETAP was developed by modifying Baaj and Mahmassani's (1990) procedure to incorporate assignment rules for coordinated transit route networks. In the latter, the waiting time implications of transfers are dramatically different from those under independently operating routes, as described below.

The assignment procedure considers one node pair at a time and identifies the corresponding competing feasible paths. The definition of competing paths for a given node pair is based on a lexicographic strategy that considers two criteria: 1) the number of transfers to reach the trip destination and 2) the trip time incurred on the alternative paths. A tripmaker is assumed to always attempt to complete his or her trip by following the path that involves the fewest possible number of transfers. When several paths have the same (minimum) number of transfers, passengers are assumed to use paths with travel times within a threshold of the least time path. Completing paths are identified using efficient path search routines implemented in the LISP computer language (Baaj and Mahmassani, 1990).

Trips for each demand pair are assigned to competing routes at the origin and further assigned at transfer nodes if transfers are needed to complete the trip. At each terminal, the split among competing paths is computed according to rules that account for both uncoordinated and coordinated operations. At terminals with uncoordinated operations, a "frequency share" rule is applied. This rule stipulates that a competing route carries a proportion of the flow equal to the ratio of its frequency to the sum of the frequency of all competing routes. At coordinated operations terminals, several competing routes may be coordinated to arrive within a preset time window. In this case, a "vehicle-availability" rule is first applied to determine the probability that buses along a certain combination of competing routes are available. For a certain combination of competing routes, a "least downstream travel cost" rule is employed. Deviation of the probability that a certain combination of competing routes is available is described in Shih (1994). Once a given amount of demand is assigned to each competing path, network descriptors and system performance measures pertaining to user costs are updated. The trip assignment model sequentially considers each demand pair until all demand pairs are assigned.

As a result of the assignment, demand pairs can be classified into 0-transfer, 1-transfer, 2-transfer, and unsatisfied (more than two transfers) demand pairs. The corresponding numbers and fractions of total demand satisfied with no, one, or two transfers, as well as unsatisfied, are computed, as these constitute essential figures of merit to evaluate a particular network-route design.

Once the trip assignment procedure is complete, the resulting network descriptors are applied to determine the vehicle size and route frequency for each bus route. The joint frequency setting and vehicle sizing procedure computes the optimal vehicle size via a mathematical formulation which minimizes the total cost (operator cost and user cost) of each individual route. The required route information for the computation of the optimal vehicle size is determined via the trip assignment procedure. Once the vehicle size on each bus route is determined, the frequency of service is set to achieve an applicable maximum allowed load factor. Details of the frequency setting and vehicle sizing procedure has been presented in a separate paper (Shih and Mahmassani, 1994). For timed-transfer design, the resulting frequencies (set by the above definition) of routes that meet at the same transit center are adjusted to the same or multiple integer values for coordinated operation. Then the resulting route frequencies are compared to the input frequencies for each iteration. If frequencies of two consecutive iterations converge, the NETAP calculates system performance measures pertaining to operator cost, level of service, fuel consumption, and system utilization, and determines the network shape. If the revised frequencies are significantly different from the input values, the NETAP reiterates with the revised frequencies as the input frequencies. While this heuristic procedure is not guaranteed to converge, computational experience to date with test networks has always resulted in convergent behavior.

The Transit Center Selection Procedure (TCSP)

The TCSP incorporates several criteria that reflect commonly used guidelines in the transit industry for selecting transit centers. First, transit centers need to provide good transfer opportunities. Transfer opportunities for a demand node are measured by the number of potential routes that serve that node, or can reach it within a certain travel time. The latter ensures that the demand node can be inserted into potential routes without incurring a high insertion cost. Second, sufficient originating trips should be generated within the feeder bus service area of each transit center. Third, each transit center should be located at a major activity center that generates high total node demand (including originating, terminating, and transferring trips). In addition, each transit center should be separate from other centers at a minimum travel time to avoid service overlap. The transit centers generated according to the above criteria are supplied to the designer so that infeasible centers violating geometric, economic, and other considerations can be eliminated. Figure 4 shows the flow chart of the TCSP; details of this procedure are described hereafter.

The TCSP starts with a screening process that eliminates all nodes that do not meet certain requirements in terms of transfer opportunities, originating demand covered by the service areas, and total node demand. All nodes with a number of potential routes (NPR) exceeding a prespecified value form the set of candidate nodes (S) for transit centers. Candidate nodes with insufficient originating demand within their service areas and candidate nodes with low total node demand are then eliminated.

A sequential selection process is then used to identify transit centers one by one by checking the minimum separation travel time requirement. The order of center selection is based on total node demand, with the highest total demand node (among all candidate nodes) considered first. Initially, the set of transit centers (TC) is empty. In each iteration, the procedure checks the node with the highest total demand and removes it from the set of candidate nodes. If the travel time along shortest path between the node under consideration and each center already in the set TC is greater than a prespecified minimum, this node is added to TC. The selection process continues until the set of candidate nodes (S) is empty. In the final step, the selected transit centers are supplied to the designer in order to eliminate infeasible nodes due to other factors such as accessibility, land availability, and geographical limitations.

The Network Improvement Procedure (NIP)

The network improvement procedure (NIP) seeks to improve the set of routes generated by the RGP. It incorporates the modifications suggested by Baaj and Mahmassani (1991), which act on either the transit system coverage level or the route structure level. These include: 1) discontinuation of service on routes that suffer from low ridership or are too short or both, 2) merging of low ridership routes with other medium to high ridership routes, 3) splitting of routes where desirable, and 4) branch exchange of routes in such a way that the number of transfers at the intersection nodes of two routes is reduced. The NIP also includes a special route splitting modification for timed-transfer system design. This modification splits routes with unequal loading on two separate segments of the routes divided by a transit center. With this split, better resource allocation can be achieved; the negative effect of the transfers induced by the route split is reduced by the provision of coordinated service at the transit center. In addition, the NIP provides the option of demand responsive service (DRS) instead of fixed route operations. Under the DRS option, ineffective routes are discontinued, and unsatisfied system demand is served by DRS oriented around transit centers and associated service areas. These network improvement modifications are implemented as modules which can be used individually or in any sequence desired by the designer.

Illustrative Application: Austin Case

In this section, the Austin, Texas, urban area serves to illustrate our solution framework. The existing transit network consists of 40 uncoordinated routes with fixed schedules, operated by the Capital Metropolitan Transit Authority (Capital Metro, for short). Based on this transit system, required inputs for this application were generated, including data on the network connectivity and a peak hour transit O-D demand matrix. A total of 177 nodes are selected from the major bus stops of the existing transit system to describe the service area and associated network connectivity. The network connectivity consists street links that connect these 177 nodes and are suitable for bus operations. The generation of the demand matrix is based on Tsyganlnitzky's fluid analogy model (Tsyganlnitzky, 1979) using daily boarding and alighting data for the 40 routes provided by Capital Metro.

The application illustrates the design and analysis of a coordinated bus system with variable vehicle sizes. The minimum system directness level for the application is set at 60%, and the system coverage level is set at 80%. The maximum demand per minimum route length increase insertion strategy (MDML) is followed in the route expansion process. The shortest paths are used in the layout of initial skeletons. Other information required for the route generation, including the initial number of skeletons (M), the maximum operational bus frequency, the transfer penalty, and the maximum load factor on all routes are set to 25, 30 buses/hour, 5

minutes of in-vehicle travel time, and 1.25, respectively. In addition, the following parameter values are used in the network analysis procedure. The coefficients, a and b , in the operator's cost function are derived from the operator costs associated with different bus sizes which were provided by Capital Metro; they are equal to 2.96 and 0.0078, respectively. The values of waiting time and in-vehicle travel time are set to $9 and $3 per hour. Three commercially available vehicle sizes with 37, 27, 15 seats and with fuel efficiency coefficients of 3, 6, 9 miles per gallons, respectively, are considered in this application.

Figure 5 shows nine transit centers generated by the transit center selection procedure. These centers include a downtown transit center (node 2); two shopping malls in the north (nodes 19 and 36), a shopping mall in the south (node 84), a commercial center surrounded by major residential areas in the far north (node 9), two shopping centers located at major arterial intersections in the south (nodes 73 and 78), and two shopping centers in the east (nodes 42 and 101).

A total of 27 routes are generated by the route generation procedure with 60.30% of demand satisfied directly and 87.14% of the total demand satisfied, meeting the required minimum system directness level of 60% and the required minimum system coverage level of 80%, respectively. Information for all 27 routes is shown in Table 5.1, which includes route round trip time, service frequency, load factor, vehicle size, and nodal composition. Other information, including the number of buses required, the operator cost, the waiting cost, and the link flows for each route are also computed, but not listed in the table. One set of 21 coordinated routes resulting from the design consists of R1, R2, R4, R5, R6, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R20, R22, R23, R24, R26 and R27.

Table 2 provides a summary of system performance measures, including the demand information, user cost, operator cost, fuel consumption, system utilization, and network shape. The bus system satisfies 60.30% of demand without transfer, 23.96% with one transfer, 2.87% with two transfers; 12.86% of demand is unsatisfied. The total in-vehicle travel time is 149,769 minutes which is equivalent to $7,488.44, and the total out-of-vehicle waiting time is 43,229 minutes which is equivalent to $6,478.02. The network transfer penalty is equal to 8,590 minutes. The network total user cost, which is the sum of the total in-vehicle travel time, total out-of-vehicle waiting time, and network transfer penalty, is equal to 201,588 minutes. The system operates at a cost of $6,593.59, requires 91 15-seat buses, 52 27-seat buses and 16 37-seat buses, and consumes 289.53 gallons of fuel per peak hour. The system utilization is equal to 0.56. All figures are for the peak hour. The network is classified as a three-nucleus spinal network.

The application illustrates the solution approach using a given set of design parameters. Extensive computational testing on the overall solution framework as well as each individual procedure using Austin data were also performed by Shih (1994).

Concluding Remarks

In this paper, a solution methodology for the design of transit networks with coordinated operations is presented, including a description of its four major components: 1) route generation procedure (RGP), 2) network analysis procedure (NETAP), 3) transit center selection procedure, and 4) network improvement procedure (NIP). This model contains the following principal features: 1) ability to identify transit centers, 2) ability to configure the route network around the transit center concept, 3) design of coordinated timed-transfer systems, 4) trip assignment model for both coordinated and uncoordinated transit systems, 5) design of variable vehicle sizes which provides the transit operator with an additional choice dimension in designing the service configuration to better meet user needs and desired service level, and 5) design of integrated bus systems which serves high density demand with fixed-route, fixed schedule service and low density demand with demand responsive service.

Ongoing and future research focuses on the development of a graphic display ability to delineate the resulting transit network, and extensive numerical testing of the overall solution approach as well as of each individual procedure on other real transit networks. In addition, incorporating other service choice dimensions to improve the performance of transit systems will enhance the realism and usefulness of the model.

References

Baaj, M. H. (1990), The Transit Network Design Problem: An AI-Based Approach, Ph.D. thesis, Department of Civil Engineering, University of Texas, Austin, Texas.

Baaj, M. H. and Mahmassani, H. S. (1990), "TRUST: A Lisp Program for the Analysis of Transit Route Configuration," Transportation Research Record 1283, Transportation Research Board, Washington, D.C., p 125-135.

Baaj, M. H. and Mahmassani, H. S. (1991), "An AI-Based Approach for Transit Route System Planning and Design," Journal of Advanced Transportation Vol. 25, No. 2, p 187-210.

Baaj, M. H. and Mahmassani, H.S., (1994), "A Hybrid Route Generation Algorithm for the Design of Transit Networks," submitted to Transportation Research.

Ceder, R. B. and Wilson, N. H. (1986), "Bus Network Design," Transportation Research, Vol. 20B, No. 4, p 331-344.

Dubois, D., Bell, G., and Llibre, M. (1979), "A Set of Methods in Transportation Network Synthesis and Analysis," Journal of Operations Research Society 30, p 797-808.

Hasselstrom, D. (1981), Public Transportation Planning -- A Mathematical Programming Approach, Ph.D. thesis, Department of Business Administration, University of Gothenburg, Sweden.

Israeli, Y. and Ceder, A. (1991), "Transit Network Design," Presented at the 70th Annual Meeting of Transportation Research Board, Washington, D.C..

Lampkin, W. and Saalmans, P. D. (1967), "The Design of Routes, Service Frequencies and Schedules for a Municipal Bus Undertaking: A Case Study,". Operation Research Quarterly 18, p 375-397.

Newell, C. E. (1979), "Some Issues Related to the Optimal Design of Bus Routes," Transportation Science 13, p 20-35

Rea, J. C. (1971), "Designing Urban Transit Systems: An Approach to the Rout-Technology Selection Problem," PB 204881, University of Washington, Seattle, WA.

Shih, M. C. (1994), A Design Methodology for Bus Transit Route Networks with Coordinated Operations, Ph.D. thesis, Department of Civil Engineering, University of Texas, Austin, Texas.

Shih, M. C. and Mahmassani, H. S. (1994), "A Vehicle Sizing Model for Bus Transit Systems," presented at the 73th Annual Meeting of the Transportation Research Board, Washington, D.C., and forthcoming in Transportation Research Record.

Silman, L. A., Barzily, Z., and Passy, U. (1974), "Planning the Route System for Urban Buses," Computers and Operations Research 1, p 201-211.

Tsygalnitzky, S. (1979), Simplified Methods of Transportation Planning, Master Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Van Nes, R., Hamerslag, R., and Immer, B. H. (1988), "The Design of Public Transport Networks," Transportation Research Record 1202, Transportation Research Board, Washington, D.C., p 74-83.

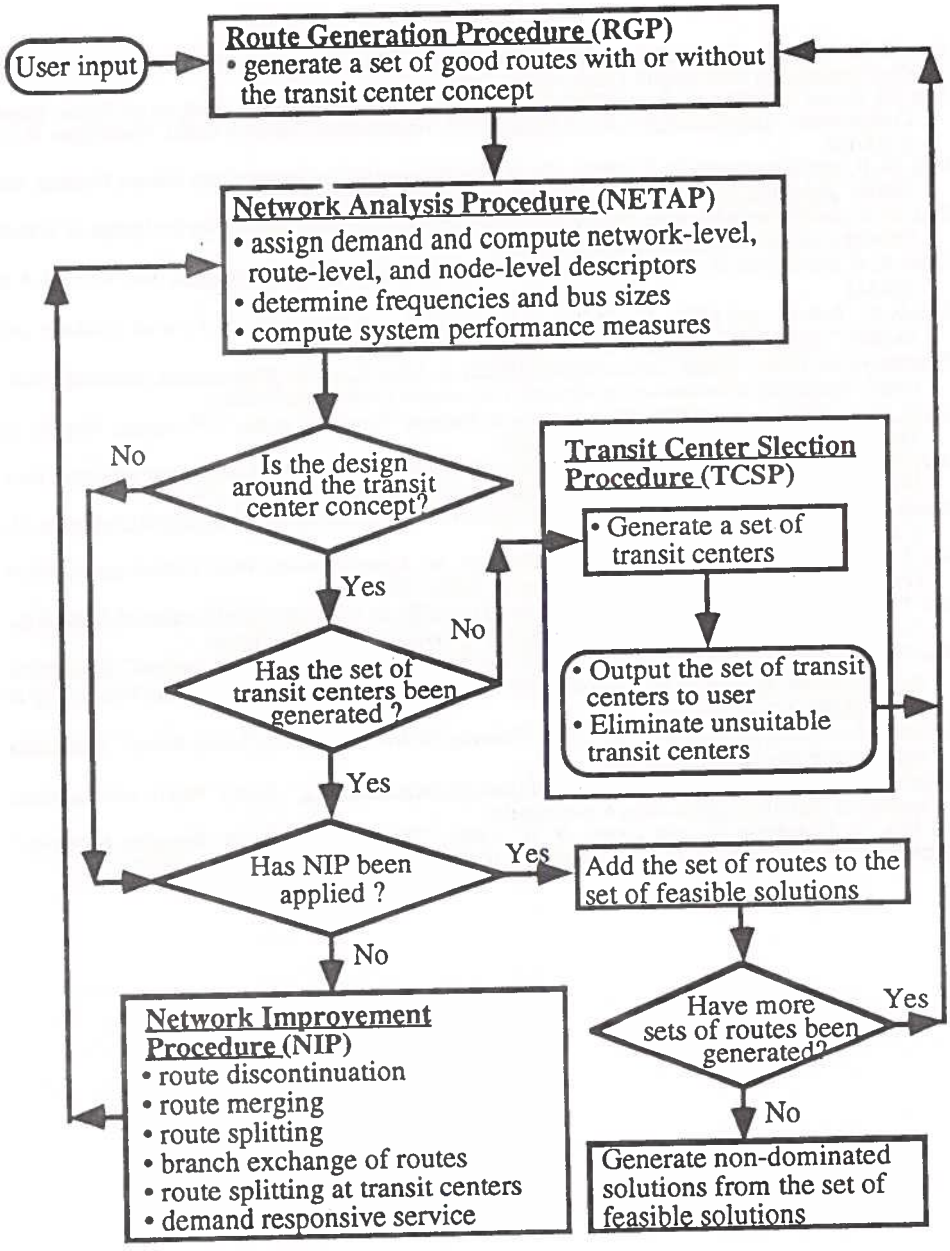


Figure 1 Solution Approach

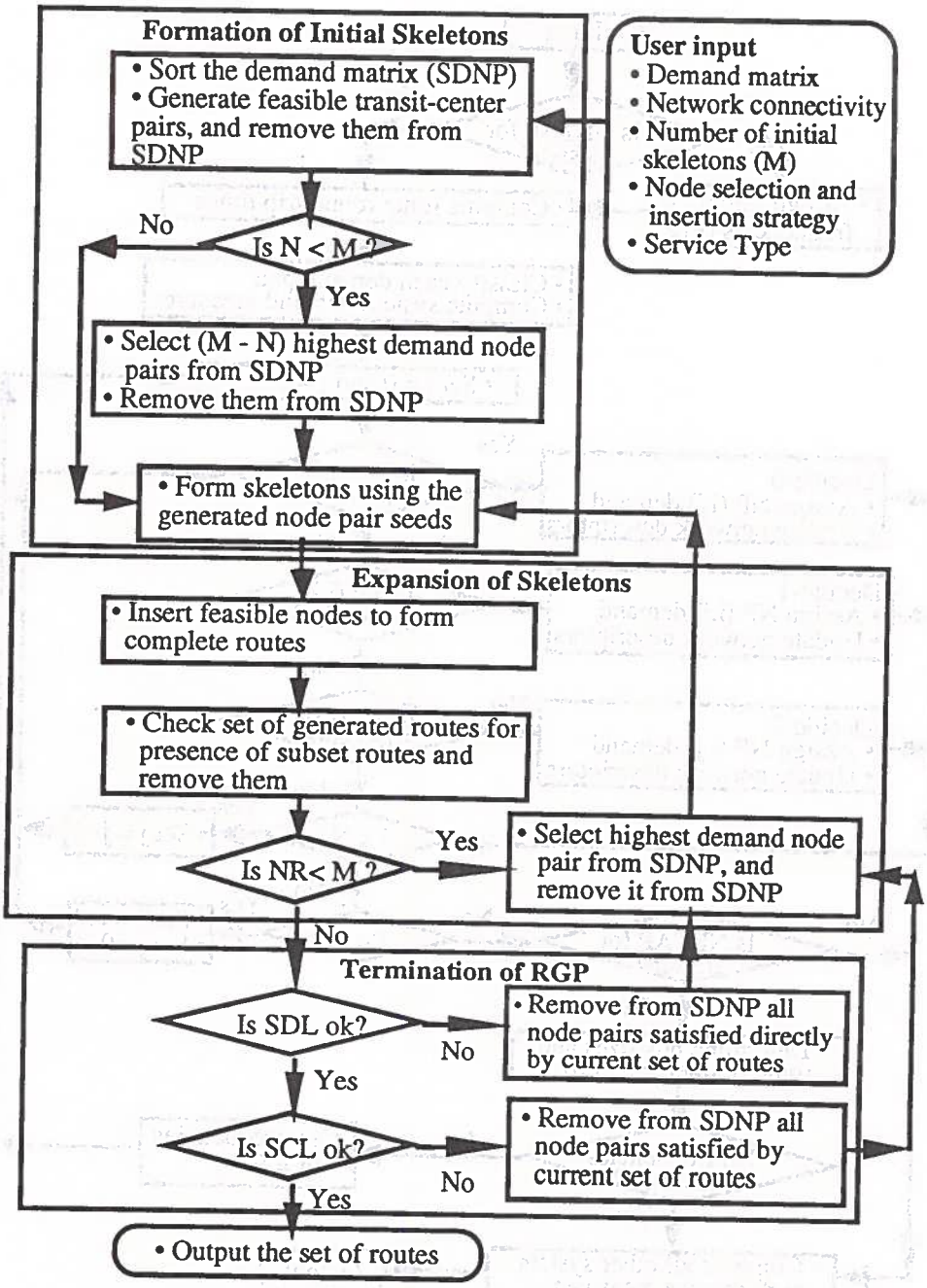


Figure 2 Route Generation Procedure (RGP)

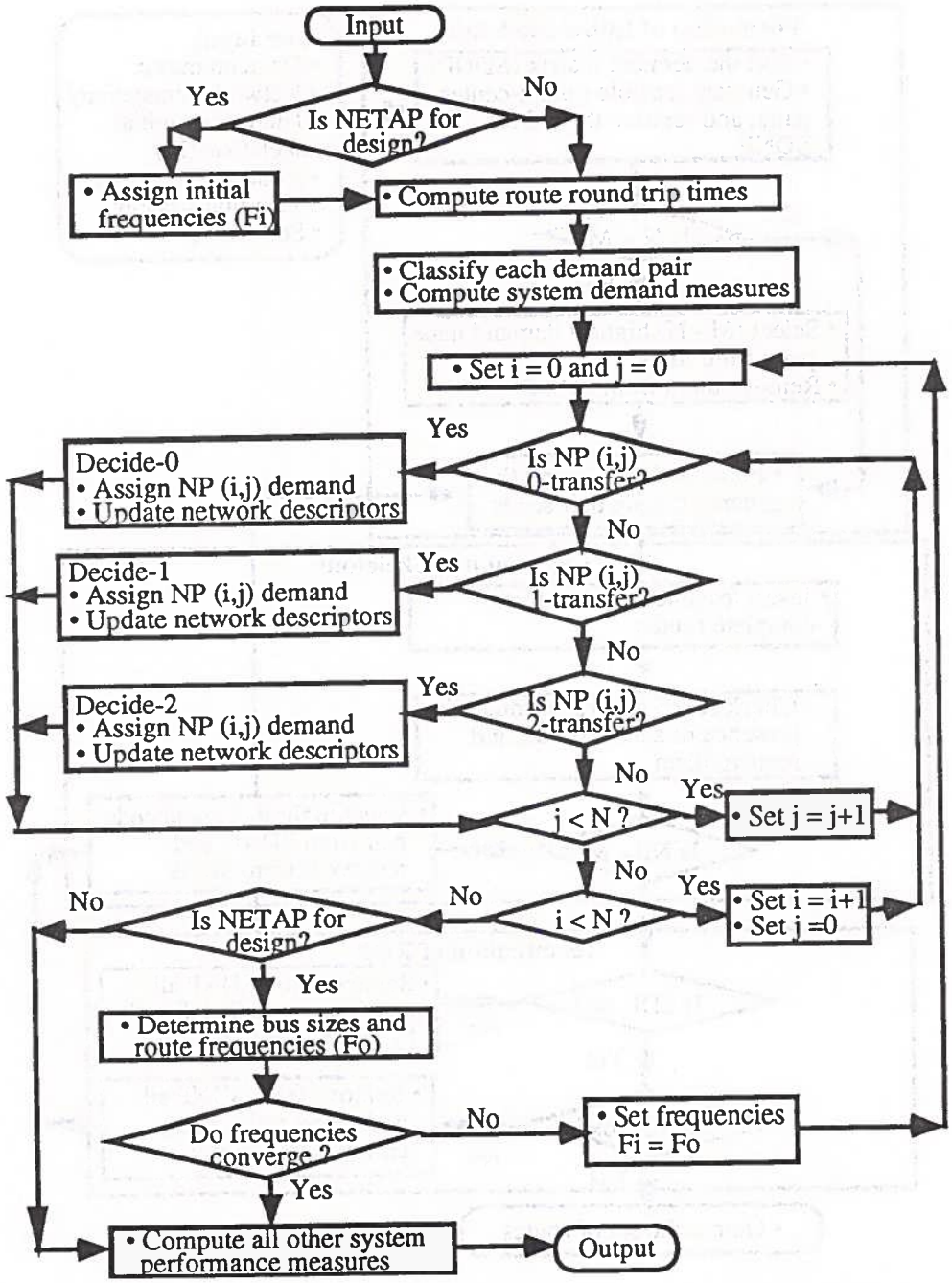


Figure 3 Network Analysis Procedure (NETAP)

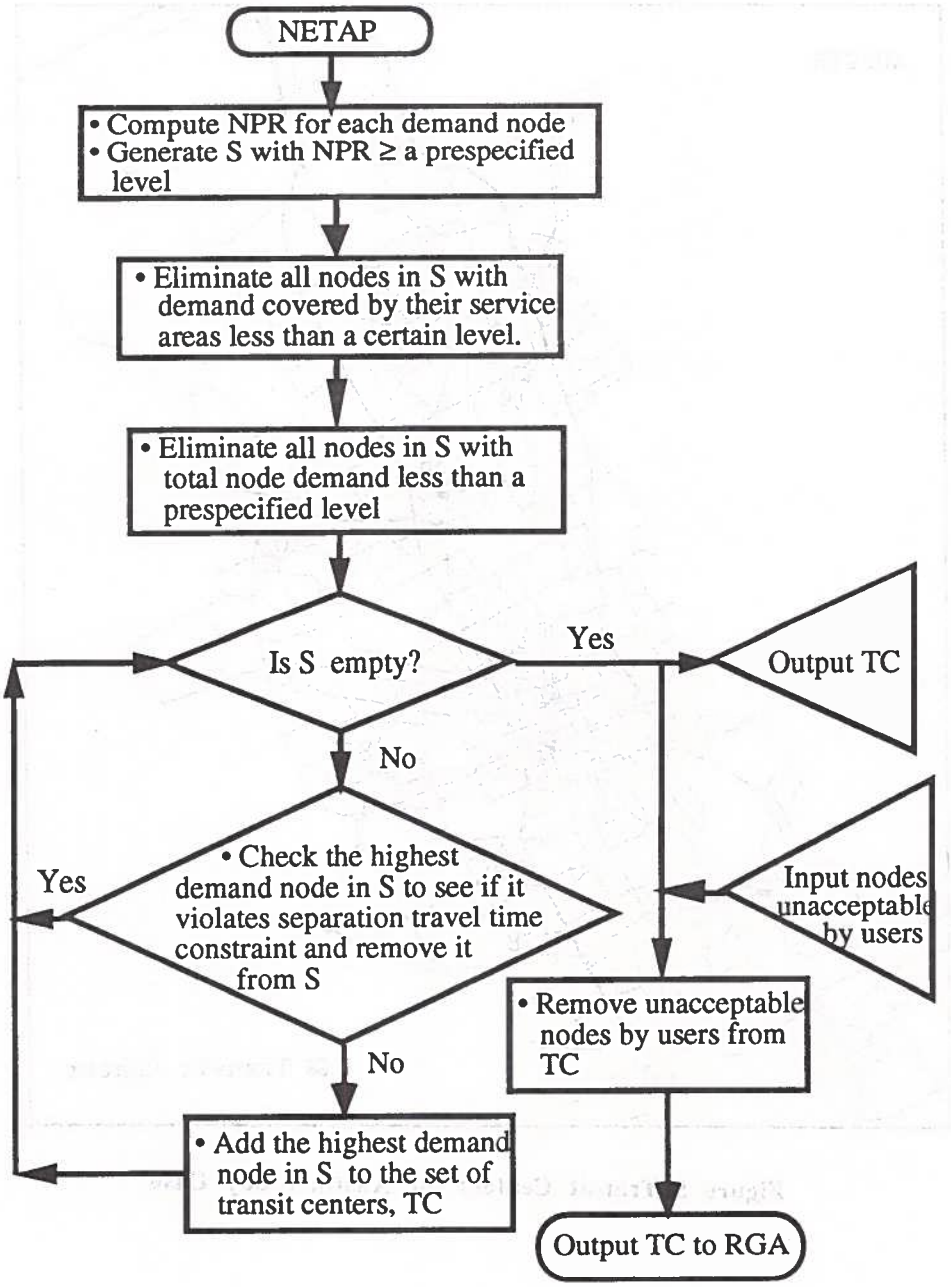


Figure 4 Transit Center Selection Procedure (TCSP)

