A "sums and shares" mixture model to study pedestrian flows within a multimodal transport hub

P. De Nailly[a,b,*], E. Côme[a], L. Oukhellou[a], A. Samé[a] and J. Ferriere[b]

[a] Université Gustave Eiffel, Marne La Vallée, France
etienne.come@univ-eiffel.fr, allou.same@univ-eiffel.fr, latifa.oukhellou@univ-eiffel.fr
[b] Régie Autonome des Transports Parisiens, Paris, France
paul.denailly@ratp.fr, jacques.ferriere@ratp.fr
[*] Corresponding author

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

January 13, 2022

Keywords: Mixture model, Sums and shares models, Multimodal transport hub, Mobility data

1 INTRODUCTION

This work focuses on a multimodal transportation hub that attracts an average of 140,000 people on a typical week day (2019). Passenger flows tend to form transportation routes towards areas of interest. However, depending on the time of day, the period of the year, or local events, these flows do not necessarily go to or transit through the same places. Several sensing systems positioned at different count locations in the transport hub allow multivariate count data to be collected. Extracting information from these large sets of highly noisy count series is challenging since these series share common dynamics in response to specific events, but may also have their own dynamics due to more localized events.

Our goal in this paper is to model these mobility multivariate time series through a set of mobility patterns, to obtain a synthetic vision which helps understanding how the transport hub operates, or to use as a basis for prediction work. Because of the non-stationary nature of the data, the modeling is subdivided into several time segments (Truong *et al.* (2020)). At the same time, covariates add distinctive features to these patterns (Zhong *et al.* (2015)). Overdispersion and correlations (Winkelmann (2008)) between series are characteristics frequently encountered with count data. To take these phenomena into account, we drew on the "Negative binomial sums and Pólya shares" strategy presented in Jones & Marchand (2019). With this approach, periods are considered homogeneous if, conditional on covariates and segments, both the totals (i.e., "sums") of people observed in the transport hub and their spatial distribution (i.e., "shares") among several locations are similar. Combining multiple mobility data sources is a valuable addition to this kind of study, as it enriches the data available on observed flows in the multimodal transport hub. To this end, this study relies on counts obtained from both ticketing logs collected by automated fare collection systems and stereo camera sensors.

2 SEGMENTATION MODEL STRUCTURE

The writing of the segmentation model first go through the elicitation of the regression model before adapting it into a mixture model, able to detect regime changes in the time series.

2.1 Negative binomial sums and Pólya shares regression model

We consider hereafter that Y is a L-vector of counts across L locations : $Y = (Y_l)_{l\in(1,..,L)}$. To introduce it, let $V = \sum_l Y_l$ be the sum of these counts; the proposed strategy first models V as a Negative binomial distribution. Then conditionally on $V = v$, the count vector Y follows a Dirichlet-multinomial (i.e. Pólya) distribution. This model can take into account overdispersion as well as correlations, but these will always be either positive or negative (Jones & Marchand (2019)). The model may be written as follows:

$$V|\mathbf{x} \sim \mathcal{NB}(\exp(\mathbf{x}^T\boldsymbol{\gamma}), r)$$
$$Y|\mathbf{x}, v \sim \mathcal{DM}(v, (\exp(\mathbf{x}^T\boldsymbol{\xi}_l))_{l\in 1,...L})$$

where $\mathcal{NB}$ and $\mathcal{DM}$ as the Negative Binomial and Dirichlet Multinomial distributions, respectively. We note $\mathbf{x}$ as a $D \times 1$ vector of D exogenous factors. r is the scale parameter of the $\mathcal{NB}$ distribution. $\boldsymbol{\gamma}$ is the vector $(D \times 1)$ of regression parameters of Negative binomial regression, and $\boldsymbol{\xi_l}$ the vector $(D \times 1)$ of Dirichlet multinomial regression coefficients.

2.2 Negative binomial sums and Pólya shares mixture model

Let us now consider $Y_{j,t}$ as the L-vector of counts for time slot t of day j. We assume that, on day j, the counts at each time slot t can be associated with the dynamics of one segment among S possible segments, with a certain probability. If we adapt the Negative binomial sums and Pólya shares regression model to a mixture model framework, we end up with a generative model which includes a set of indicator variables, denoted by Z_j ($Z_j \in \{0,1\}^s$) and encoding the segment membership of the days, with $s \in \{1, ..., S\}$. The number of segments S is chosen a priori. The following generative model is assumed for the observed data:

$$Z_j|j \sim \mathcal{M}(1, (\pi_s(j;\alpha))_{s\in 1,...,S})$$
$$Y_{j,t}|Z_j = s, \mathbf{x}_{j,t}, \boldsymbol{\zeta}_s \sim \mathcal{D}(Y_{j,t}|\mathbf{x}_{j,t}, \boldsymbol{\zeta}_s)$$

with

$$\pi_s(j;\alpha) = \frac{exp(\sum_{m=1}^{M+4}\alpha_{s,m}a_m(j))}{\sum_h exp(\sum_{m=1}^{M+4}\alpha_{h,m}a_m(j))}$$

with $\alpha_{s,m}$ a weight to be estimated and

$$a_m(j) = j^{m-1}, m \in \{1, ..., 4\}$$
$$a_{m+4}(j) = (j - \kappa_m)^3, m \in \{1, ..., M\}$$

