



# Multivariate count time series segmentation with “sums and shares” and Poisson lognormal mixture models: a comparative study using pedestrian flows within a multimodal transport hub

Paul de Nailly<sup>1,2</sup> · Etienne Côme<sup>1</sup> · Latifa Oukhellou<sup>1</sup> · Allou Samé<sup>1</sup> · Jacques Ferriere<sup>2</sup> · Yasmine Merad-Boudia<sup>2</sup>

Received: 1 August 2022 / Revised: 22 March 2023 / Accepted: 24 April 2023 /  
Published online: 29 May 2023  
© Springer-Verlag GmbH Germany, part of Springer Nature 2023

## Abstract

This paper deals with a clustering approach based on mixture models to analyze multidimensional mobility count time-series data within a multimodal transport hub. These time series are very likely to evolve depending on various periods characterized by strikes, maintenance works, or health measures against the Covid19 pandemic. In addition, exogenous one-off factors, such as concerts and transport disruptions, can also impact mobility. Our approach flexibly detects time segments within which the very noisy count data is synthesized into regular spatio-temporal mobility profiles. At the upper level of the modeling, evolving mixing weights are designed to detect segments properly. At the lower level, segment-specific count regression models take into account correlations between series and overdispersion as well as the impact of exogenous factors. For this purpose, we set up and compare two promising strategies

---

✉ Paul de Nailly  
paul.de-nailly@ratp.fr; paul.denailly@laposte.net

Etienne Côme  
etienne.come@univ-eiffel.fr

Latifa Oukhellou  
latifa.oukhellou@univ-eiffel.fr

Allou Samé  
allou-badara.same@univ-eiffel.fr

Jacques Ferriere  
jacques.ferriere@ratp.fr

Yasmine Merad-Boudia  
yasmine.merad-boudia@ratp.fr

<sup>1</sup> Cosys-Grettia, Université Gustave Eiffel, Descartes Boulevard Newton, 77454 Champs-sur-Marne, France

<sup>2</sup> EDT, RATP, 54 Quai de la Rapée, 75012 Paris, France

that can address this issue, namely the “sums and shares” and “Poisson log-normal” models. The proposed methodologies are applied to actual data collected within a multimodal transport hub in the Paris region. Ticketing logs and pedestrian counts provided by stereo cameras are considered here. Experiments are carried out to show the ability of the statistical models to highlight mobility patterns within the transport hub. One model is chosen based on its ability to detect the most continuous segments possible while fitting the count time series well. An in-depth analysis of the time segmentation, mobility patterns, and impact of exogenous factors obtained with the chosen model is finally performed.

**Keywords** Mixture model · Sums and shares models · Poisson log-normal · EM algorithm · Multimodal transport hub · Mobility data

**Mathematical Subject Classification** 62H10 · 62H30

## 1 Introduction

Multivariate count data are increasingly collected in the form of time series with the help of various sensing systems. These time series are found in many domains such as economics, meteorology, transport or bioinformatics. In the field of mobility, these series are often presented as counts of people acquired at several points in a city (Fernández-Ares et al. 2017), a transport network (Mützel and Scheiner 2021) or a multimodal hub (de Nailly et al. 2021). In a context where transportation areas are very congested due to increased use of public transport, crowd movement analysis and evacuation planning in these areas will become more and more of a topic of interest. Thus, statistical analysis of human flows has become an active research topic that can provide valuable information to urban and transport planners. These research topics improve the ability to manage and predict crowd movement. As described in detail in Cecaj et al. (2021) many applications exist to model crowds. In mobility planning, Balzotti et al. (2018) identified the pedestrian movements that most likely change the crowd density. In a similar way, event management can be helped with these models. Singh et al. (2020) use Wifi-based crowd counting to forecast density evolutions as a consequence of an event. This type of work can also be used as an aid to crowd safety management. Finally, epidemiology is another possible application. Ronchi et al. (2020) present a methodology to use crowd modeling as an aid to assess safety in confined and open spaces. The present study focuses on a multimodal transportation hub, i.e. a place designed to connect multiple modes of travel and through which large numbers of people may move. Several sensing systems positioned at different count locations in the transport hub allow multivariate count data to be collected. 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. Three elements should be taken into account when modeling people count data:

- The mobility of people is impacted by various factors that can be calendar (e.g. time and type of day, public holidays) or non-calendar, such as concerts or transport disruptions (Toqué et al. 2018; Briand et al. 2019).
- Mobility data are subject to long-time effects such as trends, seasonal effects or effects related to exceptional events such as the Covid19 pandemic (de Nailly et al. 2021). These series thus present a non-stationary aspect.
- Potential dependencies may exist between the different count locations, which is challenging to model for count data, as suggested by Singh et al. (2020).

Extracting information from a large set of highly noisy count series is a difficult task. These series share common dynamics in response to certain events, but may also have their own dynamics for more localized events. Our goal is to create a model of these multivariate count time series using a set of covariates, which is useful for understanding the operations / use of the transport hub or as a basis for prediction work. Because of their non-stationary nature, the time series must be subdivided into several segments (Truong et al. 2020). One can look at this work by considering time segmentation on multivariate count data as a way to capture regularity through mobility patterns that can be easily interpreted, while covariates add distinguished features to these patterns (Zhong et al. 2015). However, we do not use any autoregression term in this work, we rather focus on taking into account covariates which impact the counts at each time slot. Mixture models are a promising strategy for this purpose, as they allow the modeling to be subdivided into segments within which both simple and complicated distributional forms can be used for the observed data (Magidson and Vermunt 2002). Overdispersion and correlations between series are characteristics frequently encountered with count data (Winkelmann 2008). In order to take these phenomena into account, we drew on two strategies, found in the literature, capable of modeling numerous, noisy and possibly correlated count data. These two strategies, namely “sums and shares” (Jones and Marchand 2019) and Poisson log-normal models (Chiquet et al. 2021), tackle the modeling of count data according to distinct philosophies. In the first approach, periods are considered homogeneous if, conditional on covariates and segments, the totals (i.e. “sums”) of people observed in the transport hub and their distribution (i.e. “shares”) among several locations are similar. In the second one, periods are considered as homogeneous if, conditional on covariates and segments, the series have similar mean counts and interact in the same way. Both strategies are discussed in the Sect. 4. Our work is thus positioned as the search for a model that can both take into account numerous, dispersed and correlated data, and gather the most continuous periods possible.

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 two sources of people count data: (i) ticketing logs collected by automated fare collection systems quantify the number of trips to, from and between the different transport lines in the hub, (ii) stereo camera sensors count all incoming and outgoing flows of people between the transport hub and various places of work, leisure or commerce in the vicinity.

- We aim to synthesize dynamic count data into regular spatio-temporal mobility profiles. For this purpose we propose methods to flexibly detect time segments in

multivariate count series. We build “smooth” mixture models, whereby we model the transition between segments using logistic functions as mixing weights. These logistic functions integrate spline functions to ensure a temporal regularity to the segmentation, especially between proximal days. At the upper level evolving mixing weights are designed to help detecting segments. At the lower level, segment specific count data regression models handle passenger flows dynamics within each segment.

- We use “sums and shares” and Poisson log-normal regression models within the segmentation methods, well suited to handle overdispersed, correlated, multivariate count data. To the best of our knowledge, no study has been conducted to compare the two presented models.
- We conduct an in-depth analysis on real passenger flows data from the transportation hub of the Europe’s leading business district “La Défense”. The transportation hub provides access not only to the business district but also to a major shopping center and concert/conference hall. The temporal segmentation is performed at the day scale but the model also manages the hour scale. This representation draws useful information for operators and managers of urban spaces by valorizing the collected data. Segmentation on a daily basis gives an idea of the long-term dynamics of passenger flows, while modeling on an hourly basis makes it possible to study in detail the impact of calendar and non-calendar factors.

The paper is organized as follows. We first position our work within an existing literature on count data, mixture models and segmentation (Sect. 2). Then we introduce our motivating case study by including a global presentation of the multimodal transport hub, the series of counts, and the exogenous factors impacting them (Sect. 3). Section 4 details the two models and their variants. It also explains how to fit the model parameters. Section 5 compares the two models, by first applying them to simulated data, then to actual data from the case study. Finally, we present the results of segmentation and mobility patterns obtained with the preferred model. Section 6 concludes the paper.

## 2 State of the art

Count data are ubiquitous in the field of human mobility and are tracked by numerous data collection technologies. Smart cards and automated fare collection (AFC) systems in particular produce large quantities of data which are frequently used to analyze urban mobility (Briand et al. 2017; Pavlyuk et al. 2020; Wang et al. 2021). Stereo camera technology is another data collection system seen in a number of applications, and more specifically in two fields: pedestrian detection (Kristoffersen et al. 2016) and vehicle navigation (Peláez et al. 2015). The collected data can be noisy hence difficult to interpret; they should therefore be analyzed using a clustering approach. The following sections provide a state of the art of this component.

### 2.1 Analyzing count data through clustering

Finding groups of similar mobility patterns in massive and noisy human mobility behavior data falls in the field of time series segmentation, which deals with time series

that are subject to regime changes. From a statistical point of view, segmentation is found by estimating a common break in the mean and the variance of count data (Bai 2010). It can also be seen as a clustering framework that partitions the count time series into a reduced set of groups sharing common changes in regime. As mentioned by Ghaemi et al. (2017), the following three categories of mobility patterns can be analyzed:

1. Spatial patterns, which provide information on the spatial distribution of people, enabling public transport operators to optimize resource allocation (Li et al. 2020).
2. Temporal patterns, which give information on how a public transport network is used over time (Briand et al. 2017; Ghaemi et al. 2017), helping transport operators to predict affluence and understand the evolution of demand.
3. Spatio-temporal patterns, as with Pavlyuk et al. (2020), which determine dates with similar daily mobility patterns (e.g. origin–destination profiles). Transport operators need to identify the variability of travel demands that change in space and time.

Several approaches have been deployed for clustering mobility data. These include distance-based approaches, such as hierarchical ascendant classification (HAC) or K-means algorithms. These two methods are used in Agard et al. (2006) to study weekly travel behavior on buses. Summaries of weekday passenger activity are obtained with the aggregation of bus trips. HAC and K-means are applied in order to study group behavior. DBSCAN is another clustering method based on the density of data points. With smart card data that provide spatio-temporal information about trips, Manley et al. (2018) calculate the regularity of a user with DBSCAN clustering that detects the time slots of the day when the passenger uses the network most frequently. The comparison of the clusters formed among the different users allows the authors to highlight the metro stations or bus lines that are regularly used, which lines are linked or which types of populations (e.g., residential or working) are linked to particular stations or bus lines.

However, distance-based approaches present certain problems. They are hard clustering methods, i.e. each data belong to only one class, contrary to soft clustering methods (Baid and Talbar 2016). Moreover they can't be extended to include exogenous variables (Magidson and Vermunt 2002).

## 2.2 Probabilistic methods for clustering count data

Probabilistic model-based methods (Bouveyron et al. 2019; McLachlan et al. 2019) involve clustering data using a mixture of probability distributions. These models seem suited for our case study, due to the wide variety of distributions within each segment, which are due to the effects of exogenous factors and the interrelationships between count locations. Moreover, these models offer interpretability that is valuable to better understand temporal and spatial mobility dynamics within the transport hub (Magidson and Vermunt 2002). Since the observed data are counts that are overdispersed, the work should be oriented towards certain types of model-based methods. Poisson mixture regression models can be successfully fitted to counts, in the presence of exogenous factors (Côme and Oukhellou 2014; Mohamed et al. 2016). Nevertheless,

because of the assumption of equal dispersion, i.e. expectation and variance are equal conditionally on segments and covariates, the usefulness of these methods can be limited when data are highly overdispersed. The use of Negative binomial regressions can alleviate this issue because means and variances differ (Hilbe 2011). Mixtures of these models are used in transcriptomic analysis as with Li et al. (2021), but we did not find an application of such models in the field of mobility.