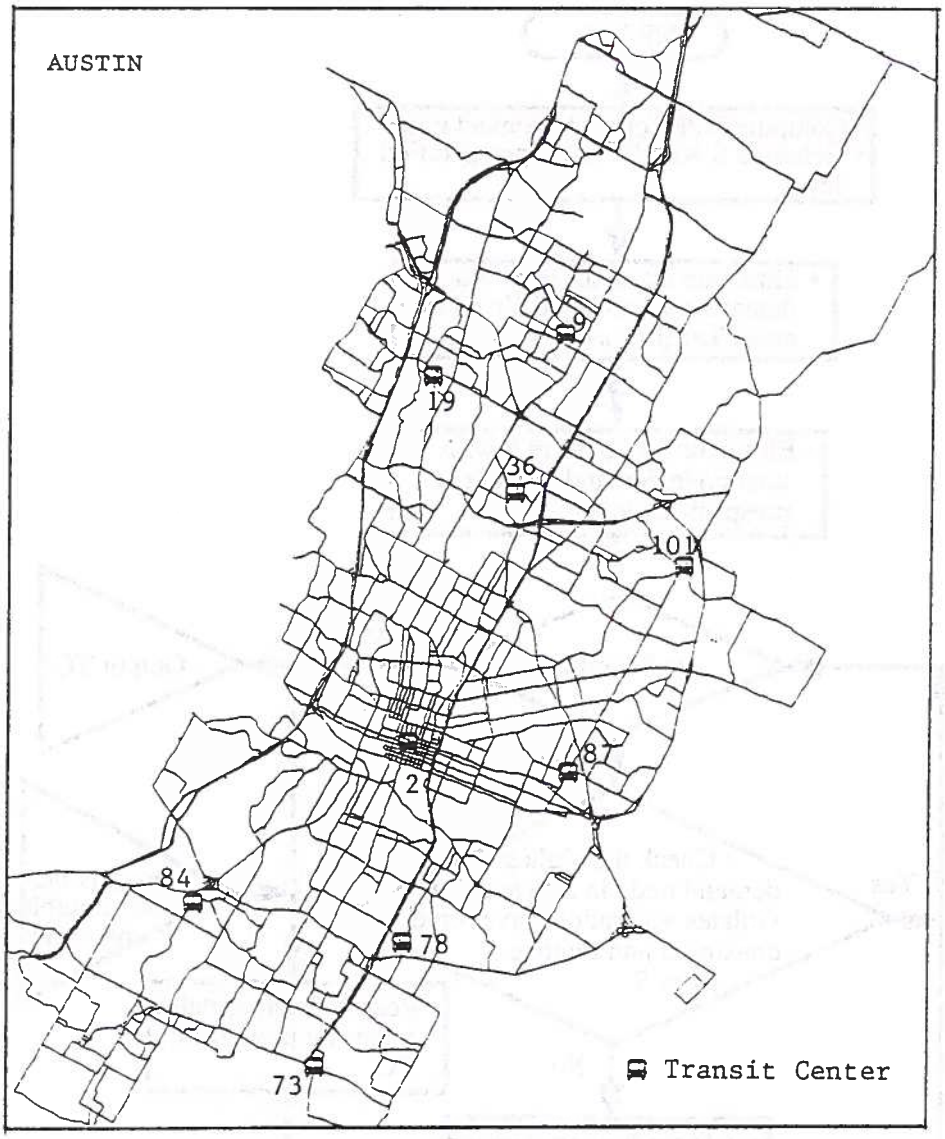


Figure 5 Transit Centers for Austin Study Case

Table 5.1 Route Information

ROUTE	RTT	FREQ	LF	BS	NODAL COMPOSITION
R1	101.0	4.0	1.25	27	(140 176 171 175 3 21 2 15 108 63 1 52 154 75 68 76 78 70 72 73)
R2	89.0	4.0	1.25	27	(3 21 2 15 108 1 52 53 54 83 66 122 84 138 139 67 57)
R3	28.6	1.0	0.27	15	(66 139 67)
R4	37.2	2.0	1.27	15	(107 140 2 108 1 52 81 64)
R5	100.4	8.0	1.23	15	(5 94 16 85 26 4 175 3 21 171 176 2 108 15 14 1 63 86 23 24 41 117 25)
R6	93.0	4.0	1.25	15	(3 2 108 1 52 53 82 71 73 121)
R7	22.0	1.33	1.25	15	(41 86 103)
R8	51.4	2.0	1.25	15	(73 78 77 40 120)
R9	114.8	8.0	1.25	15	(4 170 26 5 105 27 6 35 80 36 7 28 29 8 37 114 9)
R10	116.6	4.0	1.21	15	(90 22 15 2 108 1 52 81 53 82 129 65 84 138 137 145)
R11	117.2	4.0	1.25	27	(63 1 14 15 108 2 21 3 168 10 22 90 23 41 24 87 89 88 33 92 102 143)
R12	117.8	4.0	1.25	27	(19 29 36 80 35 61 104 60 59 20 30 11 23 24 87 41)
R13	61.6	1.0	0.6	15	(2 108 1 52 53 54 55 83 66)
R14	119.4	8.0	0.94	37	(3 2 15 168 10 74 34 62 79 104 61 35 80 36 8 37 114 9 113 155)
R15	60.4	2.0	0.93	15	(120 40 77 78 169 121)
R16	119.6	4.0	0.98	27	(154 52 1 14 10 74 34 62 79 104 35 80 36 29 164 162 161)
R17	92.8	4.0	1.25	15	(36 80 142 46 45 44 101 102 143 144)
R18	32.2	1.0	0.87	15	(67 160 159)
R19	53.6	4.0	1.24	15	(130 40 120 41 87 42)
R20	102.6	4.0	1.25	15	(34 74 141 10 15 14 1 52 53 82 71 73 121)
R21	100.6	2.13	1.25	15	(115 8 113 112 164 156 158 157)
R22	90.6	4.0	0.96	27	(3 2 108 1 52 154 75 53 68 69 70 54 55 56 57)
R23	118.4	8.0	0.96	27	(108 2 21 3 175 171 4 16 85 26 5 105 27 17 6 7 28 18 19 109 165)
R24	118.2	8.0	1.03	15	(140 108 15 22 90 10 141 62 59 20 91 100 43 44 101 102 143)
R25	60.6	1.0	1.21	15	(65 66 72 56 57 139 67)
R26	67.0	2.0	1.0	15	(19 29 8 36 46 115)
R27	74.8	2.0	1.33	15	(19 29 112 8 37 114 149)

RTT: ROUND TRIP TIME
FREQ: FREQUENCY
LF: LOAD FACTOR
BS: BUS SIZE

Table 2 Summary of System Performance Measures

******* Demand Information *******

	DEMAND	PERCENTAGE
NETWORK TOTAL DEMAND	5784.0	100
PERCENTAGE OF DEMAND SATISFIED WITHOUT TRANSFER	3488.0	60.30
PERCENTAGE OF DEMAND SATISFIED WITH 1 TRANSFER	1386.0	23.96
PERCENTAGE OF DEMAND SATISFIED WITH 2 TRANSFERS	166.0	2.87
UNSATISFIED DEMAND	744.0	12.86

******* User Cost *******

	MINUTES	PERCENTAGE
NETWORK TOTAL USER COST	201588	100
NETWORK IN-VEHICLE TRAVEL TIME	149769	74.29
NETWORK WAITING TIME	43229	21.44
NETWORK TRANSFER PENALTY	8590	4.26

FOR WAITING TIME VALUE = 9 $/hour
TOTAL WAITING COST = 6478.02 $/hour

FOR IN-VEHICLE TRAVEL TIME VALUE = 3 $/hour
TOTAL IN-VEHICLE TRAVEL COST = 7488.44 $/hour

******* Operator Cost *******

SYSTEM OPERATION COST = 6593.59 $/hour

NUMBER OF 15 SEAT BUSES REQUIRED : 91 buses

NUMBER OF 27 SEAT BUSES REQUIRED : 52 buses

NUMBER OF 37 SEAT BUSES REQUIRED : 16 buses

******* Fuel Consumption *******

TOTAL SYSTEM FUEL CONSUMPTION : 289.53 gallons/hour

******* System Utilization*******

SYSTEM UTILIZATION : 0.56

******* Network Shape*******

THREE-NUCLEUS SPINAL NETWORK

DISTRIBUTED DISPATCHING CONTROL IN PUBLIC TRANSPORT NETWORK

Andrzej Adamski

Institute of Automatics, University of Mining and Metallurgy
30-059 Cracow, Al. Mickiewicza 30/B1, Poland.

ABSTRACT

Transit on-time service in public transport has been a chronic operational problem created by the existence of unpredictible operational random events and high complexity of the public transport processes. The potential of new information and communication technologies offers great opportunities for the development of effective integrated management and control tools for improvement of transit service reliability. This paper focuses on the distributed dispatching control problem, embedded into the follow-up direct control layer in the hierarchical control system. Autonomous dispatching control actions are practically realized by vehicles equipped with on-board computers and some communication media. In the paper wide spectrum of flexible single and multicriteria dispatching control strategies are presented. Different control criteria described in terms of waiting time and control random variables moments are used as service irregularity and control cost measures. The analytical solutions with various nonlinear regulators (half-wave rectifier, relay, polynomial, elastic minimal interval types) for dispatching regularity and punctuality control are derived and numerically illustrated.

1. INTRODUCTION

Maintaining on-time transit service has been a chronic operational problem arising mainly from random operational events and high complexity of public transport processes. It is generally agreed that this is a key operational problem which affects both travellers (reliability of service) as well as operators (productivity and efficiency). Although many researchers have studied various aspects of this problem and recommended some remedial headway-based or schedule-based control actions, representative and operational control tools are still lacking. The potential of new information and communication technologies (wireless communication systems) and new existing hardware capabilities (modern AVLM systems, low-cost highly powered on-board computers, new generations of intelligent traffic selective detectors) offers great opportunities for the development of effective and flexible management and control tools for public transport. This paper focuses on the original optimal distributed dispatching control problem which is solved at the bottom layer of hierarchical system [7][12][17]. Previous research related to dispatching control problem was devoted to :

- Creation of analytical [3, 7, 8, 14, 18, 20, 32, 33], simulation [19, 21, 24, 29, 31, 36], probabilistic [35] and empirical [1, 2] models for the description of the phenomena observed on the bus lines and studying the sources of service unreliability. The analytical models frequently are based on highly simplified assumptions (e.g. uniformity of route sections, infinite capacity of buses, continuity of passengers arrivals, disregarding the influence of bus load and alighting passengers on the dwell times etc). The complex simulation models, on the other hand, pose statistical problems in the analysis of the output and, similarly as empirical models have limited transferability (these are in essence models for a given place and given set of input data). The sophisticated probabilistic model is numerically demanding and poses the representativity problems resulting from the numerical derivation of probability distributions on the basis of discrete computer representation of other (continuous) distributions.
- Implementation of the above models for developing, evaluation and in some cases experimental testing of the efficiency of simple static control actions (e.g. holding, skipping, turning, rescheduling, reserve actuation) realized on the bus lines [1, 2, 3, 12, 13, 20, 29, 32, 33, 37, 38, 39]. Most of these approaches were concentrated on simple holding threshold type (checkpoint) strategies and analytical [12, 13, 20, 32, 33], simulation [1, 31, 37, 39] attempts to optimize threshold value parameter with respect to average passenger waiting time. The analytical solutions for more sophisticated holding strategies (e.g. flexible or synchronizing type) and for multicriteria problems were presented in [12, 13, 15, 16, 17]. The general

conclusion concerning these works may be as follows: The simple control actions which are realized at appriopriate points along the route may be to a high degree profitable even when vehicles locations are roughly estimated.

- Dynamical dispatching control actions [3, 7, 12] realized in the hierarchical structures [2, 5, 7, 12, 17] or expert systems [13, 16 ,17]. This approach in general enables us to integrate sensibly in time and space simple control actions in dynamic control strategy [3].
- Public transport priority control at traffic intersections [5, 7, 12, 17, 26, 30]
- Development of hardware for vehicle location and monitoring, signal preemption, management of peripheral on-board passengers information facilities, cashless fare collection, efficient communication means between vehicles and dispatching centers etc [9 and survey references cited there].

In this paper the solution of on-line real-time adaptive, distributed dispatching control problem, embedded into follow-up direct control layer in hierarchical public transport management and control system, is presented. The main purpose of this control is the compensation of off-schedule (punctuality control), off-regular headways (regularity control) and off-transfer times (synchronizing control) transit service deviations [3, 12, 15]. It suppresses various operational disturbances and transient effects being supported by effective real-time vehicle location and communication feedback. Very important advantage which follows from the action of the dispatching control layer is the fact that for moderate and fast disturbances this layer may act autonomously because in this case the stabilization feedback ensures proper accommodation. This motivates the importance of distributed adaptive control idea developed in this paper in which the optimizing layer works periodically (repetitive scheduling) or occasionally (occasional scheduling) [3], while the autonomous dispatching control actions are practically realized by local control units i.e. vehicles equipped with on-board computers and some communication media or by human operators equipped with computer support tools [13]. At this point it should be noted that both local control equipment units have been practically realized on the basis of microcomputer system called ZPD and that the support system has controlled for several years 15 bus lines in Cracow in normal exploitation conditions. In the paper a wide spectrum of optimal dispatching control analytical solutions (punctuality, regularity, synchronizing type) for different control criteria (expressed in terms of waiting time and control random variable moments to represent the service irregularity and costs of control actions measures) are provided. The second section deals with the background of the dispatching control. The analytical solutions for a family of flexible (i.e. with recursively adjusted parameters) holding strategies, punctuality and regularity type, are derived in the next section and illustrated by numerical examples. These strategies minimize the weighted linear combination of waiting and holding times measures. As special cases the well-known results which are developed up till now (e.g. optimal "minimal headway" threshold holding strategy) are obtained. The extensions of these holding strategies to multicriterial dispatching control problems with the real-time adaptive adjustment of domination structure are also presented. The last section is devoted to the dynamic real-time on-line adaptive dispatching control problem. This on-line policy has the ability to adjust itself in real-time to suit service variations (self-tuning).

2. BACKGROUND OF THE DISPATCHING CONTROL

2.1. Standardization of notation

In order to facilitate the representation of the control plant the following standard description of the public transport line is proposed: **L (RP , GPF , PAP , BAP , PSP , CS)** where the item sequence gives the following line specifications:[12]

1. **General Route Parameters (RP)**
   - **control type**: isolated (**I**), in networks (**N**), critical (**C**)
   - **layout type**: [**length** (short (**s**), medium (**m**), lomg (**l**);
     number (**n**) and **type** (i.e. Single (SRS) / Multiple (MRS) Route]. **of stops / number of vehicles** (**m**)
   - **service frequency** (**f** : f_1 / f_2) / **service intervals** (**H:** h_1 / h_2) in the peak and of peak periods
   - scheduled **Time Buffers** (**BT**) for Terminals (T) and Timing Points (TP)
2. **General Process Features (GPF)**
   - **working conditions**: steady-state (**s - s**), dynamical (**d - s**)
   - **process characteristics**: isotropy (**iso**) / anisotropy (**aniso**)
3. **Passenger Arrival Pattern (PAP):**
   - **random (RAP), coincidential** (**CAP**), **timed** (**TAP**), **mixed (RTAP)** passenger arrivals
4. **Bus (Vehicle) Arrival Pattern (BAP):**
   - **Headways / Trip Times Distributions (HD / TD):**
     deterministic (**d**); **random** (**r**): markovian (Poisson) (M), normal (N), arbitrary (G), binomial (B)
     identical (i.d), identical and independent (i.i.d), correlated (c) distributions

* **Bus (Vehicle) Capacity (BCA): unlimited (UBCA), limited (LBCA)**
* **Vehicles stops mode: holding at all stops (a-s); on request** at selected stops (**r-s**), **mixed (ar-s)**

5. Passengers Service Process (PSP):
* **service process is : disregarded (0), dominated by boarding / alighting** passengers (**B / A**), by **both streams** of passengers (**BA**)
* **number of boarding (n^B), / alighting** (n^A) passengers **are correlated** in time and /or space (**c**)

6. Control Strategies (CS):

Information Pattern (IP):

*Real-time **Automatic Vehicle Location and Identification** system ($AVM_{l,z}$) with available information:
(l=1) full identification and location information, (l=2) incomplete information
(z=1) bus load and passengers boarding /alighting information , (z=2) incomplete information
***Statistical** information available (**s**)

Type / Kind (T / K):
* **punctuality (P), regularity (R), synchronizing (SYN)** type
* holding (**H**), skipping (**S**), rescheduling (**RSCH**), reserve (**R**), turning (**T**), buffering (**BT**)

For example the specifications of the typical Basic Route Module (BRMI) are presented in Table 1 [3, 12]. This means: The bus line segment with two single route stops working in the stationary isotropic conditions and service intervals H realized by m buses. The buffer BT is assigned to timing point TP. The passengers arrivals are random and trip times of following buses at a given bus stop and the same bus at the following stops are correlated. It is assumed that passengers board the first arriving bus (i.e. the bus capacity is unlimited). In passenger service processes the numbers of boarding passengers to consecutive buses are correlated. In order to facilitate the description of the optimization problems the following standard notation is proposed:

P $_{\min}$ **u Q** | Ω where the sequence gives the following item specifications of optimization problems:

Ω-. Set of admissible controls in the decision space U

Q. Criterion function $Q:\Omega \rightarrow Y$ where **Y** is a linear space called the criterion space

u. Vector of decision variables

P. Scalar optimization problem **(PO)** or Multicriterial optimization problem **(MOP)**. In the last case the criterion space $(Y, \leq D)$ is partially ordered by a strictly positive convex cone D. This makes it possible to find in a given set $Q(\Omega) \subset Y$ of admissible performance its D-minimal element $q=Q(u)$ which fulfils the condition $(q-D) \cap Q(\Omega) = \phi$ called the polyoptimal performance and correspondingly the control element $u \in \Omega$ called the polyoptimal solution.

2.2. Dispatching control premises

The planned passengers service determinants in public transport are described in terms of the number of operating vehicles and the service schedules: in space (routes, stops locations with assigned vehicles stoppage modes) and in time (arrival /departure times to/from timing points). The original premise for dispatching control application is the discrepancy between the planned and realized service. Traffic phenomena observed in functioning of public transport vehicles along the routes are very complex stochastic processes, due to the occurence of disturbances (e.g. unreliable operation of driver-vehicle complex), very complex interactions (from other vehicles, passengers demand and system environment), essential and diversified dynamics of traffic phenomena (e.g. structural instability called bunching). This is why the stochastic environment with high level of uncertainty and human behavioural anisotropy is a typical component of these phenomena. In this field specific to public transport the "control by opportunity" context of control problems in natural way appears [17]. It was practically observed and analyzed in detail during the dispatching control experiments conducted on several bus lines in Cracow. This real-time control feature may be roughly presented as follows. The possibilities of profitable (in the sense of present and future quality of service measures) control actions on the bus line appear and disappear. In the first case they also dynamically evolve in the time-space context along the route. This means that (if it is at all possible) we have at a given moment the opportunity to realize very profitable control actions only at a specified control point (points) on the route. The difficulty concerns the fast identification of these control points along the route and application in real-time effective dispatching control actions. The public transport system is a complex large scale multigoal and functionally non-homogeneous system with essential (in some operating ranges) non-linearities and high sensitivity of quality measures on parameters fluctuations [7, 26, 29, 31]. In the control of such a system we can for example postulate the overall goal to be, in a very general sense, the efficient utilization of resources (e.g. energy, manpower, vehicles, environment, capital) in the passenger service satisfying some quality specifications (e.g. service reliability) consistent with goals and constraints which may be imposed by society.

However, we are then concerned with the broad spectrum of decision-making and control functions (e.g. process control, scheduling [22, 28], planning etc) which play role in the effective operation of the system with respect to its transportation goals. The important control premise is concerned with the total integration of the control plant by control structure (horizontally from supply resources to final product i.e. passenger service, vertically from global system management to single vehicle control). The control problem in this generalized context is extremely difficult to handle in the traditional way. The proposal of new dispatching control methods (off-line and on-line real-time type) embedded into an integrated hierarchical public transport management and control system was widely presented and deeply motivated in [17].. These control methods to a high degree overcome and relieve the above-mentioned difficulties by proper integration of new control philosophies (e.g. multilayered functional hierarchical structure, multicriteria management approach, on-line real-time adaptive and distributed control actions) with the presently existing hardware capabilities [17]. In this hierarchical structure the optimization layer is directly connected with dispatching control. It solves the transit planning optimization problem aimed at efficient (from the user and operator viewpoint [22][28]) transport of O-D transit passengers. In general it contains two main interrelated elements i..e. Network Route Design (setting frequencies) and Service Schedule (timetable development, bus and driver scheduling) [22]. At this layer a high spectrum of operational requirements have to be fulfilled by the solution e.g. very high efficency and productivity, high temporal and spatial flexibility [different operational modes (line-haul, route deviation, demand-actuated), different service variants throughout the day, schedule and route modifications], robust features, reduced solution preparation time. To cope with these difficulties and to overcome those presented in preceding section, several mechanisms were used at this layer. First of all, the solution task of this problem is split into the direct dispatching control layer which realizes regulatory (follow-up) control on the lines and provides a stable steady-state operation of the plant, and optimization layer realizing optimizing (transit planning set-point) control of the steady-state operated processes [2, 7, 25]. This ensures that in the transit planning problem the steady-state approach and a simplified (to essential nonlinearities and dominant disturbance effects) model are well motivated. Additionally, this makes it possible to effectively use feedback to increase robust features (relative independency of stabilization task on bus line model accuracy and errors of the optimizer) and operational flexibility e.g. real-time on-line repetitive scheduling [3]. At this point it should be emphasized that without the existence of adaptive (O-D matrix updating) and dispatching control (stable, steady-state guarantor) layers it is very difficult to formulate a reasonable transit planning optimization problem (e.g. due to uncertainty in fixed demand assumption and in steady-state trip cost estimates). To our knowledge no approach tackles this problem explicitly. The output from this layer includes the set of bus routes and schedules (which may be in time-space context interpreted as planned bus trajectories). At the direct control layer the optimal dispatching control problem is solved. The main purpose of this control is stabilization of bus trajectories (punctuality control) or headways (regularity control) around schedule trajectories / headways and consequent counteraction to random off-schedule deviations [2, 3, 12]. The multilevel hierarchical structure makes it possible to cope with control task complexity. It makes full use of the system structure i.e. partition (decomposition) of this task into a collection of smaller local problems (e.g. bus lines or its modules) and their integration (via coordination instruments e.g. to fulfil the inter-line transfers requirements) in order to achieve the overall control objective. The benefits result from the fact that the labour demand increases nonlinearly with the problem dimension. In this case the system is decomposed not only according to its natural (geographical) structure but also along lines of weak interactions, which gives significant reduction in the required frequency of coordination actions [17]. The efficient solution in this approach is conditioned by the high control plant description flexibility in the sense of various degree of aggregation: in space (stop, part of line, line) in the population of vehicles (one vehicle, several vehicles, all vehicles on the line) and in time (single events, dynamical states, steady-states). For this purposes the line elementary unit operational modules (EU) are introduced (see Figs. 1 and 2). It contains:the inter-stop line segment (a+b) and two points, one service point (e.g. bus stop) and second timing or disturbance point at which control actions (e.g. at signalized traffic intersection) or disturbamces can be introduced. These modules may be described on various levels of aggregation in terms of PAP, BAP, PSP processes. The whole line operation, being aggregation of passengers and vehicles processes can be flexibly represented by a connection of the EU modules. Using the simple EU modules aggregation rules we can conveniently describe the line (or its parts) by means of connection of EU modules (of the same type) and two standard (in general substitute) modules BRMI/BRM (Fig. 2). The aggregation in the vehicles population can be realized by assignment of the PAP_i, BAP_i, PSP_i to individual vehicles. Similarly, the aggregation in time can be realized by introduction of different component processes for various modules. As can be seen the parallel connection of modules can represent the overlapping segments of various lines (Fig 3.). The control plant module-based representation together with the simple connection algebra is advantageous because it is a

flexible computer-oriented tool for representation of various network problems and enables us to concentrate on the solution of the distributed "small" control problems for modules [12] (see Table 1 and Figs. 1-4). Different stochastic module characteristics and aggregation /disaggregation tools are developed in [12]. For illustration the following formula for the i-th bus stop time at the j-th stop on the homogenous bus line modules and assumed linear PSP model [14] (see Table 1: $\tau = C + kH$ where C- is dead time, H- service interval) is presented:[3][12].

$$\tau_{ij} = Cr^{j}\sum_{k=1}^{i-1}\binom{j+k-2}{k-1}s^{k-1} + \binom{i+j-2}{j-1}r^{j-1}s^{i-1}\tau_{11} - r^{j-1}\sum_{k=2}^{i}\binom{i+j-k-1}{j-1}s^{i-k}H_{k1} +$$

$$+ s^{i}\sum_{k=2}^{j}\binom{i+j-k-1}{j-k}r^{j-k}\overset{\bullet}{\tau}_{0k} + \sum_{k=1}^{i}\sum_{l=2}^{j}\left\{\binom{i+j-k-l-1}{j-l}r - \binom{i+j-k-l-1}{i-k}s\right\}r^{j-l}s^{i-k}T_{kl-1}$$

where

$r=1/(1-k)$; $s=1-r$; $\tau_{0k}^{*}=\tau_{0k}+T_{0k}$ corresponds to the reference trajectory (e.g. the bus ahead of the first bus)

Proof:

For $i = 1$ the τ_{1j} formula can be easily checked for $j = 1, 2$. Assuming its validity for $j = k$ we obtain $\tau_{1k+1} = r\tau_{1k} + s\tau^{*}_{0k+1}$ which confirms its validity for $j=k+1$. Similarly for $j=1$ the τ_{i1} formula can be easily checked for $i=1,2$ and assuming its importance for $i=k$ we obtain $\tau_{i1} = s\tau_{i-11} + r\tau^{*}$ which confirms its importance for $i=k+1$. In the general case for fixed i, using the well-known formula $\binom{N}{j-1}=\binom{N-1}{j-2}+\binom{N-1}{j-1}$ we obtain $\tau_{ij} = r\,\tau_{ij-1}+s\tau_{i-1j}$ which confirms the formula validity for arbitrary fixed i .Similarly for fixed arbitrary j-th we obtain $\tau_{ij} = r\,\tau_{ij-1}+s\tau_{i-1j}$. Hence the formula is valid for arbitrary indices i and j. Thus the proof is completed. ♦

Remarks:

1. This formula may be used for aggregation purposes (i.e. for generation of substitute characteristics of the aggregated modules , parts of lines and lines at whole.
2. It is also a very convenient tool for illustration and analysis of the disturbances propagation along the route (e.g. as special cases the well-known Newell-Potts (1964) and Chapman -Mitchel (1978) formulae can be easily obtained [12].
3. Assuming the practically verified random representations for quantities in this formula the adequate distributions for modules , parts of line and lines can be derived.

In the next section this approach will be used for the derivation of optimal dispatching control actions.