Here $\mathcal{D}$ is a Negative binomial sums and Pólya shares regression model with $\boldsymbol{\zeta}_s = (r_s, \boldsymbol{\gamma}_s, \boldsymbol{\xi}_s)$ the set of parameters controlling the conditional distributions within segment s. The variable Z_j follows a multinomial distribution of parameter $\boldsymbol{\pi}$ (i.e the vector of association weights). The association weights follow a logistic transformation of cubic spline functions with M nodes and therefore change according to day j. The idea behind this scheme is to help the model detect regime changes that are otherwise difficult to detect from the data. Each cubic spline is a piecewise cubic polynomial with knots at κ_m, $m \in \{1, ..., M\}$ ($M = 25$ knots here). The parameters of the model are estimated with the Expectation Maximization (EM) algorithm (Dempster *et al.* (1977)). The model was built in the R environment using the glm.nb function in the MASS (Ripley *et al.* (2013)) package and the MGLM (Kim *et al.* (2018)) package.

3 EXPERIMENTATION ON REAL-WORLD DATA

The present analysis is based on hourly aggregated count data from sensors and ticketing systems provided by the Paris public transport operator (i.e. Régie Autonome des Transports Parisiens or RATP) between April 2019 and September 2020. This study focus strictly on working days between 7 am and 1 am (the following day), to avoid introducing variables for non-working days, and thus simplify the model. Exogenous factors integrate the $\mathbf{x}_{j,t}$ vector and include the 1-hour time slots. We will focus in the following section on the results associated with the model with $S = 10$ segments according to the BIC criterion.

The first result that can be obtained with this model is the period segmentation, as shown in Figure 1. We can observe a richness of segments induced by various context changes such as maintenance work, strikes, or health measures against the Covid19 pandemic. This diversity of segments, with few returns, underlines the need for urban operators to adapt to a regularly changing situation. This result highlights that, at time of writing, the regular use of the hub (i.e. "Normal 2019" and "Closure of an exit" in Figure 1) has not returned since the beginning of the Covid19 pandemic.

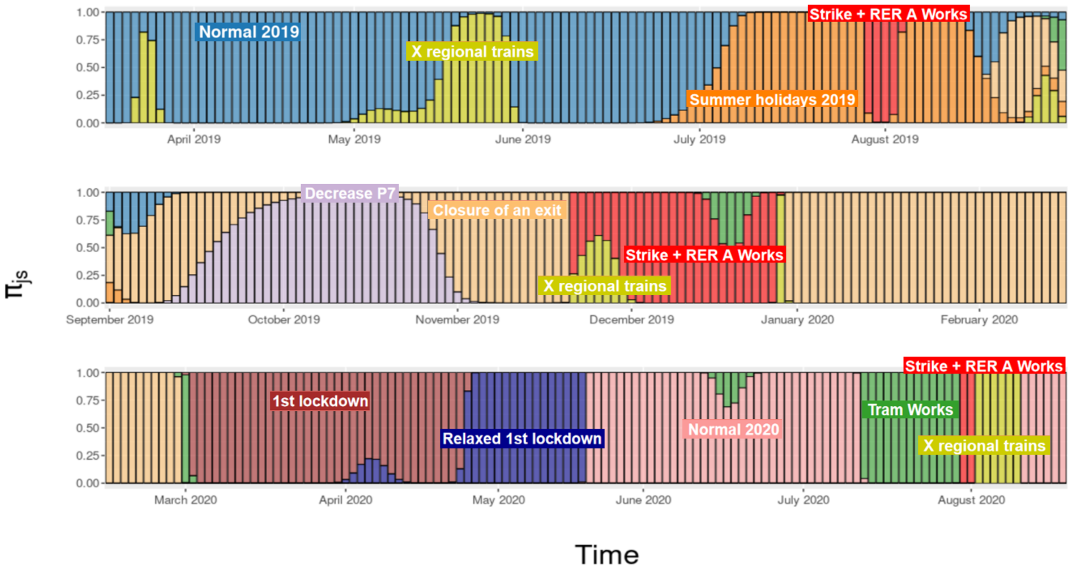

Figure 1 – *Bar plot representation of time segmentation. Each day is associated with the probabilities of belonging to each segment. All 10 segments each have their own colors and labels.*

From a spatial point of view, one can study the characteristics of people flow distribution in the transport hub within each segment. Indeed, each segment s is associated with a set of typical distributions between the L locations, as displayed in Figure 2. The location codename is labeled "O" for outgoing flows and "I" for incoming flows. Depending on the segments encountered, there is an overuse (in red) or an underuse (in blue) at certain locations compared to "Normal 2019". For example, in segment "Strike + RER A Works" there is an overuse of accesses to and from the metro and an underuse of accesses to and from the RER, highlighting an expected transfer of users to the metro line when the RER traffic is disrupted.

