Partitioning of urban networks for MFD applications

S. F. A. Batista[a,*], D.M. Bramich[a], J. Balsa-Barreiro[a] and Mónica Menéndez[a]

[a] Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
sergio.batista@nyu.edu
* Corresponding author

Extended abstract submitted for presentation at the 11th Triennial Symposium on Transportation Analysis conference (TRISTAN XI)
June 19-25, 2022, Mauritius Island

March 18, 2022

Keywords: Network partitioning; Gaussian Mixture models; Nested sampling; Urban networks; Aggregated traffic models.

1 INTRODUCTION

Traffic modeling is a useful tool for developing policies aimed at mitigating congestion problems that many cities face worldwide. Aggregated traffic models based on the Macroscopic Fundamental Diagram (MFD) (Geroliminis & Daganzo, 2008) present promising prospects in this direction. The application of these models requires the partitioning of urban networks, represented as a sequence of directed links, into a set of connected regions where all vehicles travel at similar average speeds. The MFD encapsulates the relationship between the average travel speed and the accumulation of vehicles in each region during a given time interval. MFD-based models mimic the dynamics as exchange flows between adjacent regions.

An important question is how to best partition urban networks into regional networks. On one hand, the partitioned regions should have reasonable sizes (i.e. it is undesirable to have a region with 10 intersections next to one with 1000 intersections), and be topologically fully connected, compact and well separated (i.e. non overlapping regions). On the other hand, traffic conditions within each region should be approximately homogeneous (i.e. congestion should be approximately homogeneous over the region). Several authors have discussed different methodologies for partitioning urban networks for MFD applications. Ji & Geroliminis (2012) were the first to implement a Normalized Cut algorithm to partition urban networks for MFD models. Saeedmanesh & Geroliminis (2016) proposed a model based on the "snake's similarity". The model starts with a single road, and then iteratively adds roads with close similarity values to the "snake". Then, Symmetric Non-negative Matrix Factorization is used for partitioning the urban network. This method ensures fully connected regions, but not necessarily regions that are compact and well separated. Lopez *et al.* (2017) showed that the K-means algorithm results in lower within-cluster variances than the Normalized Cut and DBSCAN methods to partition urban networks, therefore offering better performance in comparison with the two previously introduced methods. Casadei *et al.* (2018) proposed a spatiotemporal algorithm that ensures the consistency of travel times between the urban and regional networks. Ambühl *et al.* (2019) used real traffic data. They did the partition using random walks along the urban network selecting the one that minimizes the scatter on estimating the regions' MFDs. While some of these approaches ensure fully connected regions, they fail in most cases to deliver regions of reasonable sizes, compact and not overlapping.

Recently, Batista *et al.* (2021) utilized Gaussian Mixture Models (GMMs) to partition urban networks by clustering a data set consisting of the Cartesian coordinates of the nodes. GMMs assume that the underlying density distribution of the data can be modeled as a (convex) linear combination of K Gaussian components. In this application of GMMs, each Gaussian component represents a network region. As part of the clustering problem, the optimal number of components K (i.e. regions) needs to be determined. The authors applied the GMMs in two steps to partition the whole metropolitan area of Munich (Germany). First, the authors fit their data with multiple different GMMs covering all possible values of K between 1 and 200. The authors then computed the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for each fit, and selected the value of K so as to minimize either the AIC or BIC. The final partitioning was determined by the GMM fit for the selected value of K. This partitioning approach enables the extraction of regions that respect the appropriate topological features (i.e. in terms of size, compactness, connectivity, and separability) for the MFD applications. However, we have since found that this methodology for selecting the optimal K is confounded by the many symmetries in the likelihood function because of the unidentifiability of the model components (e.g. the "label switching problem" is one type of unidentifiability). Specifically, the regularity conditions that validate the AIC and BIC approximations do not hold for GMMs as they require the model components/parameters to be identifiable (see Chapter 7 in Frühwirth-Schnatter *et al.* (2018)), which explains why Batista *et al.* (2021) found that neither the AIC or BIC achieve a global minimum for a reasonable range of K. This paper proposes an alternative methodology that performs model selection by computing the Bayesian evidence for each value of K, and selects the K that yields the greatest evidence. This naturally penalises models with more components (and parameters). We also present some preliminary results from applying the proposed methodology to partition the network of Innsbruck (Austria).