3. FLEXIBLE DISPATCHING CONTROL TOOLS

Traditionally, the dispatching control and supervision is run by human dispatchers (located mainly at control centers, terminals, control points) and equipped with some hadware and communication media). Experience has shown that control decisions realized by human operators are characterized by:

- limited range and repertory depending on individual skills, professional experience and intuition .
- considerable amount of subjective heuristic logic
- diversified level of efficiency, violently reduced with elapsed working hours or superposition of several disturbances at the same time (e.g. at multibus line terminals often the selection of appropriate decision in complex peak hour traffic situation exceeds the capabilities of human dispatchers).

In the light of high sensitivity of control plant with respect to irrational control decisions, the improvements of efficiency of dispatching actions, especially in difficult rush hour periods are of paramount importance. This is why the support tools for dispatching control purposes at control centers, timing points and at vehicles are developed and will be presented in this section.

3.1. Flexible holding strategies

Let us consider a bus line realizing service with high/medium frequency so that the random passengers arrival pattern is well motivated [1, 9, 29, 32, 39]. We assume that the dispatching holding control strategy may be used to an early vehicle at any point along the route and is mainly directed to compensate vehicle headways (in the case of regularity control)/ scheduled times (in punctuality control) variability. This means that control impacts on the mean scheduled values can be negligible). The control actions u(e) are based on

off-schedule deviations (e) which are random variables having normal type N(0,σ) distribution [1, 36] with recursively estimated in real-time parameters [9]. The idea of the flexible holding strategy with the half-wave rectifier type regulator is presented in Fig. 5. The flexibility of the presented holding strategies result from the fact that for a wide family of nonlinear u(e) dependencies (i.e. types of regulators see Table 4) the optimal analytical solutions are derived, in which parameters can be adaptively adjusted. The general optimal dispatching holding problem may be stated as follows:

$$\mathbf{PO_{min}} \quad u \quad S(u) = \gamma\ Q_1(u) + (1-\gamma)\ Q_2(u) \quad | \ u \geq 0 \qquad (1)$$

where weighting by constant $\gamma \in [0,1]$, $Q_1(u)$ and $Q_2(u)$ components represent in general holding costs (i.e disutility measure of on-board passengers in the held bus) and disutility measure of arriving passengers in control zone respectively. In the papers [2, 12, 20, 36, 38], for example, the average amount of time for a bus to be held $Q_1(u)=E(u)$ and average wait time expressed by the Welding formula $E(w)=E(h)/2 + Var(h)/2\ E(h)$ were considered. In order to facilitate intuitive explanations let us start the presentation of the flexible holding strategies from this case and the half-wave rectifier type regulator $u(e) = \max [0, Pe]$; $P \in [0,1]$ described by the control parameter P (Fig. 5, Table 2). The optimal solution for regularity control and $\gamma \neq 1$ has the simple form

$$P^* = 1 - d\,v/\sigma \qquad (2)$$

where $v = \gamma H_S /(1-\gamma)\sigma$, H_S- is scheduled headway, σ- is standard deviation of e, and $d = E|e| = \sqrt{2/\pi}$

Remarks: 1. For $\gamma=1$ and $\gamma=0$ the obvious intuitive solutions $P^*=0$ and $P^*=1$ are obtained.

2.The special case ($P=1$, $e=h_{min}-h$) of this control strategy is the well-known "**minimal-headway**" threshold holding strategy (in which threshold value $h_{min} \in [U, H_S]$ where $U= (0.65 - 1)H_S$ is used in practice). As can be easily seen the optimal value of the threshold parameter is equal to $h_{min} = \max [U, (1-2\gamma)H_S/(1-\gamma)]$ which confirms the results presented in [36, 38]

3. The sensitivity and robust features of the solution may be conveniently illustrated by the analytical solution of bicriterial Min-Min problem with Pareto preference ordering [17]. The set of compromise solutions AB, in the criteria space is shown in Figs. 7.1.and 7.2. For a fixed value of σ parameter the relative sensitivity of Q1 and Q2 criteria with respect to variability of the γ parameter is well illustrated by the tangent at the point P* to the compromise set (Fig. 7.1). As can be seen the increase of scheduled headway H_S improves the sensitivity feature. For a fixed H_S value and decreasing σ parameter the criteria tend to be more cooperative (i.e. from random service $\sigma=H_S$ to full cooperation for regular service $\sigma=0$) (Fig. 7.2).

4. The adaptive implementation of this strategy with recursive adjustment of regulator parameter is presented in Fig. 5. For real-time recursive estimation of the time-varying parameter v several RLS methods may be used (e.g. RLS with exponential forgetting) [12].

The possible extensions of the approach illustrated on this holding strategy may be conducted in a few (not necessary disjoint) directions for other types of: **performance criteria (see Tables 2), regulators (Table 4) and strategies e.g. punctuality, synchronizing type (see Table 4 and Fig. 6).** Various passengers disutility functions may be used as objective functions in our optimization problems. As is well-known, the common choice in literature is the mean waiting time. The attractiveness of such a measure however can be misleading because it penalizes long and short waiting times with the same weight and does not take into account some service reliability aspects. This is why the variance and second order moment of control and wait time random variables (which more penalizes long than short waiting times) may be more attractive disutility measures. In particular, the optimal analytical solutions of optimization problem (1) for very general cases i.e. Q1 equal to $E(u^k) = a_k\ (k-1)!!\ (P\sigma)^k/2$ where ($a_k=1$ for even and $a_k = d/\sigma$ for uneven k-th), $E^2(u)$, and Var(u) and Q2 equal to E(w), $E(w^2)$, $I=Var(h)/2H_S$ and $I=Var(w)/\sigma^4$ are also derived and are presented in Table 2.1. To illustrate these results the corresponding numerical examples to four bus lines with various parameters of service intervals distribution are also presented in Table 2..2. At this point it should be noted that all considerations concerning the multicriteria solutions and adaptive implementations may be applied to these very general cases. In Fig.7.3. for illustration the Set of Compromises for various criteria (second item in Table 2.1) are presented. As can be seen for fixed σ and H_S the Compromise Set range increases for higer order moments.

Example 1: Let us consider the bus lines with scheduled intervals H_S and intervals standard deviations σ :
A. $H_S = 10$ (min), $\sigma = 1$ (min); B. $H_S = 10$ (min), $\sigma = 2.5$ (min); C. $H_S = 5$ (min), $\sigma = 1$ (min);
D. $H_S = 5$ (min), $\sigma = 1$ (min) ; (see Tables 2.1 and 2.2)

Remarks:

1. For the second case, formally when k increases then $P^* \rightarrow A$ ($A=\pi/(\pi-1)$). Hence delaying the vehicles for a longer period then the supplement to H_S (i.e. for $e=H_S - h$) seems to be beneficial.

2. From the 4-th case in the table it can be easily seen that when we take into account the higher control moments then the longer holding is beneficial due to lower importance of these moments with respect to the irregularity service measure.

In Table 4 the optimal analytical solutions for a wide spectrum of nonlinear types of regulators (relay, polynomial, elastic minimal interval) for regularity and punctuality control cases are presented. The relay type regulator $\mathbf{u(e) = a\ 1\ (e - e_m)}$ where $\mathbf{1\ (\bullet)}$ is Heaviside unit function and two control parameters $\mathbf{a}$ and $\mathbf{e_m}$ represent the magnitude of the holding time (a) and threshold value (e_m) of the regulator insensitivity zone respectively. The two types i.e. A. $\mathbf{a \neq e_m}$ and B. $\mathbf{a = e_m}$ of this regulator will be applied in derivation of optimal parameters of this holding strategy. Denote the probability density and distribution functions of the standard normal distribution N (0,1) by f(u) and F(u) respectively. Defining $\mathbf{C(u)= f(u)/[1-F(u)]}$ we can derive the $Q_1= E(u^k) = a^k\ f(u)/\ C(u)$ where $\mathbf{u = e_m / \sigma}$ and $Q_2=E(w)$ [12] and solve the optimal control problem (1). The optimal values of the control parameters for two types (A and B) relay regulator and approximation of the C(u) function are presented in Table 3.1 and numerically illustrated in Table 3.2. The flexible minimal headway strategy can be realized by the regulator $\mathbf{u(e)= P\ (e - e_m)\ 1(e - e_m)}$ where similarly as in relay regulator the control parameters $P \geq 0$ and $e_m= H_s - h_{min}$ represent the regulator gain (P) and insensitivity zone (e_m) coefficients respectively. As the special case (P=1) of this strategy corresponds to the well-known minimal headway strategy we consider this case first and in the sequel generalize it to the $P \geq 0$ case. The general problem is formulated as in (1): $PO_{min}\ S(u)= \gamma\ E(u^k) + (1-\gamma)\ E(w)$ where

$$\mathbf{E\ (u^k) = f\ (u)\ C_{k+1} / \sigma} \qquad (3)$$

and $C_1 = \sigma / C(u)$; $C_2 = \sigma^2\ [1-u / C(u)]$; and for $k \geq 3$ the following recurrence relation hold

$$\mathbf{C_k = \sigma^2\ (k-2)\ C_{k-2} - \sigma\ u\ C_{k-1}} \qquad (4).$$

Remarks:**1**. The following property for C_k can be inductively proved: $\mathbf{d\ C_k / d\ u = -\ C_{k+1}/\sigma}$ (5)

Proof:

For k=1 and k=2 this property can be checked by differentiation of C_1 and C_2. We assume the validity of (5) for k= n-1 and n and prove it for k= n+1. From the recurrent relation (4) for C_k it follows that $dC_{n+1}/du = \sigma^2\ (n-2)\ dC_{n-1}/du - \sigma\ C_n - \sigma\ u\ dC_n/du$. Replacing the both derivatives in the right hand side according to (5)and using (4) we obtain $- C_{n+2}/\sigma$. Thus , the proof is completed. ♦

2. For $\gamma=0$ (i.e. Waiting Mode (WM)) the optimal solution is $e_m=0$ i.e. $h_{min}=H_s$. For $\gamma=1$ (i.e. Control Mode (CM)) the optimal solution $e_m = - C_{k+1}/C_k$ can be derived from the $C_{k-1}=0$ condition.

3. For $\gamma \neq 1$ the general optimal solution is equal to $e_m= \sigma\ v\ k\ C_k / C_1$ where $v = \gamma H/ (1-\gamma)\sigma$.

Proof.

$dS(u)/du = (-v/\sigma^2)[C_{k+2}+\sigma\ u\ C_{k+1}] + u\ C_1= 0$ and from (4) the expression in square brackets can be replaced by $\sigma^2 k\ C_k$ so that the solution of this equation is equal to $u = e_m / \sigma = vk\ C_k / C_1$ ♦

The analytical closed form solution for k=1,2,3 (with convenient analytical representation of the C(u) function derived in [12]) are presented and numerically illustrated in Tables 3. The general case with $P \geq 0$ can be considered in a similar way [12]. The analytical solutions and illustrative examples are also presented in Tables 3. The family of flexible polyoptimal types regulators which can adjust control parameters to particular lines , control points, time periods etc. can be represented by $\mathbf{u\ (e) = a\ e^n\ 1\ (e)}$. In this case $Q_1=\mathbf{E(u^k)=a^k\ \sigma A(nk-1)}$ where $A(m)=[\ (2/\pi)^i\ m\ !!\ \sigma^m\ /2]$ and $i=[1-(-1)^m]/4$ and so the optimal value (in the sense of problem (1)) of the control parameter a is the solution of the following algebraic equation: $v\sigma k\ A(nk-1)\ a^{k-1} + A(2n-1) - A(n) = 0$. The solution for k=1 is presented in Table 4. In the special cases: $\gamma=0$: $\mathbf{a=A(n)B(n)/\sigma^n}$; $\gamma=1$: $\mathbf{a=0}$. As can be easily checked for k = n =1 the solution (2) is obtained. The examples of equivalents of punctuality and synchronizing types of strategies are presented in Table 4 and Tables 5. Consider a simple holding punctuality strategy for transfer synchronization [15] between C and N buses: buses depart immediately if they are late (with respect to scheduled departure times t^S) and at their scheduled departure time otherwise. For this case the transfer wait function can be compactly expressed as:$w(x,y)=x+\max\{0,\ v_C\}$ if $x+\max\{0,v_C\}>0$ or $w(x,y)= y+\max\{\ 0,\ v_N\}$ otherwise, where $v_C=t_C^S-t_C$, $v_N=t_N^S-t_N$ are off-schedule deviations of C and N buses at the transfer point, and $u_1=\max\{\ 0,\ v_C\}$, $u_2=\max\{\ 0,\ v_N\}$ are control variables. Analytical formulas for different disutility functions of the transfer waiting time and assumed Shifted truncated Exponential distribution of bus arrivals times to the transfer point are derived in [15]. The comparison of control efficiency (relative to no-control case) in terms of values of various disutility functions and offset times control variables Δ, are presented in Tables 5 . It is easy to see that the application of control at the transfer point is highly profitable. The idea of the multicriteria approach in this case with adaptive adjustment of the domination structure according to dynamically evolving "arrivals windows" on both lines. is presented in Fig. 6. The preference ordering is introduced by means of convex preffered cones e.g. Pareto preference is introduced by $D=R^2_+$ cone. To obtain the adaptive control actions we relate the domination structure with thecurrently detected transfer situations. To achieve this, it is proposed

to represents domination structure in adaptive self-organizing way by dominated cones generated in real-time by preference directions v^i of the "arrival windows " modes which are most suitable for the current transfer situation. In Fig. 6. the domination cone $D_{\alpha\beta}$ dependent evolution of Compromise Set is illustrated. It is easy to see that for the D_{BA} domination cone the compromise set corresponds to the set of points AM-BM, from which final solution will be selected. The polyhedral D_{BA} cone can be considered as convex cone generated by preference directions $\{v^1,v^2\}$ connected with arrival windows modes $D_{BA}=\{\sum_{i=1}^{2} a_i v^i \mid a_i \geq 0\}$. In a similar way the other operational modes may be represented.

Example 2: Let us consider the bus lines with scheduled intervals H_s and intervals standard deviations σ : A. H_s = 10 (min), σ = 1 (min); D. H_s = 5 (min), σ = 1.5 (min) ; (see Table 3.1 and 3.2)

4. DYNAMIC DISPATCHING CONTROL

The present hardware capabilities and costs make it possible to realize dispatching control actions in a fully decentralized hardware structure, by means of autonomous actions (e.g. signal pre-emption, holding, rescheduling, skiping etc.) of individual vehicles, equipped with on-board computers and efficient communication means [5]. In this section the automatic dispatching control aspects connected directly with protection of scheduled service are considered. For illustration we take dynamical model of the bus line [3][5], in the form:

$$x_{j+1} = A_j^1 x_j + A_j^n x_{j-n} + B_j u_j + A_j^1 z_j \qquad (4.1)$$

where (see Figs 1 and 4) i = 1,., m the bus /j = 1,..,n of possible control and /or service points along the route (e.g. bus stops, traffic intersections, terminals etc.) indices

x_{ij} - state variables representing off-schedule deviations ($x_j \in R^m$), u_{ij} - control variables ($u_j \in R^r$)

z_{ij}- disturbance variables representing off-schedule deviations in travel times and driver behaviour. ($z_j \in R^m$)

$A_j^1 \in R^{m \times m}$; $B_j \in R^{n \times r}$ are lower triangular matrices with elements resp. $a_{ik} = b_{ik}/r_k$; $b_{ik} = (-1)^{i+k}$ $\sum_{l=k+1}^{i} k_l/r_l$ (k and r=1-k characterize passenger arrival to loading rate at bus stops). $A_j^n \in R^{m \times m}$ is a matrix with non-zero elements in the last column equal to $c_{im}=(-1)^i \prod_{l=1}^{i} k_l/r_l$

As shown in [3] for the bounded (to the polyhedron in R^m) disturbance vectors, the system is controllable and stabilizable (i.e. any state $x_j \neq 0$ may be transferred in finite interval by means of some input u_j to the state x_k=0). The uncontrolled system is structurally unstable in the Lapunov sense which makes it possible to describe the bunching effect.. The basic model (4.1) make it possible to represent different aggregated models in the set of vehicles and bus stops. In [12] aggregated very flexible analytical solutions of the basic model are given in the set of vehicles and bus stops,hence the fast control actions may be realized on various aggregation levels of the bus route and by individual vehicles. The main purpose of the control is the proper (i.e. with not too high level and rate of change of interventions) elimination of the off-schedule deviations of the buses. The quadratic performance measures of the off-schedule deviations and control costs were used in [2,3,5,15]as the appropriate representation of the passengers and transit operators costs. For example proposed in [16] version, which also incorporates the penalties imposed on the rate-of -change of control seems to be beneficial.

$$J_{Tr-j} = x_{Tr} Q_{Tr} x^T_{Tr} + \sum_{k=j}^{Tr-1} [x_k^T Q_k x_k + u_k^T R_k u_k + (u_k - u_{k-1})^T S_k (u_k - u_{k-1})] \qquad (4.2)$$

where Q_k, R_k, and S_k are symmetric nonnegative definite weighting matrices of reals such that R_k+S_k positive definite. The first and the second term in (4.2) may be regarded as the off-schedule deviations penalties at transfer point (Tr) and all precedent bus stops. The last two terms penalize both the weighted sum of squares of control actions and the weighted sum of squares of the first backward differences in control actions. These terms represent the costs of the magnitude of control actions (e.g. time losses of the on-board passengers due to holding buses at control points) and adaptive control smoother features. The state and controls may be additionaly bounded by hard inequality constraints $x_k \in [L,U]$; $u_k \in [U,V]$. The optimal control problem may then be formulated as follows:select the control sequence $\{u_k\}$ k= j,...,Tr-1 to minimize (4.2) subject to (4.1). We are especially interested in solution of this optimal control problem in decentralized adaptive control scheme (i.e. by autonomous control actions of individual vehicles equipped with on-board computers). This control concept was widely presented and efficiently solved in [12]. With the control scheme presented in this paper, the indirect adaptive controller [34] applied at a given bus consists of an identifier, a controller and an observer. A recursive identification algorithm is used to obtain a model of the subsystem (whichi consists of three consecutive buses) "seen" by the autonomous control unit, and state-space linear-quadratic (LQ) control strategy based on the "seen" subsystems is used to form the controlled input [12]. The decentralized observer is used at each control unit to estimate subsystem states for LQ

feedback gains. Control actions are communicated between consecutive buses so that the identifiers can accurately model the subsystems as well as decentralized observers can give accurate estimates of subsystem states in the presence of their interconnections. The overlapping effects of various autonomous control units are taking into account by application of Siljak transformations which are based on the inclusion principle. The theoretical framework for derivation of adaptive decentralized control scheme and practical control software requirements are widely presented in [12, 16, 17].

5. CONCLUSIONS

1. New hardware capabilities make it possible to design distributed hierarchical management and control systems which have highly profitable features for application in the traffic control.

2 It is the attractiveness of the presented control idea that many criteria, types of regulators and control strategies can be implemented and that it is computer-oriented and network tool applicable to wide spectrum of demands met at real traffic situations. It makes it also possible working in not exactly known traffic environment and to starting with simple traffic model which is adjusted during system operation to the real traffic situations (i.e. it has on-line real time dynamical modelling and model-tracking features).

3. The presented results constitute an important step to generation of highly efficient robust adaptive multicriterial solutions for the networks problems. The closed form of analytical solutions enables us to deal with the full representation of the solution which is crucial for solution sensitivity (robustness) analysis, as well as for representation of the Set of Compromise solutions in the multicriterial optimization.

4. .The high improvements in the quality of service may be obtained on the synchronization way, by proper integration of scheduling and dispatching control actions.

The future works will be concentrated on the development of network related analytical and numerical results for a wide range of alternative control actions, and network criteria.

REFERENCES

1. ABKOWITZ M., A. EIGER., I. ENGELSTEIN: Optimal control of headway variation on transit routes. Journal of Advan. Transp. 20 1, 73-88, 1986.
2. ADAMSKI A.,RUDNICKI A.:Public Transport as a Control Plant, VII Conf. of AUTOMATICS, vol. 1, 722-730, Rzeszow 1977. (in Polish)
3. ADAMSKI A.: Optimal Dispatching Control of Bus Lines.Proc. 9th IFIP, Conference on Optimization Techniques, Lecture Notes in Control and Infor. Science Springer Verlag, No. 23, Part 2, 334-344, 1980
4. ADAMSKI A., P GORCZYCA., K HAJDUK., A KORNECKI., A KORYTOWSKI. Hierarchical Traffic Control System for the City of Cracow. SYSTEM SCIENCE VI, .Inst. Cyber. Tech. No. 57/ 21. 1979.
5. ADAMSKI A.:Multilayer control system for public transport. ZNAGH no 131,101-11, 1980 (in Polish)
6. ADAMSKI A ,KORYTOWSKI A: Optimal Structures of Hierarchical Traffic Control Systems. SYSTEM SCIENCE, vol.7, No. 1, pp. 110-116, 1981.,
7. ADAMSKI A.:Optimal Dispatching Control of Bus Lines. Proc.of the 4th IFAC/ IFIP/ IFORS Inter. Conference on CONTROL in TRANSPORTATION SYSTEMS Baden-Baden pp. 67-72, 1983.
8. ADAMSKI A.:A method of optimal dispatching control for bus lines. Proc. of the 1st I.A.S.T.E.D Symposium on APPLIED INFORMATICS. Lille, vol.3, pp. 267-270, 1983.
9. ADAMSKI A.,RUDNICKI A.: Control systems in the public transport. (Chapter 6 in a book: Computer Traffic Control Systems: Application Examples,) pp. 104-148, WKi L 1984.
10. ADAMSKI A: Optimal dispatching control of bus lines in hierarchical system. Proc. 4th Conf."Science and Practice of Transport". vol 4, pp. 1-11, 1985.
11. ADAMSKI A.: Polyoptimal integrated control of the traffic on an intersection, Proc. 5th IFAC/ IFIP/ IFORS Inter. Conference on CONTROL in TRANSPORTATION SYSTEMS July 8-11, Vienna 1986.
12. ADAMSKI A: Optimal Dispatching Control in Public Transport. Hab. Thesis. Scien. Bulletins of AGH, AUTOMATICS, Bulletin no. 50, 1989.
13. ADAMSKI A: Expert system for dispatching control of the public transport vehicles at terminals. Proc. International Conference TRANSSYSTEM-89, vol 2, 1989.
14. ADAMSKI A.: Probabilistic Models of Passengers Service Processes at Bus Stops. Transportation Research., vol. 26B, no. 4, pp. 253-259, 1992.
15. ADAMSKI A: Transfer Optimization in Public Transport. Sixt Inter Workshop on Computer-Aided Scheduling in Public Transport Lisboa, 1993.
16. ADAMSKI A.: Real-time Computer-Aided Control in Public Transport from the point of view of Schedule Reliability. Sixt International. Workshop on Computer-Aided Scheduling in Public Transport, Lisboa, 1993.
17. ADAMSKI A.: Optimal adaptive dispatching control in an integrated public transport management system. Second Meeting of the EURO Working Group on Urban Traffic and Transportation 15-17 Sept., Paris 1993.

18. ADEBISI O:A Mathematical Model for Headway Variance of Fixed Route Buses. Tran.Res. 20B,59-70,1986
19. ANDERSON P, A. HERMANSSON, E. TENGVALD. and G. SCALIA-TOMBA: Analysis and simulation of an urban bus route . Transportation Research, vol. 13A, pp. 439-466, 1979.
20. BARNETT A.: On controlling Randomness in Transit Operations. Transp. Sci., 8/2, pp. 102-116, 1974.
21. BLY P.H., R.L. JACKSON.: Evaluation of Bus Control Strategies by simulation. TRRL Rep. 637, 1974.
22. CEDER A,. N.H.M. WILSON: Bus Network Design. Transp. Research, vol. 20B, no. 4, pp. 331-44, 1986.
23. Chong C.Y. Hierarchical estimation. Proc. 2nd MIT/ONRC Workshop Monterey July 1979.
24. DORAS J,.GIRARDOT M,. HEURGEN E: Practice of simulation at RATP. Automatisme 23/5,124-32, 1978.
25. FINDEISEN W: Multilevel control systems (in Polish) PWN 1974.
26. GORDON R.L.: Control of bus headways by traffic signal timing method, IEEE Trans. Veh. Technology, vt-27 , no.2 , pp. 69-76, 1978 .
27. HASSAN M.F.,G. SALUT.,M.G. SINGH ., A. TITLI: A decentralized computational algorithm for the global Kalman filter. IEEE Trans. Aut. Control vol. 23, pp. 262-268, 1978.
28. ISRAELI Y,. A. CEDER: Transit route design using scheduling and multiobjective programming techniques. Sixt Inter. Workshop on Computer-Aided Scheduling in Public Transport Lisboa, 1993.
29. JACKSON R:Evaluation by simulation of control strategies for a high frequency bus service, TRRL807,1977
30. JACOBSON J.Y.,SHEFFI Y: Analytical Model of Traffic Delays under Bus Signal Preemption:Theory and Application. Transportation Research vol. 15B (2), 127-38, 1981
31. KOFFMAN D.:A simulation study of alternative realtime bus headway control strategies. Trans. Res. Record no. 663, pp. 41-46, 1978.,
32. NEWELL G.F.: Control of Pairing of Vehicles on Public Transportation Route, Two Vehicles, One Control Point. Transportation Science, 8/3, 1974.
33. OSUNA E.E, G.F. NEWELL: Control Strategies for an Idealized Public Transportation System. Transp. Science, 6 , pp. 52-72, 1972.
34. OSSMAN K.A. :Indirect adaptive control for interconnected systems IEEE Trans., AC-34, p. 908-911, 1989.
35. POWELL W.B.,SHEFFI Y.:A Probabilistic Model of Bus Route Performance. Trans.Sci. 17,376-404,1983
36. TURNQUIST M.,BOWMAN L.: Control of Service Reliability in Transit Networks. US.DOT Rep.5/ 1979
37. TURNQUIST M.A.:Strategies for Improving Bus Transit Service. Final. U.S. DOT Rep. no. 27, 1982.
38. TURNQUIST M.A., S.W. BLUME: Evaluation the Potential Effectiveness of Headway Control Strategies for Transit Systems. Trans. Res. Record 746, 25-29, 1980.
39. VANDERBONA U.,RICHARDSON A.J.:Effect of checkpoint control strategies in a simulated transit operation. Transp.Research vol. 20A/6, 429-36,1986,

$S(u) = \gamma Q_1 + (1-\gamma) Q_2$

Table 2.1

	Q_1	Q_2	$u = \max\{0, P^* e\}$, $P^* \in [0,1]$; $v = \gamma/(1-\gamma)\, C_H$; $C_H = \sigma/E(H)$; $d = E\lvert e\rvert = \sigma\sqrt{2/\pi}$
1	$E(u^k)$, $k=1$ $k=2$ ($B=1$) $E^2(u)$; ($B=1/\pi$) $Var(u)$;($B=1-1/\pi$ $k=3$ $k\to\infty$	$E(w)$	$P^* = 1 - d\,v/\sigma$ ($v \le \sigma/d$) $P^* = 1/(1 + 2B\sigma v)$ ($v \ge 0$) $P^* = 2/\left[1+\sqrt{1+24dv\sigma}\right]$ $P^* \to 1$
2	$E(u^k)$ $k=1$ $k=2$ ($B=A$ $E^2(u)$; ($B=A-$ $Var(u)$ ($B=1$) $k=3$ $k\to\infty$	$I=Var(h)/2H_s$	$A=\pi/(\pi-1)$ $P^* = A(1-2v)$ ($v \in [1/2\pi, 1/2]$) $P^* = A/(1 + 2B\sigma v)$ ($v \ge 0$) $P^* = 2A/\left[1+\sqrt{1+24d\sigma vA}\right]^2$ $P^* \to A$ ($P^* = 1$)
3	$E(u^k)$ $k=1$ $k=2$; ($B=1$) $E^2(u)$; ($B=1/\pi$) $Var(u)$ ($B=1-1/\pi$ $k=3$ $k\to\infty$	$E(w^2)$	$A = (H_s-d)/(H_s-2d)$ $P^* = A\{1 - v(A-1)/(2A\sigma)\}$ ($v \in [2\sigma, 2A\sigma/(A-1)]$) $P^* = A/[1 + B\sigma v(A-1)/d]$ ($v \ge d/B\sigma$) $P^* = 2A/\left[1+\sqrt{1+8\sigma vA(A-1)}\right]$ $P^* \to 1$
4.	$E(u^k)$ $k=1$ $k=2$; ($B=1$) $E^2(u)$; ($B=1/\pi$ $Var(u)$;($B=1-1/\pi$ $k=3$ $k\to\infty$	$I = Var(w)/\sigma^4$	$A=(1-C_H^2)/(1-2C_H^2)$ $P^* = A[1 - dv\sigma H_s(A-1)/2A]$ ($v \le 2A/(d\sigma H_s(A-1))$ $P^* = A/[1 + 2B\sigma^3 vH_s(A-1)]$ ($v \ge 1/2B\sigma^3 H_s$) $P^* = 2A/\left[1+\sqrt{1+16dvHA(A-1)\sigma^3}\right]$ $P^* \to 1$

Table 2.2.