4 CONCLUSION

This paper sets up a model to segment multidimensional mobility time series, whose dynamics evolve according to characterized periods. This strategy was based on the "sums and shares" models found in literature. We chose to apply a Negative binomial and Pólya shares mixture model to analyze mobility data collected at a major transport hub. The regression coefficients of these models are dependent on the segments to which they belong. This work reveals how various restrictions to combat the Covid19 pandemic significantly affected pedestrian flow dynamics in the transportation hub. These restrictions were not the only events that impacted hub patronage over the long term. We found that given situations may lead to specific over- and under-use of particular locations. This type of study can be applied to any situation where a large set of count

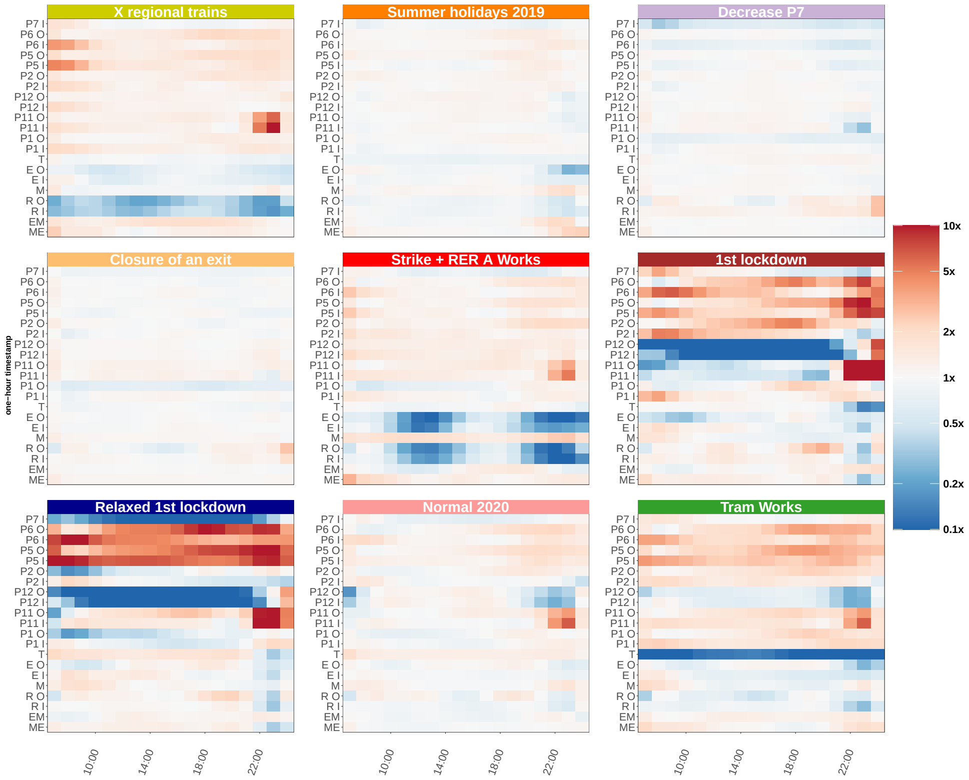

Figure 2 – *Typical spatial distributions between the L (=21) locations found in each segment s. Each cell of the heat maps corresponds to a time slot at a given location. For a given cell, the color reflects the log ratio between the proportion of flows in the current segment and that in the Normal 2019 reference segment. Colors thus reflect the differences between the proportions of flows in each segment and those in the reference segment, with regards to spatial distribution.*

data is available and there is an interest in synthesizing information from typical spatio-temporal profiles from distinct periods. In the field of mobility still, these models can be applied to the study of a public transport network or a whole city, in which the characterization of human travel patterns is of great importance. This study leads the way to more advanced clustering or prediction modeling work. In particular, it allows periods with variable flow dynamics to be distinguished, which can be helpful when predicting patronage in specific contexts.

References

Dempster, Arthur P, Laird, Nan M, & Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Jones, MC, & Marchand, Éric. 2019. Multivariate discrete distributions via sums and shares. *Journal of Multivariate Analysis*, **171**, 83–93.

Kim, Juhyun, Zhang, Yiwen, Day, Joshua, & Zhou, Hua. 2018. MGLM: an R package for multivariate categorical data analysis. *The R journal*, **10**(1), 73.

Ripley, Brian, Venables, Bill, Bates, Douglas M, Hornik, Kurt, Gebhardt, Albrecht, Firth, David, & Ripley, Maintainer Brian. 2013. Package 'mass'. *Cran r*, **538**, 113–120.

Truong, Charles, Oudre, Laurent, & Vayatis, Nicolas. 2020. Selective review of offline change point detection methods. *Signal Processing*, **167**, 107299.

Winkelmann, Rainer. 2008. *Econometric analysis of count data*. Springer Science & Business Media.

Zhong, Chen, Manley, Ed, Arisona, Stefan Mueller, Batty, Michael, & Schmitt, Gerhard. 2015. Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science*, **9**, 125–130.