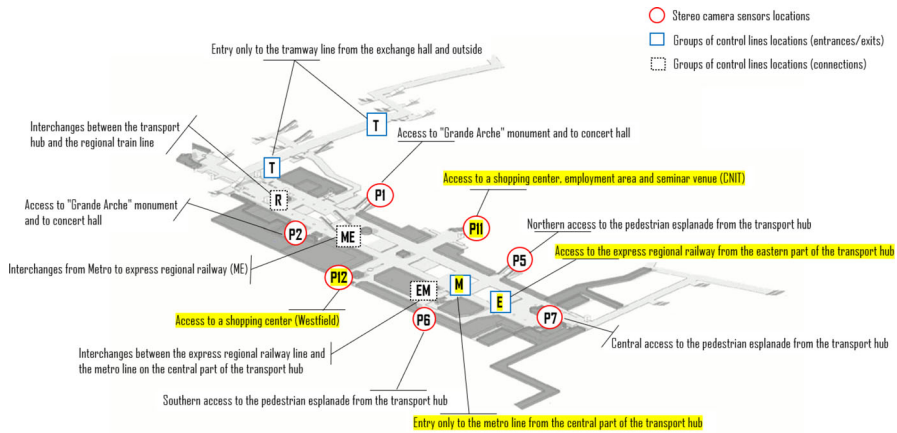
A problem occurring when working with multivariate count data are the dependence relationships between the count series. A challenge is thus to quantify these relationships. Multinomial and Dirichlet multinomial regressions are the models usually applied to multivariate count data (Zhang et al. 2017). However, these models cannot handle total counts summed over all series. As stated by Peyhardi et al. (2021), this induces a constraint in terms of dependencies between the count series, as any series is deterministic when all the other series are known. A first solution is to write the count distributions as “sums and shares” distributions, which allows dependencies between series to be taken into account. This model assumes that total counts follow a univariate distribution (e.g. Poisson, Negative binomial), used as a parameter of a multivariate distribution (e.g. Multinomial, Dirichlet multinomial) to model their separation into different categories. For example it is possible to obtain a distribution by compounding a negative binomial with a multinomial distribution as did Sibuya et al. (1964). Here we were inspired by the work presented by Jones and Marchand (2019). “Sums and shares” models are explained and developed in Sect. 4.1.1. A second way not to assume independence between series is to use the multivariate Poisson log-normal distribution first proposed by Aitchison and Ho (1989), which is also well explained by Chiquet et al. (2021), and interestingly used in a model-based clustering method in the work of Silva et al. (2019). In this model the dependence structure between the series is taken into account with a covariance matrix that is estimated within a hidden layer. “Poisson log-normal” models are explained in Sect. 4.1.2. The different models introduced here differ in terms of how complex the treatment of noisy data is. A model that is able to take into account very noisy data could indeed not guarantee that the observations are always similar within the segments because of the large variances. When considering the example of the Poisson log-normal mixture model, an encoding with diagonal covariance matrices, as opposed to that with unconstrained covariance matrices, should give more sensitive segments, as the former cannot be characterized by the dependence structure of the data.

Both methods appear to be suitable for modeling multivariate, overdispersed and correlated count data. In the following section we present the data associated with our case study and show that they meet these different characteristics.

### 3 Case study

#### 3.1 The “La Défense” multimodal transport hub

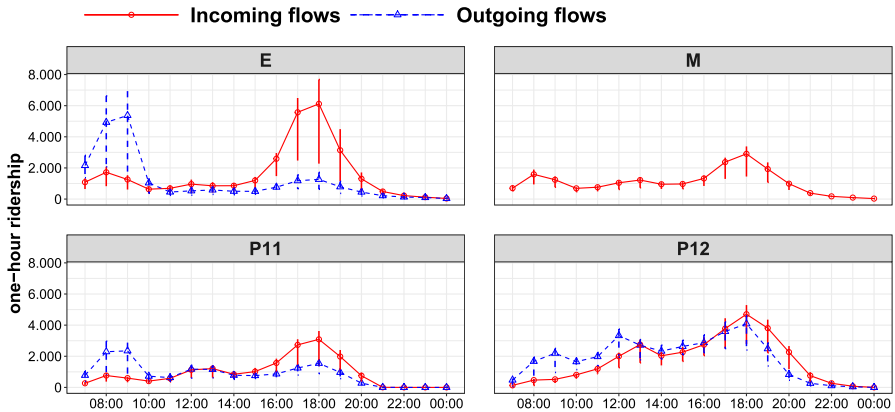
This work focuses on the “La Défense” hub, a major tertiary center in the Paris region, to which a large number of workers commute daily, primarily (about 85%) by public transportation. In addition to being a business zone with ca. 180,000 jobs, the “La



**Fig. 1** Location map of the transport hub with the count locations used in this study, each with a description of its vicinity. Count locations presented in Figs. 2 and 3 are highlighted here

Défense” center hosts a shopping mall and a concert hall. Thus, users travel to the “La Défense” hub daily, mainly to work, but also to shop, attend concerts or go to university. “La Défense” is served by multiple modes i.e. a metro line (i.e. “Metro 1”), an express regional railway line (i.e. “RER A”), two regional train lines (i.e. “L” and “U”) and a tramway line (i.e. “T2”). Note that the M1 and RER A lines are parallel when crossing Paris. As studied by de Nailly et al. (2021), these lines are complementary and in competition. In particular, users may take one line when the other is disrupted.

Two sources of passenger flow count data are available for our study. First, a set of stereo camera sensors, placed at all access/ egress points to/ from the hub, record the number of people entering/ leaving every minute. This count system captures various pedestrian flows to and from shopping centers, office towers, bus stations, or a concert hall. Then, ticketing logs collected by automated fare collection systems capture the volume of flows between the different transport lines of the hub. Ticketing data are presented in the following form: for each control line within the transport hub, the number of validations is aggregated by 10-minute intervals. In addition, each control line is associated with a function that provides the nature of the flow (i.e. incoming, outgoing and transit). In this study, we work with groups of control lines that share the same functions and are spatially located close to one another. The transport hub is organized in two levels: the level at which the different transport lines operate, and a level called “interchange hall” which allows users to access these lines. On the outside, above the transport hub, there is a pedestrian esplanade from which it is possible to access the different work and leisure areas of “La Défense”. Figure 1 presents a schematic view of the interchange hall with all count locations used in this study. Counts at these locations are collected either by sensors or groups of control lines. In addition, each location is attached to a codename and a description of the surroundings (ibid.).



**Fig. 2** Counts collected at locations E, M, P11 and P12. Medians, 1st and 3rd quartiles of hourly counts across the entire observation period. Note that only incoming flows are captured at metro access M

### 3.2 Characteristics of the “La Défense” count data

This study’s analysis is based on hourly aggregated count data from sensors and ticketing systems, as provided by the Paris public transport operator (i.e., Régie Autonome des Transports Parisiens or RATP) between April 2019 and September 2020. The daily patterns of use of public transport frequently alternate between working days (i.e., non-holiday weekdays) and non-working days (i.e., bank holidays and weekends). Therefore, this study will focus strictly on working days between 7 am and 1 am (the following day) in order to avoid introducing variables for non-working days, which simplifies the models. Several points characterize the count data, as we show below.

#### 3.2.1 Inter and intra-location variabilities are clearly visible

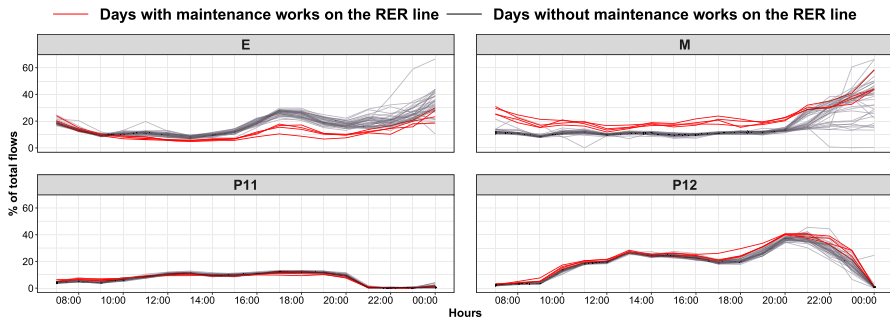
Fig. 2 illustrates this phenomenon for the count locations E, M, P11 and P12 described in Fig. 1.

It shows the median, 1st quartile, and 3rd quartile values of the incoming and outgoing flow counts recorded every hour across the whole period. These graphs suggest significant variability in the flows captured at the different hub locations, highlighting large inter-location variabilities. There is also intra-location variability, as observed through the 25% and 75% quantiles. These large differences indicate significant variability in the counts over the entire period for several locations and hours of the day.

#### 3.2.2 Changes in time series regimes must be taken into account

Fig. 3 represents the proportion of counts captured at each of the locations from Fig. 2 at each hour (in relation to the total hourly counts summed over the four locations)





**Fig. 3** Evolution of the proportion of flows that pass through each count location (i.e. E, M, P11 and P12) at each hour for incoming flows (these proportions are calculated in relation to the total hourly counts summed over the seven flows). Each day from summer 2019 is represented. Days highlighted in red occurred during maintenance works on the RER transport line

during summer 2019 in which a period of maintenance works on the RER line occurred. This figure illustrates an example of local changes in regime regarding location use, as a result of a particular period. It shows significant changes in the spatial distribution of flows between locations once the maintenance work period begins. This effect is visible for locations M and E. Ridership at location M increased during maintenance works, whereas that at location E diminished, as people used the metro line instead of the RER line.

### 3.2.3 Several covariates are likely to have a significant impact on the use of the transport hub

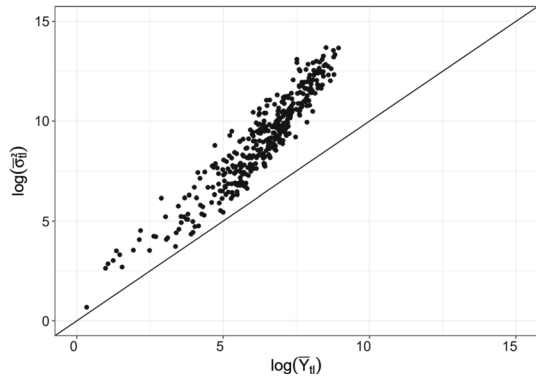
Hereafter, we specify a non-exhaustive list of these covariates which will be used in the regression parts of all the models:

1. The time of day (i.e. hours).
2. Concerts at *Paris La Défense Arena* (i.e. one of the largest concert halls in Europe with a maximum capacity of around 40,000 seats).
3. Disruptions of the RER line. This regional express railway is one of the busiest public transportation lines in the world, with about 1.4 million passengers per working day.

### 3.2.4 Overdispersion and correlations are found in the series, even when removing covariate and segment effects

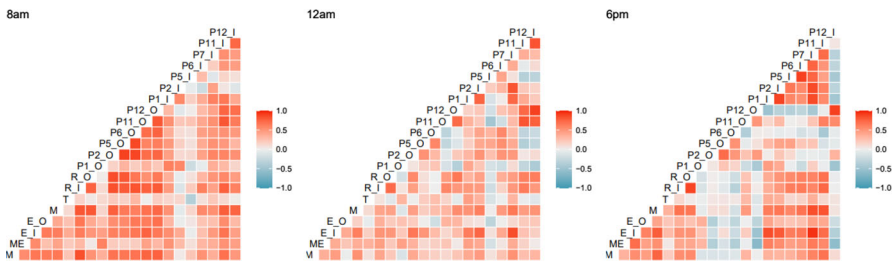
We take the example of a short normal period, i.e. April 2019, for which we removed all time slots with special events such as transport disruptions or concerts. First we visualize the overdispersion effect by computing empirical means and variances of each count location at each hour over this period. The results shown in Fig. 4 suggest that variances (y-axis) are much higher than means (x-axis), which indicates overdispersion.

**Fig. 4** Log-scaled empirical means versus empirical variances by count location and time slot, for the April 2019 period



Correlations between count series are abundant in the dataset. To highlight this point, we have plotted the correlations between the counts at different locations for three particular times in Fig. 5: one hour of the morning peak (8:00 am), the noon hour (12:00 pm), and one hour of the evening peak (6:00 pm). The location codename is associated with “O” when it is an outgoing flow and “I” when an incoming flow. Note that correlations can be positive or negative but we still see a predominance of positive correlations here.

Considering all these aspects, a model capable of properly modeling these mobility data should handle regime changes through the most continuous segments possible while being able to take into account correlations between series as well as overdispersions. The next section will introduce two promising strategies that can address this issue.



**Fig. 5** Correlation matrices between the different counting locations for three particular hours of the day. Note that the correlations seem to be more impacted by the direction of the flows (“O” or “I”) than by the geographical proximity of the locations. We see a tidal effect with “O” flows well correlated during the morning rush hour (8am) and “I” flows well correlated during the evening rush hour (6pm). Since the uses of the pole are less subject to these tidal effects outside of the peak hours, the correlations seem to be less strong at 12am

## 4 Segmentation model structures and estimation

Regression models “sums and shares” and Poisson-Lognormal are first introduced. In particular, we will explain to what degree and how these different strategies handle or not the phenomena of correlation and overdispersion. Then we will explain how we transformed these models into mixture models, able to detect regime changes in the time series.

### 4.1 Regression models for correlated and overdispersed multivariate count data

Hereafter,  $Y$  is considered as an  $L$ -vector of counts among  $L$  locations:  $Y = (Y_l)_{l \in \{1, \dots, L\}}$ . We propose to model these count data through two distinct regression models, namely “sums and shares” and “Poisson log-normal”. The general notation of these regression models is:

$$Y|\mathbf{x}, \boldsymbol{\zeta} \sim \mathcal{D}(\mathbf{x}, \boldsymbol{\zeta}), \quad (1)$$

with  $\boldsymbol{\zeta}$  a set of parameters controlling the conditional distributions  $\mathcal{D}$  that we are looking for. We note  $\mathbf{x}$  as a  $D \times 1$  vector of  $D$  exogenous factors. In the following we introduce the different regression models we wish to compare.

