Neurocomputing 141 (2014) 124–138

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Clustering the Vélib' dynamic Origin/Destination flows using a family of Poisson mixture models

Andry Njato Randriamanamihaga [a,*], Etienne Côme [a], Latifa Oukhellou [a], Gérard Govaert [b]

[a] Université Paris-Est, IFSTTAR, COSYS-GRETTIA, F-77447 Marne-la-Vallée, France
[b] UMR CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, F-60205 Compiègne, France

ARTICLE INFO

*Article history:*
Received 28 June 2013
Received in revised form
11 November 2013
Accepted 6 January 2014
Available online 26 April 2014

*Keywords:*
Generative model-based clustering
Dynamic Origin/Destination matrix
Count time series clustering
Urban mining
Human mobility analysis
Vélib' bike sharing system

ABSTRACT

Studies on human mobility, including Bike Sharing System Analysis, have expanded over the past few years. They aim to give insight into the underlying urban phenomena linked to city dynamics and generally rely on data-mining tools to extract meaningful patterns from the huge volume of data recorded by such complex systems. This paper presents one such tool through the introduction of a family of generative models based on Poisson mixtures to automatically analyse and find temporal-based clusters in Origin/Destination flow-data. Such an approach may provide latent factors that reveal how regions of different usage interact over time. More generally, the proposed methodology can be used to cluster edges of temporal valued-graphs with respect to their temporal profiles and is thus particularly suited to mine patterns in dynamic Origin/Destination matrices commonly encountered in the field of transport. An in-depth analysis of the results of the proposed models was carried out on two months of trips data recorded on the Vélib' Bike-Sharing System of Paris to validate the approach.

© 2014 Elsevier B.V. All rights reserved.

## 1. Introduction

Over the past 10 years, new urban mobility policies have been deployed as an answer to the soaring population, urban densification and the need for safe, clean and sustainable public transport. New strategies that aim to reduce traffic congestion and car dependence have therefore been adopted. One of them is the promotion of soft modes of transport in urban areas such as walking and cycling, the latter as part of a Bike Sharing System.

A Bike Sharing System, often denoted BSS, is a public and individual transport scheme which provides users with rented bikes, available 24/7 all year long. The bikes are available throughout the urban areas in strategically located stations. A typical use of the system corresponds to a user taking a bike at a station close to their starting point and returning it later close to their destination. This particular usage can be seen as an individual trip, or can serve a multi-modal journey – frequent in megacities – as a vital *last-mile* connection.

Since the system is fully automated, obtaining digital footprints that describe the dynamics of people's moves has become a relatively easy task. Fine-grained, informative and sizeable Origin/Destination (denoted O/D) trips or station states are made available and have motivated the study of such systems over the past few years. Some studies focus on resource redistribution [22] while others concentrate on usage patterns [10]. In the latter case, the authors usually focus on capturing the underlying urban phenomena linked to city dynamics. Automatic algorithms that give a clear, concise and reliable view of system-usage in relation with the city, of the *how*, *when and why people move*, are consequently being developed.

This paper presents such a family of automatic algorithms adapted to the specificities of the Vélib' usage, based on statistical models that automatically clusters the observed Origin/Destination flows between BSS stations. The approach that has been adopted, based on count-time series clustering, follows and refines the work initiated in [9,10], which relies on temporal profile analysis to cluster BSS stations and find the usages which best describe the *pulse of the city*. It also shares the main objectives highlighted by [32], whose aim is to find functional areas in a city through the mining of taxicab mobility data. The methodology proposed in this paper differs slightly from the latter as we introduce a family of generative dedicated models of count time-

* Corresponding author at: Components and Systems (COSYS) Departement, Engineering of Surface Transportation networks and Advanced Computing (GRETTIA) Laboratory, IFSTTAR - Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, 14 - 20 Boulevard Newton Cité Descartes, CHamps-sur-Marne, F - 77447 Marne-la-Vallée, Cedex 2. Tel.: +33 1 81 66 83 98.
*E-mail addresses:* andry.randriamanamihaga@ifsttar.fr (A.N. Randriamanamihaga), etienne.come@ifsttar.fr (E. Côme), latifa.oukhellou@ifsttar.fr (L. Oukhellou), gerard.govaert@utc.fr (G. Govaert).

http://dx.doi.org/10.1016/j.neucom.2014.01.050
0925-2312/© 2014 Elsevier B.V. All rights reserved.

series analysis, using adapted Poisson mixtures which describe the O/D flow dynamics between pairs of stations and find temporal-based clusters over the Vélib' large-scale transit data. An *Expectation–Maximization* (denoted EM) algorithm is derived to learn the model parameters and to cluster the O/D pairs. Such O/D records may highlight both common and unfamiliar interactions and exchanges between the pairs of stations over time. The crossing of the model results with sociological data is carried out to this end, and shows the close links between the neighbourhood of the stations and their associated usage profiles. The proposed approach is finally validated on the O/D data collected from the Vélib' Bike Sharing System of Paris.