$\gamma \backslash P$	k = 1	k = 2	k = 3	$E^2(u)$	Var (u)	k = 1	k = 3	k = 2	E^2 (u)	Var(u)
CASE 1	A: H = 10 (min), σ = 1 (min)					B: H=10, σ=2.5		C: H=5 (min), σ=1 (min)		
0	1	1	1	1	1	1	1	1	1	1
0.1	0.11	0.31	0.70	0.60	0.40	0.76	0.80	0.47	0.73	0.56
0.5	0	0.05	0.09	0.13	0.07	0	0.29	0.09	0.24	0.13
0.8	0	0.01	0.02	0.04	0.02	0	0.15	0.02	0.07	0.04
1	0	0	0	0	0	0	0	0	0	0
γ_{max}	0.11	1	1	1	1	0.24	1	1	1	1
CASE 2	B: H = 10 (min), σ = 2.5 (min)					D: H= 5 (min)		σ = 1.5 (min)		
0.1	0.16	0.34	0.25	0.72	0.46	0.38	0.40	0.55	0.96	0.69
0.3	0	0.18	0.13	0.29	0.15	0	0.22	0.20	0.49	0.28
0.5	0	0.05	0.09	0.14	0.07	0	0.15	0.10	0.26	0.13
0.8	0	0.01	0.05	0.04	0.02	0	0.08	0.02	0.07	0.04
1	0	0	0	0	0	0	0	0	0	0
γ_{max}	0.11	1	1	1	1	0.13	1	1	1	1

CASE 3	B: H = 10 (min), σ = 2.5 (min)					D: H= 5 (min) σ = 1.5 (min)				
0.1	1	1	0.82	1	1	1	0.98	1	1	1
0.3	1	0.78	0.53	1	0.90	1	0.66	0.80	1	0.94
0.5	1	0.50	0.37	0.83	0.62	0.69	0.47	0.50	0.90	0.63
0.8	0.27	0.18	0.20	0.43	0.24	0	0.25	0.17	0.42	0.23
1	0	0	0	0	0	0	0	0	0	0
γ_{max}	0.55-083	≥0.166	≥0.05	≥0.39	≥0.21	0.47-074	≥0.01	≥0.193	≥0.43	≥0.23
CASE 3	B: H = 10 (min), σ = 2.5 (min)					D: H= 5 (min) σ = 1.5 (min)				
0.1	1	1	0.82	1	1	1	0.98	1	1	1
0.3	1	0.78	0.53	1	0.90	1	0.66	0.80	1	0.94
0.5	1	0.50	0.37	0.83	0.62	0.69	0.47	0.50	0.90	0.63
0.8	0.27	0.18	0.20	0.43	0.24	0	0.25	0.17	0.42	0.23
1	0	0	0	0	0	0	0	0	0	0
γ_{max}	0.55-083	≥0.166	≥0.05	≥0.39	≥0.21	0.47-074	≥0.01	≥0.193	≥0.43	≥0.23
CASE 4	B: H = 10 (min), σ = 2.5 (min)					A: H= 10 (min) σ = 1 (min)				
0.1	0.29	0.100	0.15	0.27	0.140	0.97	0.78	0.826	0.94	0.88
0.3	0	0.030	0.08	0.08	0.040	0.68	0.57	0.540	0.78	0.63
0.5	0	0.012	0.05	0.03	0.018	0.60	0.42	0.330	0.60	0.42
0.8	0	0.003	0.02	0.01	0.004	0	0.245	0.110	0.29	0.15
1	0	0	0	0	0	0	0	0	0	0
γ_{max}	0.01-013	≥0.0008	≥0.0002	≥0.0003	≥0.002	0.02-071	≥0.003	≥0.005	≥0.004	≥0.007

$S(u)=\gamma Q_1+(1-\gamma)Q_2$

Table 3.1

	Q_1	Q_2	$u=P(e-e_m)\,1(e-e_m)$; $e_m=H-h_{min}\in[0,H]$; $v=\gamma/(1-\gamma)$ C_H; $C_H=\sigma/E(H)$
A 1.	$E(u^k)$ $k=1$ $k=2$ $E^2(u)$; $k=3$	E(w)	$P^*=1$, $h_{min}=\max[0, h_{min}^*]$ $h_{min}^*=(1-2\gamma)H_s/(1-\gamma)=H_s-\in[1/2\pi,1/2]$) ($\gamma\in[0, 0.5]$) $h_{min}^*=H_s-\sigma^2 v/(0.62+0.341\,\sigma v)$ $\sigma v\le 0.62/(C_H-0.341)$ $h_{min}^*=H_s-\alpha\sigma[A-\sqrt{A^2-1}]$; $\sigma v\in[0, 1.22]$; $\alpha=2$, $A=1.1+1.25/\sigma v$ $\sigma v\in[1.22,6]$; $\alpha=1.25$, $A=0.92+1.08/\sigma v$; $\sigma v\in[6,26]$; $\alpha=1.6$, $A=0.97+0.85/\sigma$ $h^*_{\min}=H_S-\sqrt{A^2+3.67\sigma^2}+A$; $A=1.4627\sigma+0.611/v\sigma$
B	$E(u^k)$ $k=1$	E(w)	$P^*\ge 0$ Elastic minimal headway strategy [12] 1. $P^*=1$ $e_m=\sigma v$ 2. $P^*=11.9-3.5\,v$; $e_m=3.16\,\sigma$ special cases: $\gamma=0$: $P=1$, $e_m=0$; $\gamma=1$: $P\to\infty$, $e_m=3.7\sigma$, $u(e)=e1(3-3.7\sigma)$
	Q_1	Q_2	$u=a\,1(e-e_m)$; $e_m\in[0,H_s]$; $a\in[0,H_s]$; $v=\gamma/(1-\gamma)$ C_H; $C_H=\sigma/E(H)$
A	$E(u^k)$ $k=1$ $k=2$;	E(w)	$a\neq e_m$ $a=\sigma v\,[0.7A-0.6]$ where $A=\sqrt{1+3.3/v^2}$ $e_m=0.35\,\sigma v\,[2+A]$ $a=1.3\sigma/(1+2\sigma v)$; e_m $0.65\,\sigma$; special cases: $\gamma=0$: $a=1.3\,\sigma$, $e_m=0.65\sigma$; $\gamma=1$: $e_m\to\infty$, $a=0$

B	$E(u^k)$	E(w)	a = e_m :special cases: γ=0 : a= 0.886 σ, γ=1: a→∞,
	k = 1		$a = \left\{2.15\sqrt{(\nu-0.2)^2+0.4}-0.5\right\}\sigma$ for v ≤1
			$a = \sigma[2.14\nu-0.5]$ for v≥0.9
	k = 2;		$a = \sigma\sqrt{B+\sqrt{B^2+2}}$; B=1/(2vσ-1)

Table 3.2.

γ \ P	k = 1	k = 2	k = 3	$E^2(u)$		k = 1	k = 3	k = 2	E^2 (u)	
CASE A 1	A: H = 10 (min), σ = 1 (min)					D: H= 5 (min) σ = 1.5 (min)				
0	10	10	10	10		5	5	51	5	
0.1	8.9	8.84	9.23	9.52		4.4	3.96	4.48	4.7	
0.3	5.7	7.93	9.10	9.3		2.86	2.62	4.34	4.25	
0.5	0	7.51	9.05	8.86		0	1.77	4.316	3.88	
0.8	0	7.19	9. 0	7.41		0	0.96	4.29	2.6	
1	0	0	7.1	7.1		0	0	0.61	0.65	
CASE A 2	A: H = 10 (min), σ = 1 (min)					D: H= 5 (min) σ = 1.5 (min)				
γ→	0	0.1	0.3	0.5	0.8	0	0.1	0.3	0.5	0.8
k=1	10	9.94	9.76	9.49	0	5	4.91	4.78	4.48	2.25
CASE B	A: H = 10 (min), σ = 1 (min)					D: H= 5 (min) σ = 1.5 (min)				
γ→	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
k=1 e_m/P	1.11/ 1	2.5/ 1	4.28/ 1	6.67/ 1	10/ 1	0.55/ 1	1.25/ 1	2.14/ 1	2.22/ 1	5/ 1
k=1 P/e_m	8 / 3.16	3.16/3.2				10/ 4.7	9 / 4.7	7/ 4.7	4.1/ 4.7	0.25/4.7

Specifications of the standard BRMI module — Table 1

R P		GPF	PAP	B A P		PSP	C S	
CT	I	s-s iso	RAP	HD	T_{ij} i.d	BA n^B_{ij} c n^B_{i-1j}	IP l z	
n/m	SRS2/m			TD	T_{i-1j} c T_{ij} c T_{ij+1}			
f	H			BCA	UBCA		T / K	
BT	TP			SR	a-s			

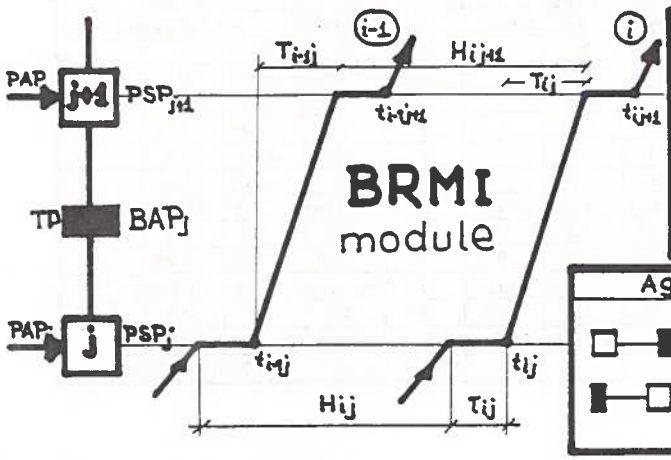


Fig. 1 BRMI standard module

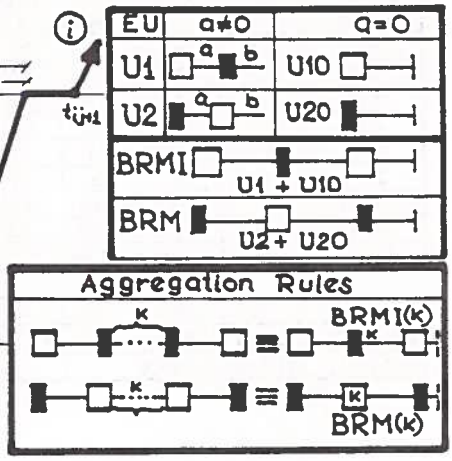


Fig. 2. Elementary and standard modules

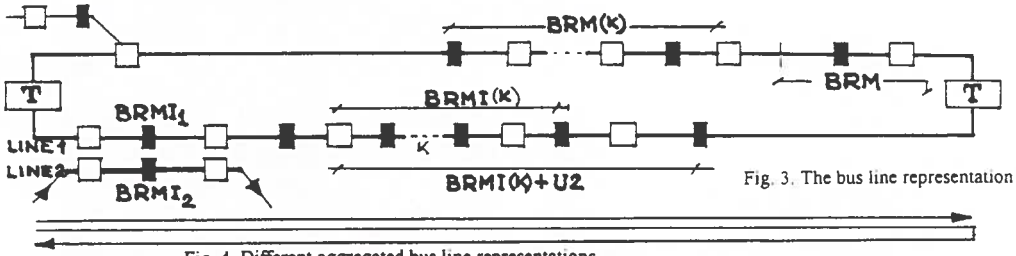
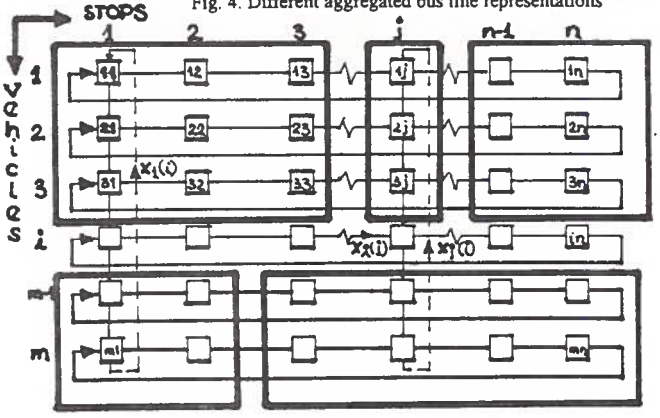


Fig. 3. The bus line representation

Fig. 4. Different aggregated bus line representations



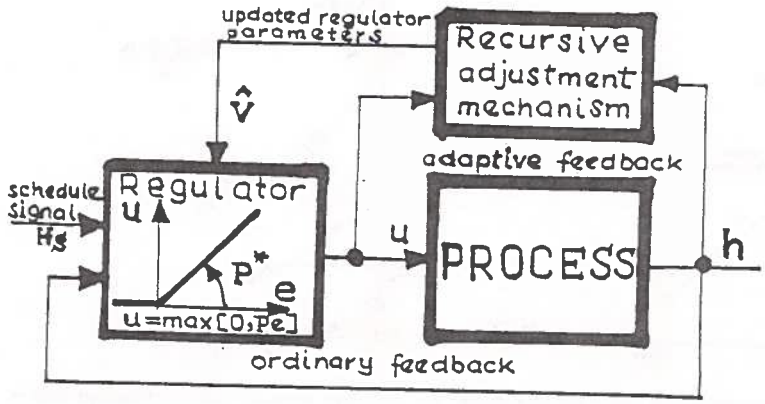
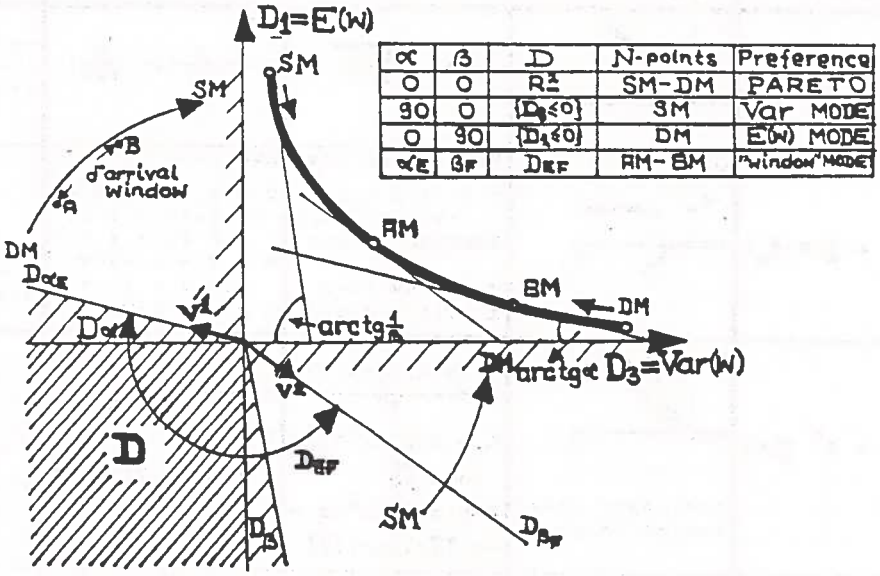
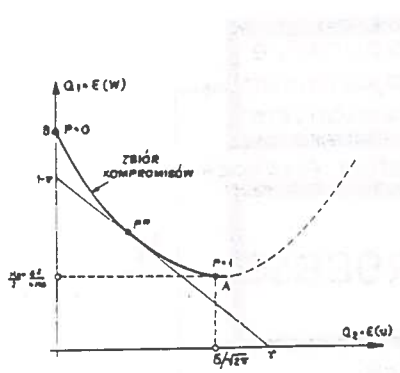


Fig.5 Block digaram of an Flexible Holding strategy



α	β	D	N-points	Preferenca
0	0	R^2	SM-DM	PARETO
90	0	$\{D_3 \le 0\}$	SM	Var MODE
0	90	$\{D_1 \le 0\}$	DM	E(W) MODE
α_E	β_F	D_{EF}	RM-BM	"window" MODE

FIG. 6. ADAPTIVE ADJUSTMENT OF THE DOMINATION STRUCTURE



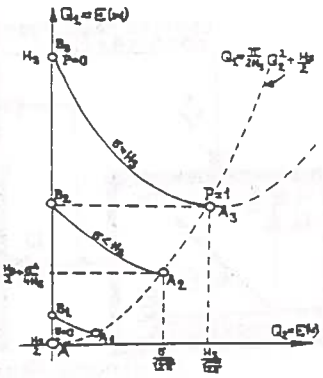
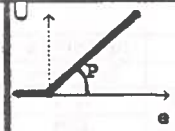
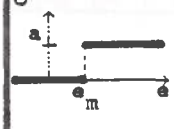
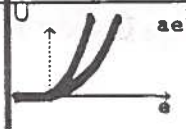
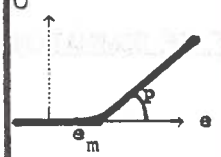


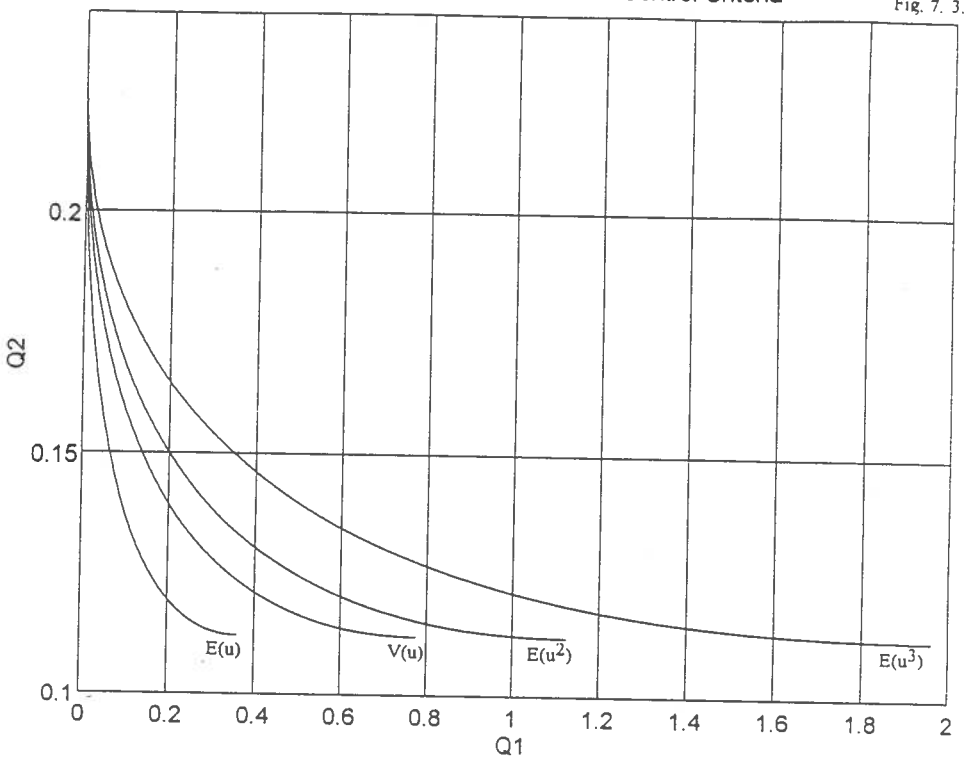
Fig. 7.1. Compromise Set AB in criteria space. Fig. 7.2. Evolution of Compromise Set with σ

FLEXIBLE CONTROL STRATEGIES TABLE 4

	REGULATOR		STRATEGY		REMARK
	FORMULA	Characteristics	REGULARITY	PUNCTUALITY	
1	$u = \max[0, Pe]$ $P \in [0,1]$		$P^* = 1 - \sqrt{\frac{2}{\pi}}\, v$	$P^* = \frac{\pi}{\pi - 1}\left[1 - \frac{v}{\sqrt{2\pi}}\right]$	ch-cs linear detect
2	$u = a\, 1(e - e_m)$		Version A: $a = e_m$ $u = -0.5 + 2.15 \cdot (V - 0.2)^2 + 0.4$ Version B: $a \neq e_m$ $u = 0.35V\,[2 + X]$ $a = v[0.7\,X - 0.6]$	$a = e_m$: for $V \in (0,3]$ $u = 0.058334V^2 + 0.0916V + 0.85$ for $V \in (3, 6.75]$ $u = V - 0.89/V$ for $V > 6.75$ $u = V - 1/V$	$X = 1 + 3.3/V^2$
3	$u = a\, e^n\, 1(e)$	 nonlinear polynomial ch-cs.	$a^* = \frac{B(n)[A(n) - v]}{\sigma^n}$ $A(n) = (2/\pi)^{\alpha} n!!/(n-1)!!$ $B(n) = (2/\pi)^{\beta}(n-1)!!/(2n-1)!!$	$a^* = \frac{B(n)[2A(n) - v]}{\sigma^n D(n)}$ $D(n) = 2 - B(n)(2n-1)!!$	$\alpha = (-1)^n/2$ $\beta = [1 - (-1)^n]/4$
4	$U = P\,(e - e_m)$ $1(e - e_m)$ where $P \geq 0$ $e_m = H_s - h_{min} \geq 0$		1. Minimal interval: $P=1$, $u=v$ $e_m = \gamma H_s/(1-\gamma)$; $h_{min} = (1-2\gamma)/\gamma e_m$ 2. Elastic minimal interval: $P = 11.9 - 3.5v$; $e_m = 3.16\,\sigma$ 3. PREFOL : $(e = H_{i+1} - H_i - u_{i-1})$ $P = 1/2$; $e_m = \gamma\, E(H)/(1-\gamma)$ 4. HOLDBACK / PUSHOUT : $(\gamma = 0)$ $P = 1/2$; $e_m = 0$		Elast. minim. interval strat.

The Compromise Set for Different Control Criteria

Fig. 7. 3.



Tables 5

TABLE 1			
Δ	E (w)	E(w) no control	Gain (%)
4	4.38	4.00	- 9.5
3	3.50	3.33	- 5
2	3.00	3.49	+ 14
1	3.53	5.00	+29.4
0	5.91	10.00	+40.9

TABLE 2			
Δ	E(w^2)	E(w^2) no control	Gain (%)
4	21.20	18.00	- 17.70
3	16.70	19.38	+ 13.70
2	23.72	77.80	+ 69.50
1	52.80	90.90	+ 41.90
0	116.81	201.79	+ 42.11

TABLE 3			
Δ	Var(w)	Var(w) no control	Gain (%)
4	2.06	1.9	- 8.42
3	4.31	8.32	+ 48.19
2	14.63	52.90	+ 72.34
1	40.40	67.67	+ 40.30
0	81.98	101.90	+ 19.50

TABLE 4			
Δ	F(w^*)	F(w^*) no control	Gain (%)
4	0.1735	0.2100	+ 17.40
3	0.0176	0.0814	+ 78.40
2	0.1455	0.164	+ 11.35
1	0.2000	0.200	+ 0. 00
0	0.3696	0.500	+ 26.00
- 2	0.88	1.00	+ 12.00
- 4	1.00	1.00	0.0

Rural Public Transport Optimization

Jan Cerny, Professor, Chair of Transportation and Communication,
Institute of Management, University of South Bohemia,
Jindrichuv Hradec, Czech Republic

1. State of Art

The standard optimization problem P1 of rural public transport can be formulated in a verbal mode as follows:

Given a network G which can be used for the transport, a passenger demand B with origins/destinations in G and finally a vehicle fleet F to be used for to cover the demand B. To find the set of journeys J on G, the assignment ` of journeys J to the demand B (precising which journeys will be used by the given elementary demand bnB) and another assignment a of vehicles F to the journeys J such that the following constraints hold:

C1: Each passenger is transported.

C2: Each passenger starts/finishes his trip in the given time window (if given).

Ç3: Any vehicle is in any time assigned to not more than one journey.

C4: If two journeys j1 and j2 (immediately one after another) are assigned the same vehicle, it is feasible (i.e. a vehiccle having finished the journey j1 is able to catch the departure of j2).

The objective expresses the passengers total time loss and it is to be minimized.

The more general problem P2 can be obtained if not only the journeys J and the assignments ` , a are submitted for optimization, but the optimization deals also with the network G and the fleet F.

Moreover, a problem P3 of the precise demand estimation (i.e. "where we can get the demand B from?") in the rural transport theory is often confronted.

We have to note that similar collection of problems is met in the urban transit theory as well. But there is a big difference: the urban theory (characterized by the continuous demand model and by the periodic service on the routes) is incomparably higher developed than the rural one (characterized by the discontinuous demand and an infrequent service in the form of individual journeys). One can say

that the ratio between numbers of papers published in both direction is n:1, where n>10.

As concerns the problem P1 we can mention - a method proposed by the reseearch team from the University of Zilina (not yet published, but successfully applied to practice). The method works in "man-machine" mode. It finds weak points of the transport service and proposes the improvements. The main misadvantage of this approach is that it proposes "local improvements" only, not dealing with "global situation".

- the model by S. Nguyen and S. Pallottino [9], which is better theoretically founded. It is very inspirative and we know it was successfuly applied to practice, but no paper with the description in details is available. We are afraid it may be complicated.

Therefore we propose an original approach which seems to be hopeful.

Other authors studied only some part of the problem P1. The main attention was paid to the vehicle sheduling problem, i.e. to the optimization of a for the given J,F. We can mention for instance [1,2,4,6,9,11].

Because of its complexity, no models and methods of a "direct" solution of P2 are available, except of [7], where hewever only the radial type of fixed network is supposed and the fleet F is optimized. Naturally, the problem P1 can be decomposed into several parts, each of them solvable without big problems (several of them belonging also to P1). Hence we try to propose a "more direct" approach.

On the other hand the solution of P3 is quite natural, using the "standard" methods of demand estimation, as described e.g. in [3]. Therefore this problem has not been studied by us.

Finally we must to mention the fact, that the formulation of the problems P1 and P2 expresses only the aspect of "pure transportation", not concerning the impact of other views as ecology, land use etc., as mentioned (but only in the verbal form) in [8].

2. Models and Methods

2.1. Solution of P1. The main purpose is to study the problem P1 directly, without any decomposition (which looses the optimality of solution). Our original approach consists of the application of the partition/aggregation/concatenation principle:

2.1.1. To assign any b to G and to divide it into segments on individual edges (= a partition).

2.1.2. To form "multi-batches" from the group of segments on the same edge in the same (or almost the same) times of arrivals and departures, but not exceeding the vehicle capacity by the number of passengers in a multi-batch (= an aggregation).

2.1.3. To cover the multibatches by vehicles i.e. to construct journeys - the sequences of multibatches (= a concatenation).

2.2. Solution of P2. Another target is to propose a mathematical model for the problem P2. Since it is not fully satisfactory, we try to find a method reflecting the results of 2.1 in the optimization of transport infrastructure.

2.3. The final target is to discuss the possibility to include the requirements of ecology and land development theory into the optimization models.