#### 4.1.1 Sums and shares regression models

The first strategy draws inspiration from the work presented by Jones and Marchand (2019). Let  $V = \sum_l Y_l$  be the sum of the counts; the proposed strategy models the multivariate counts as follows:

1. The sum  $V$  follows a distribution  $\mathcal{G}(\mathbf{x}, \boldsymbol{\zeta})$ ;
2. Conditionally on  $V = v$ ,  $Y$  follows a distribution  $\mathcal{H}(v, \mathbf{x}, \boldsymbol{\zeta})$  on the simplex defined by  $\{0, \dots, V\}^L$ .

Consequently, the probability mass function is given by

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\zeta}) = h(\mathbf{y}|v, \mathbf{x}, \boldsymbol{\zeta})g(v|\mathbf{x}, \boldsymbol{\zeta}) \quad (2)$$

with  $g$  and  $h$  are the densities of  $\mathcal{G}$  and  $\mathcal{H}$  respectively. In the following we will focus on two sums and shares models: Poisson-multinomial and Negative binomial-Dirichlet multinomial.

*Poisson sums and multinomial shares.* This first option applies a Poisson distribution for the sum distribution  $\mathcal{G}(\mathbf{x}, \boldsymbol{\zeta})$  and a multinomial distribution for the joint distribution of counts  $\mathcal{H}(v, \mathbf{x}, \boldsymbol{\zeta})$ . As stated in Lemma 4.1 of Zhou et al. (2012), this joint distribution is the same as the one produced by  $L$  independent Poisson variables with parameters  $r_1 = \lambda u_1, \dots, r_L = \lambda u_L$ , where  $\lambda$  is the parameter of the Poisson distribution and  $u_1, \dots, u_L$  are the parameters of the multinomial distribution. Thus this model can be seen as a baseline in our study since the components  $Y_1, \dots, Y_L$  are

independent. The mean and variance of  $Y_l$  are  $\mathbb{E}(Y_l|\mathbf{x}) = \mathbb{V}(Y_l|\mathbf{x}) = \lambda u_l$ ; the covariance between  $Y_l$  and  $Y_{l'}$  is  $\text{Cov}(Y_l, Y_{l'}|\mathbf{x}) = 0$ . It can be seen from the mathematical expression of the moments that neither overdispersion nor correlation are handled.

*Negative binomial sums and Pólya shares.* This second option introduces correlations between count series and models overdispersed counts. It is possible to mix  $\lambda$  and  $u_1, \dots, u_L$  over distributions for random variables  $\Lambda > 0$  and  $U_1, \dots, U_L \in (0, 1)$  such that  $U_1 + \dots + U_L = 1$ . Here  $\Lambda$  follows a gamma distribution and  $U_1, \dots, U_L$  a Dirichlet distribution so that  $\mathcal{G}(\mathbf{x}, \boldsymbol{\zeta})$  is a Negative binomial distribution ( $\mathcal{NB}$ ) and  $\mathcal{H}(v, \mathbf{x}, \boldsymbol{\zeta})$  a Dirichlet-multinomial (or Pólya) distribution ( $\mathcal{DM}$ ). With this specification, the model may be written as follows:

$$V|\mathbf{x} \sim \mathcal{NB}(\exp(\mathbf{x}^T \boldsymbol{\gamma}), r) \tag{3}$$

$$Y|\mathbf{x}, V \sim \mathcal{DM}(V, (\exp(\mathbf{x}^T \boldsymbol{\xi}_l))_{l \in 1, \dots, L}), \tag{4}$$

with  $r$  the shape parameter and  $\boldsymbol{\gamma}$  the vector ( $D \times 1$ ) of regression parameters of the  $\mathcal{NB}$  regression.  $\boldsymbol{\xi}_l$  is the vector ( $D \times 1$ ) of  $\mathcal{DM}$  regression coefficients linked to exogenous effects  $\mathbf{x}$ . Note that  $\boldsymbol{\zeta} = (r, \boldsymbol{\gamma}, \boldsymbol{\xi})$  here.

**Properties 1** The moments of  $Y$  from Negative binomial sums and Pólya shares model as described by Jones and Marchand (2019) are written as follow:

- $\mathbb{E}(Y_l|\mathbf{x}) = \frac{r q_l}{\bar{k} q} = k \frac{q_l}{q}$
- $\mathbb{V}(Y_l|\mathbf{x}) = \frac{r q_l}{\binom{\bar{k}}{l}^2 q^2 (1+q)} [q \{r + 1 + (1 + q) \frac{r}{\bar{k}}\} + (q - r) q_l]$
- $\text{Cov}(Y_l, Y_{l'}|\mathbf{x}) = \frac{r(q-r)q_l q_{l'}}{\binom{\bar{k}}{l}^2 (q.)^2 (1+q)}$ ,

with  $k = \exp(\mathbf{x}^T \boldsymbol{\gamma})$ ,  $q_l = \exp(\mathbf{x}^T \boldsymbol{\xi}_l)$  and  $q. = \sum_{l=1}^L q_l$ .

From the variance expression we can see that overdispersion is taken into account. Moreover the signs of the covariances are the same for all  $l, l'$  and depend on the sign of  $q. - r$ .

### 4.1.2 Poisson–Lognormal model

The second strategy consists of using a two-layer hierarchical model, with one observation layer modeling the count data and one hidden layer that estimates the dependencies between the different location counts  $Y_l$ . The multivariate Poisson log-normal distribution addresses this by modeling count data with a Multivariate Gaussian latent layer, which is exponentiated before being used for parametrizing independent poisson distributions. This model is well explained by Chiquet et al. (2021). The advantages lie in the fact that there is no need to assume independence between the series, and any overdispersion can be taken into account. The equations are:

$$\theta|\mathbf{x} \sim \mathcal{N}(\mathbf{x}^T \boldsymbol{\rho}, \boldsymbol{\Sigma}) \tag{5}$$

$$Y|\theta \sim \mathcal{P}(\exp(\theta)), \tag{6}$$

where  $\mathcal{N}$  is a Gaussian distribution. Each  $Y$  is modeled via a Gaussian latent vector  $\theta$ . The mean of the latent vector is a combination of covariates  $\mathbf{x}$  and of the  $(D \times L)$  regression-parameter matrix  $\boldsymbol{\rho}$ . The  $(L \times L)$  covariance matrix  $\boldsymbol{\Sigma}$  describes the underlying structure of the dependencies between the  $L$  items. We consider two cases in this study concerning  $\boldsymbol{\Sigma}$ . In the first case,  $\boldsymbol{\Sigma}$  is diagonal, which means that only the variances are estimated, and the covariances are assumed to be null. In the second case,  $\boldsymbol{\Sigma}$  is estimated without restrictions, i.e. all the covariances are estimated. Note that  $\boldsymbol{\zeta} = (\boldsymbol{\rho}, \boldsymbol{\Sigma})$  here.

**Properties 2** The moments of  $Y$  from the Poisson-Lognormal model as described by Chiquet et al. (2021) are written as follow:

- $\mathbb{E}(Y_l|\mathbf{x}) = \exp(\mu_l + \frac{1}{2}\Sigma_{l,l})$
- $\mathbb{V}(Y_l|\mathbf{x}) = \mathbb{E}(Y_l|\mathbf{x}) + \exp(\mu_l + \frac{1}{2}\Sigma_{l,l})^2(\exp(\Sigma_{l,l}) - 1)$
- $\text{Cov}(Y_l, Y_{l'}|\mathbf{x}) = \mathbb{E}(Y_l|\mathbf{x}) \mathbb{E}(Y_{l'}|\mathbf{x})(\exp(\Sigma_{l,l'}) - 1)$ ,

with  $\boldsymbol{\mu} = \mathbf{x}^T \boldsymbol{\rho}$ .

From these equations, we see that this model accounts for overdispersion. It also supports negative and positive correlations when there is no restriction on the covariance matrix.

### 4.1.3 Summary of the models

Table 1 lists all the models compared within this paper. Each model is associated with an acronym used in the rest of the paper. Elements on the number of estimated parameters and overdispersion/correlation handling are also displayed. Correlations between the series are handled in distinct ways. For the NegPol model the correlations are captured by the dirichlet-multinomial parameter vector and their sign is governed

**Table 1** Mixture models studied with acronyms<sup>b</sup>, numbers of parameters, correlation and overdispersion handling

Acronym	Model	Number of param	Overdisp. handling	Corr. handling
PoiMult	Poisson multinomial	$b1 = D + (D \times (L - 1))$	✗	✗
NegPol	Negative binomial Pólya	$b2 = D + 1 + (D \times L)$	✓	✓ <sup>a</sup>
PLNdiag	Poisson lognormal diagonal $\boldsymbol{\Sigma}$	$b3 = D \times L + L$	✓	✗
PLNfull	Poisson lognormal full $\boldsymbol{\Sigma}$	$b4 = D \times L + L^2/2 + L$	✓	✓

Note that  $b1 < b2 < b3 < b4$  if  $D < L + 1$  and  $b1 < b3 < b2 < b4$  if  $D > L + 1$  with  $D$  the number of exogenous factors and  $L$  the number of count locations

<sup>a</sup>All correlations either positive or negative

<sup>b</sup>Acronyms are preceded by an “s” (e.g. sPoiMult) to designate the smoothed mixture models, as explained in Sect. 4.2

by the negative binomial parameters. For the PLNfull model, it is the directly estimated covariance matrix that captures the correlations.

## 4.2 Mixture models

We now consider  $Y_{j,t}$  as the L-vector of counts for time slot  $t$  of day  $j$ . We assume that each day  $j$  can be associated with the dynamics of one segment among  $S$  possible segments, with a certain probability. Associating dynamics at the day scale rather than at the time slot scale is, in our opinion, a good way to synthesize information over long periods such as years. If we adapt the previously presented regression models to a mixture model framework, we end up with generative models which include a set of indicator variables denoted by  $Z_j$  ( $Z_j \in \{0, 1\}^S$ ) encoding the segment membership of the days, with  $s \in \{1, \dots, S\}$ . The number of segments  $S$  is chosen a priori. The following generative model is assumed for the observed data:

$$Y_{j,t} | Z_{j=s}, \mathbf{x}_{j,t}, \boldsymbol{\zeta}_s \sim \mathcal{D}(\mathbf{x}_{j,t}, \boldsymbol{\zeta}_s), \quad (7)$$

with  $\boldsymbol{\zeta}_s$  the set of parameters controlling the conditional distributions within segment  $s$ . This generative model assumes that knowing the segment of the day and current values of covariates, the counts at each timestamp  $t$  follow a distribution of parameters specific to each segment. The variable  $Z_j$  follows a multinomial distribution ( $\mathcal{M}$ ) of parameter  $\boldsymbol{\pi}$  (i.e the vector of association weights). We will compare two ways of modeling segment memberships  $Z_j$ , as detailed below:

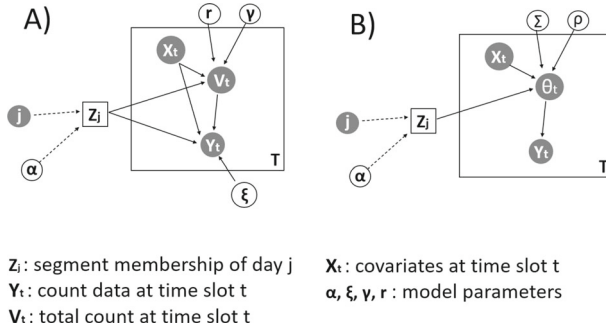
- In the first scheme, the association weights  $\pi_s$  are fixed and will not change during the days.

$$Z_j \sim \mathcal{M}(1, (\pi_s)_{s \in \{1, \dots, S\}}) \quad (8)$$

- In the second scheme denoted as “smooth”, the association weights evolve with days  $j$  as they follow a logistic transformation of cubic spline functions with  $M$  nodes. The idea behind this scheme is to help the model to detect regime changes that are difficult to detect from the data, by taking into account a relation between proximal days through spline functions. Moreover the cubic spline functions can help to detect separate periods that are similar to each other.

$$Z_j | j \sim \mathcal{M}(1, (\pi_s(j; \boldsymbol{\alpha}))_{s \in \{1, \dots, S\}}) \quad (9)$$

with



**Fig. 6** Graphical models of the sums and shares mixture model **A** and Poisson-Lognormal mixture model **B**. Grey circles are observed data. White circles are parameters to be estimated.  $Z_j$  is the hidden layer of the generative model. Dashed links represent the potential intervention of the smoothed association weights  $\pi_s(j; \alpha)$ .  $T$  is the total number of time slots for day  $j$

$$\pi_s(j; \alpha) = \frac{\exp(\sum_{m=1}^{M+4} \alpha_{s,m} a_m(j))}{\sum_h \exp(\sum_{m=1}^{M+4} \alpha_{h,m} a_m(j))}, \tag{10}$$

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

$$a_m(j) = j^{m-1}, m \in \{1, \dots, 4\} \tag{11}$$

$$a_{m+4}(j) = (j - \kappa_m)^3, m \in \{1, \dots, M\}. \tag{12}$$

Each cubic spline is a piecewise cubic polynomial with knots at  $\kappa_m, m \in \{1, \dots, M\}$ . Note that in the mathematical notations we will use the “smoothed” association weights  $\pi_s(j; \alpha)$  in order to include the formalism of the estimation of the  $\alpha$  parameters.