2 METHODOLOGICAL FRAMEWORK

GMMs fall into the class of unsupervised machine learning models, that can also be used for clustering data. Let us denote our data set of N pairs of Cartesian coordinates, with one pair for each node in the urban network, as $\mathbf{D} \equiv \{(x_j, y_j)\}_{j=1}^{j=N}$. Next, for a GMM $\mathcal{M}_K$ with K components, let $f_k(x, y)$ represent the (density of the) k-th Gaussian component with parameter vector $\boldsymbol{\theta}_k$. Then the probability density function (PDF) of a bivariate random variable G distributed as $\mathcal{M}_K$ can be written as:

$$f_G(x, y) = \sum_{k=1}^{K} \phi_k \, f_k(x, y) \tag{1}$$

where ϕ_k is the weight for the kth component. The weights must satisfy $\phi_k \geq 0$ for all k, and $\sum_{k=1}^{K} \phi_k = 1$. We gather the full set of parameters for $\mathcal{M}_K$ into the single vector $\boldsymbol{\Theta}_K = (\phi_1, ..., \phi_K, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$.

Assuming that the data $\mathbf{D}$ are all independently drawn from the PDF of G, then the likelihood function $\mathcal{L}(\boldsymbol{\Theta}_K)$ can be written as:

$$P(\mathbf{D} \,|\, \boldsymbol{\Theta}_K, \mathcal{M}_K) = \mathcal{L}(\boldsymbol{\Theta}_K) = \prod_{j=1}^{N} \sum_{k=1}^{K} \phi_k \, f_k(x_j, y_j) \tag{2}$$

Via Bayes Theorem, the posterior distribution of $\boldsymbol{\Theta}_K$ may be computed from:

$$P(\boldsymbol{\Theta}_K \,|\, \mathbf{D}, \mathcal{M}_K) = \frac{P(\mathbf{D} \,|\, \boldsymbol{\Theta}_K, \mathcal{M}_K) \; P(\boldsymbol{\Theta}_K \,|\, \mathcal{M}_K)}{P(\mathbf{D} \,|\, \mathcal{M}_K)} = \frac{\mathcal{L}(\boldsymbol{\Theta}_K) \; \Pi(\boldsymbol{\Theta}_K)}{Z_K} \tag{3}$$

where $\Pi(\boldsymbol{\Theta}_K)$ and Z_K are the prior parameter distribution and the Bayesian evidence, respectively. The Bayesian evidence is computed as:

$$Z_K = P(\mathbf{D} \,|\, \mathcal{M}_K) = \int \mathcal{L}(\boldsymbol{\Theta}_K) \; \Pi(\boldsymbol{\Theta}_K) \; \mathrm{d}\boldsymbol{\Theta}_K \tag{4}$$

Finally, again using Bayes theorem, the posterior probability of model $\mathcal{M}_K$ from a set of models $\{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_Q\}$, where $1 \leq K \leq Q$, is given by:

$$P(\mathcal{M}_K \,|\, \mathbf{D}) = \frac{P(\mathbf{D} \,|\, \mathcal{M}_K)\; P(\mathcal{M}_K)}{P(\mathbf{D})} = \frac{Z_K\, \Psi_K}{\sum_{i=1}^{Q} Z_i\, \Psi_i} \tag{5}$$

where $\Psi_i = P(\mathcal{M}_i)$ is the prior probability of the ith model. The optimal value of K then corresponds to the model that maximises $P(\mathcal{M}_K \,|\, \mathbf{D})$.

The integral in Equation 4 can only be performed numerically, and even then it is very difficult. Nested Sampling (Speagle, 2020), unlike other Monte Carlo Markov Chain sampling methods, is able to perform these integrations and compute the Z_K.

For this, we need to specify the model fully and adopt appropriate priors on the model parameters. We model each component of the GMM as a radially symmetric Gaussian profile with free parameters $\boldsymbol{\theta}_k = (x_{\mathrm{c},k}, y_{\mathrm{c},k}, \sigma_k)$. Since we are not using any prior information to guide the network partitioning, we adopt non-informative priors on all parameters. For the "location" parameters $x_{\mathrm{c},k}$ and $y_{\mathrm{c},k}$, we adopt independent Uniform priors, and for the "scale" parameters σ_k, we adopt independent Uniform priors on $\ln \sigma_k$. With a similar motivation, we use a flat Dirichlet prior for the ϕ_k parameters. We are currently investigating the use of sorted priors for the ϕ_k to help remove multi-modality from the posterior (Buscicchio *et al.*, 2019). Finally, we assume a Uniform prior on the Ψ_i (while we are also investigating an exponential prior on the Ψ_i to further promote sparseness in the selected model).

3 PRELIMINARY RESULTS AND DISCUSSION

This section discusses some preliminary results. In particular, we discuss a preliminary validation of the full Bayesian treatment from Section 2 against the GMM fitting method described in Batista *et al.* (2021). For a low number of components, both approaches should provide consistent estimates of the parameters of the Gaussian components (i.e. weights, locations, and sigmas). For this purpose, we set $K = 5$ components and determine the static partition of the urban network representing the city of Innsbruck (Austria). This network was retrieved from OpenStreetMaps (OpenStreetMap contributors, 2020), and consists of 1992 nodes and 4448 links. Table 1 lists estimates of the weights, locations, and sigmas of each of the five Gaussian components, for the GMM fitting method described in Batista *et al.* (2021) and for our proposed Bayesian approach. Figure 1 shows the Innsbruck network partitioned into the 5 components/regions.