REFERENCES

[1] Bodin, L.- Golden, B.- Assad, A.- Ball,M.: Scheduling of vehicles and crews. Pergamon Press, Oxford etc., 1983.

[2] Carpaneto, G.- Dell'Amico, M.- Toth,P.: A Branch and Bound Algorithm for the Multiple Depot Scheduling Problem. Networks 19(1989), 531-548.

[3] Cascetta, E.: Metodi quantitativi per la pianificazione dei sistemi di Trasporto. Cedam, Padova, 1990 (a monograph)

[4] Cerny, J.: Optimal Solution of the Transport Scheduling Problem. Zbornik prac VSDS Zilina 6(1967), 215-222.

[5] Cerny, J.- Kluvanek, P.: Foundation of Mathematical Theory of Transportation (a monograph in Slovak). Veda, publ.house of Slovak Acad. of Sci. Bratislava, 1991.

[6] Cerny, J.- Paluch, S.: Bus Scheduling with Flexible Journeys. Trans.Int.Conf. of Oper.Res.Soc of Italia, Capri, September 1993, pp. 31-35.

[7] Cerny, J.: Supply/Demand Relation on the Single Bus Line. Trans. of Int. Conf. on Modern Transport. Zilina, Sept. 1993.

[8] Deiters, J.: Der OPNV aus der Sicht der Wissenschaft. In: Offentlicher Personennahverkehr im landlichen Raum, Kommunal u. Reginalstudien No.17 (1992), 25-40.

[9] Maze, T.Z.- Knasnabis, S.: Bus Garage Location Planning with Dynamic Vehicle Assignments. Transp.Res. B, 9b.(1985), 1-13.

[10] Nguyen, S. - Pallottino, S.: Demand Estimation and Passenger Assignment. An invited lecture to the Conference AIRO'90 in Sorrento, Italy, 1990.

[11] Paluch, S.: Optimization of Bus and Crew Scheduling in CSFR. Trans.Int.Conf."Hexagonal Countries Transp. Systems Development", VSDS Zilina, May 1992.

Algorithms for special classes of sparse directed rural postman problems

Michael O. Ball, College of Business and Management and Institute for Systems Research, University of Maryland, College Park. MD 20742
Hediye Gun, Applied Mathematics Program, University of Maryland, College Park. MD 20742

Abstract: Rural postman problems are a principal component of the solution of arc routing problems which arise in applications such as route layout for meter readers, postmen and sanitation vehicles. The directed rural postman problem (DRPP) is defined over a directed network, G=(N, A U R), where N is the node set, A is the set of non required arcs and R is the set of required arcs. A cost is associated with each arc in A. The output is a closed walk, called a postman tour, that traverses each arc in R exactly once and minimizes the total cost of the arcs traversed in A. If the network induced by R is connected the DRPP can be solved in polynomial time using network flow techniques. On the other hand, if this network is not connected then the DRPP becomes NP-hard.

For the applications mentioned above, the DRPP must be solved over large, sparse street networks. In this paper, we present several results related to the solution of DRPP's over sparse network classes. The results are of two types. The first type is related to the solution of DRPP using a strong cutting plane approach. The second type provides efficient exact algorithms for solving DRPP over special network classes,

In most polyhedral approaches to graph related problems, the input is assumed to be in the form of a complete graph. We feel that due to the problem size involved as well as the inherent sparcity of the problems, it is important to directly deal with sparsity in the solution of DRPP. With this in mind, we defined two sparse network models: for the first, we assume that whenever arc (ij) appeared in R, arc (ij) also appeared in A; for the second, we assumed that the node set associated with each component in the graph induced by R induced a strongly connected network relative to (N. A.). The first class arises naturally in the study of transportation networks. The underlying assumption is simply that any street segment on which service is required can also be used for out-of-service operations. The second class is arises naturally when formulating the travelling salesman problem as an instance of DRPP.

In the most natural formulation of the DRPP the requirement that each of the required arc components be connected by non-required arc paths can be stipulated using so-called postman cut constraints. A postman cut is a set of arcs directed from a node subset M to the node subset N-M, where i.) all required arc components are completely contained in either M or N-M and ii.) both M and N-M contain at least one required arc component. The postman cut constraints require that at least one arc in each postman cut be used in any postman tour. For each of the network classes described in the previous paragraph, we determine the dimension of the DRPP polyhedron and give conditions under which a postman cut constraint induces a facet of the convex hull of integer solutions. Furthermore, we give an efficient algorithm for solving the associated separation problem. These results provide the key ingredients for constructing a strong cutting plane approach to solving DRPP over sparse networks using postman cut constraints. Computational studies by ourselves and others indicate that formulations using postman cut constraints are very strong so that this approach should be very effective.

The second set of results characterizes certain solvable cases of the DRPP. In particular, we analyze four network classes. These are the classes where the undirected version of (N,A) is i.) a simple path, ii.) a simple cycle, and iii.) a tree and the case iv.) where (N,R U A) is a special type of grid network. We show that for cases i.), ii.) and iv.), DRPP can be solved exactly, in polynomial time, using the balance and connect algorithm of Ball and Magazine. Fore case iii.) we modify the Ball and Magazine algorithm so that it requires the solution of a minimum steiner tree problem rather than the solution of a minimum spanning tree problem. This new algorithm finds the optimal solution and is polynomial whenever the underlying steiner tree problem is polynomially solvable.

In addition to presenting these results, we will discuss their potential usefulness within systems for solving general arc routing problems.

The Traveling Salesman Problem Under Accessibility Constraints

Frédéric Semet

The purpose of the Traveling Salesman Problem (TSP) is to find the minimum cost of a trip with the salesman starting at any arbitrary city (his home), visiting every customer exactly once, and returning to the original city. In this case, the tour is performed by a single vehicle driven by the salesman.

In the Traveling Salesman Problem under Accessibility Constraints (ACTSP), the tour is not performed by a single vehicle any more but by a combination of two, e.g., train and truck or truck and trailer. Therefore, we have to solve the problem bearing in mind customer restrictions, as some customers are accessible with both kinds of vehicle whereas others only with one of them.

For the sake of clarity and simplicity, we call NAC customers (Non Accessibility-Constrained customers) those accessible with both vehicle types and AC customers (Accessibility-Constrained customers) those accessible with only one vehicle type. It is assumed that, for all AC customers, the same vehicle type has to be used for the delivery. Thus, the set of customers is not homogeneous and can be partitioned in two subsets.

A solution comprises a main tour originating from and terminating at an arbitrary customer (e.g., the central depot) and subtours beginning and finishing at a customer on the main tour (excluding the central depot) (see Figure 1). All customers located on the main tour belong to the subset of NAC customers, whereas customers inserted on subtours can belong to either subset. The customer at the origin of a subtour is called the root of the subtour.

The vehicle performing the main tour is called the NAC Vehicle (VNAC), while the vehicle visiting the customers located on subtours will be called the AC Vehicle (VAC). Generally, the VAC is carried by the VNAC or draws the VNAC on the main tour. Thus, the VAC is forced to return to the root after having visited all the customers on the subtour.

Note that two other constraints govern the route configuration. First, a single subtour may be initiated from a customer located on the main tour. Hence, a NAC customer will be visited at most twice. Second, the number of customers served on a subtour is bounded by the capacity of the VAC. For each customer i, demand d_i is the load to be delivered to that customer. The capacity constraint states that the total load delivered by the VAC on a subtour cannot exceed the VAC capacity D. There are no capacity restrictions on the VNAC.

The contents of the VNAC include the goods to be delivered to the customers on the main tour and the goods to be reloaded in the emptied VAC. We assume that if the VAC is partially (or entirely) empty after a completed subtour it can be reloaded from the VNAC with goods needed for the next subtour.

Note that a customer at the root of a subtour will be served only once by the VNAC even if it is visited twice.

The objective of the ACTSP is to determine a lowest cost solution whereby each customer will be visited either on the main tour or on a subtour. The cost of a solution consists in the main tour cost and the costs of the subtours connected to the former.

Figure 1 presents a solution of the ACTSP. The rectangle represents the starting and ending customer (central depot), the triangles are the NAC customers and the circles the AC customers.

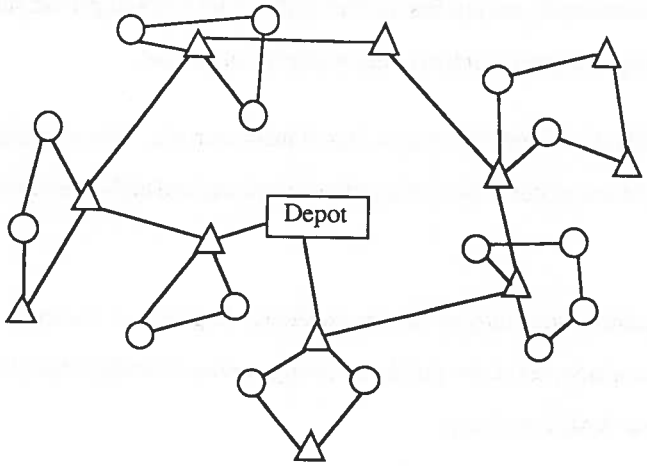


Figure 1: A solution of the ACTSP.

Whereas the vehicle routing problem (VRP) and the TSP have been studied intensively, very little work has been reported on the ACTSP to our knowledge (see Semet and Taillard (1993)).

Since the TSP can be transform polynomially in the ACTSP, the ACTSP is an NPhard problem. Therefore, much attention is given to the development of three heuristic algorithms as the existence of a polynomial time algorithm to solve the problem exactly is unlikely.

The first algorithm is an insertion based method which is similar to those used to solve the VRP. First the main tour is constructed by inserting all NAC customers successively on it and solving the TSP. Next, the subtours are expanded simultaneously by applying the insertion criterion presented above until all AC customers are inserted. At each iteration only feasible insertions are considered, i.e., insertions such that the VAC capacity is never exceeded on the subtours. After each insertion, a TSP is solved to improve the modified subtour.

The second routine is an adaptation of the two-phase heuristic due to Fisher and Jaikumar (Fisher and Jaikumar (1981)) to the ACTSP. As a solution for the VRP is obtained by solving first a generalized assignment problem (GAP) and then TSPs, the proposed

heuristic method consists in solving first a GAP and second TSPs to get the subtours, and a generalized traveling salesman problem to determine the main tour.

The last method is based on the tabu search metaheuristic. This adaptation is mainly characterized by the use of constraints' relaxation techniques and embodies an intensification scheme.

Finally, a comparison study of these procedures is given on the basis of numerical results obtained on a large set of test problems having various characteristics (VAC capacity, numbers of AC and NAC customers).

References

Fisher, M.L. and Jaikumar, J. (1981), "A Generalized Assignment Heuristic for Vehicle Routing", *Network*, Vol. 11, pp. 109-124.

Semet, F. and Taillard, E. (1993),"Solving Real-life Vehicle Routing Problems Efficiently Using Taboo Search", Annals of Operations Research, Vol. 41, pp. 469-488.

INITIAL RESULTS OF THEORETICAL AND EXPERIMENTAL RESEARCH CONCERNING ROAD TRAFFIC FLOW ANALYSIS AND MODELS.

V. Torrieri[1], *V. Astarita*[1], *D. Gattuso*[2], *A. Vitetta*[2]

Abstract

The road traffic safety improvement and a more efficient use of infrastructure for economical and environmental needs, implies the continuous development of the technics for control of circulation, which in turn needs a deeper theoretical and experimental knowledge about traffic flow behaviour. The activity follows two ways: a theoretical research attempting to provide models able to simulate traffic situations in states of transition; a variety of experimental activities in order to study flow behavioural aspects, that today is impossible to observe by the usual traffic sensors. In this report we intend to present: the calibration of the theoretical model on the basis of the comparison between theoretical data and experimental data; the estimation of the main parameters and the checking of the V.T.I.P. algorithm.

Introduction

In this paper some results of research carried out on vehicular traffic on motorways are presented with particular reference to the temporary condition, with the aim to improve the design methods and the management of the infrastructure. The research is carried out by some researchers from the "Dipartimento di Trasporti" from the University of Naples, of the "Laboratorio di Analisi dei Sistemi di Trasporto" from the University of Reggio Calabria and by the "Società Tangenziale" of Naples.

The approach of the research is theoretical-experimental: the survey of the behavioural characteristics of traffic flow through innovative instruments, like the Image Processing (IP), supplies new opportunities of analysis and is oriented to the acquisition of statistics parameters, useful for the calibration of simulation models. The results, obtained until now, regard:

— a freeway traffic flow model;
— an algorithm for the automatic detection of traffic flow parameters.

In this report the results of a pilot experiment looked at to test the possibilities offered by the survey instrument and the models for in-depth study on the traffic flow behaviour is presented. After an outline of the traffic flow modelling and some considerations about the measure instruments that constitute fundamental support for traffic analysis, the attention is focused on the traffic flow model on freeway, developed to simulate the traffic phenomena also for temporary states. Afterwards an algorithm to measure data from traffic video-record is recalled; the algorithm offers opportunities of different and significant measures; some of them, as reported in section four, are used specifically to favour the analysis and to improve the characteristics of the model.

[1] University of Naples
[2] University of Reggio Calabria

1. Traffic flow and survey equipment.

The road vehicular traffic development in the course of the time has determined new problems and has steered a lot of research to the description, the comprehension and the prediction of mobility phenomena. Studies regarding road vehicular traffic in the last quarter of century are numerous; many deterministic models have been proposed and they can be traced back to two classes:

- — microscopic theories, based on single vehicular behaviour; among of these, the car following theory (Herman and others, 1959; Edie, 1960), referring to the circulation of vehicles moving in a line, on a single lane;
- — macroscopic theories, tending to describe the traffic phenomena on analogy with physical phenomena already studied (fluid-flow analogy, mechanical-analogy); for example the hydrodynamic analogy, which recalls some aspects of hydraulic flows (Lightill and Whitham, 1955; Richards, 1956; Greenberg, 1959).

More recently the traffic streams were recognised and studied as stochastic processes (Miller, 1960); an universal classification of stochastic theories does not exist yet, but it is already commonly accepted to recognise a general difference between "pure" models not based on the behavioural hypotheses relative to drivers and models which refer to such types of hypotheses.

Some theories combine different approaches, giving rise to "mixed" models, In this last class are the so-called "Boltzmann models", which originate from the kinetic gases theory (mechanical analogy), but include driver behavioural features, typical of microscopic theories, and distribution of aleatory variables.

The theories, deterministic, stochastic or mixed, recall many traffic variables, sometimes related to a single vehicle, sometimes related to traffic stream. Between the first variables are the headway, the intervehicular space, the time speed and the space speed, the acceleration; among the second, the volume, the density, the occupation rate, the mean speed, the density, the travel time on a road section. In fact a variety of the other variables exists, which can mark a traffic stream: the overtaking rate, the lane changing rate, the accident rate or disturbing rate, the distribution of flow over the lanes, the surface density, and others.

Many tools are available able to offer direct or indirect measures of traffic vehicular variables. Such instruments can be used to efficiently detect a limited set of information and, generally, the accuracy of measurement is strictly linked to the specialisation of the detector.

The most resent technological progress in the field of traffic detection, particularly those linked to the film recording and automatic IP (Blosseville and others, 1990; Briquet, 1990; Hoose, 1991), allow us to foresee the possibility for new development of traffic flow analysis, with reference to the largest information offered, aid to the continuity and automatization of measures; especially:

- — for monitoring and measurement;
- — for application in traffic control (driver information; facilities and service management; prevention, detection and management of accident, etc.);
- — for the development of the research (theoretical and experimental models).

2. A freeway simulation traffic model

The freeway traffic model proposed here could be included among the Boltzmann models already mentioned. It is developed from the Kinetic Theory of Vehicular Traffic already used by Prigogine and Herman (1971) who developed this theory from a mathematical statistical approach similar to that of the classical theory in statistical physics, applied successfully to the Kinetic Theory of Gases. In this theory, while the hydrodynamic analogy is preserved, a distribution function of speed $f(x, v, t)$ is supposed where x is the space, v the speed and t the time and the wishes if the drivers are introduced through a distribution function $f_0(x,v,t)$ of desired speed. The mathematical formulation of the kinetic model can be expressed through the following kinetic equation:

$$\frac{df}{dt} = \left[(1-P)\cdot k\cdot(\bar{v}-v)\cdot f\right] - \left[\frac{f-f_0}{T}\right]$$

with: $\bar{v}$, local mean speed;
k, local density;
P, probability of passing;
f_0 distribution function of desired speed;
T, relaxation time for the tendency of f towards f_0.

The two right terms express respectively the amount of interaction between vehicles and the tendency of f towards f_0.

For a homogeneous and stationary situation (df/dt=0 and df/dx = 0) the solution is:

$$f = \frac{f_0}{1 - k \cdot T(1-P)(\bar{v} - v)}$$

From this solution and, along with some other hypotheses on P and T, it is possible to plot flow-concentration diagrams that may be qualitatively compared with experimental data. This theory also explains the propagation of kinematics waves and the existence of instability phenomena.

The problem of the practical application of Kinetic Theory lies in the difficulties raised by an attempt to solve the equation for situations different from homogeneous and time independent flow. Another problem is that of spontaneous lockup of flow after temporary saturation of street capacity, as emphasised by Alvarez Brey and Casado. P. Ross (1988) indicated this problem as one of the commonest limitations of some macroscopic simulation models: many models, especially those based on a deterministic density-speed relation, may cause a total stop and when such a situation is reached there is no way for the queue to dissipate.

Recently this kind of problem has been explained in terms of a poor numerical description, and it would appear that like other macroscopic models the Kinetic Theory model is able to dissipate a queue. Unfortunately the difficulties involved in the solution of kinematic equations make proof difficult to obtain. However, our model (based on kinetic theory) is able to dissipate queues.

Alvarez, Brey e Casado (1990) proposed a simulation model based on the ideas of Prigogine and Herman which is capable of overcoming the difficulties of the Kinetic Theory calculation. The simulation that have been carried out with the Alvarez model are steady state situation and situations evolving towards steady state. The expression for the probability P of passing is $P=1-k/k_S$ according to the general expression proposed by Prigogine and Herman:

$$P = 1 - \left(\frac{k}{k_s}\right)^{\alpha}$$

Global density is used here, but local density can change considerably from site to site when a queue is growing or in any transitory situation due to some perturbation of stationary flow. The expression chosen for P should have been such as to reflect local congestion. So there are two main problems with this model:

1) the expression for T and P depends on global density and not on local density;
2) the space ruling proposed by the model which is an over-simplification of the traditional well-known car following models can lead to too high a value of local density so that a formed queue would never be able to dissipate.

These problems have been addressed and solved in the formulation of our model. Our model was developed expressly to deal with transitories in dishomogeneous road conditions, which are characteristic of perturbed flow (Astarita, 1994). This can be accomplished by assuming K to be the local density. Therefore, P is no longer a characteristic of all the simulated line but depends on the position of the vehicle, allowing the model to consider automatically changes in road and traffic conditions. It is also introduced in the model a minimum spacing which depends on jam density (K_S) and on the geometrical street conditions.

A comparison between this model and other well-known macroscopic model (Ross model and Lightill and Whitham model) is proposed for a test situation (Ross 1988) with a road system in over-saturated condition due to a sudden reduction (10 minutes time) in capacity from 4.000 to 2.000 veh/h with queue formation and dispersion (fig. 1). The following characteristics of model are used:

- total simulated time 30';
- total simulated road length 8 Km;
-free flow speed 30 m/s;
- capacity 4.000 veh/h except between abscissas x=4.000 m and x=4.400 m where capacity is reduced to 2.000 veh/h from time 7'30" to 15';
- for the model based on a deterministic flow relationship (Lightill and Whitham 1955) the following Greenshields equations between speed (v) and density (k expressing in veh/m) are used:
  capacity=2.000 veh/h: v(k)=30(1-k/0.074074)
  capacity=4.000 veh/h: v(k)=30(1-k/0.148148);
- for all macroscopic models the following steps for calculation were used: dx=80 m, dt=1 s.

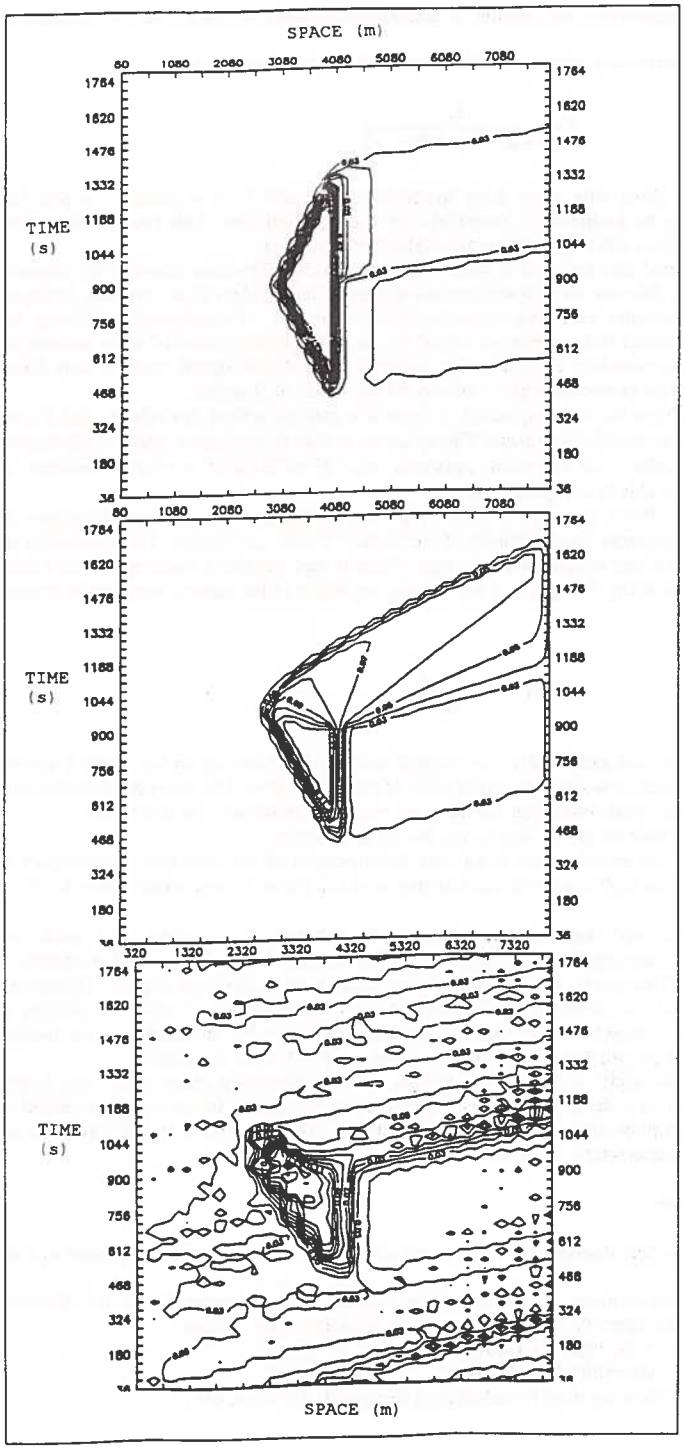


Figure 1
Density evolution for test situation with the following models (contour lines) compared:

Ross

Deterministic (Lighthill and Whitham)

Microscopic model proposed

Results show analogies between our microscopic model, the deterministic model and the Ross model: a reduction of flow out of the bottleneck and the formation of a queue. With the recovering of initial conditions the output flow is able to dissipate the queue. In the Ross simulations which works by limiting speed so as to achieve a flow equal to capacity, there are exactly 2.000 veh/h coming out of the bottleneck and 4.000 veh/h at the end of perturbation.

The level countour lines shows the formation and dissipation of the queue (Fig. 1). With the microscopic model the shape is similar to that obtained with the deterministic model: when recovering of initial capacity is set the frontal part of the queue goes back and the queue reaches areas not initially involved. Ross's model shows a queue shape that dissipates without moving back in space. This because of infinite wave speed propagation as explained by Newell (1988).

3. An algorithm to automatic measurements of traffic parameters from filmed scenes.

The technics of IP enough appear promising for applications connected with traffic management (measurement, monitoring, control, support for the modellistic analysis).

Relief is accessible to traffic measures such as surface density, queue length, lane change (and many other variables associated with the movement of vehicles) which nowadays are impossible or very problematic; the greatest advantage is the fact that all the measures, classical and innovative, are obtained using only one instrument.

The use of camera and the use of classical sensors, especially magnetic loops, are combined nowadays, in the surveillance of traffic conditions; some applications are directed to make systems for automatic detection of accidents. The video-sensors appear able to offer remarkable developments also for traffic control; the modern programme of the traffic control, both in the links and in the junctions, needs in fact qualitative indicators, that can only be obtained by using video-sensors: for example queue length, surface occupation rate of link, etc.

The support for the modellistic analysis is made through:

— the opportunity of experimental verification of theoretic models;
— the identification of new elements able to characterise the phenomena;
— the determination of modellistic instruments more adherent to representate phenomena.

In the Laboratorio di Analisi dei Sistemi di Trasporto from University of Reggio Calabria, an algorithm oriented for the automatic detection of traffic parameters from filmed scenes has been implemented; the algorithm (*V.T.I.P.-: Vehicular Traffic Image Processing;* Vitetta, 1993) has been applied to traffic image sequences recorded using a professional video-camera.

The processing occurs on images that have been previously transformed into numeric form using an analogic-numeric conversion process. The IP is subdivided in a set of co-ordinated operations, reported in every frame of film sequence. The operations are relative to a superficial domain pre-pointed on the monitor by operator at the beginning of global procedure; it generally has a trapezoidal form, with two sides put on the edges of the roads and the other two sides which individualise the entry section (in the background) and the exit section (in the foreground) of approaching vehicles under observation.

The operations in a single image are of morphologic-mathematics kind and they are oriented to:

— correct the distorting perspective effects;
— identify the vehicles present in the domain;
— associate a spatial-temporal coordinates to the identified vehicles.

An opportune procedure therefore intervenes to link together the information about the sequence of images in a rational way and to reconstruct the vehicles trajectories and to raise the traffic flow parameters.

The algorithm offers good results for non-congested flows and for normal light conditions (daytime, fog absence or other visual impediments). The filming has to be recorded with the video-camera fixed both in height and angle-shot, and the camera has to frame a strip of road section preferably between 40 and 80 metres in length.

4. Results of automatic image processing designed to improve the traffic flow model.

A specific application of *V.T.I.P.* procedure has been addressed to the experimental analysis of some significant parameters for the accuracy of the traffic flow model already described; after some specifications concerning the state of free or conditioned traffic, the interesting parameters are defined and the results, obtained by automatic processing of two-lanes motorway traffic images, are presented.

4.1. Free and conditioned vehicles on the motorway.

On the two-lane straight motorway, it is possible to define "perfectly free" a vehicle (understood: "driver") as having full freedom, in the sense that it can travel with a desired speed and in the lane of choice. Such a condition is typical for an isolated vehicle; the driver will travel in the near-side lane of the road, but often he shifts to the centre of carriageway in order to enjoy better driving comfort (greater visibility, smoother trajectory on a bend, a greater sense of safety as regards lateral obstacles, etc.).

The vehicle is not "perfectly free" when other vehicles travel in the same direction and are present in its visibility field (front and/or behind). In this context it can be defined "caught in a queue" if the headway between it and the vehicle in front is inferior to a threshold value (h_{MAX}).

A vehicle can be called "free" if it can travel at the speed desired, even though it is eventually conditioned in the choice of lane; a typical situation is that of a vehicle in the right-lane which overtakes a slow-moving vehicle in the same lane. On the contrary, the vehicle will be "conditioned". The conditioning is generated firstly as regards the maintenance of speed run, and secondary as regards the freedom to change lane.