In the rest of the paper we use the acronym of the models as defined in Table 1 preceded by an “s” to designate the smoothed versions. These models require the additional estimation of  $(S - 1) \times M$  parameters  $\alpha$ . The graphical models for mixture of sums and shares models and mixture of Poisson log-normal models are shown in Fig. 6.

### 4.3 Parameter estimation

The parameters of the model are estimated by the maximum likelihood method solved by the Expectation Maximization (EM) algorithm (Dempster et al. 1977) as explained with Algorithm 1 (text in blue stands for smoothed model estimations). Details about the different steps for the parameter estimation are developed in “Appendix A” for the “sums and shares” mixture models and in “Appendix B” for the Poisson log-normal

mixture model. For each estimation of the S-segment models, we ran five trials, each with a segment initialization as explained with Algorithm 2. The way parameters are initialized is crucial for the ability of the EM algorithm to converge faster and provide acceptable solutions. The same initialization procedure was implemented for both models, using a hierarchical ascendant clustering (HAC) (i.e. Algorithm 2). HAC method builds a hierarchy of clusters from the bottom-up: it first puts each day in its cluster before repeatedly indentifying the closest two clusters and combining them into one cluster until all the days are in a single cluster. The number  $S$  of clusters is then selected. HAC requires a distance matrix computed between the  $J$  days; we used euclidian distance here. We use Ward's linkage method to determine how close two clusters are. This non-random initialization was chosen in particular with respect to the Poisson log-normal mixture model since a random initialization would have implied very large covariance computations  $\Sigma_s$  on heterogeneous data, which would have made it difficult to identify homogeneous segments. For each run, we use a set of five randomly selected days per segment for parameter initialization since we consider that each experiment should not have exactly the same starting point, so as to allow the mixing models to potentially find different solutions (Lashkari and Golland 2007). Because of this potential variable finding of different optima, depending on the segment initialization, the EM-run may lead to the disappearance of a segment during the E step. This is why running several simulations for each run is valuable here. Moreover we added a "hot restart" step in the EM framework in order to handle segment disappearance. This step consists in resetting the disappeared segment by using days from other segments with the smallest values of *a posteriori* membership probabilities  $(\tau_{j,s})$  i.e. days which are least likely to belong to those segments. Each run is stopped when the difference in decay between two successive log-likelihoods is below a given threshold, which we set at  $10^{-6}$ . The model with the highest log-likelihood is finally chosen. Selecting an optimal number of segments  $S$  is crucial. In the context of mixture models and the EM algorithm, a natural choice for model selection is to use the Bayesian Information Criterion (BIC, Schwarz (1978)). The optimal number of segments is thus selected by means of the search for the minimum value of this criterion. All models were built in the R environment using the glm function in the stats package, glm.nb function in the MASS (Ripley et al. 2013) package, multinom function in the nnet (Ripley et al. 2016) package and also MGLM (Kim et al. 2018) and PLNmodels (Chiquet et al. 2021) packages. Time segmentation is obtained by updating at each step of the EM algorithm the conditional expectation of membership of the days to the segments  $(\tau_{j,s})_{s=1,\dots,S}$ .



**Algorithm 1** EM algorithm to estimate parameters  $\zeta$ 

**Inputs:**  $Y$  tensor ( $J$  days  $\times T$  time slots  $\times L$  locations),  $X$  tensor ( $J$  days  $\times T$  time slots  $\times D$  covariates), number of segments  $S$ . **Outputs:** Estimated parameters ( $\zeta$ ), posterior probabilities  $\tau_{j,s}$

- 1: **Initialization**
- 2: Initialize  $\zeta^{(0)}$ ,  $\tau^{(0)}$  and  $\alpha^{(0)}$  ▷ see Algorithm 2 for details
- 3:  $c \leftarrow 0$
- 4: **repeat**
- 5:   **E step:** compute the *a posteriori* probabilities
- 6:   **for** each segment  $s \in \{1, \dots, S\}$  **do**
- 7:     Compute  $\tau_{j,s}^{(c)}$  each day  $j \in \{1, \dots, J\}$  ▷ see Equations A.6 and B.17
- 8:   **end for**
- 9:   **for** each segment  $s \in \{1, \dots, S\}$  **do**
- 10:     **if** number of days in segment  $s = 0$  **then**
- 11:       Process hot restart
- 12:       break
- 13:     **end if**
- 14:   **end for**
- 15:   **M step:** update parameters
- 16:   Compute  $\alpha^{(c+1)}$  ▷ see Equations A.7 and B.16
- 17:   **for** each segment  $s \in \{1, \dots, S\}$  **do**
- 18:     Compute  $\zeta_s^{(c+1)}$  ▷ see Equations A.7 and B.16
- 19:   **end for**
- 20: **until convergence**

**Algorithm 2** Initialization for parameters  $\zeta^{(0)}$  and posterior probabilities  $\tau^{(0)}$ 

**Inputs:**  $Y$  tensor ( $J$  days  $\times T$  time slots  $\times L$  locations),  $X$  tensor ( $J$  days  $\times T$  time slots  $\times D$  variables), number of segments  $S$ . **Outputs:** Initialized parameters ( $\zeta^{(0)}$ ), posterior probabilities  $\tau^{(0)}$

- 1: Reshape  $Y$  into  $Y_{lin}$ , a tensor ( $J$  days  $\times TL$  time slots/locations)
- 2: Find  $S$  clusters within the  $J$  days based on euclidian distances. Apply HAC on  $Y_{lin}$  in order to associate each day  $j$  (each row) to one of the  $S$  clusters (segments).
- 3: **for** each segment  $s \in \{1, \dots, S\}$  **do**
- 4:   Sample randomly 5 days  $j$  from segment  $s$
- 5:   Compute  $\zeta_s^{(0)}$  based on these days through a pass in the M step
- 6:   Calculate  $\tau_{j,s}^{(0)}$  with uniform a priori on  $\pi_s(j; \alpha)$  for each day  $j$  ▷ see Equations A.6 and B.17
- 7: **end for**
- 8: Compute  $\alpha^{(0)}$

## 5 Numerical experiments

This section compares the different models on simulated and actual data from the case study. On simulated data, the goal is to evaluate the capacity of each model to classify days well in controlled settings. We will also highlight the abilities of these models in their smoothed or unsmoothed versions. We will then apply the models to the real data set. First, we will compare the different models based on their fitting capacity when varying the number of segments. Then for the chosen model, three results will be detailed: the segmentation of the total period into  $S$  time segments, the analysis of typical spatio-temporal flow patterns within these segments, and the impact of exogenous variables.

### 5.1 Experiments using simulated data

The purpose of working with simulated data is to comfort us in the ability of our models to correctly detect segments in a controlled setting. Notably, we seek to:

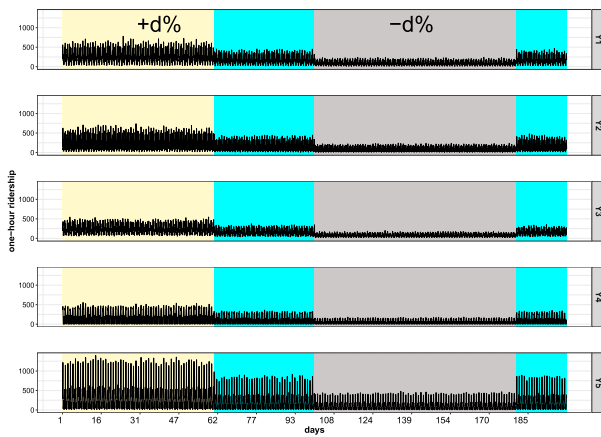
- evaluate the capacities of the models to correctly classify days coming from time series subject to controlled global or local regime changes;
- study the impact of the number of knots  $M$  for smoothed models. We thus test models with  $M = 5, 20, 50$  and  $80$  knots.

The experiments conducted here are carried out in a simple case and are therefore not generalizable; they do, however, give us some insights about the detection capabilities of the models. The data generation protocol is as follows: we create  $L = 5$  series of counts, with 3 segments during a period of two hundred days. These series are generated from a PLNdiag or a PoiMult model (see Table 1) with parameters specified a priori (i.e.,  $\mu$  for PLNdiag,  $\lambda$  and  $\mathbf{u}$  for PoiMult). For the PLNdiag simulation model we limited the variance values  $\Sigma_{l,l'}$  to  $1 \times 10^{-3}$ . The data generating mechanism relates to the models' assumptions in the following way: both PLN and PoiMult generate independent count series respecting the assumptions of the segmentation models, although the data generated with PLN are slightly overdispersed. Regime changes are generated by increasing or decreasing the counts with an average  $d\%$ . Note that this rate is slightly different from one time series to another. As shown in Fig. 7, we generate three segments according to the following protocol:

1. Segment 1 includes days 1 through 60 and is characterized by an average increase of  $d\%$  of counts.
2. Segment 2 includes days 61 to 100 and 181 to 200 and is not impacted by any change.
3. Segment 3 includes days 101 to 180 and is characterized by a loss of  $d\%$  of counts.

The regime change can impact all the time series (i.e., global impact) or a single series (i.e., local impact). One example of simulated count data generated according to a PoiMult model is shown in Fig. 7.

The following set of experiments studied the impact of the rate of count change  $d$  on the segmentation capacities of four models: PoiMult, sPoiMult, PLNdiag and sPLNdiag (see Table 1). At first, the impact of regime changes will be global and then



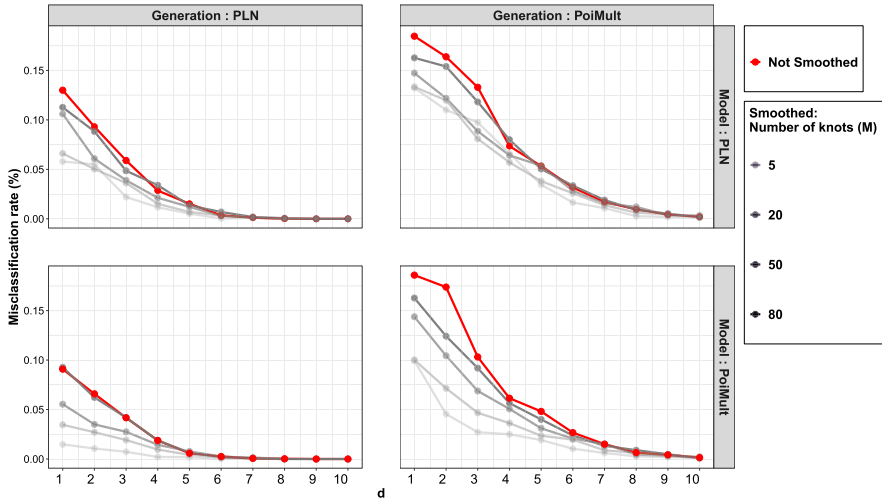
**Fig. 7** Simulated data over 3 segments with  $d = 50$  and PoiMult model. The  $(Y_t)_{t \in \{1, \dots, 5\}}$  series are the simulated data. The difference between the  $Y_t$  is derived from an a priori specified difference of the  $\lambda$  and  $u$  sets at each  $t$ . The colors yellow, cyan and grey correspond respectively to segments 1, 2 and 3 (color figure online)

we will test the local case. Each experiment consists in detecting segments in a set of simulated data with a specific rate  $d$ . Each experiment is repeated 20 times with each time a new generation of simulated data and a task of segment detection. Within each experiment, the search for segments is performed with  $S=2, 3$  or 4 segments, with the expectation that the number of segments  $S=3$  used to generate the simulated data performs better.