The paper is organized as follows. Section 2 surveys previous work on transit data analysis, using data-mining approaches with a focus on BSS. Related work on Poisson mixture models for count-series clustering is also discussed in this section. Section 3 presents the key characteristics of the Vélib' Bike Sharing System of Paris, and discusses the features to be taken into account in Section 4, which introduces the proposed family of statistical model based upon count-series clustering, thereafter applied on the Vélib' transit data. The results are presented and discussed in Section 5.

## 2. Related work

As the result of greater use of Information and Communication Technologies in everyday life, more and more people carry passive urban mobile sensors such as mobile phones, GPS traces and ticketing data in public transport systems; this has led to a huge increase in the quality and quantity of data related to human mobility. This kind of data, not initially designed for modelling purposes, can be used to develop new approaches to urban mobility studies based on data mining. Unlike the traditional studies on mobility mainly based on travel surveys, these alternative approaches benefit from both the temporal dimension of the data – which is collected within a long temporal window – and the spatial resolution of this kind of information across the city. In the following, we position our work within the context of transport and data mining. After an overview of the relevant literature concerning the mining of transport data, specific count-time series based clustering models, applied on Bike Sharing System data, are given.

### 2.1. Mining of transport data

In recent years, studies using numerical traces to obtain an insight into the pulse of a city have attracted more and more attention. The availability of such data in the field of transport is becoming a privileged way to propose new frameworks that analyse how citizens move around. In the area of road transport, GPS traces of taxicabs have for example been used to estimate real-traffic information [11] or to detect traffic issues [33]. Another case study is presented by [34], where GPS trajectories of taxicabs have been used to detect flawed urban planning in a city.

In the context of public transport, the widespread use of automated fare collection systems has similarly been used in innovative studies on human mobility. The most frequent issues concern user characterization, user classification, network planning and demand forecasting [1,23]. Advanced analysis of the large volume of trips generated by these systems helps not only to identify cultural and geographic aspects of the city, such as its polycentric structure [26], but also to detect urban mobility patterns [17]. Finally, some applications, such as those proposed by [14], which rely on a better understanding of the travellers behaviour, are more user-oriented and can be used to provide personalized traveller information.

### 2.2. Count-series clustering

We now introduce a short survey of the relevant literature concerning shared bicycles, urban sensing and count-series clustering. Different issues have been addressed in the various case studies of Bike Sharing Systems, including the concerns of *users* (who want to find available bikes/docking points), *transport operators* (who have to deal with the problem of balancing load across stations) and *urban planners* (who have to decide how to design social space). The first topic investigated in [2,6,22,16] is that of the optimization of bike redistribution policies. Other studies address the problem of forecasting stations or, more globally, network usage in either the short term or the long term, such as in [10,3,12,20,30].

The third topic, which our work takes on board, is clustering. The aim is to uncover spatio-temporal patterns in the system usage, thus highlighting the relationships between time of day, location and usage. Clustering can be done on the basis of station occupancy statistics such as the number of available bicycles and free slots throughout the day for every station, such as in [10,20,12,14]. It can also be done, as is the case here, from trips data (or count data) taking the form of Origin/Destination information of anonymized individual users [3–5,31,30].

In the following paragraph, although we are interested in finding distinct usages of the system using count data, the proposed approach presents notable differences from previous studies. One of the key differences concerns the nature of the data. Whereas [10,14] use station occupancy, which is basically the number of available bicycles and the number of vacant parking slots, we compute displacement flows based on the collection of O/D trips generated at each use of the system, for each hour of each day. Our second contribution concerns the specification of our main model, which directly handles the differences in behaviour observed during weekdays and at weekends. It differs from [4,30], who also handle this temporal trend, but through data preprocessing and feature construction. Finally, the key difference with most of these previous studies resides in the objective of the clustering since many of these previous studies deal with station clustering while this paper will focus on O/D pair clustering.

From a methodological point of view, the proposed model based on Poisson Mixtures also differs from the literature on this topic, such as [28,13]. The proposed model family incorporates individual effects as in [25] where Poisson Mixtures with individual effects are used for clustering high-throughput sequencing data, but also includes categorical covariates which describe known information on observed days. The technical elements of these models will be given in Section 4. The Vélib' system is described and the data used in this work is presented through general statistics and figures.