It is possible to distinguish between the vehicles moving on the right-lane (normal) or on the left-lane (overtaking-lane). A vehicle on the normal lane (see Figg.2 and 3) is conditioned when it see the temporal gap, as regards the previous vehicle, under critical threshold (h_{1a}) and, meanwhile the conditions of a sufficient temporal interval (h_{1p}) as regards the first vehicle travelling, in rear position, on this lane. A vehicle on the overtaking is "conditioned" instead when it sees the temporal gap between it and the previous vehicle in the same lane decreasing under a critical threshold (h_{2a}), it tends to slow down expecting the other to go back into the right-lane. It can also be influenced by the perception of a reduction of the temporal gap, as regards the following vehicle, on the same lane, under a critical threshold (h_{2p}); it then tends to move into the normal lane, only doing so, if there is a sufficient temporal gap (h_{1a}). In this last case, the vehicle does not, however, suffer any direct constraints on its speed and therefore cannot really be defined conditioned.

The scheme reported in Fig. 3 is on the basis of a classification of free or conditioned vehicles, using a procedure of IP; it is indeed possible to observe the vehicular behaviour on a section, even if it is of limited length, and it is possible to conclude that:

- a vehicle will be free on the normal lane if $h_{1a} \geq h_{1MAX}$; it will be caught in a queue if $h_{1a} < h_{1MAX}$; it will be caught in a queue and conditioned if $h_{1a} < h_{1MAX}$ and it will not have a sufficient temporal gap on the overtaking lane;
- the vehicle will be free on the overtaking lane if $h_{2a} \geq h_{2MAX}$, it will be caught in a queue and conditioned otherwise.

Figures 4A, 4B, 4C indicate the free and conditioned vehicular flow distribution among the carriage freeway lanes, as regards traffic density; in particular Fig. 4A, indicates the percentage distribution of vehicular flow between four possible states, obtained from lanes (1=normal lane; 2=overtaking lane)-freedom (F=free, K=conditioned) associated with an increasing density on the global route. It is qualitatively similar to the distribution obtained by J.Rorbech (1974), but quantitatively different. Moreover it consents to define the different levels of infrastructure service in a simple way, because the density is an indicative primary parameter, as is reported by the Highway Capacity Manual. Figures 4B e 4C give instead the distributive rate between free and conditioned vehicles on the single lane.

4.1.1. Movements along lanes and overtaking manoeuvres.

The probability of overtaking in Kinetic Theory of Prigogine and Herman may have different interpretations. It is good to specify that on 4 lane freeways (two lanes each direction), a lane changing movement is not always a passing manoeuvre. In fact a lane changing movement can be accomplished independently from the need of passing, and also a passing manoeuvre could be accomplished without changing lane.

It is possible to give a definition for the probability P of passing regarding the afore mentioned definitions of free and conditioned vehicle; for a given density and in a prefixed time interval we have

$$P = \frac{\text{nunber of free vehicles}}{\text{number of total vehicles}}$$

Fig. 5, that has also been obtained with camera automatic vehicle detection image processing express as the probability P depending on density and allows us to point out some remarks:

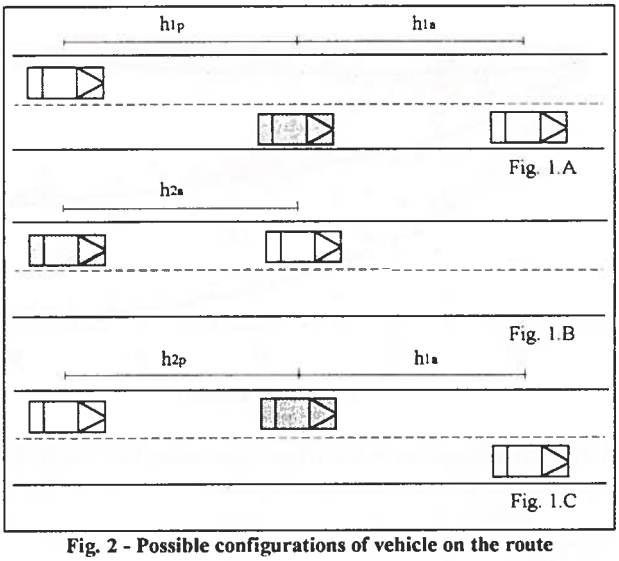


Fig. 2 - Possible configurations of vehicle on the route

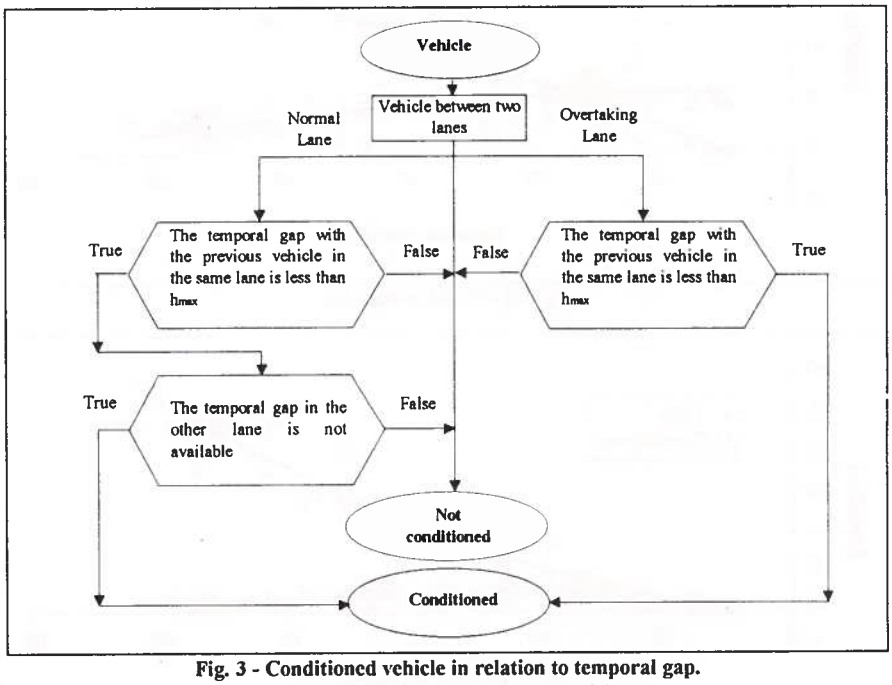


Fig. 3 - Conditioned vehicle in relation to temporal gap.
Hypothesis: constant speed per vehicle.

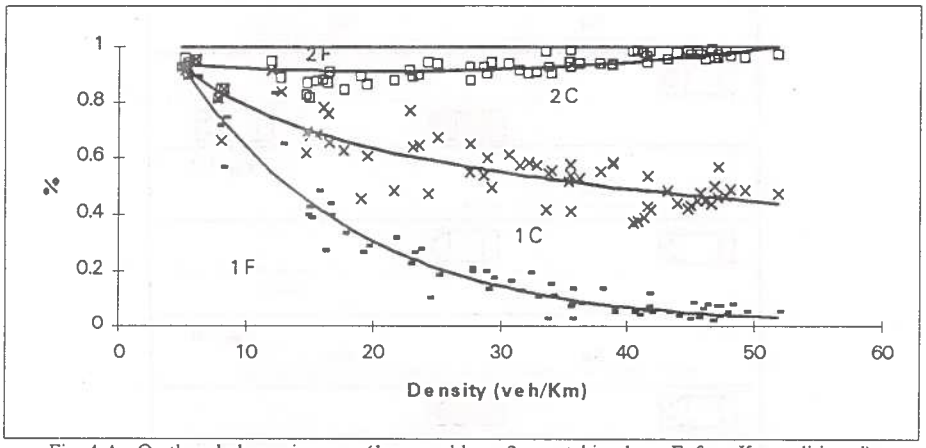


Fig. 4.A - On the whole carriageway (1=normal lane, 2=overtaking lane, F=free, K=conditioned)

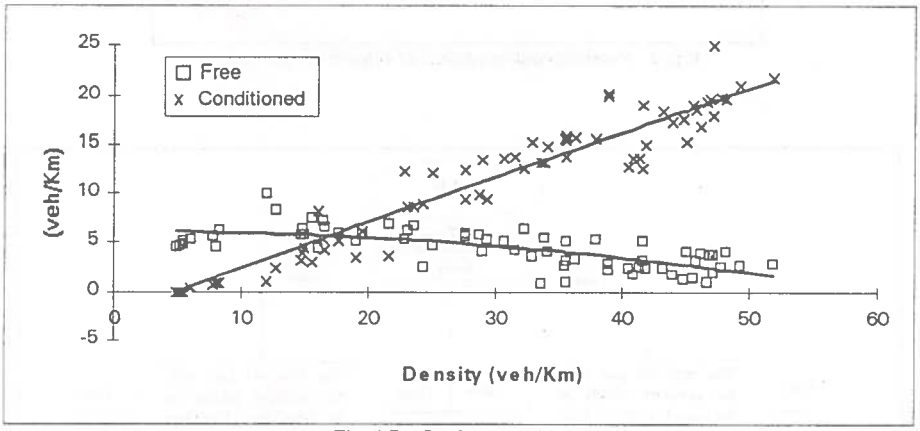


Fig. 4.B - On the normal lane

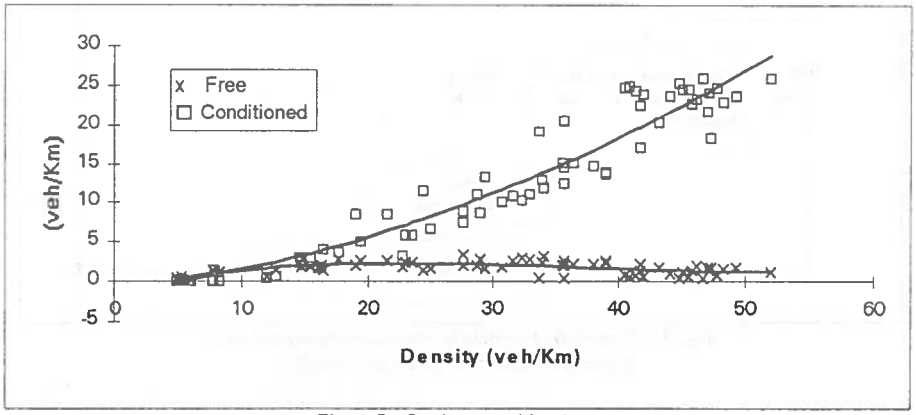


Fig. 4.C - On the overtaking lane

Fig. 4 - Distribution of vehicle over the four state possible as a function of the traffic density

— for low value of traffic density probability P is constant and equal to 1;
— with increasing density P decreases in a way that can be expressed by the function:

$$P = 1,378 \cdot \exp\left[-0,083 \cdot k_T^{0,96}\right]$$

which assures a good fitness of experimental data ($R^2 = 0.95$);
— P value is almost zero for density over 60 veh/km.

This last result is a confirmation of that already presented in a previous work (Torrieri, Gattuso, Vitetta, 1994). In that paper there was also shown a tridimensional distribution of surface density from which it is possible to extract the flat profile showed in Fig.6, which is significant because it is indicative of the distribution of frequency of lane changes depending on vehicular flow. This frequency is zero for free flow, and increases gradually with flow, reaches a maximum and then decreases for higher flow values reaching a value of zero at 3200 veh/h.

4.1.2. Traffic flow distribution between lanes of a two lanes per direction freeway.

Automatic camera traffic detection has made it possible to reconstruct the curves in Fig.7; in which the distribution of flow between the two freeway lanes is shown as a function of density k_T; these two curves, that have been obtained by regression, indicate the tendency that the left lane has, for increasing density, in attracting a growing number of cars; in over 30 veh/km of density the left lane becomes the more used of the two.

4.1.3. Implication of experimental analysis on traffic model.

The definition of free and conditioned vehicle the knowledge of their distribution between lanes at various density values, the specification of an experimental model for probability of passing depending on density, and all the other results, lead to an evaluation of some parameters of the simulation model reducing the number of parameters to estimate by direct traffic observation and so making easier a validation of model simulations. In addition this experimental data and all the information obtained can also lead to a change carry on to a change in the structure of the model; for example it is possible not to use a generic function any more for probability of overtaking, and through the immediate application of experimental data build a simulation model closer to reality. A synoptic table (Tab. 1) is presented, in conclusion, showing the mathematical functions afore mentioned.

Function		Parameters				Figure	R^2
y	$F(K_t)$	a	b	c	d		
Free on the lane 1	$c*EXP(-a*Kt^b)$	0.08	0.9649	1.37833		4.A	0.95
Conditioned on the lane 1	$c*EXP(-a*Kt^b)$	0.45	0.3021	1.9358		4.A	0.77
Conditioned on the lane 2	$a+b*K_t+c*K_t^2$	0.95	-0.0038	9.16E-05		4.A	0.35
Free density in lane 1	$a+b*K_t+c*K_t^2$	6.03	0.0063	-0.0018		4.B	0.47
Conditioned density in lane 1	$a+b*K_t+c*K_t^2$	-2.34	0.4775	-0.0003		4.B	0.92
Free density in lane 2	$a+b*K_t+c*K_t^2+d*K_t^3$	-1.12	0.3495	-0.0109	9.75E-05	4.C	0.47
Conditioned density in lane 2	$a+b*K_t+c*K_t^2$	-1.07	0.1811	0.0076		4.C	0.92
Overtaking probability	$c*EXP(-a*Kt^b)$	0.03	1.1404	1.2549		5	0.95
Flow on the lane 1	$a+b*K_t+c*K_t^2$	133.57	31.3132	-0.1923		7	0.86
Flow on the lane 2	$a+b*K_t+c*K_t^2+d*K_t^3$	-25.05	3.7706	1.3553	-0.01471	7	0.94

Tab. 1 - Experimental functions $y=F(K_t)$ with K_t=total density in the route. ($K_t \in [5, 52]$ veic/Km)

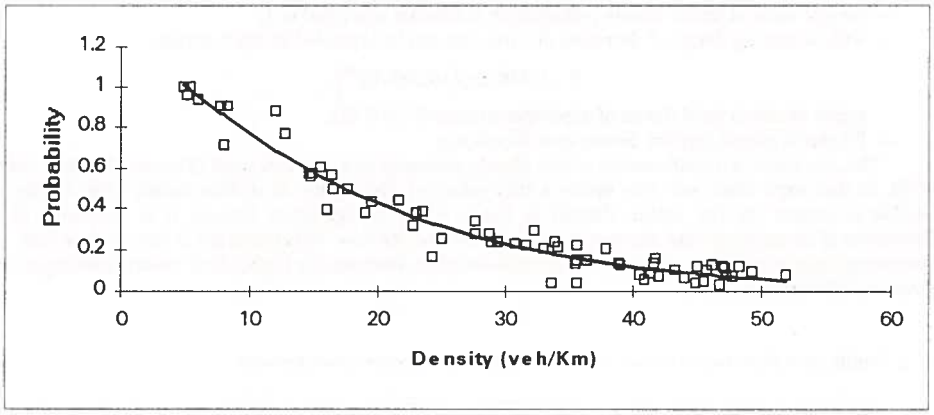


Fig. 5 - Overtaking probability as a function of the traffic density

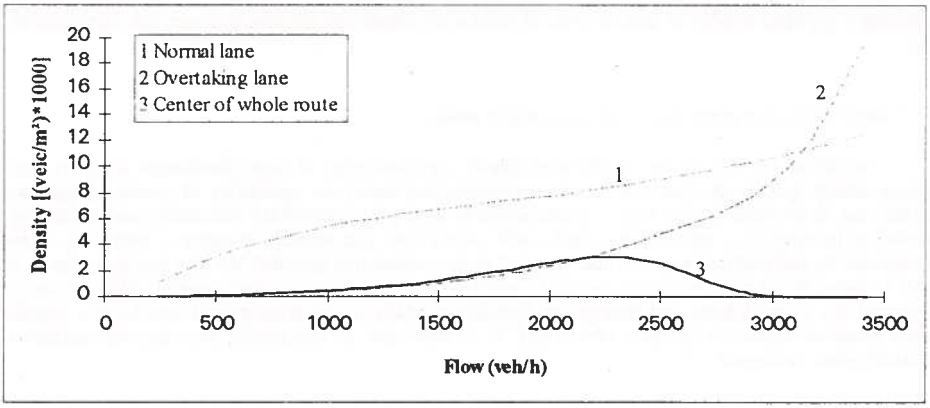


Fig. 6 - Vehicular superficial density as a function of the total traffic flow in a two lane freeway

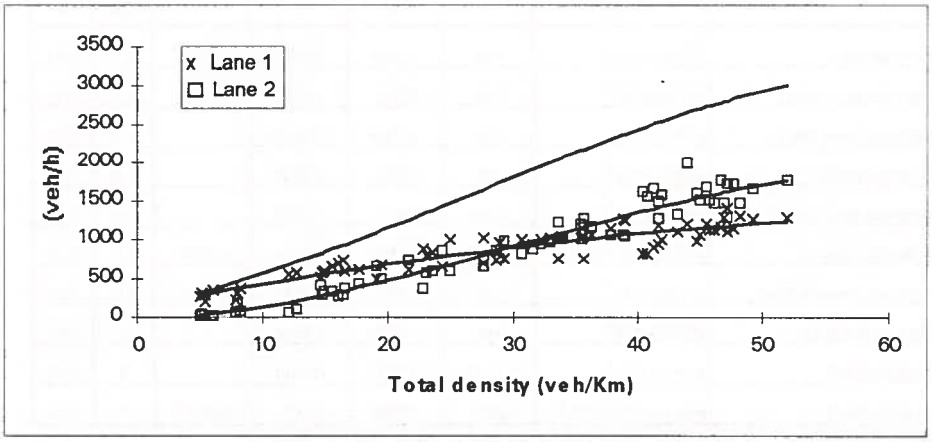


Fig. 7 - Distribution of traffic flow between two lanes as a function of the traffic density

5. Conclusions and orientations for the research course.

Further progress are expected from the combination of theoretical studies and experimental research; some depth of study is needed to improve the traffic model structure, to enlarge the wealth of information extractable by automatic image processing, to acquire new cognitive elements concerning the traffic stream on multi-lane roads.

Bibliography

Alvarez A., Brey J., Casado J. (1990): A simulation model for traffic flow with passing. *Transpn Res., 24 B.*

Astarita V. (1994): A stochastic microscopic simulation traffic model based on the kinhetic teory of Prigogine and Herman. *Proc. of the second meeting of the EURO Working Group. INRETS, Arcueil. France.*

Blosseville J.M., Lenoir F., Beucher S. (1990): Traitement d'image appliqué à la mesure du trafic. *Journée specialisée INRETS, Arcueil.*

Briquet P. (1990): Grandeurs et servitudes de "P.A.T.R.I.C.I.A." ou la "Visionique" appliquée au trafic routier et urbain. *Rapport de recherche LPC n. 156. Laboratoire Central des Ponts et Chaussées, Paris.*

Edie L.C., Foothe R.S. (1960): Experiments on a single-lane flow in tunnels. *Proc. 1st Int. Symp. on the Theory of Traffic Flow. Elsevier. Amsterdam.*

Greenberg H. (1959): An analysis of traffic flow. *Oper. Research, 7.*

Herman R., Montroll E.W., Potts R.B., Rothery R.W. (1959): Traffic dynamics: analysis of stability in car-following. *Oper. Research.,7.*

Hoose N. (1991): Computer image processing in traffic engineering, *Research Studies Press LTD.*

Lightill M.H., Whitham G.B. (1955): On kinematics waves II: a theory of traffic flow on long crowded roads. *Proc. of Royal Society. Series A, Vol.229.*

Miller A.J. (1960): Traffic flow treated as a stochastic process. *Proc. 1st Int. Symp. on the Theory of Traffic Flow. Elsevier. Amsterdam.*

Newell G.F. (1989): Comments on traffic dynamics. *Transpn Res., 11.*

Pigogine I., Herman R. (1971): Kinetic theory of vehicular traffic. *Elsevier, New York.*

Rallis T. (1967): The application of Erlang's theory to the calculation of road traffic capacity. *Proc. 3rd Int. Symp. on the Theory of Traffic Flow. Elsevier. New York.*

Richards P.I. (1956): Shock waves on the highway. *Oper. Research, Vol.4.*

Rorbech J. (1974): The multi-lane traffic flow process. *Ministry of Public Work. Road Department. Copenaghen.*

Ross P. (1988): Traffic dynamics. *Transpn Res., 22B.*

Schach S. (1970): Markov models for multi-lane freeway traffic. *Transpn. Research, Vol.4.*

Torrieri V.,Gattuso D., Vitetta A. (1994): La distribuzione spaziale del traffico veicolare su un tronco autostradale. *Terzo convegno S.I.D.T. Università "La Sapienza". Roma.*

Vitetta A. (1993): Un algoritmo per l'identificazione automatica delle traiettorie spazio-tempo di veicoli stradali da immagini filmate (V.T.I.P.). *Atti Seminario su rilievi e modellizzazione del traffico veicolare. Dip. Trasporti. Napoli*

A DSS for the urban path planning integrating parking management, public and private transportation systems[1]

Stefano Bertocchi(+), Anna Sciomachen(++)

(+) CNR - National Research Council - Dept. Information Sciences, University of Milano, Italy

(++) Dept. Information Sciences, University of Milano, Italy

EXTENDED ABSTRACT

In this paper we face the problem of guiding a private user to a parking facility. A driver in an urban area has to reach the required destination and to choose a parking space with the aim of minimizing the total travelling time. Parking facilities close to the destination may not be available due to saturation or traffic restriction. In order to make the correct choice, the driver has to know the real availability of a parking space within a "walking distance" neighborhood of the destination: if it is not the case, the driver has to know the availability of other parking spaces that make possible to reach the destination using the public transportation system.

The main idea is to suggest to the driver, starting from the origin node, a suitable path towards the required destination node while looking for an available parking space, in such a way that a set of additional constraints (e.g. time constraints, transit constraints or user transit requirements) is satisfied. We develop a model that allows the integration of the private and public transportation network. Such a model involves a great variety of different constraints: the Constraint Satisfaction Problem (CSP) methodology is the most suitable one for properly process these different constraints.

A wide range of practical problems can be solved using CSP methodology. Constraint satisfaction problems represent the intersection point between Operation Research techniques and Artificial Intelligence (knowledge representation methods in particular). The CSP's importance is due

[1]This work is partially supported by CNR - National Research Council, "Progetto Finalizzato Trasporti 2".

to the capability of giving a single representation formalism whichever the constraint formulations may be: in fact, constraints may be specified in a great variety of form such as procedures, logical expressions, equations or inequalities and so on.

Constraint satisfaction problems have three components: *variables*, *values* and *constraints*. The goal is to find assignments of the variables to their candidate values such that all the constraints are satisfied. More formally, the canonical constraint satisfaction problem involves a set

$$X = x_1, x_2, \ldots, x_n$$

of n variables x_i, each one taking its value from the associated finite domain d_i of m_i candidates, and a set

$$C = c_1, c_2, \ldots, c_k$$

of k constraints. Each constraint c_j involves a specific set $X_j \subseteq X$ of argument variables, and specifies which values of the argument variables $x_i \in X_j$ are compatibles and hence satisfy the constraint. In practice a constraint may be specified in a great variety of forms such as logical expression, algebraic equation, inequality or table listing the feasible values explicitly. A given variable can take only a finite number of candidate values, but values may be of arbitrary type, numerical or non-numerical. CSP is thus a more general problem than the related problems of linear or integer programming. No optimal solution is sought in CSP: the goal is to find the whole *feasible region*. This is possible only because the domains are restricted to being finite: linear and integer programming are more general in allowing non-finite domains.

We deal with the problem of guiding a private user to a parking facility representing the urban network by a constraint graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, i.e. a directed graph where the node set **V** consists of properly selected relevant locations in the urban area, and the arc set **E** consists of possible road connections between different locations. The model considers as variables the vectors

$$X = (x_1, y_1, x_2, y_2, \ldots, y_{n-1}, x_n)$$

of possible paths, i.e. the sequences of nodes x_i and arcs y_j connecting the origin node x_1 to the destination one x_n: in this way, nodes and arcs represent candidate values for the elements of the *path vector* X. A given origin node x_1 is necessarily a node from which the user leaves by car; a destination node x_n is a node where the user must go to, possibly by car; finally, a transit node x_i is

any node, selected by the path planning procedure, between nodes x_1 and x_n. The constraints are modeled as relations among the elements of the variable X. The model consider 7 constraint types:

C1) Tour-elimination contraint;

C2) Node-avoiding constraint;

C3) Arc-avoiding constraint;

C4) Node-forcing constraint;

C5) Arc-forcing constraint;

C6) Parking-capacity constraint;

C7) Total-travelling-time constraint.

Rising from the main idea of modelling the multimodal trip planning problem by CSP methodology, we develop a DSS (Decision Support System) model: in order to integrate the private and public transportation systems in a single transportation network (represented by the digraph **G**), we have to model two different transportation networks. The first directed graph $\mathbf{G'} = (\mathbf{V'}, \mathbf{E'})$ represents the private transportation network, while the second graph $\mathbf{G''} = (\mathbf{V''}, \mathbf{E''})$ represents the public transportation network, with $\mathbf{G'} \cup \mathbf{G''} = \mathbf{G}$ and $\mathbf{V'} \cap \mathbf{V''} \neq \varnothing$. The directed graph **G** thus results superimposing one graph on the other one. Each graph has its own incidence matrix and the same arc (i.e. two arcs with the same pair of origin and destination nodes) may be labeled by different times in **E'** and **E''**.

The user asks the system for the desired path specifying the origin node x_1, the destination node x_n, the maximum time T_{max} available for the trip and, if it is the case, some transit or vehicle requirements (the user can explicitly ask to use only the public transportation system, for example), while the transit constraints are imposed by the system supervisor. The DSS then analyses different *travelling modes*: first of all the system tries to find the optimal (x_1, x_n)-path by using only the private transportation network, searching for an available parking space in the required destination or within a "walking distance" neighborhood of the destination. Secondly the system searches for the optimal path using only the public transportation network. Finally the DSS tries to find the path using both the private and the public transportation network, searching for a parking space in an intermediate node connected to the destination by the public network.

Prolog is considered to be a very convenient language for programming CSP applications. Prolog is a non-procedural language: therefore it is not necessary to specify the procedure leading to solution. It is enough to describe facts and rules relevant for the problem: the Prolog internal search algorithm will provide the correct solution. A Prolog program is simply a finite set of predicates and clauses representing the total available knowledge. Given a predicate/clause set, the solution is a computation that, starting from the initial goal, generates new goals until the stop instruction is reached. A given *selection rule* allows to choose univocally the predicate/clause to be evaluated at each step. If there are many different clauses with the same predicate, a *search rule* allows to choose the alternative arrangement. The alternative computation search may be represented in a tree form, called *AND-OR tree*: each path on the tree (i.e. a nodes sequence from root to a terminal node) represents a possible computation. The root represents the initial goal, a non-terminal node represents one or more goal to be satisfied, while each terminal node represents a different computation result: the initial goal may be reached or not. If the computation fails, Prolog uses the *backtracking* method (i.e. to go back to the last unexplored choice) to try different computation.

Dealing with a transportation network, facts (predicates) are the graph elements used to describe the urban area, while rules (clauses) are the restriction imposed on the origin/destination path such that all the contraints are satisfied.

The path generation clause is a "generate and test" algorithm. The Prolog standard search strategy is the "depth first": each branch of the AND-OR tree is explored till the maximum depth before going up again to search new goals (i.e. an unfinished path is extended by adding nodes/arcs till a constraint is not satisfied or the destination node is reached). It is necessary to limit the maximun depth and to avoid infinite branch (typically infinite loops) construction: the bound is represented by the total travelling time (it cannot exceed the maximum time value: see the C7 constraint) and by the tour-elimination constraint (see C1). However the computation time grows with the network size, particularly if the total feasible solution space is searched: the "generate and test" algorithm first generates a candidate path set, then it tests this set in order to reduce its size. Some checks are made during the "generate" stage: if a node/arc is not valid (i.e. it belongs to the avoiding node set) then all the subsequent connected paths are not valid (see C2 and C3). Other

checks are made in the "test" section: the elements of the user required node set must be also elements of the suitable paths (see C4 and C5).