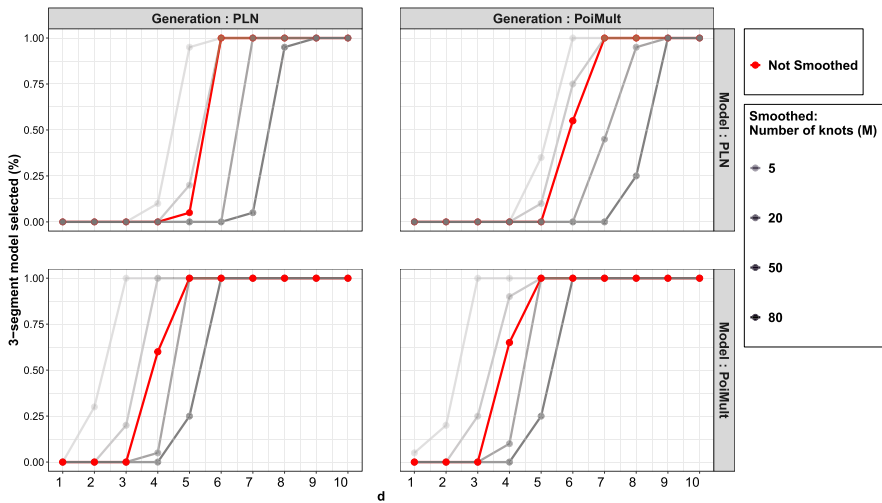
Two criteria are estimated to compare the model capacities: the mean misclassification rate of days, and the percentage of times the 3-segment model was the best fit according to the *Bayesian Information Criterion* (BIC). For experiments with global impact, results are shown in Figs. 8 and 9.

Figures 8 and 9 show that higher rates  $d$  are, as expected, associated with lower rates of misclassification of days and with more frequent selection of 3-segment models. The segments being more distinguishable, the segmentation task is easier, but not in the same way for the two generation models. It is to be noted that misclassification rates are slightly lower for data generated with PLN model. We suppose this may be due to the fact that data generated with PLN are slightly higher than data generated with PoiMult. This can be seen from the calculation of the expectation for PLN model which incorporates the variance term  $\Sigma$ . The difference between the segments would then be slightly more pronounced with the data generated with PLN.

Then it seems that the smoothed versions with a low number of knots ( $M = 5, 20$ ) obtain better results for both criteria. This result was obtained in the context of low-noise and low-dimension simulated data, but nevertheless indicates a potential advantage of smoothed models with few knots compared to unsmoothed versions. For experiments with local impact, we chose a rate  $d = 10$ . Results are shown in Table 2. In this hard-to-detect case, only the sPoiMult models with fewer knots (i.e., 5 or 20) seem to succeed in considering the three-segment version as best. “Sums and shares”



**Fig. 8** Misclassification rate for PoiMult model and PLN models with the 3-segment models. The graphics show the evolution of the impact of  $d$ : once  $d$  drops below 10%, the misclassification rate increases (differently depending on the model)



**Fig. 9** Percentage of times the 3-segment model was the best fit according to the BIC, function of  $d$

models seem to have an advantage over the Poisson lognormal models in this situation when considering both criteria. Moreover, smoothed models with fewer knots seem to classify the days better.

All of these results obtained on simulated data demonstrate the value of considering “sums and shares” mixture models and smoothed versions to properly categorize count data subject to regime changes. Studying the models on real data will deepen

**Table 2** Experimentation with a local impact  $d = 10$ 

Model	Generation : PoiMult		Generation : PLNdiag	
	Misclass (%)	3-segments sel. (%)	Misclass (%)	3-segments sel. (%)
PoiMult	0.05	0	0.03	0.15
sPoiMult (5)	<b>0.01</b>	<b>1</b>	<b>0</b>	<b>1</b>
sPoiMult (20)	<b>0.01</b>	<b>1</b>	<b>0.01</b>	<b>1</b>
sPoiMult (50)	0.03	0.05	0.02	0
sPoiMult (80)	0.04	0	0.03	0
PLN	0.08	0	0.05	0
sPLN (5)	0.06	0	0.02	0
sPLN (20)	0.04	0	0.02	0
sPLN (50)	0.08	0	0.04	0
sPLN (80)	0.08	0	0.04	0

sPoiMult models with fewer knots perform better than the other models

The two best values per metric are in bold

these conclusions, valid in a simple low-noise case. The source code for the R script and application on simulated data is available at [https://github.com/pdenailly/segmentation\\_models](https://github.com/pdenailly/segmentation_models).

## 5.2 Experiments on real-world data

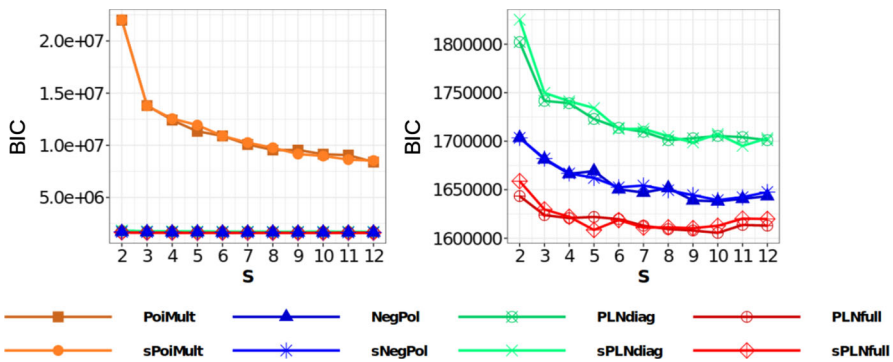
This section compares the different regression mixture models (see Table 1), using real data from the case study that motivated this work. The goal is to explore the models’ capacity on two bases: (i) their ability to fit the count data through the BIC criterion and (ii) their capacity to detect contiguous time segments, i.e. where temporally close days belong to the same segments, through the entropy criterion Celeux and Soromenho (1996). The criteria are computed for each model for a given number of segments  $S$ . The same exogenous factors  $\mathbf{x}_{j,t}$  are used for all models, they are listed in Table 3. As discussed in the section on simulated data, smoothed versions of each model will be introduced. These smoothed versions will consider  $M = 25$  knots, which is small compared to the number of days in the case study.

### 5.2.1 Comparison of fitting capacities

The BIC criterion computed for the different models is displayed in Fig. 10. As expected, the PLNfull and NegPol models, in their smoothed and unsmoothed versions, reach lower values of BIC criterion because they handle overdispersion as well as covariance (see Table 1), and should therefore be preferred. The PLNfull and sPLNfull models, however, appear to be better than NegPol and sNegPol by this criterion calculated over the entire period from Fig. 10. A second way to analyze the results of these four models is to calculate the criteria on periods that are known presumably

**Table 3** Explanatory variables

Position	Name	Description
$x_{j,t}^{1,\dots,8}$	Hour $_{j,t}$	bspline with 8 degrees of freedom on the 1-hour time slots from $t = 7$ am to $t = 12$ pm
$x_{j,t}^9$	ConcOut $_{j,t}$	Dummy variable for after a concert: 1 if there is a concert on day $j$ and $t$ is 11 pm
$x_{j,t}^{10}$	Concln $_{j,t}$	Dummy variable for before a concert: 1 if there is a concert on day $j$ and $t$ is between 4 pm and 10 pm
$x_{j,t}^{13}$	DisturbanceRERmorn $_{j,t}$	Log transformation of total duration of the RER disruption, during the morning peak (7 am to 9 am)
$x_{j,t}^{14}$	DisturbanceRER $_{j,t}$	Log transformation of total duration of the RER disruption, after the morning peak (9 am to 12 pm)



**Fig. 10** BIC criterion calculated for all mixed-membership models on the La Défense count dataset and for  $S \in (2, \dots, 12)$ . The figure on the left includes all the models. The figure on the right does not include the baseline models (PoiMult and sPoiMult) in order to better visualize the BIC of other models

homogeneous. Results about likelihood, BIC, and the number of segments detected are summarized in Table 4. The NegPol and sNegPol models are better here except for the noisy strike period. The likelihoods computed by the different models are close, but NegPol and sNegPol have smaller BIC values (except for the noisy strike period) because, as shown in Table 1, they require fewer parameters. In addition, fewer segments are needed with these models.

This last point brings us back to our objective of identifying a “cleaner” segmentation, i.e. with segments that are as continuous as possible. To measure this continuity, we introduce the entropy criterion (Celeux and Soromenho 1996) on the association weights  $\pi_s(j, \alpha)$ , which is computed as:

$$- \sum_s \sum_j \pi_s(j, \alpha) \log(\pi_s(j, \alpha)). \tag{13}$$

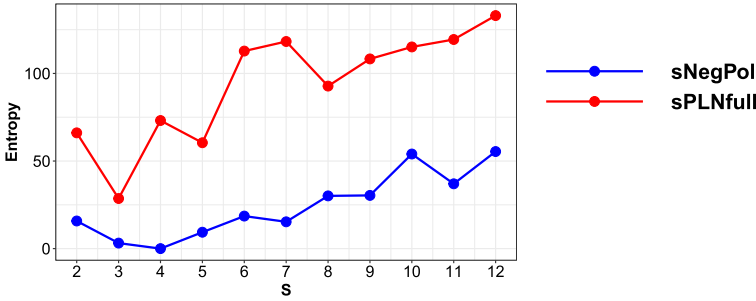
**Table 4** Log-likelihood, BIC criterion and number of segments detected on four well-known periods

Period	PLNfull		
	#seg	LL	BIC
Normal	5	-141565	316611
Strike	1	-21598	<b>49893</b>
Lockdown	2	-21256	55905
Tram works	2	-24432	62257
Period	sPLNfull		
	#seg	LL	BIC
Normal	5	-143043	321339
Strike	2	-20424	54948
Lockdown	2	-21618	57336
Tram works	2	-25407	64915
Period	NegPol		
	#seg	LL	BIC
Normal	4	-143829	304286
Strike	2	-22031	52376
Lockdown	2	-22163	<b>52641</b>
Tram works	1	-28007	60171
Period	sNegPol		
	#seg	LL	BIC
Normal	2	-144822	<b>298430</b>
Strike	2	-21732	52251
Lockdown	2	-22152	53091
Tram works	1	-27733	<b>59860</b>

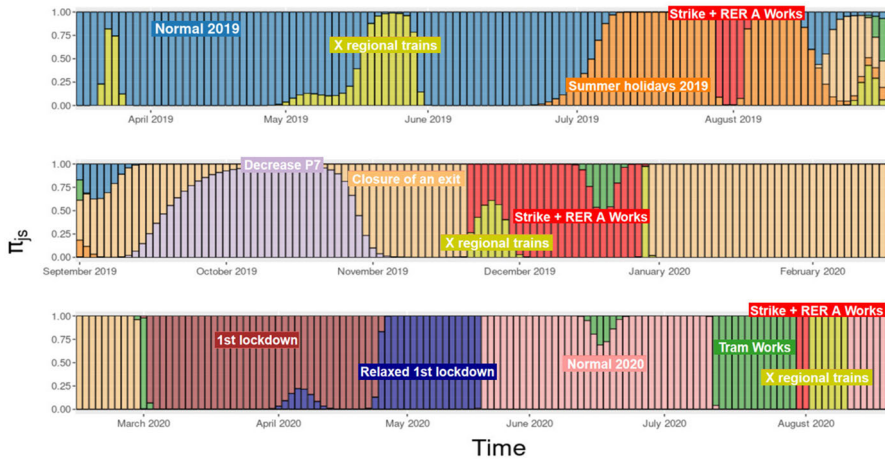
These periods are: a so-called normal period at the beginning of 2019, the period of the first lockdown against Covid19, a period of strikes and a period of works on the T2 tramway line  
The smallest BIC per period is in bold

Note that we can only calculate this criterion on smoothed models because of the dynamic evolution of  $\pi_s(j, \alpha)$ . The entropy values are displayed in Fig. 11 for the sNegPol and sPLNfull models. Entropies are smaller for the sNegPol model which highlights the capacity of this model to detect more continuous segments than the sPLNfull model.

Poisson log-normal models, by estimating variances within segments, allow more separate days to fall in the same segments, not necessarily continuous. Compared to Poisson log-normal models, “sums and shares” models seems to have a better



**Fig. 11** Entropy criterion calculated for sNegPol and sPLNfull mixed-membership models on the La Défense count dataset and for  $S \in (2, \dots, 12)$



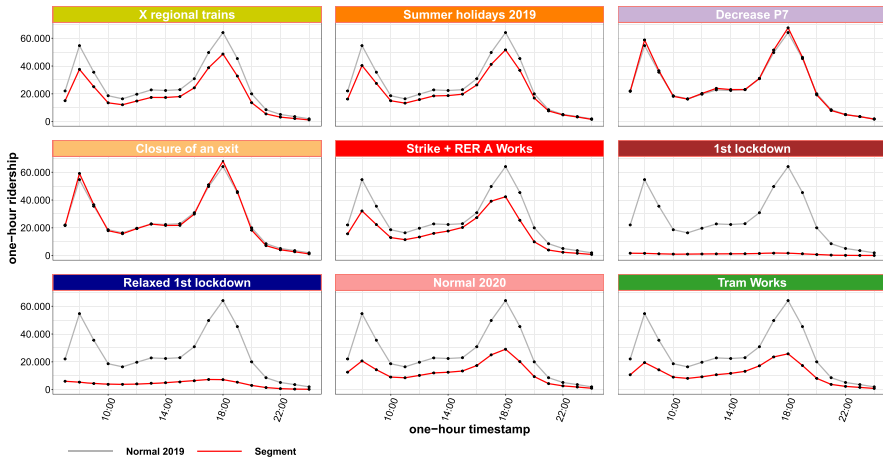
**Fig. 12** Bar plot representation of time segmentation. Each day is associated with the probabilities of belonging to each segment. Each of the  $S (= 10)$  segments has its own color and label

ability to summarize the data into continuous segments. NegPol and sNegPol provide a reasonable trade-off between the abilities to detect continuous segments, and to handle overdispersed and correlated count data. For these reasons and the several advantages found in the smoothed versions with simulated data, we will focus in the following section on the results associated with the smoothed mixture of Negative binomial sums and Pólya shares model (sNegPol) with  $S = 10$  segments (according to the BIC criterion).