As expected, from Table 1, we can observe that both approaches provide similar estimates of the region locations $(x_{\mathrm{c,k}}, y_{\mathrm{c,k}})$, showing their consistency. This happens because for a low number of components of $K = 5$, we have a multi-modality of $5! = 120$[1]. This means that we still have a (relatively) low level of complexity in the topology of the log-likelihood surface, and the Expected-Maximization algorithm used for maximizing the log-likelihood function in the fits described in Batista *et al.* (2021) converges, in general, to the global maximum. However, the number of local maxima increases with at least $K!$. This increases the topological complexity of the log-likelihood function, and the Expected-Maximization algorithm gets stuck more often in sub-optimal solutions. The full Bayesian approach that we propose with Nested Sampling does not suffer from this limitation, since it explores the full posterior probability surface. We will discuss these advantages of our modelling procedure in the full paper. The partitioning in Figure 1 depicts that the subnetwork representing region 4 is not fully connected. In the full paper, we will also discuss how geographical features of the network (e.g. the presence of rivers which can act as natural borders) should be utilized as a first partitioning level before applying our proposed model. In this step, we can adopt for example priors based on those features (e.g. geographical features).

[1]The notation "!" refers to the factorial of 5, i.e. $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Table 1 – *Parameter estimates determined for each Gaussian component, using the GMM fitting method described in Batista et al. (2021) and the full Bayesian treatment.*

Region	GMM (Batista *et al.*, 2021)				GMM Bayesian			
	ϕ_k	$x_{\mathrm{c,k}}$	$y_{\mathrm{c,k}}$	σ_k	ϕ_k	$x_{\mathrm{c,k}}$	$y_{\mathrm{c,k}}$	σ_k
1	0.242	683222	5237910	956	0.114	684005	5238256	1084
2	0.170	678967	5236958	814	0.179	679007	5236953	821
3	0.335	681464	5237387	969	0.397	681695	5237427	988
4	0.139	685247	5239092	925	0.262	685728	5239891	415
5	0.114	675817	5236282	624	0.049	675818	5236278	623

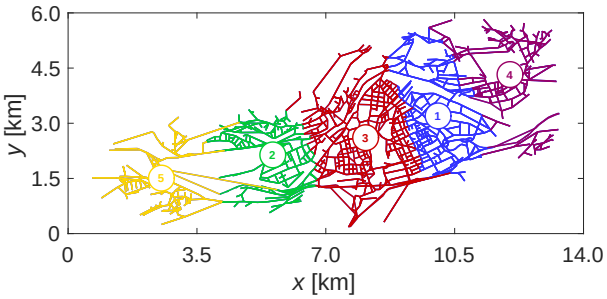

Figure 1 – *Innsbruck (Austria) city network partitioned into 5 regions.*

ACKNOWLEDGEMENTS

The authors acknowledge support by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001.

References

Ambühl, L., Loder, A., Zheng, N., Axhausen, K. W., & Menendez, M. 2019. Approximative network partitioning for MFDs from stationary sensor data. *Transportation Research Record*.

Batista, Sérgio F. A., Lopez, Clélia, & Menéndez, Mónica. 2021. On the partitioning of urban networks for MFD-based applications using Gaussian Mixture Models. *Pages 1–6 of: 2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*.

Buscicchio, R., Roebber, E., Goldstein, J.M., & Moore, C.J. 2019. Label switching problem in Bayesian analysis for gravitational wave astronomy. *Physical Review D*, **100**, 084041.

Casadei, G., Bertrand, V., Gouin, B., & Canudas-de-Wit, C. 2018. Aggregation and travel time calculation over large scale traffic networks: An empiric study on the Grenoble City. *Transportation Research Part C: Emerging Technologies*, **95**, 713–730.

Frühwirth-Schnatter, S., Celeux, G., & Robert, C.P. 2018. *Handbook of Mixture Analysis*. Chapman and Hall/CRC.

Geroliminis, N., & Daganzo, C. 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, **42**, 759–770.

Ji, Y., & Geroliminis, N. 2012. On the spatial partitioning of urban transportation networks. *Transportation Research Part B: Methodological*, **46**, 1639–1656.

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., & van Lint, H. 2017. Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. *Scientific Reports*, **7**, 1–11.

OpenStreetMap contributors. 2020. *Innsbruck dump retrieved from https://planet.osm.org* .

Saeedmanesh, M., & Geroliminis, N. 2016. Clustering of heterogeneous networks with directional flows based on "Snake" similarities. *Transportation Research Part B: Methodological*, **91**, 250–269.

Speagle, J.S. 2020. DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *MNRAS*, 3132–3158.