Once the path generation section is completed, the optimal path clause chooses the shortest path among the feasible path list generated by the path generation clause. The proposed DSS model considers only the travelling time as key parameter in classifying suitable paths. However other cost function may be considered as well.

The DSS software prototype has been tested on different networks representing urban areas. The DSS answering time is very sensitive to the total available time optionally specified by the user: if the user provides a maximum time much greater than the shortest path time, the "branch and bound" techniques are not able to limit effectually the feasible region size (and the DSS answering time, too). We develop an heuristic in order to reduce the user maximum time (if it is the case): when the user asks for a certain path (i.e. a certain origin/destination pair), the system already knows the shortest path time (the travelling time in absence of constraints). Comparing the user maximum time and the shortest path time (appropriately augmented if there are some user requirements or transit constraints), the DSS computes a new maximum time and tries to find a suitable path. If the search fails (constraints may be too tight) the maximum time is multiplied by a certain coefficient and the system makes a new attempt, and so on until the user maximum time is reached.

Different techniques are tested in order to limit the feasible region size and to reduce the answering time. Moreover a user-friendly graphic interface has been developed. The interface allows to support the user in finding the proper location of the reserved parking space.

ITALIAN RESEARCH ON MODELLING TRANSPORT/ENVIRONMENT INTERACTIONS

Maurizio Bielli and Pierfrancesco Reverberi
IASI-CNR, Viale Manzoni 30, 00185 Roma

1. INTRODUCTION

Transportation sector plays a fundamental role in the post-industrial society organisation, that is definitely directed towards a social and economic integration of the different countries achievable through more and more wide and dense communication networks.

Though the substantial benefits that a well-developed transportation network is capable to provide to the system of human activities are highly appreciated, there is a growing awareness of the potential significant inconveniences which may be attributed to its considerable growth rate registered in all recent years. Actually, economy and trade expansions encouraged by a more comprehensive transport system than the past determine a rise in transport demand which is reflected in traffic congestion due to increased mobility.

At present, transportation sector burden far exceeds the carrying capacity of the environment and the risk exists that ecological sustainability is seriously endangered by economic development. The social costs of road transport are estimated to amount to approximately 5% of Gross Domestic Product (GDP) in most developed countries; these costs are largely due to air and noise pollution caused by vehicle circulation and the lack of safety of the road network (Himanen, Nijkamp and Padjen, 1992). Traffic congestion is rapidly becoming one of the most serious problems affecting urban areas. Since road traffic in Europe is forecast to increase by between 80% and 150% by the year 2025, considerable adverse effects on the environment are anticipated, causing atmospheric and acoustic pollution levels in most cities to overcome the fixed government standards, intended to protect the large number of affected population.

It goes without saying that comprehensive and yet articulated strategies are needed in order to keep transportation sector growth within the bounds imposed by a sustainable development. Scientific research is highly contributing to pursue this aim through the activation of a large number of projects that study the whole set of transportation/environment interactions. These projects are expected to introduce significant innovations both in the technological context (e.g. by producing electric vehicles, alternative fuels, catalytic conveters, efficient and silent engines) and in the strategical decision-making process (including policy and economy measures directed towards transport demand modal and space-time restriction and redistribution, as well as transport supply operational improvement).

Environmental management is a fairly recent research area, so that the determination of the appropriate integration of monitoring and transportation networks with the land system of activities needs further investigation, if we aim at optimizing the overall development plan for the considered region (Bielli, 1991).

This paper is structured as follows. Section two reviews international research activities in the field of transport and environment interactions, focusing particularly on air and noise pollution prediction models as well as on existing tools for socio-economic evaluation of traffic impacts on the environment. The main traffic management and control actions devised by different countries for alleviating these impacts are also outlined. Next, section three analyses researches developed in the framework of the Second Finalized Project on Transportation sponsored by the National Research Council (CNR) of Italy, which coordinates and financially supports research institutes, universities and

industrial companies for the purpose of maintaining the country at the international scientific standard level. The primary aim of the paper is to classify in an understandable and, as far as possible, exhaustive way the activities performed by the Research Groups (R.G.s) working on the Project (in particular on Subprojects 1 "Mobility management and planning strategies" and 4 "Urban and metropolitan transport"), in order to characterize the most relevant research topics, on which it is necessary to concentrate the efforts if it is expected the community welfare to be achieved. A list of the major contributions expected from the Project activities is reported in the concluding remarks of section four.

2. STATE OF ART REVIEW

Recent research trends aim at coordinating and integrating the efforts of scientific communities of the different countries in a comprehensive joint framework, in order to avoid waste of resources and exploit the resulting synergies.

Among the most relevant research projects funded by the Commission of the European Communities (CEC) it is to recall DRIVE, PROMETHEUS, SAVE, THERMIE, COST, CORINE, EURET, STEP and LIFE. As regards Italy, the National Research Council developed in the years 1981/1985 a first Finalized Project on Transportation and then, in 1992, started its obvious continuation, the Second Finalized Project on Transportation; activities of both projects are closely connected and yet complementary to those of the aforementioned European programs (Bianco and Bielli, 1990).

On the whole, these programs dedicate large sections and fundings to analyse transportation system impacts on the environment, which they attempt to limit through the definition of strategies and tools for transport demand management and control. In particular, new communication and information technologies show significant potential for implementing and enforcing measures which are primarily intended to reduce congestion, enhance safety and improve operational efficiency (Shladover, 1993).

Transportation system impacts on the environment are of various nature and strength depending on the different transport modes. Thus, the negative effects of the latter should be assessed and compared according to the following general environmental quality criteria (Commissione delle Comunita' Europee, CCE, 1992):

-operational pollution, originating directly from vehicular traffic, which greatly affects air, water and vegetation quality and produces noise and vibrations;
-soil intrusion by transportation infrastructures, that often determine permanent and irreversible detriments;
-traffic congestion, which is a recurring phenomenon of increasing significance in urban and metropolitan areas, not only for road- but also for rail-, air- and navigation transport;
-hazardous freight transport risk of possibly devastating accidents.

It has been assessed that road transport is the cause of the most severe impacts of operational pollution, soil intrusion, traffic congestion and hazardous freight transport; therefore, a large amount of research project resources have been employed to improve its performances. Results achieved by these projects and their socio-economic implications largely influence transport policy measures adopted by the different countries.

It has been rapidly recognised that the technological progress contribution, which consists of efficient and silent engines, catalytic converters, alternative fuels and electric vehicles aiming at reducing vehicle emissions below acceptable levels imposed by the authorities, is not enough by itself

to curb the adverse environmental effects due to the rapid transport and mobility demand growth (Himanen, Nijkamp and Padjen, 1992; DRIVE, 1990). This means that it must be supported by the adoption of a comprehensive strategy including policy and economy measures directed towards transport demand modal and space-time restriction and redistribution, as well as transport supply operational improvement, that may be listed as follows (CCE, 1992; Himanen, Nijkamp and Padjen, 1992; DRIVE, 1990; Nishioka, 1989):
-traffic restraints, regulating the level and composition of vehicle movement through a specific area during peak hours (using capacity and/or regulatory mechanisms);
-car pooling encouragement actions (e.g. preferential lanes and/or permitted access to central areas for high-occupancy vehicles);
-regulation and control on vehicle parking in congested areas (through duration and/or availability requirements);
-provision of adequate alternatives to private car use, in the form of enhanced public transport quality and efficiency as well as safer movements on foot and by cycle;
-adoption of fiscal policies providing incentives for clean vehicle and public transport use (road pricing, driving licence,car ownership and/or vehicle mile traveled taxes, fuel consumption and pollutant emission taxes, parking fares);
-implementation of advanced communication and information technologies in areas such as traffic control, route guidance, parking management and information, fleet management, public transport, automatic debiting, travel and tourist information.

Furthermore, in the medium/long term:
-making of park-and-ride sites, tunnels, highways and/or removal of intersections and crossings;
-existing transport supply improvement, including urban and interurban rail- and roadway management and rehabilitation for vehicle speed and noise reduction (e.g. antinoise barriers and screens);
-comprehensive planning policies integrating transportation and urban management, for the purpose of reducing mobility through physical and functional interrelationships of housing, employment and community service facilities;
-application of Environmental Impact Assessment (EIA) procedures in the planning context, especially for infrastructural projects.

Due to the wide range of possible interventions, it is necessary to set up a methodology for the identification of the proper transport policy measures, on the basis of a correct forecast of their connected impacts. Therefore, it is essential to resort to Decision Support Systems (DSS) and simulation models, in order to determine optimal planning, regulatory and fiscal actions. In particular, appropriate models have to be used for each of the following areas of interest:
a) transport system planning and management;
b) air and noise pollution level prediction;
c) socio-economic evaluation of traffic impacts on the environment.

Transport system planning and management are based on two main classes of models (Meyer and Miller, 1984):
-demand/supply interaction models, which simulate passenger and car flows on each link of the considered network at different times, by suitably matching demand and supply data (that is, O/D trip matrix and transport system performance indicators);
-transport system/land use interaction models, which simulate the effects of network space-time accessibility on location choices of residential and economic activities.

Air and noise pollution monitoring and prediction and socio-economic evaluation of traffic pollution will be dealt with in next subsections; attempts to define an overall environmental quality index will also be discussed.

2.1 Air pollution modelling

Pollutant concentrations at any given point depend on total vehicle emissions and their dispersion properties. More specifically, they can be computed as the sum of a local contribution, strictly originating from neighbouring sources, and a background contribution, originating from the whole set of sources located in the study area. In general, pollutant concentrations rapidly decrease within distances of tens of metres from the source.

Each contribution can be assessed by using a modelling system which includes combined modules for the assessment of emission and dispersion phenomena (Nuzzolo, Festa and Nicotina, 1993; Matzoros and Van Vliet, 1992; Bruno and Improta, 1992). The emission module usually includes three sub-models:

a) an emission model, providing the amount of each pollutant exhausted by the different vehicle classes (identified according to type and age of vehicles) depending on operating mode (distinguished in cruise, acceleration, deceleration and idle);

b) a queue model, calculating queue lengths and waiting times at signalized junctions, that represent relevant parameters to be included in the computation of pollutant emission levels;

c) a downflow model, determining vehicle operating mode at any point of each link of the network, on the basis of queue model results.

The dispersion module also includes three sub-models:

a) a dispersion model for local contribution calculation, providing detailed and accurate results for emissions originating from adjacent areas;

b) a dispersion model for background contribution calculation, giving a less precise account of emissions originating from more distant areas;

c) a model for determining meteorological conditions, providing essential input data required by the dispersion models.

Emissions are usually computed as a function of vehicle type, age, weight, speed and operating mode, traffic volume and composition. Emission models can be classified into some broad categories. In particular, it is possible to distinguish between aggregate models, considering vehicle fleet composition at a given moment, and disaggregate models, computing specific emissions for individual vehicle type and age. Moreover, instantaneous models differ from average-speed models. Actually, instantaneous models relate pollutant emissions to the time history of speed and acceleration values and are suitable for studies dealing with individual intersections or road sections for which detailed information can be obtained reliably. On the other hand, average-speed models estimate pollutant emissions from average speeds of trips and are useful for assessing total consumption in large road networks for which detailed information relative to individual trips is not readily available. These models can be further distinguished depending on the fact that the reference average-speed value is an ideal constant speed value or the result of standardized driving cycle patterns which comprise different acceleration phases; it has to be noted that constant speed-based emissions turn out to be lower than driving cycle-based emissions, average speed being equal (Nuzzolo, Festa and Nicotina, 1993; Matzoros and Van Vliet, 1992; Bruno and Improta, 1992).

Valuable emission models have been recently proposed by INRETS in France and CORINAIR EEC project; in USA, the models MOBILE and CALINE have been developed and periodically updated in order to take account of the changing vehicle fleet composition. The major pollutants emitted from motor vehicles and considered in the aforementioned models include (DRIVE, 1990):

-Carbon monoxide (CO), that is produced as a result of incomplete combustion within the cylinders of a gasoline or diesel engine;

-Carbon dioxide (CO_2), that is the major product of the combustion process and a major contributor to the greenhouse effect;
-Nitrogen oxides (NO_x), that also contribute to the global warming of the planet and the greenhouse effect;
-Sulphur oxides (in primis, sulphur dioxide SO_2) and particulates, that are primarily due to diesel engines and represent the main cause for acid rain damages to buildings and vegetation;
-Hydrocarbons (HC), that are produced by an incomplete fuel combustion process;
-Lead (Pb), that is a cumulative poison emitted in particulate form by the exhausts of gasoline engines.
Hydrocarbons and nitrogen oxides, in the presence of sunlight and under stable meteorological conditions, may react to form photochemical smog, a derived (or 'secondary') pollutant which can inflict damage on people, wildlife and buildings and can significantly reduce visibility. The usual indicator of smog level is the concentration of ozone, that is one of its main components.

Data on pollutant emissions of road transport vehicles of different weight and size are reported and discussed in: CCE, 1992. On the whole, road traffic has been recognized as the most important source of both air and noise pollution in urban areas; growing awareness of the problem has resulted from psychological and physical disorders affecting people living close to busy roads. In particular, it is responsible for an average of 75% carbon monoxide, 80% lead, 55% nitrogen oxides, 30% hydrocarbons and 3% sulphur oxides emissions as well as 65% noise annoyance cases in the EEC countries.

The dispersion process is greatly influenced by local meteorology and topography, air mass turbulence and chemical reactions among pollutants. Because of its complexity, however, it is not possible to find a complete solution to the equations of motion and many models have arisen based to varying degrees on rigorous theoretical concepts, semiempiricism and in some cases engineering 'rules of thumb', but none of these has been shown to be clearly superior to the others and most have persisted in common use (Jakeman and Taylor, 1993).

Dispersion largely depends on wind speed and direction, but the variations of these parameters with time of day and season of the year are even greater factors in influencing pollutant concentrations. The change in air temperature with pressure, and hence with altitude, is an important element in considering the incidence of air pollution; the heat emitted from traffic streams can change the local atmospheric stability structure near the roadway, causing an increse in vertical diffusion (DRIVE, 1990).

Buildings play a fundamental role in the process of pollution dispersion in urban areas. The higher the buildings on each side of a street and the more they form a continuous canyon facade, the more difficult is to remove traffic pollutants, especially on days when the air is still. Complex aerodynamic effects must be taken into account in the modelling process of these patterns; research work in the field is mainly based on wind tunnel results, field measurements and numerical models (DRIVE, 1990).

Mechanical mixing and turbulence induced by passing vehicles will also have an effect on air movements. During calm periods the mechanical mixing is expected to dominate, especially in the street canyon, while during periods of stronger winds its effect is more likely to be additive; these effects are however very hard to model. Finally, the time from the release of a pollutant in the atmosphere and its actual measurement has been found to affect its concentration value (DRIVE, 1990).

Operational air quality models can be classified into source-receptor and statistical models. Source-receptor models include both the simplest (i.e., Gaussian and K-theory models) and the most complex ones (i.e., higher-order closure). Gaussian plume models are the most widely used because of their simplicity, given that more complicated alternatives do not perform any better in common

applications. Their limitations are mainly due to the restrictive assumptions on which they are based, which require the downwind concentrations along the vertical and crosswind axes of a plume to be normally distributed, the plume to travel in a straight line, wind speed and direction to be those defined at the source, diffusion in the along-wind direction to be neglected and the plume to be reflected from the upper and lower boundaries of a box which includes it (Carras et al., 1993). The state of the art includes various kinds of Gaussian models of different complexity and characteristics, among which the most relevant for evaluating dispersion phenomena in urban areas are CALINE 4 (see, e.g.: Benson, 1989) and HIWAY 2 (EPA, 1986), both developed by the Environmental Protection Agency (EPA) of USA.

K-theory (or Eulerian, or first-order closure) models are oriented to the calculation of pollutant concentrations in each tridimensional space position at subsequent times. Given initial conditions as well as wind and turbulence components, diffusion equation terms are to be solved on a fixed grid in order to determine the concentration values; unfortunately, the complexity of turbulence terms does not permit a practical use of this kind of models, especially when higher-order correlations among variables are considered (Bruno and Improta, 1992).

The Lagrangian approach gives a probabilistic description of the motion of pollutant particles in the atmosphere on the basis of an appropriate continuous Markov process and estimates average concentration at a given point by counting the number of particles in a certain volume including that point. Pollutant concentration at the moving point is assumed to remain equal to its initial value, if the effects of molecular diffusion are neglected. If stationary homogeneous turbulence conditions are assumed, the application of the Langevin equation describing the process reproduces results available through the simpler Gaussian plume modelling; otherwise, extension to more general turbulence fields is a slow and faltering process (Sawford, 1985).

A certain degree of uncertainty is associated with model predictions, comprising a reducible component, due to inadequacies of the model formulation and input data, and a natural irreducible component, due to the random nature of turbulence in the atmosphere. The extent of this natural variability and the consequent prediction uncertainty decreases with increasing averaging time, but it is rarely less than a factor of two. Actually, air quality models predict a mean value, while an observation represents one possible realization of pollutant concentration about the mean value; thus, variance between observations and predictions has to be expected and can be assumed as a measure of the uncertainty inherent to model predictions (Jakeman and Taylor, 1993).

A number of studies show that air quality models have good performances when they are asked to estimate the magnitude and frequency of the highest concentration values over a year. This fact justifies the usual application of such models for regulatory aims, since close temporal and spatial agreement for observed and predicted values is not necessarily demanded. Nevertheless, the greatest challenge for model builders is currently that of simulating the plume behavior in the environmental conditions producing these highest concentrations (Carras et al., 1993).

On the other hand, statistical models rely on the identification of the distributional form describing pollutant concentrations at a receptor, which estimates of mean and extreme values may be derived from. Although no "universal distribution" exists for data relative to any pollutant, knowing the form it takes in specific cases is essential for assessing the environmental impact of pollution. In particular, when parameter values may be related to factors such as the prevailing meteorology or emissions, model results can be used for determining emission-regulating strategies, as a function of properly selected threshold levels for each pollutant (Carras et al., 1993).

Data provided by suitably located monitoring stations measuring pollutant concentrations in the atmosphere should be analysed and compared to fixed government standards, which impose regulatory limits on some specified statistics computed from the distribution of values taken by the

various pollutants over different time intervals. For instance, it is estabilished that the 8-hour mean and the 1-hour mean value of carbon monoxide concentration distribution cannot exceed 10 mg/m^3 and 40 mg/m^3 respectively. On the other hand, for sulphur dioxide the median of the 24-hour mean concentration distribution over a year cannot exceed 80 $\mu g/m^3$, while 250 $\mu g/m^3$ is the threshold for the 98th percentile of the same distribution. Other pollutant concentrations are regulated similarly. Individual vehicle emission rates according to type have also been subject over the years to more and more stringent EEC directives (DRIVE, 1990).

2.2 Noise pollution modelling

Noise level measured at a reception point is a function of distance from the source, distinctive source properties (in the road traffic case, road type, gradient and surface as well as traffic flow volume and composition), local meteorology (in particular, air temperature, humidity, wind speed and direction) and topography (that is, presence of crossings and/or signalized intersections). It is possible to distinguish three main classes of traffic noise forecasting methods (DRIVE, 1989):
a) manual methods, based on tables or simple analytical equations, which are suitable for application to smooth morphological conditions, in order to provide a first approximation analysis;
b) on-scale simulation of physical models, which reproduce in detail complex land patterns, at the expense of a greater amount of time and other resources;
c) numerical simulation by automatic computation codes, which permit to overcome the main limits of the aforementioned methods, although the imposed computational burden on the available hardware could influence considerably model formulation and solution procedures.

At present, a large number of traffic noise forecasting programs have been built by research Institutes all over the world on the basis of experimental surveys attained in measurement campaigns that considered the most typical road traffic situations observed in the different countries. In particular, models for noise level estimation in urban areas are based on network traffic condition simulation, which gives each vehicle type, position, speed and acceleration rate at any given moment. On the basis of the obtained data it is possible to derive, by means of a regression analysis, each vehicle class noise levels, that can be combined according to traffic volume and composition as well as distance from measurement point to obtain a comprehensive evaluation of noise pollution. Some advanced models include correction factors taking into account the acoustic reflection and diffraction phenomena caused by some physical barriers that may influence noise propagation mechanisms, such as vegetation, hills, anti-noise screens or parallel rows of dwellings (Nuzzolo, 1990; DRIVE, 1989).

Among the most commonly used software codes there are STAMINA, developed in USA by the Federal Highway Administration (FHWA), and CETUR, developed in France. Both programs evaluate the sound level at a reference point located at a conventional distance from the source; noise propagation mechanisms are taken into account by including correction factors in the model, generally determined on an empirical basis. The influence of meteorological conditions is not considered explicitly in the programs; rather, the range of possible weather conditions is specified within which model results can be assumed to be reliable. Three categories of vehicles (passengers, light vehicles and trucks) are considered in STAMINA, while CETUR evaluates separately the contribution of light and heavy vehicles. A common flaw of these models lies in the fact that noise emission factors for the different vehicle classes are assigned invariably based on the average characteristics of the vehicular fleet of the country where the programs have been developed. It follows that, due to the peculiarities of USA vehicular fleet, emission factors computed by STAMINA are very different from the corresponding values calculated by CETUR for the European reality (DRIVE, 1989).

Noise pollution levels cannot overcome threshold values which reflect the nature of the affected areas and the reference time period (day or night). Most countries agree both on limit values and area classification criteria; the former are measured in terms of "Continuous Equivalent Levels" (Leq) in dB(A), which give the probability distribution of noise levels at a site, while the latter roughly include 'protected areas', 'residential areas', 'heterogeneous areas', 'busy areas', 'mainly industrial areas' and 'exclusively industrial areas'.

In short, it can be stated that noise annoyance cases occur when the 55 Leq dB(A) threshold for residential areas and the 65 Leq dB(A) threshold for busy areas are overcome; it is worth noting that in Europe the affected population rate ranges from 34% (Denmark) to 74% (Spain), if the reference (overcome) limit is 55 dB(A), and from 4% (BENELUX) to 23% (Spain), if the reference (overcome) limit is 65 dB(A). Noise emissions for individual vehicles are also restricted according to type by government standards; although these vary from country to country, typical levels are generally recommended (DRIVE, 1990).

2.3 Environmental impact analysis

The application of the discussed composite modelling system to environmental impact analysis is based on the premise that, although the actual absolute levels of pollution may be affected by many other factors besides those included in the component models, the modelling system can reasonably detect relative differences in levels of pollution among alternative sets of traffic load distributions deriving from different traffic management schemes, which can accordingly be ranked on a number of environmental quality objectives (Anderson and Taylor, 1988).

The concept of 'environmental capacity' may be used to indicate the level of possible conflict between traffic and living activities in a street or urban area; it is defined as the maximum number of vehicles (and associated 50th percentile speed and proportion of heavy vehicles) that may pass along the street in a certain time period and under fixed physical conditions without causing environmental detriment (Immers and Oosterbaan, 1991).

A comprehensive 'Environmental Sensitivity Index' may be compiled as an overall rating of the street in question, on the basis of inventories of land use, road network and traffic variables as well as an assessment of the levels of sensitivity of these variables for the street (Anderson and Taylor, 1988). The accuracy and reliability of such an index is heavily dependent on the nature of the indicators which combine to form it, that should be based on objective criteria, preferably measured on a quantitative scale, and the selected weighting system, which should reflect preferences and priorities of the affected community.

Environmental quality may be seen as the inverse measure of environmental degradation. It has been stated that safe and annoying levels of a range of pollutants have been set down by international bodies such as the World Health Organisation and the Organisation for Economic Cooperation and Development, and national, state or regional environmental protection agencies. Given that several pollutants may be present simultaneously in differing concentrations, an overall air quality level may be defined according to one of the following classes (Anderson and Taylor, 1988):

-Level One, when all pollutants are below acceptable levels;
-Level Two, when one or more pollutants exceed their acceptable annoyance levels;
-Level Three, when one or more pollutants exceed their detrimental annoyance levels;
-Level Four, when one or more pollutants exceed their detrimental health effect levels.

Level One represents clean air, while Level Four fails to meet any of the specific air quality standards and the other intermediate Levels may still satisfy the overall air quality objective; a particular sample of air is assigned to a Level on the basis of its rating for each pollutant. A similar

approach may be adopted for noise pollution, with reference to fixed thresholds for the Leq measured during day- and night-time. In such a way, an assessment method for overall environmental quality evaluation is defined on the basis of multi-dimensional criteria.

At present, control intervention is likely to be limited to the application of gating strategies which restrict traffic entry to a specific area when preset pollution thresholds are actually (or likely to be) exceeded. Comprehensive traffic restraint policies are expected to shorten fuel consumption by cars with at least 10% and exhaust emissions with more than 40% (CO), 50% (HC) and 85% (NO_X), even in the face of a forecast 40% increased car ownership (compared to the late 1980s) for the next 20 to 30 years (Himanen, Nijkamp and Padjen, 1992).

It should be noted that the reduction of the different kinds of pollutants may not be achieved simply by the implementation of any single technique. This is due to the fact that the mechanisms behind the production of each pollutant are different. For instance, the highest levels of carbon monoxide emissions occur when a vehicle is cold or when the engine is idling. Generally, they tend to decrease with increasing vehicle speed, and so do emissions of hydrocarbons; on the contrary, those of nitrogen oxides can be expected to increase. On the other hand, carbon dioxide emissions are approximately proportional to the amount of fuel consumed. Since fuel and emission optimization strategies, which encompass road transport informatics technologies, are expected to result in reductions of both fuel consumption and exhaust emissions, they must be properly adjusted in order to take into account the wide range of considered elements and phenomena (DRIVE, 1990).

2.4 Socio-economic evaluation

Socio-economic evaluation of traffic pollution is necessary to assure equity between polluters and those suffering pollution by determining the amount of transfer price for a proper incentive policy. An important market principle is the internalization of external effects of the transport system (primarily due to traffic accidents, air pollution, noise annoyance and recurring congestion) through pricing of the external costs for society. If these effects are not considered properly, an allocation of resources will result that produces and consumes too much transportation, favouring transport modes that create many negative effects at the expense of those that cause few negative effects. Studies conducted in Sweden demonstrate that at present only private cars with catalytic converters, rail traffic on the main line system and domestic aviation pay according to the current evaluation of external costs (Hansson, 1991).

The problem is that environmental impacts can neither be estimated nor evaluated and a cost responsability consequently determined. Even if an optimal charge cannot be defined, society is compelled to establish restrictions, regulations and economic incentives to reduce detrimental effects, in light of the fact that nature cannot manage current pollution level. Once an environmental objective is formulated, e.g. as a specified limitation for an emission, the question is how to reach this goal; well, the pricing of external effects can be used as an economic instrument to reach any specified limitation as cost efficiently as possible. Without a charge, there are only costs for measures reducing the emissions, while, with emission charges, there would be counteracting reductions in the total charges paid for emissions; the levels of the charges will determine whether it is favorable or not in specific cases to take measures to reduce emissions (Hansson, 1991).

Imposing taxes and fares on infrastructure use, miles traveled, fuel consumption, pollutant emissions, car ownership, driving licence and/or encouraging the purchase of unleaded-fueled cars with catalytic converters may represent effective means to achieve the estabilished government standards for emission levels. Thus, environmental constraints may be capable of stimulating

technological progress. Some examples of in force or proposed charges for road, rail, air and navigation traffic emissions are reported in: Hansson, 1991.

It has been stated that external costs are not easy to measure according to economic-financial tools (e.g. cost/benefit analysis); nevertheless, social costs of road traffic in several countries have been assessed in a OECD report (1989). It has been obtained that noise annoyance and air pollution costs, including both damage costs (e.g. productivity losses, health-care costs, decline in property values) and abatement costs (e.g. adjusted vehicle technology and antinoise screens or barriers), added to the damage and recovery costs resulting from incidents due to lack of safety of the road network, range from 2,5% to 3,2% of GDP. This is, however, only a conservative estimate: if congestion and land consumption (e.g. road and parking space building) would also be included, the non-internalized social costs of transport might amount to about 5% of GDP in most developed countries.