### 5.2.2 Segmentation results on the chosen model

**Temporal segmentation** The temporal segmentation obtained with the model is shown in Fig. 12. We can observe a richness of segments induced by various context changes such as maintenance works, strikes, or health measures against the Covid19





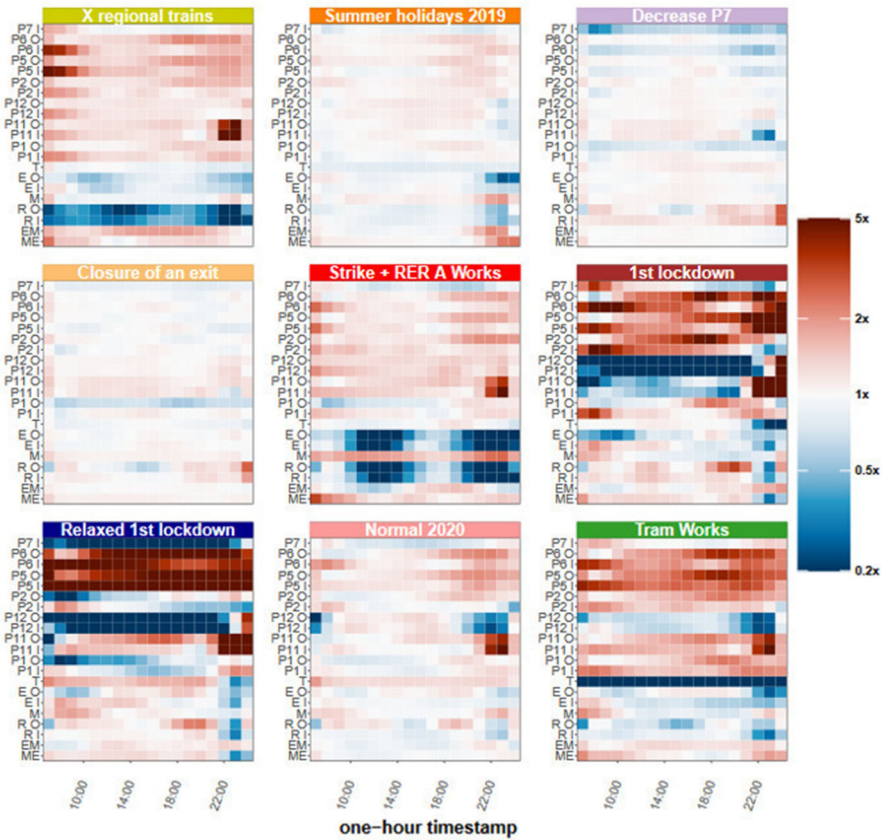
**Fig. 13** Typical profiles of total flows found in each segment  $s$ . Each profile is compared to that of *Normal 2019* i.e. the reference segment (in grey)

pandemic. We detail each segment in Table 5 of “Appendix C” (see the *Period characteristics* column)<sup>1</sup>. This diversity of segments, with few “returns”, underlines the need for urban operators to adapt to a regularly changing situation. We associate these segments with typical total flows in the hub  $\lambda_{j,t}^{(s)}$  in Fig. 13. Understandably, total flows have largely decreased since the beginning of the Covid19 pandemic, which is visible in all segments beyond the *First lockdown* segment. This result highlights that at the time of writing this paper, use of the hub has not returned to normal (i.e., “Normal 2019” and “Closure of an exit” in Fig. 12) since the beginning of the Covid19 pandemic.

**Typical distributions among the L locations** From a spatial point of view one can study the characteristic distribution of the flows of people in the transport hub within each segment. Indeed each segment  $s$  is associated with a set of typical distributions among the  $L$  locations  $\mathbf{u}_{j,t}^{(s)} = ((u_{j,t,l}^{(s)}))_{l \in \{1, \dots, L\}}$ , displayed in Fig. 14<sup>1,1</sup> Depending on the segments encountered, there is an overuse (in red) or an underuse (in blue) at certain locations compared to that of “Normal 2019”. For example in the 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 is stopped.

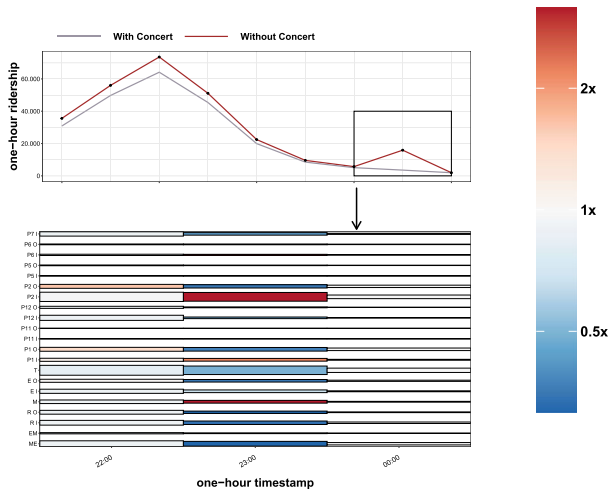
Considering both total flows (Fig. 13) and spatial distributions (Fig. 14), one can see that time segments are generally recognizable through both total profiles and spatial distributions. Two of them (i.e. *Decrease P7* and *Closure of an exit*), however, appear to be distinguished primarily through spatial distributions only. A description of the spatial distributions is presented in Table 5 in “Appendix C”.

<sup>1</sup> For these sections, only  $\mathbf{x}_{j,t}^{1, \dots, 8}$  (see Table 3) are used for computing the results, in order to exclude non-calendar effects. Thus, the total profiles and spatial distributions section are invariant by day.



**Fig. 14** Typical spatial distributions among the  $L (=21)$  locations found in each segment  $s$ . Refer to Fig. 1 for a description of the locations. As a reminder, “O” corresponds to an outgoing flow and “I” to an incoming flow. 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 of the reference segment, with regards to spatial distribution

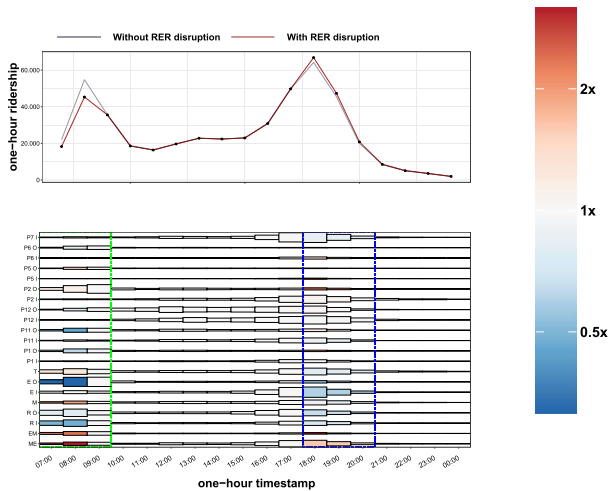
**Impact of non-calendar factors** The impact of exogenous factors is analyzed through a with- and without-factor comparison of typical profiles and spatial distributions. Note that we prioritized understanding the impact of these factors under normal conditions, i.e., within the *Normal 2019* segment. For each exogenous variable, we will study the standard profile of total flows with and without the impact of the atypical factor. For all variables, a heatmap is produced to compare, in the same segment, the distributions  $\mathbf{u}_{j,t}^{(s)}$  of the model built with the atypical factor and the  $\mathbf{u}_{j,t}^{(s)}$  of the model without that factor, using log-ratios. In these heatmaps, the height of each cell is proportional to the mean count at the corresponding location and time slot across the entire study period.



**Fig. 15** Typical profiles and spatial distribution. Comparison with and without a concert, within the *Normal 2019* segment

The impact of a concert on the use of “La Défense Grande Arche” station is displayed in Fig. 15. As expected, typical profiles show an increase in total ridership during the afternoon, as people arrive at the transport hub in anticipation of the concert. When a concert is held, there is also a peak of entries at 11pm, which represents people leaving the concert and taking public transportation. From a spatial point of view, entries to the hub close to the concert hall (P1 I and P2 I) are favored. The entry point to the metro line (M) is also more used, compared to the reference, which is not the case for the access to the express regional railway line (E I). We surmise that people leaving the concert who wish to take the express regional railway line will prefer another station (“Nanterre-Préfecture”), which is closer to the concert hall. People wishing to take the metro line have no alternative to the “La Défense Grande Arche” station, as it is the terminus, and are therefore more likely to be detected in our study.

Transport disruptions on the express regional railway line have a substantial effect on the transport hub use as shown in Fig. 16. The selected model highlights the delay phenomenon for morning arrivals (i.e., lower ridership on morning time slots). As expected, this phenomenon is not visible on a disrupted evening peak, as people are already present at the transport hub. There is an impact on transfers between the express regional railway and metro lines (i.e., ME and EM count locations). Disruption during the morning peak period tends to increase pedestrian flows between metro and RER (i.e., people arrive by metro, then exit the transport hub through the RER station). It also increases RER-to-metro access as people shift to the metro line to leave the hub and go to Paris. The RER line is strongly impacted, especially in the morning, due to the absence of all the users who failed to reach the transport hub.



**Fig. 16** Typical temporal profiles and spatial distributions. Comparison with and without express regional railway disruptions, within the *Normal 2019* segment. Two separate disruptions are modeled here: one during the morning peak (green rectangle) and one during the evening peak (blue rectangle) (color figure online)

## 6 Conclusion

This paper sets up a statistical model to segment multidimensional mobility time series, whose dynamics evolve according to characterized periods. Two strategies inspired from the literature, namely “sums and shares” models and “Poisson log-normal” models, are compared for this task, both in terms of likelihood and segment consistency. Each strategy has advantages and disadvantages. For example, the Poisson-Multinomial model cannot take into account overdispersions, nor correlations, which is not the case for the other models. The Negative Binomial - Dirichlet Multinomial model can take correlations into account, but they will always be positive or negative. The Poisson log-normal models seem to be more flexible and fit the observed data better (when only considering likelihoods). The “sums and shares” models seem to detect continuous segments better, which is more consistent with the reality of our case study. Moreover, there are benefits to using logistic regressions of spline functions to express the probability of each day belonging to which segment. This encoding seems to provide the model with a better capacity for detecting localized and/or low impact events. We chose to apply a smoothed 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. Furthermore, a set of atypical events was incorporated in the model for their impacts to be studied: we have thus considered concerts and public transport disruptions.

Operationally, 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 use of the hub over the long term. The study of the impact of atypical factors reveals how pedestrian flows react accordingly. We found that given situations, whether a time segment or an exogenous factor, may lead to specific over- and under-use of particular locations. This type of study is replicable to any situation where a large set of count data is available and where the aim is to synthesize information from typical spatio-temporal profiles from distinct periods. We are thinking of the field of mobility, extended to the study of a public transport network or a city where the characterization of human travel patterns is of great importance. For example, in the case of road traffic, this type of model could be used to search for typical traffic situations within a city, to help set up a traffic control system adapted to each type of situation. This type of problem can also emerge in ecology, genomics, or others. In the field of genomics, this could help to segment the expression dynamics of a group of genes and to identify within these segments subgroups of over or under expressed genes. In ecology, we could imagine an extension to the work proposed in Chiquet et al. (2021) by adding a temporal dimension. The models we used could segment the time into homogenous periods during which communities of different species are distributed similarly between various sites. For example, we might observe interdependent prey-predator communities whose abundances between different sites vary over a year. These models can be helpful to study the impact of covariates, isolated from the rest and potentially within different periods.

Further investigations are required to overcome some of the limitations of this type of modeling, including:

- An exogenous factor can either be explicitly coded in the model or left to be found by the model. This decision induces variability when constructing segments.
- It is necessary to have a sufficient quantity of exogenous data. Increasing the number of segments means that the exogenous factors are modeled in increasingly specific contexts, for which fewer data is available.

Furthermore, additional work could be done to include an autoregressive term in these models. This inclusion could help account for the intrinsic variability of each day in the model, potentially changing the allocation of days to segments (Ren and Barnett 2020). This study paves the way for 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 ridership in specific contexts.

**Code availability** R-code and simulated data used are available in the GitHub repository [https://github.com/pdenailly/segmentation\\_models](https://github.com/pdenailly/segmentation_models)

## Declarations

**Conflict of interest** The authors declared that they have no conflicts of interest.

### Appendix A: Mixture of sums and shares model estimation

Given  $z_{j,s} = 1$  and  $\mathbf{x}_{j,t}$ , the series  $\mathbf{y}_{j,t}$  are distributed according to the following mixture model:

$$p(\mathbf{y}_{j,t}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, r, \boldsymbol{\xi}) = \sum_{s=1}^S \pi_s(j; \boldsymbol{\alpha}) g(v_{j,t} | \mathbf{x}_{j,t}, \boldsymbol{\gamma}_s, r_s) h(\mathbf{y}_{j,t} | v_{j,t}, \mathbf{x}_{j,t}, \boldsymbol{\xi}_s), \tag{A1}$$

with  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_s)_{s=1, \dots, S}$ ,  $r = (r_s)_{s=1, \dots, S}$  and  $\boldsymbol{\xi} = (\boldsymbol{\xi}_s)_{s=1, \dots, S}$ . The parameters of the model are estimated with the Expectation Maximization (EM) algorithm (Dempster et al. 1977) which requires a complete data log-likelihood maximization. The complete data log-likelihood can be written:

$$\mathcal{L}_c(\boldsymbol{\alpha}, \boldsymbol{\gamma}, r, \boldsymbol{\xi}) = \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T z_{j,s} \log(\pi_s(j; \boldsymbol{\alpha}) g(v_{j,t} | \mathbf{x}_{j,t}, \boldsymbol{\gamma}_s, r_s) h(\mathbf{y}_{j,t} | v_{j,t}, \mathbf{x}_{j,t}, \boldsymbol{\xi}_s)). \tag{A2}$$

Given the initial value of the parameters  $\boldsymbol{\xi}^{(0)}$ ,  $\boldsymbol{\gamma}^{(0)}$ ,  $r^{(0)}$  and  $\boldsymbol{\alpha}^{(0)}$ , the following two steps are repeated until convergence.

- *Expectation step (E)* The expectation of the completed log-likelihood is evaluated knowing the observed data  $Y$  and the set of current parameters:  $\boldsymbol{\xi}^{(c)}$ ,  $\boldsymbol{\gamma}^{(c)}$ ,  $r^{(c)}$  and  $\boldsymbol{\alpha}^{(c)}$ .

$$Q(\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\xi}^{(c)}) = \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T E_{\boldsymbol{\xi}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\alpha}^{(c)}} [z_{j,s} | Y] \tag{A3}$$

$$\log \left( \pi_s(j; \boldsymbol{\alpha}^{(c)}) g(v_{j,t} | \mathbf{x}_{j,t}, \boldsymbol{\gamma}_s^{(c)}, r_s^{(c)}) h(\mathbf{y}_{j,t} | v_{j,t}, \mathbf{x}_{j,t}, \boldsymbol{\xi}_s^{(c)}) \right), \tag{A4}$$

where

$$E_{\boldsymbol{\xi}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\alpha}^{(c)}} [z_{j,s} | Y] = \tau_{j,s}^{(c)} \tag{A5}$$

$$= \frac{\pi_s(j; \boldsymbol{\alpha}^{(c)}) \prod_T g(v_{j,t} | \mathbf{x}_{j,t}, \boldsymbol{\gamma}_s^{(c)}, r_s^{(c)}) h(\mathbf{y}_{j,t} | \mathbf{x}_{j,t}, v_{j,t}, \boldsymbol{\xi}_s^{(c)})}{\sum_{s'} \pi_{s'}(j; \boldsymbol{\alpha}^{(c)}) \prod_T g(v_{j,t} | \mathbf{x}_{j,t}, \boldsymbol{\gamma}_{s'}^{(c)}, r_{s'}^{(c)}) h(\mathbf{y}_{j,t} | \mathbf{x}_{j,t}, v_{j,t}, \boldsymbol{\xi}_{s'}^{(c)})}. \tag{A6}$$

The a posteriori probabilities that each day  $j$  belongs to segment  $s$ ,  $\tau_{j,s}^{(c)}$ , are updated at each iteration of step E.

- *Maximization step (M)* Parameters  $\boldsymbol{\xi}^{(c+1)}$ ,  $\boldsymbol{\gamma}^{(c+1)}$ ,  $r^{(c+1)}$  and  $\boldsymbol{\alpha}^{(c+1)}$  that maximize  $Q(\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\xi}^{(c)})$  are calculated. It is possible to rewrite this quantity as:

$$Q(\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\xi}^{(c)}) = Q_1(\boldsymbol{\alpha}^{(c)}) + Q_2(\boldsymbol{\gamma}^{(c)}, r^{(c)}) + Q_3(\boldsymbol{\xi}^{(c)}) \tag{A7}$$

where

$$Q_1(\boldsymbol{\alpha}^{(c)}) = \sum_{s=1}^S \sum_{j=1}^J \tau_{j,s}^{(c)} \log(\pi_s(j; \boldsymbol{\alpha}^{(c)})) \tag{A8}$$

$$Q_2(\boldsymbol{\gamma}^{(c)}, r^{(c)}) = \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T \tau_{j,s}^{(c)} \log(g(v_{j,t} | \mathbf{x}_{j,t}, \boldsymbol{\gamma}_s^{(c)}, r_s^{(c)})) \tag{A9}$$

$$Q_3(\boldsymbol{\xi}^{(c)}) = \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T \tau_{j,s}^{(c)} \log(h(\mathbf{y}_{j,t} | \mathbf{x}_{j,t}, v_{j,t}, \boldsymbol{\xi}_s^{(c)})). \tag{A10}$$

The maximisation of  $Q_1$  consists in solving a weighted multinomial logistic regression. New values of  $\boldsymbol{\alpha}$  can be found using iterative procedures such as iteratively reweighted least squares (IRLS) (Holland and Welsch 1977). This problem is solved with the multinom function of the nnet package (Ripley et al. 2016).  $Q_2$  is the log-likelihood corresponding to a negative binomial generalized linear model. Its maximisation is solved through an alternating iteration process provided by the glm.nb function in the MASS package (Ripley et al. 2013). Within each segment  $s$ , for a given value of  $r_s^{(c)}$  the linear model is fitted using an IRLS method. Next for fixed found  $\boldsymbol{\gamma}_s^{(c)}$  parameters, the  $r_s^{(c)}$  parameter is estimated with score and information iterations. The two steps are alterned until convergence and  $\boldsymbol{\gamma}_s^{(c+1)}$  and  $r_s^{(c+1)}$  are found. Note that  $\tau_{j,s}^{(c)}$  are here used as prior weights in the fitting process. The criterion  $Q_3$ , which is associated to a weighted Dirichlet multinomial regression model, is solved with the MGLM package (Kim et al. 2018). Because Dirichlet multinomial distribution does not belong to the exponential family, IRLS method is not used, as the expected information matrix is difficult to calculate. The method used here combine the minorization-maximization (MM) (Lange et al. 2000) algorithm and the Newton’s method. MM and Newton updates are computed at each iteration and the one with the higher log-likelihood is chosen.

### Appendix B: Poisson log-normal mixture model estimation

The series  $y_{j,t}$  are distributed according to the following mixture model:

$$p(\mathbf{y}_{j,t}; \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\Sigma}) = \sum_{s=1}^S \pi_s(j; \boldsymbol{\alpha}) \int_{R^P} \left[ \prod_{p=1}^P g(\mathbf{y}_{j,t,p} | \theta_{j,t,p}) \right] m(\theta_{j,t} | \boldsymbol{\rho}_s, \boldsymbol{\Sigma}_s) d\theta_{j,t}, \tag{B11}$$

with  $\boldsymbol{\rho} = (\boldsymbol{\rho}_s)_{s=1,\dots,S}$  and  $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_s)_{s=1,\dots,S}$ .  $g$  is a Poisson distribution and  $m$  a Gaussian distribution function. The EM algorithm can be used for parameter estimation but finding the expected value of the complete data log-likelihood requires estimating the conditional expectations  $\mathbb{E}(Z_{js} \theta_{j,t} | \mathbf{y}_{j,t}, \boldsymbol{\rho}_s, \boldsymbol{\Sigma}_s)$  and  $\mathbb{E}(Z_{js} \theta'_{j,t} | \mathbf{y}_{j,t}, \boldsymbol{\rho}_s, \boldsymbol{\Sigma}_s)$

which are intractable. These conditional expectations can be calculated with an EM algorithm coupled with a Markov chain Monte Carlo (MCMC-EM) algorithm as presented in the work of Silva et al. (2019), which however comes with a heavy calculation load. We refer instead to the work presented by Chiquet et al. (2019) that uses variational approximation which is an approximate inference technique. The idea behind variational inference is to use Gaussian densities and approximate complex posterior distributions by minimizing the Kullback–Leibler divergence between the true  $p(\theta)$  and approximating densities  $q(\theta)$ . The marginal log-likelihood for  $\mathbf{y}_{j,t}$  can be written as

$$\log p(\mathbf{y}_{j,t}) = F(q(\theta_{j,t}), \mathbf{y}_{j,t}) + D_{KL}(q(\theta_{j,t})|p(\theta_{j,t})), \tag{B12}$$

with  $D_{KL}(q(\theta_{j,t})|p(\theta_{j,t}))$  the Kullback-Leibler divergence between  $p(\theta_{j,t})$  and  $q(\theta_{j,t})$ .  $F(q(\theta_{j,t}), \mathbf{y}_{j,t})$  is the expression of the variational lower bound of the log-likelihood. This is the criterion that we aim to maximize in the parameter estimation process. In the case of the Poisson-Lognormal model,  $q$  is assumed to be a Gaussian distribution:

$$q(\theta_{j,t}; \mathbf{m}_{j,t}, \mathbf{S}_{j,t}) = \mathcal{N}(\theta_{j,t}; \mathbf{m}_{j,t}, \mathbf{S}_{j,t}), \tag{B13}$$

with  $\mathbf{m}_{j,t}$  and  $\mathbf{S}_{j,t} = \text{diag}(\mathbf{S}_{j,t})$  the variational parameters associated with sample  $\mathbf{y}_{j,t}$  at day  $j$  and time slot  $t$ . To minimize the Kullback–Leibler divergence, the variational lower bound has to be maximized. The complete data log-likelihood can be written as follows:

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, \mathbf{m}, \mathbf{S}) = & \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T z_{j,s} \log(\pi_s(j; \boldsymbol{\alpha})) + \\ & \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T z_{j,s} [F(q^{(s)}(\theta_{j,t}), \mathbf{y}_{j,t}) + D_{KL}(q^{(s)}(\theta_{j,t})|p^{(s)}(\theta_{j,t}))], \end{aligned} \tag{B14}$$

where  $D_{KL}(q^{(s)}(\theta_{j,t})|p^{(s)}(\theta_{j,t}))$  is the Kullback–Leibler divergence between  $p(\theta_{j,t}|\mathbf{y}_{j,t}, z_j = s)$  and  $q^{(s)}(\theta_{j,t})$  with  $q^{(s)}(\theta_{j,t}) = \mathcal{N}(\mathbf{m}_{j,t}^{(s)}, \mathbf{S}_{j,t}^{(s)})$ . And the variational lower bound of the log-likelihood for each observation  $\mathbf{y}_{j,t}$  is

$$\begin{aligned} F(q^{(s)}(\theta_{j,t}), \mathbf{y}_{j,t}) = & \frac{1}{2} \log |\mathbf{S}_{j,t}^{(s)}| - \frac{1}{2} (\mathbf{m}_{j,t}^{(s)} - \mathbf{x}_{j,t}^T \boldsymbol{\rho}_s)' \boldsymbol{\Sigma}_s^{-1} (\mathbf{m}_{j,t}^{(s)} - \mathbf{x}_{j,t}^T \boldsymbol{\rho}_s) - \text{tr}(\boldsymbol{\Sigma}_s^{-1} \mathbf{S}_{j,t}^{(s)}) - \\ & \frac{1}{2} \log |\boldsymbol{\Sigma}_s| - \frac{P}{2} + (\mathbf{m}^{(s)})'_{j,t} \mathbf{y}_{j,t} - \sum_{p=1}^P (\exp(m_{j,t,p}^{(s)} + \frac{1}{2} s_{j,t,p}^{(s)}) + \log(y_{j,t,p}!)). \end{aligned} \tag{B15}$$

The EM algorithm is used to estimate the parameters and the following two steps are repeated until convergence.