Scientific research is currently attempting to obtain significant results in the building process of proper models for each of the discussed reference areas. Although it could be interesting to analyse in detail research projects developed in the field of transport and environment interactions by the different countries all over the world, space limitations impose to focus the attention in this paper to the Italian Finalized Projects on Transportation, attempting to characterize the most relevant research topics on which it is necessary to concentrate the efforts if it is expected the community welfare to be achieved. This is the subject of next section.

3. ITALIAN RESEARCH

At present, two Finalized Projects on Transportation (PFT 1 and PFT 2) have been carried out by National Research Council of Italy, the first in the years 1981/1985, while the second, started in 1992, is currently being developed.

Researches undertaken in the field of transportation and environment interactions have evolved from a specific transport mode-oriented approach (PFT 1), that analyses independently the environmental impacts of each transport mode, to an integrated approach (PFT 2), that tries to elicit the complex set of interrelationships between the transport and the environmental systems as well as within the two systems. At the same time, the attention of researchers has focused on urban and metropolitan areas, given that these are currently experiencing recurring traffic congestion phenomena which cause significant detrimental effects on mobility, accessibility, safety and efficiency of the transportation system and on the environment (Bianco and Bielli, 1990).

The relevance of the problem is emphasized by the fact that in European countries passenger transport volumes have expanded at an average rate of 4 percent in the last decade and this trend is predicted to continue in such a way that road traffic is forecast to increase by between 80% and 150% by the year 2025. Moreover, the transportation sector accounts for 18-23% of primary energy use in the high-income OECD countries and, in particular, urban travel accounts for approximately 8-10% of national energy use; owing to the anticipated demand growth, energy use for travel is expected to increase as well.

In order to classify in an understandable and, as far as possible, exhaustive way the activities performed by the Research Groups (R.G.s) working on PFT 2 (in particular on Subprojects 1 "Mobility management and planning strategies" and 4 "Urban and metropolitan transport"), a matrix has been conceived that lists on the rows the environmental resources subject to possible impacts due to the transport system, whose main features and actions are listed on matrix columns (see table 1). The generic matrix entry consists of two elements. The first one is the number of R.G.s (twenty-six in

all) partially or, more rarely, fully engaged in the study of the effects of action j on resource i; the second is a weight that measures the R.G. overall involvement in the considered topics on a tentatively determined qualitative three-valued scale (s=strong; m=medium; w=weak), according to the financial resources and the number of man-months invested in the Project, as provided by the triennal plan 1992/1994.

Matrix data analysis enables us to draw some interesting remarks.

As it could be expected, one of the most important research topics is related to monitoring and modelling road traffic dynamics and vehicle pollutant emission and dispersion processes in urban areas, for the purpose of realizing an integrated traffic control system including socio-economic and environmental constraints, which strongly influence the decision-making process inherent to the adoption of optimal traffic management strategies.

Experimental studies have been carried out in order to evaluate the performances of air and noise pollution simulation models in urban and extraurban areas. In particular, Italtel Telesis (1993) has developed a lagrangian particle model for studying the dispersion of primary pollutants in an urban topography and has performed sensitivity analysis and first validation trials within specific street canyons. Obtained results show that the proposed model (called GEM) is at least as reliable as the commonly used USA-EPA APRAC-3 (Simmon et al., 1981) and CPBM (Yamartino and Wiegand, 1986) models. The APRAC-3 box model represents a first step towards the adequate representation of the so-called "canyon effect" within air pollutant dispersion modelling in urban areas, particularly when road width is comparable to building height. CPBM is a multibox model which adopts a Gaussian approach in order to compute pollutant concentration with respect to each of the different boxes in which a street canyon has been previously divided. However, the GEM model cannot be used currently as a suited tool to perform on-line selection of traffic management and control actions devised for protecting the environment, due to the fact that its implementation and solution still require an excessive computational effort.

Fraternali and Sozzi (1993) discuss the potentialities of a proper calibration of the APRAC-3 model to the concentration values of road traffic-related pollutants (such as CO) measured by an air quality monitoring system covering some reference area. This is done in light of obtaining reliable short-term predictions, with particular reference to warning conditions for the environment.

Moreover, a numeric lagrangian particle dispersion model has been developed by Cenedese, Leuzzi and Monti (1993) for computing air pollutant concentrations in a close area to high-flow urban or arterial roads under variable meteorological conditions. The Langevin equation describing the dispersion process reproduces atmospheric turbulence conditions typical of Gaussian models, given that higher-than-second order correlations among variables are neglected. Intensive laboratory-testing has been conducted that simulates propagation mechanisms corresponding to different traffic network and land use patterns; further refinements of the model are under study in order to enhance the quality of the achieved results.

On the other hand, Brambilla (1993) has carried out an extensive survey based on a large amount of noise level data, collected in the proximity of some Italian freeways, and aimed at the development of a reliable traffic noise forecasting model. Regression analysis is the statistical tool which leads the model-building process and permits to achieve promising results if compared with existing prediction models, such as those developed by the USA-FHWA.

A research area that appears to be somewhat overlooked in the Project deals with socio-economic evaluation of transport policy impact on the environment. Even if external costs such as air and noise pollution or lack of safety are not easy to measure by common economic-financial tools, it is recognised that the pricing of external effects can be used as an effective way to reach some

environmental objectives, as the standard levels fixed for the different pollutants by international agencies and organizations require.

The availability of a complete and reliable set of information represents an essential requirement for the correct formulation of any simulation or prediction model. Nevertheless, the logistic and economic problems associated with the setting of a monitoring system that may guarantee the real time control of traffic conditions and environmental pollution on the whole network impose to search for the optimal location of measurement points, in order to achieve the maximum possible amount of information from a minimal set of detectors.

In order to obtain significant, coherent and reliable data from the monitoring network, measuring stations must be strategically placed over the area of interest and must be continuously functioning over the reference time interval, given that international standards for some pollutants are set with respect to certain statistics computed from their 24-hour mean concentration distribution over a year. Since point measures provided by these stations could be greatly influenced by local traffic and meteorological conditions, they can hardly be assumed as representative of the mean pollution level in the surrounding area (Cirillo, 1992).

Thus, one of the most relevant open problems is that of determining the optimal location and type of monitoring stations to be used in the analyses. Researches are currently being developed in the framework of the Transportation Project for the purpose of determining the optimal fusion of data provided by monitoring systems with data derived from simulation models, concerning sites where no sensors are located. As recognized by Bielli et al. (1994), an adequate knowledge of traffic flow patterns over the whole reference network, derived on the basis of flow values measured by a minimal set of properly located detectors, is essential for achieving this purpose.

Due to the fact that there exist a number of different pollutants and some of them are correlated, the problem must be faced of which pollutants can be assumed as proper indicators of the pollution level produced by traffic in the reference area. Field experiments carried out by National Health Institute monitoring station in Rome indicate that CO has to be considered the best traffic-generated pollution level indicator, since the distribution of its values matches very well within-day traffic variations and, in particular, highest concentration levels are associated with peak-time urban transport demand, which generally causes congested situations to occur. Furthermore, since CO standards are established with respect to its 8-hour and 1-hour mean concentration values, reliable predictions are allowed in the very-short time.

Pollution monitoring is of primary importance in order to achieve the following objectives (CEC, 1993; Mizoguchi, 1993):

a) building databases that allow to define proper relationships between traffic conditions and environmental pollution;

b) controlling traffic so as to maintain pollution levels below acceptable thresholds;

c) informing drivers of the implications of their decisions and encouraging them to modify their long term behaviour, including mode choice.

Environmental Impact Assessment (EIA) is defined as an effective procedure to prevent and/or mitigate the harmful effects of human activities on the environment. It allows the parallel evaluation of alternative infrastructural or technological projects, in view of their optimal integration in comprehensive strategic development plans for the region of interest (Bresso, Russo and Zeppetella, 1988).

R.G.s working on PFT 2 are engaged in a wide range of studies on this topic, including a critical examination of current directions and regulations, common evaluation tools and case-study analyses, that is expected to bring out their inherent shortcomings and, consequently, to yield significantly improved methods and techniques. However, we have to point out that the above

mentioned researches tend to focus on one environmental resource at a time, in spite of the systemic and interdisciplinary nature of EIA.

Problems arising from the inherently subjective nature of certain steps of the evaluation process are adequately investigated. These steps include the standardization and normalization of environmental indicator values, in view of their subsequent aggregation in overall synthetic indices of environmental impact, and the identification of a weighting system that could properly reflect the relevance of the different environmental components affected by the considered activities.

A wide range of interventions could be evaluated in the framework of the EIA procedure if this is applied at the upper levels of the decision-making process relative to land use planning and management; actually, it becomes possible to exploit fully its considerable potentialities, that allow not only to assess the effects of technological or dimensional alterations to previously proposed projects, but also to look for their best site location or even to discard them and select different strategies for the development of the considered region.

EIA shows several interesting peculiarities, that make possible to overcome some of the main theoretical limitations of the most common investment appraisal economic technique, that is, cost/benefit analysis. In particular, they could be listed as follows (Bresso, Russo and Zeppetella, 1988):

-definition of a comprehensive and heterogeneous system of indicators that could exhaustively represent the whole set of investigated phenomena, thus including incommensurable effects that cannot be evaluated in monetary units;

-identification of the system of priorities and preferences of the affected population, to be adequately taken into account in the decision-making process relative to the realization of a certain project, in such a way that the process itself meets some clarity and transparence requirements;

-evaluation of long-term impacts of proposed projects and plans, in order to assess current decision inheritances to future generations.

EIA consists of a series of steps requiring appropriate techniques of analysis. Impact evaluation is based on a comparison between the anticipated environmental system dynamics in the case in which the considered project is realized or not. Static methods (checklists, matrices, mappings) or dynamic methods (coaxial matrices, networks) may be used in the impact identification step. Then, when the different classes of impacts have to be weighed in order to obtain an overall synthetic index, it is possible to resort to one of the typical multicriteria analysis quantitative methods, such as dominance analysis, concordance analysis, conjunctive or disjunctive methods, lexicographic methods, linear assignment or simple additive weighting methods (Bresso, Russo and Zeppetella, 1988).

Due to some lack of information or to the fact that some attributes related to the ranking-criteria for the assessed project alternatives are measurable only on an ordinal scale, it is not always possible to assign a precise quantitative value to the attributes themselves. Hence, Rostirolla (1993) has proposed an approach to environmental decision-making problems that is capable of dealing with fuzzy data sets, by introducing transformation curves that convert them in "certain" numbers; obviously, the need exists to standardize the conversion procedure adopted in the defined approach, in order to make possible the comparison among results of different studies.

Although the impacts of infrastructural and technological projects generally involve the whole mass of environmental resources of a certain area, R.G.s working on PFT 2 tend to focus their attention on particularly sensitive components, identified according to the nature of the assessed projects and the affected areas; for instance, in urban and metropolitan areas these components are clearly traffic, population and socio-economic activities, air and noise.

A DSS for evaluating the environmental impact (mostly air and, to a lesser extent, noise pollution) of new infrastructural projects in urban transportation systems has been developed by Busillo, Passaro and Martini (1993). Its functional components include:

a) a database of current directions and regulations on environmental matters;

b) a geographical information system (GIS) handling spatially referenced socio-economic and environmental data and producing cartographies of simple understanding and manipulation;

c) an environmental monitoring system collecting data through a properly located set of sensors;

d) a modelling system which comprises combined modules for the assessment of emission and dispersion phenomena (as discussed in section 2), that are calibrated on the basis of the input data provided by the subsystems listed at the preceding points, including road network topology and structure.

In particular, the air pollutant concentration distribution is determined by using the USA-EPA software HIWAY 2, which is suited particularly to short-term predictions giving the amount of inhert pollutants (such as SO_2 and NO_X) that affect surrounding areas to high-speed urban arterials.

It has to be noted that spatial scale is both a direct consequence of the nature and significance of the investigated phenomena and a relevant factor for the choice of the proper methodological approach for their study. For example, there exist different classes of models appropriate to the different possible extensions of the region of interest for the dispersion phenomenon. Hence, we have to distinguish local scale models (where emphasis is given to dispersion within a region of a few tens of kilometres), mesoscale models (within hundreds of kilometres) and macroscale models (as far as thousands of kilometres). PFT2-R.G.s' main concern is with local scale models, that include microscale models operating within a distance of a few hundreds of metres and allowing the evaluation of the effects of traffic control and management actions adopted at the intersection level.

4. CONCLUSIONS

Scientific research plays a fundamental role in the solution process of the serious problems associated with the transport system impacts on the environment. In the expected scenario of a considerable increase of mobility and traffic congestion, it is essential to identify proper tools and methodologies aiming at the restriction both of non-renewable resource consumption and pollutant emissions by the transport sector, in order to preserve the carrying capacity of the environmental system and keep within the bounds imposed by a sustainable development.

In light of this objective, international research projects tend to identify standard strategies, actions and tools aiming at operational pollution reduction, transport demand growth and congestion control, available roadspace use optimization and network safety improvement (with particular attention to hazardous freight transport risk of accidents). Since the technological progress contribution is not enough by itself to protect the environment, it must be supported by the adoption of a comprehensive strategy including policy and economy measures directed towards transport demand modal and space-time restriction and redistribution and/or transport supply improvement.

According to the analysis carried out in this paper, it is possible to state that there is still much to be investigated in the field of transport and environment interactions. In this framework, there is an urgent need for software programs that may take into account the potentialities of new information technologies and the new socio-economical requirements, such as energetic and environmental needs. In fact, theoretical and empirical models have been developed both to include route guidance and information system effects in the optimal assignment procedures for estimating travelers' behaviour and to evaluate changes in pollution levels with respect to alternative driving cycles on the basis of

microscopic recording of time and speed in floating car measurements (see, e.g.: Bell et al., 1992; Ben-Akiva, Koutsopoulos and Mukundan, 1992; Mahmassani, Hu and Jayakrishnan, 1992; Bruno and Improta, 1992). At present, however, no satisfactory models have been produced for general use, due to the great complexity of the considered problems.

In short, researches developed in the framework of Finalized Project on Transportation by CNR of Italy are expected to achieve the following results:

-mathematical models and software for air and noise pollution prediction;

-simulation models for traffic situation, meteorological condition and pollutant concentration correlation analysis;

-implementation of an advanced multifunctional traffic regulating system, capable of reducing the harmful environmental effects of road traffic by the identification of optimal management and control strategies;

-methods and models for determining the optimal number and sites of traffic and pollution monitoring detectors;

-realization of a database relative to the pollutant emission levels associated with different classes of vehicles, to be carried out by laboratory measurements and field experiments at pilot test-areas;

-mathematical models and software for optimal service and facility location, according to integrated urban planning intervention strategies;

-realization and implementation of a DSS for transport/environment planning at different spatial scales;

-methods, models and software for the EIA of significant projects.

These contributions are intended to provide an adequate theoretical and metodological basis that can be exploited effectively in applications to real-world problems. For this purpose, PFT 2 research activities include field experiments to be carried out in order to validate the developed models and compare them with those built in the framework of the most relevant research projects currently undertaken all over the world.

REFERENCES

Anderson M. , Taylor M.A.P. : " Estimating the environmental impacts of road traffic at the local level : a PC-based modelling system ", in : Newton P.W. , Taylor M.A.P. , Sharpe R. , Desktop planning : microcomputer applications for infrastructure and services planning and management , Hargreen , 1988.

Bell M.G.H. , Busch F. , Shield C.M. , Anderson J.M. : " Assignment in the integration of urban traffic control and dynamic route guidance ", Proceedings of the Second International Seminar on 'Urban Traffic Networks', Capri, Italy, 1992.

Ben-Akiva M. , Koutsopoulos H.N. , Mukundan A. : " A dynamic traffic model system ", Proceedings of the Second International Seminar on 'Urban Traffic Networks', Capri, Italy, 1992.

Benson P. : "Caline 4- A dispersion model for predicting air pollution concentrations near roadways", Department of Transportation, State of California, USA, 1989.

Bianco L. , Bielli M. : " Pianificazione dei trasporti e ricerca scientifica " , Convegno URBANIA , Padova , 1990.

Bielli M. : " Prospettive della ricerca sul monitoraggio ambientale ed il controllo del traffico ", 47.a Conferenza del traffico e della circolazione, ACI, Stresa, Ottobre 1991.

Bielli M. , Lioi A. , Reverberi P. , Talamo M. : " Minimum number of traffic counting points and their locations for estimating or updating network flows and O/D trip matrices " , Submitted to European Journal of Operational Research, 1994.
Brambilla G. : "Caratterizzazione del rumore da traffico autostradale italiano per lo sviluppo di un modello previsionale", Proceedings of the First National Conference of Finalized Project on Transportation 2, Rome, 1993.
Bresso M. , Russo R. , Zeppetella A. : " Analisi dei progetti e valutazione di impatto ambientale " , F. Angeli , Milano , 1988.
Bruno G. , Improta G. : " Traffic control under pollutant emission constraints ", Proceedings of the Second International Seminar on 'Urban Traffic Networks', Capri, Italy, 1992.
Busillo M. , Passaro M. , Martini P. : " Un DSS per la valutazione dell'impatto ambientale nell'ambito dei sistemi di trasporto urbani ", Proceedings of the First National Conference of Finalized Project on Transportation 2, Rome, 1993.
Carras J.N. , Ross D.G. , Sawford B.L. , Simpson R. : " Air quality modelling " , in : Young P. , Concise Encyclopedia of Environmental Systems , pp. 53 / 58 , Pergamon Press , 1993.
Cenedese A. , Leuzzi G. , Monti P. : "Dispersione di inquinanti in prossimita' di un rilevato stradale", Proceedings of the First National Conference of Finalized Project on Transportation 2, Rome, 1993.
Cirillo M. : "I modelli di simulazione della qualita' dell'aria", in: 'Inquinamento dell'aria e patologia umana', Ed. F. Folini, 1992.
Commission of the European Communities, CEC, : "Integrated Urban Traffic Management", Advanced Transport Telematics Project, Area 3 Report, 1993.
Commissione delle Comunita' Europee, CCE : Libro verde: "L'impatto dei trasporti sull'ambiente", 1992.
DRIVE Project V1005 : " PREDICT : Pollution reduction by information and control techniques " , Working paper , Deliverable n.1 , 1990.
DRIVE Project V1036 : " EVA (RTI)- Evaluation process for road transport informatics " , Working paper 3.4 on 'Operating cost, air and noise pollution' , 1989.
EPA : " Guideline on air quality models ", USA, 1986.
Fraternali D. , Sozzi R. : " Monitoraggio e simulazione a scopo preventivo per la qualita' dell'aria di arterie urbane a forte traffico " , in: Consorzio Milano Ricerche, 'Analisi e controllo del traffico urbano: tecnologie informatiche e metodologie di pianificazione', Consiglio Nazionale delle Ricerche, Progetto Finalizzato Trasporti Due, 1993.
Italtel Telesis : " Sistemi di controllo del traffico basati su misure dei parametri di inquinamento ambientale ", Working Report, Consiglio Nazionale delle Ricerche, Progetto Finalizzato Trasporti Due, 1993.
Hansson L. : " Pricing of air pollution in the Swedish transport policy " , Transportation Research Record 1312 on " Energy and environmental issues " , TRB , USA , 1991.
Himanen V. , Nijkamp P. , Padjen J. : " Environmental quality and transport policy in Europe", Transportation Research A , vol. 26 A , n. 2 , pp. 147 / 157 , 1992.
Immers B.H. , Oosterbaan N.G.J. : " Model calculation of environment-friendly traffic flows in urban networks " , Transportation Research Record 1312 on " Energy and environmental issues " , TRB , USA , 1991.
Jakeman A.J. , Taylor J.A. : " Air pollution : uncertainty of model predictions " , in : Young P. , Concise Encyclopedia of Environmental Systems , pp. 49 / 53 , Pergamon Press , 1993.
Kainuma,Y., Shiozawa,K., and Okamoto,S. : "Study on the optimal location of ambient air monitoring stations", Atmospheric Environment, vol. 24B, n. 3, pp. 395 / 406, 1990.

Mahmassani H. , Hu Y. , Jayakrishnan R. : " Dynamic traffic assignment and simulation for advanced network informatics ", Proceedings of the Second International Seminar on 'Urban Traffic Networks', Capri, Italy, 1992.
Matzoros A. , Van Vliet D. : " A model of air pollution from road traffic , based on the characteristics of interrupted flow and junction control " , Transportation Research A , vol. 26 A , n. 4, pp. 315 / 355 (parts I and II) , 1992.
Meyer M.D. , Miller E.J. : "Urban transportation planning", McGraw-Hill, 1984.
Mizoguchi T. : " Air pollution : monitoring systems " , in : Young P. , Concise Encyclopedia of Environmental Systems , pp. 46 / 48 , Pergamon Press , 1993.
Nishioka S. : " Traffic pollution : control policy and research trends " , Transportation Research A , vol. 23 A , n.1 , pp. 73 / 81 , 1989.
Nuzzolo A. , Festa, D. , Nicotina, E. : "Simulazione dell'inquinamento da traffico nelle aree urbane", Sistemi di Trasporto, n.4 , pp.15 / 24 , 1993.
Nuzzolo A. : "Inquinamento da traffico ed ambiente urbano", Convegno URBANIA, Padova, 1990.
Organisation for Economic Cooperation and Development, OECD : "The social costs of land transport", Environment Directorate, Paris, 1989.
Rostirolla P. : "Un approccio multicriteriale per problemi decisionali con dati fuzzy", Proceedings of the First National Conference of Finalized Project on Transportation 2, Rome, 1993.
Sawford B.L. : "Lagrangian statistical simulation of concentration mean and fluctuation fields", Journal of Climatics and Applied Meteorology, vol. 24, pp. 1152 / 1166, 1985.
Shindo J. , Oi K. , Matsumoto S. : "Considerations on air pollution monitoring network design in the light of spatio-temporal variations of data", Atmospheric Environment, vol. 24B, n. 2, pp. 335 / 342, 1990.
Shladover S. : " Potential contributions of Intelligent Vehicle/Highway Systems to reducing transportation's greenhouse gas production " , Transportation Research A , vol. 27 A , n. 3 , pp. 207 / 216, 1993.
Simmon P.B. , Patterson R.M. , Ludwig P.L. , Jones L.B. : " Emissions and diffusion modelling ", User's Manual for the APRAC-3/ Mobile 1, EPA, USA, 1981.
Yamartino R.J. , Wiegand G. : " Development and evaluation of simple models for the flow, turbulence and pollutant concentration fields within an urban street canyon ", Atmospheric Environment, vol. 20, n. 11, 1986.

INDEX OF AUTHORS

Adamski A., 917
Ali A.I., 645
Anily S., 859
Asakura Y., 225
Assad A., 689
Astarita V., 945
Avella P., 461
Averbakh I., 639

Babic O., 693
Badra H.O., 481
Ball M., 689, 939
Barceló J., 17, 239, 373
Barnhart C., 73
Belenky A.S., 393
Bell M.G.H., 109, 227
Bellei G., 387
Ben-Akiva M., 19, 577, 579
Berka S., 109
Berman O., 639
Bernstein D. II , 23
Bertocchi S., 957
Bertsimas D., 661
Bianco L., 591, 747
Bielli M., 387, 407, 963
Boyce D. E., 119
Braca J., 877
Bramel J., 201, 877
Brcheret L., 109
Brotcorne L., 365

Campbell J.F., 881
Cangalovic M., 693
Cantarella G.E., 495, 839
Carotenuto P., 407
Carraresi P., 43
Carrese S., 623
Carvalho T.A., 185
Cascetta E., 9, 495, 577
Catanas F., 533
Ceder A., 423, 781
Ceria S., 547
Cerny J., 935
Chabini I., 475

Chen Y., 887
Christofides N., 481, 643
Codina E., 373
Cohen S., 549
Cole M.H., 151
Crainic T.G., 825, 833

de Cea J., 517
De Wolfe D., 365
Dell'Olmo P., 591, 805
de Palma A., 579
Desrosiers J., 361
Di Febbraro A., 561, 603
Di Gangi M., 495
Di Noto D. , 561
Dogan K., 151
Dumas Y., 361

Elloumi N., 135
Ezhkova I., 591

Farkas A., 235
Farvolden J. M., 825
Federgruen A., 219
Fernández J.E., 517
Ferrari P., 305
Fischetti M., 539
Florian M., 7, 475, 887
Freling R., 319
Friedrich B., 819
Friesz T.L., 23

Gastaldi M., 407
Gattuso D., 945
Gendreau M., 33, 61, 213, 711, 825
Gendron B., 833
Gittoes D.H., 335
Godfrey G.A., 185
Goetschalckx M., 151
Gori S., 623
Grandinetti L., 465
Guerriero F., 465
Gun H., 939
Gunn H., 577

Hadjiconstantinou E., 643
Haj-Salem H., 135
Hamacher H.W., 37
Henry J.J., 109
Hertz A., 33, 61
Hohlacov D. J., 693
Hurley W.J., 255, 267

Inaudi D., 899
Israeli Y., 423

Jaillet P., 395, 853

Kaysi I., 579
Keller H., 227
Kim D., 73
Kohl N., 763
Kolokolov A., 179
Koutsopoulos N.H., 235
Kwan R.S.K., 543

Labbe' M., 365
Langevin A., 881
Laporte G., 33, 61, 207, 213
Louveaux F., 207

Madsen O.B.G., 763
Magirou V., 281
Mahmassani H.S., 25, 903
Malucelli F., 43
Maniezzo V., 747
Mannino C., 547
Marcotte P., 891
Martello S., 539
Martín R., 239
Mesquita M., 749
Mifflin T., 689
Mihnea S., 61
Milot L., 617
Mingozzi A., 643, 747
Mirchandani P.B., 805
Mosheiov G., 859
Musmanno R., 465

Nassos G.C., 281
Nellas G.J., 281
Nemhauser G., 151
Nijkamp P., 731
Nobili P., 547
Noon C., 779
Nuzzolo A., 715

Odoni A.R., 661
Osman I.H., 67

Paixão J., 319, 533, 749
Pallottino S., 43
Papageorgiou M., 135
Parisini T., 561
Peeta S., 25
Petersen E.R., 255, 267
Picano T., 623
Postner B., 877
Powell W.B., 3, 185, 335
Psaraftis H.N., 281
Purtov A. M., 401

Ratliff D.H., 13
Recagno V., 561, 603
Reshetnik I., 781
Reverberi P. , 963
Ricciardelli S., 747
Rinaldi G., 823
Russo F., 715

Sachse T., 227
Sacone S., 561, 603
Saidi S., 711
Salhi S., 67
Sanso' B., 617
Sassano A., 547
Schumacher A., 37
Sciomachen A., 957
Seguin R., 213
Semet F., 941
Sforza A., 461
Shaptsev V.A., 401
Sharaiha Y. M., 481
Shield C.M., 109, 227
Shin M.C., 903
Shmerling S., 645
Shulman V.B., 401
Simao H.P., 185
Simchi-Levi D., 201, 877
Solomon M., 361
Song M., 853

Soriano P., 825
Soumis F., 361, 711
Stone L., 689

Ta-Yin H., 25
Thompson T., 689
Tokarev Y.P., 401
Torrieri V., 945
Tosic V., 693
Toth P., 27, 539
Truemper K., 823

van den Bergh J., 731
Van hamme L., 207
van Ryzin G., 219
Velardi V., 899
Vigo D., 27
Vitetta A., 839, 945

Wei R., 151
Wren A., 543
Wu J.H., 497

Yu G., 853

Zang X., 151
Zhu D., 891
Ziliaskopoulos A., 25
Zoppoli R., 561

Finito di stampare giugno 1994 - Tipografia ESAGRAFICA s.r.l.
(00184) Roma - Via della Polveriera, 12/13/13a - Tel. 4885197-4882678