- *Expectation step (E)* The expectation of the completed log-likelihood is evaluated knowing the observed data  $Y$ , the set of current parameters  $\boldsymbol{\rho}^{(c)}$ ,  $\boldsymbol{\Sigma}^{(c)}$  and  $\boldsymbol{\alpha}^{(c)}$  and variational parameters  $\mathbf{m}_{j,t}^{(c)}$ ,  $\mathbf{S}_{j,t}^{(c)}$ .

$$\begin{aligned}
 Q(\boldsymbol{\rho}^{(c)}, \boldsymbol{\Sigma}^{(c)}, \boldsymbol{\alpha}^{(c)}, \mathbf{m}^{(c)}, \mathbf{S}^{(c)}) &= \sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T \tau_{j,s}^{(c)} \log(\pi_s(j; \boldsymbol{\alpha}^{(c)})) + \\
 &\sum_{s=1}^S \sum_{j=1}^J \sum_{t=1}^T \tau_{j,s}^{(c)} E_{\boldsymbol{\rho}^{(c)}, \boldsymbol{\Sigma}^{(c)}, \boldsymbol{\alpha}^{(c)}, \mathbf{m}_{j,t}^{(c)}, \mathbf{S}_{j,t}^{(c)}} [F(q^{(s)}(\theta_{j,t}), \mathbf{y}_{j,t}) + \\
 &D_{KL}(q^{(s)}(\theta_{j,t}) | p^{(s)}(\theta_{j,t}))],
 \end{aligned}
 \tag{B16}$$

with  $\tau_{j,s}^{(c)} = E_{\boldsymbol{\rho}^{(c)}, \boldsymbol{\Sigma}^{(c)}, \boldsymbol{\alpha}^{(c)}, \mathbf{m}_{j,t}^{(c)}, \mathbf{S}_{j,t}^{(c)}} [z_{j,s} | Y]$ . The variational lower bound of the log-likelihood is used to approximate  $\tau_{j,s}^{(c)}$ :

$$\tau_{j,s}^{(c)} = \frac{\pi_s(j; \boldsymbol{\alpha}^{(c)}) \prod_{t=1}^T \exp(F(q^{(s)}(\theta_{j,t}), \mathbf{y}_{j,t}))}{\sum_{h=1}^S \pi_h(j; \boldsymbol{\alpha}^{(c)}) \prod_{t=1}^T \exp(F(q^{(h)}(\theta_{j,t}), \mathbf{y}_{j,t}))}.
 \tag{B17}$$

Note that this approximation is used in the R package `PLNmodels`.

- *Maximization step (M)* The maximization step is divided into two parts:
  - Conditionally on  $\boldsymbol{\rho}_s$  and  $\boldsymbol{\Sigma}_s$  and given  $\tau_{j,s}$ , variational parameters  $\mathbf{m}_{j,t}^{(c)}$  and  $\mathbf{S}_{j,t}^{(c)}$  are updated. Because  $F(q^{(s)}(\theta_{j,t}), \mathbf{y}_{j,t})$  is strictly concave with respect to  $\mathbf{m}_{j,t}^{(c)}$  and  $\mathbf{S}_{j,t}^{(c)}$ , it is possible to obtain  $\mathbf{S}_{j,t}^{(c+1)}$  with the fixed-point method and  $\mathbf{m}_{j,t}^{(c+1)}$  with Newton’s method.
  - Knowing  $\tau_{j,s}^{(c)}$ ,  $\mathbf{m}_{j,t}^{(c+1)}$  and  $\mathbf{S}_{j,t}^{(c+1)}$  parameters  $\boldsymbol{\rho}^{(c+1)}$ ,  $\boldsymbol{\Sigma}^{(c+1)}$  and  $\boldsymbol{\alpha}^{(c+1)}$  are obtained.

### Appendix C: Description of the time segments

See Table 5.

**Table 5** Time segmentation

Name	Period characteristics	Total flows and spatial distribution
Normal 2019	Early 2019 excluding school holidays. Reference period	
X regional trains	Brings together periods during which the exchanges between express regional railway and regional train lines decrease strongly. A slope failure in June 2019 interrupting the “U” line, a strike in december 2019 and extensive maintenance works on the RER line in August 2020	Strong loss of incoming and outgoing flows to and from regional trains (R) access
Summer holidays 2019	July and August 2019	Decrease in total flows. No major differences in spatial distributions
Closure of an exit	Closure of an exit from regional trains to esplanade	No changes in total flows. For spatial distributions the differences are barely visible except for a decrease in the number of passages through the P1 exit and some modifications to the R access
Decrease P7	The conditions here are the same as those of the <i>Closure of an exit</i> segment	No differences for total flows. Unexplained decrease in the use of exit P7; this may be a period during which the counting sensor had a technical problem. The other locations suffer the same consequences as in the <i>Closure of an exit</i> segment
Strike + RER A Works	Period characterized by a massive mobilization against the reform of the French pension system and some periods during which maintenance works occurred on RER line in Paris	Strong decrease in express regional railway use (i.e. incoming, outgoing) and regional train use due to their partial operation or to lack of interchanges
First lockdown	First lockdown period due to the Covid19 pandemic	Almost total loss of total flows. Westfield (P12) access closed and not used as a result. Because of the very small number of people who visited the hub during this period, it is difficult to link strong changes in the use of other accesses with actual changes in route choice behavior

**Table 5** (continued)

Name	Period characteristics	Total flows and spatial distribution
Relaxed first lockdown	First stage of lifting lockdown, with restrictions	Almost total loss of total flows. Total loss at the Westfield mall (P12) and central pedestrian esplanade (P7) access points. Consequent increase in the use of both northern (P5) and southern (P6) access points to the esplanade
Normal 2020	Period without curfew or lockdown in 2020, a “back-to-normal” period	Strong loss of total flows. Decrease in use of Westfield shopping mall (P12) accesses in the evening
Tram Works	Periods of maintenance works on the tramway line	Total loss of incoming flows to tramway (T) access

Description of typical profiles. Descriptions are made in comparison of the *Normal 2019* reference profile

## References

- Agard B, Morency C, Trépanier M (2006) Mining public transport user behaviour from smart card data. *IFAC Proc Vol 39*(3):399–404
- Aitchison J, Ho C (1989) The multivariate poisson-log normal distribution. *Biometrika* 76(4):643–653
- Bai J (2010) Common breaks in means and variances for panel data. *J Econom* 157(1):78–92
- Baid U, Talbar S (2016) Comparative study of k-means, gaussian mixture model, fuzzy c-means algorithms for brain tumor segmentation. In: International conference on communication and signal processing 2016 (ICCASP 2016), Atlantis Press, pp 583–588
- Balzotti C, Bragagnini A, Briani M et al (2018) Understanding human mobility flows from aggregated mobile phone data. *IFAC-PapersOnLine* 51(9):25–30
- Bouveyron C, Celeux G, Murphy TB et al (2019) Model-based clustering and classification for data science: with applications in R, vol 50. Cambridge University Press, Cambridge
- Briand AS, Côme E, Trépanier M et al (2017) Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transp Res Part C Emerg Technol* 79:274–289
- Briand AS, Come E, Khoudjia M, et al (2019) Detection of atypical events on a public transport network using smart card data. In: European transport conference 2019 Association for European Transport (AET)
- Cecaj A, Lippi M, Mamei M et al (2021) Sensing and forecasting crowd distribution in smart cities: Potentials and approaches. *IoT* 2(1):33–49
- Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. *J Classif* 13(2):195–212
- Chiquet J, Robin S, Mariadassou M (2019) Variational inference for sparse network reconstruction from count data. In: International conference on machine learning, PMLR, pp 1162–1171
- Chiquet J, Mariadassou M, Robin S (2021) The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Front Ecol Evol* 9:188
- Côme E, Oukhellou L (2014) Model-based count series clustering for bike sharing system usage mining: a case study with the vélib’ system of paris. *ACM Trans Intell Syst Technol(TIST)* 5(3):1–21
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Ser B (Methodol)* 39(1):1–22
- Fernández-Ares A, Mora A, Arenas MG et al (2017) Studying real traffic and mobility scenarios for a smart city using a new monitoring and tracking system. *Futur Gener Comput Syst* 76:163–179
- Ghaemi MS, Agard B, Trépanier M et al (2017) A visual segmentation method for temporal smart card data. *Transp A Transp Sci* 13(5):381–404
- Hilbe JM (2011) Negative binomial regression. Cambridge University Press, Cambridge

- Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. *Commun Statistics-theory Methods* 6(9):813–827
- Jones M, Marchand É (2019) Multivariate discrete distributions via sums and shares. *J Multivar Anal* 171:83–93
- Kim J, Zhang Y, Day J et al (2018) Mglm: an r package for multivariate categorical data analysis. *R J* 10(1):73
- Kristoffersen MS, Dueholm JV, Gade R et al (2016) Pedestrian counting with occlusion handling using stereo thermal cameras. *Sensors* 16(1):62
- Lange K, Hunter DR, Yang I (2000) Optimization transfer using surrogate objective functions. *J Comput Graph Stat* 9(1):1–20
- Lashkari D, Golland P (2007) Convex clustering with exemplar-based models. *Adv Neural Inf Process Syst* 20
- Li J, Zheng P, Zhang W (2020) Identifying the spatial distribution of public transportation trips by node and community characteristics. *Transp Plan Technol* 43(3):325–340
- Li Y, Rahman T, Ma T et al (2021) A sparse negative binomial mixture model for clustering rna-seq count data. *Biostatistics* 24(1):68–84
- Magidson J, Vermunt J (2002) Latent class models for clustering: a comparison with k-means. *Can J Marketing Res* 20(1):36–43
- Manley E, Zhong C, Batty M (2018) Spatiotemporal variation in travel regularity through transit user profiling. *Transportation* 45(3):703–732
- McLachlan GJ, Lee SX, Rathnayake SI (2019) Finite mixture models. *Annu Rev Stat Appl* 6:355–378
- Mohamed K, Côme E, Oukhellou L et al (2016) Clustering smart card data for urban mobility analysis. *IEEE Trans Intell Transp Syst* 18(3):712–728
- Mützel CM, Scheiner J (2021) Investigating spatio-temporal mobility patterns and changes in metro usage under the impact of covid-19 using taipei metro smart card data. *Public Transp* 1–24
- de Nailly P, Côme E, Samé A et al (2021) What can we learn from 9 years of ticketing data at a major transport hub? a structural time series decomposition. *Transp A Transp Sci* 18(3):1445–1469
- Pavlyuk D, Spiridovska N, Yatskiv I (2020) Spatiotemporal dynamics of public transport demand: a case study of riga. *Transport* 35(6):576–587
- Peláez G, Bacara D, de la Escalera A, et al (2015) Road detection with thermal cameras through 3d information. In: 2015 IEEE intelligent vehicles symposium (IV), IEEE, pp 255–260
- Peyhardi J, Fernique P, Durand JB (2021) Splitting models for multivariate count data. *J Multivar Anal* 181(104):677
- Ren B, Barnett I (2020) Autoregressive mixture models for serial correlation clustering of time series data. arXiv preprint [arXiv:2006.16539](https://arxiv.org/abs/2006.16539)
- Ripley B, Venables B, Bates DM et al (2013) Package ‘mass’. *Cran r* 538:113–120
- Ripley B, Venables W, Ripley MB (2016) Package ‘nnet’. *R Package Version* 7(3–12):700
- Ronchi E, Scozzari R, Fronterre M (2020) A risk analysis methodology for the use of crowd models during the covid-19 pandemic. *LUTVDG/TVBB* (3235)
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 461–464
- Sibuya M, Yoshimura I, Shimizu R (1964) Negative multinomial distribution. *Ann Inst Stat Math* 16(1):409–426
- Silva A, Rothstein SJ, McNicholas PD et al (2019) A multivariate poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC Bioinf* 20(1):1–11
- Singh U, Determe JF, Horlin F et al (2020) Crowd forecasting based on wifi sensors and lstm neural networks. *IEEE Trans Instrum Meas* 69(9):6121–6131
- Toqué F, Côme E, Oukhellou L, et al (2018) Short-term multi-step ahead forecasting of railway passenger flows during special events with machine learning methods. In: CASPT 2018, conference on advanced systems in public transport and transitdata 2018, p 15
- Truong C, Oudre L, Vayatis N (2020) Selective review of offline change point detection methods. *Signal Process* 167(107):299
- Wang Z, Liu H, Zhu Y et al (2021) Identifying urban functional areas and their dynamic changes in beijing: Using multiyear transit smart card data. *J Urban Plan Dev* 147(2):04021002
- Winkelmann R (2008) *Econometric analysis of count data*. Springer Science and Business Media, Berlin
- Zhang Y, Zhou H, Zhou J et al (2017) Regression models for multivariate count data. *J Comput Graph Stat* 26(1):1–13

- Zhong C, Manley E, Arisona SM et al (2015) Measuring variability of mobility patterns from multiday smart-card data. *J Comput Sci* 9:125–130
- Zhou M, Hannah L, Dunson D, et al (2012) Beta-negative binomial process and poisson factor analysis. In: *Artificial intelligence and statistics*, PMLR, pp 1462–1471

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.