## 3. The Vélib' bike sharing system of Paris

In this section, the Vélib' system is described and the data used in this work, as well as its specificities, are presented through general statistics and figures that stress the natural features taken into account in the proposed models.

### 3.1. Architecture of the Vélib' network

The Bike Sharing System of Paris, called Vélib',[1] has been deployed since July 2007 and is operated as a concession by Cyclocity, a subsidiary of the French outdoor advertising company

[1] *Vélib'* stands for the French expression *Vélo en Libre Service*. It can also be seen as a portmanteau of the French words *Velo´* and *Liberté*.

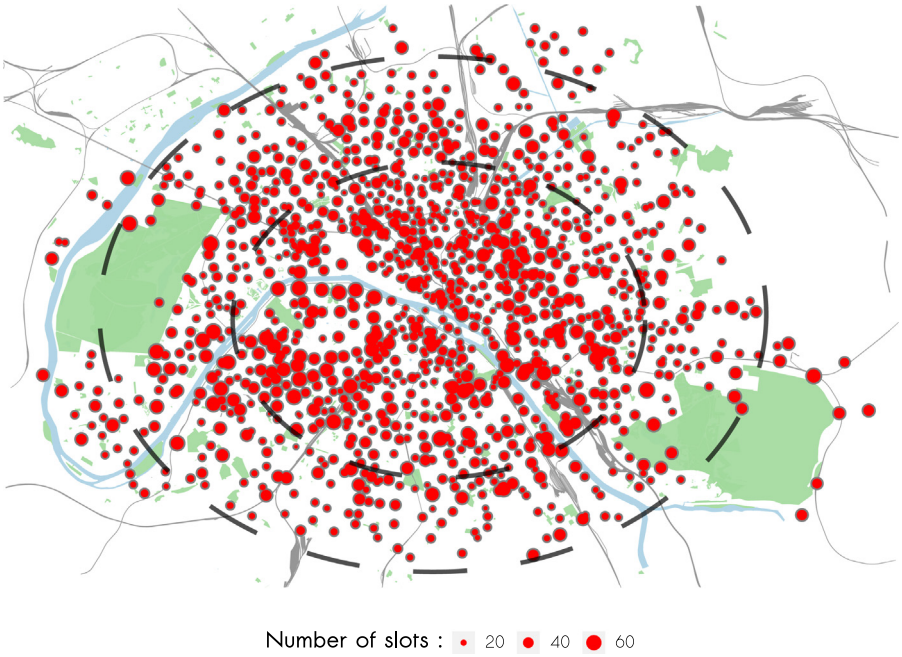


**Fig. 1.** Map of station distribution throughout the Paris region. Each station is sized with respect to the number of slots/docking points, ranging from 8 to 70, with an average value of 32 slots. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

JCDecaux. At its debut in 2007, 7000 bicycles were spread across 750 fixed stations. In six years, it has expanded to 1208 stations which hire out more than 18,000 bikes throughout the city.

Each station is equipped with an automatic rental terminal and has stands for 8–70 *docking points*. Vélib' is mainly available *intramuros* as shown in Fig. 1, which maps the position of the stations and total capacity for the Paris region. Station density is commonly higher in the city centre; the inner circle shown in Fig. 1 includes the business centres and the top tourist spots of Paris. Around the most important railway stations (drawn as grey lines), there are more Vélib' stations, and they are bigger. In contrast, the places near the outer circle, which are closer to the Paris Péripherique inner ring road, are less well equipped: visually there are more gap between the stations. However, the off-centre stations close to parks (represented by green shapes on the map) have more docking points.

Considering the number of annual subscribers, 224,000 and still rising, and the average number of 110,000 trips per day, Vélib' is large-scale and is now one of the largest Bike Sharing Systems in the world and the biggest Bike Sharing System in Europe.

### 3.2. General statistics based on April 2011 data analysis

Vélib', like any Bike Sharing System, offers a non-stop and fully automated service. In this study, the analysis is based entirely on anonymized O/D trips data, provided by *JCDecaux* and *Ville de Paris*, for two one-month periods in April and September 2011. The system recorded respectively around 2,500,000 and 3,260,000 displacements during these periods. Preprocessing the data, trips with a duration of less than 1 min and which looped at the same station were excluded, representing respectively 4.4% and 4.8% of the observed trips during April and September 2011. Fig. 2 displays the total number of recorded trips during April 2011, observed over a week, with respect to the type of subscription: annual (plotted in blue) or one day (plotted in red). The blue curve shows a repetitive but distinct pattern depending on the type of day. Weekdays (Monday to Friday) are marked with peaks at the commutes (8am and 6pm) and during the lunch break, whereas the highest volume usage at weekends (Saturday and Sunday) is evenly distributed throughout the afternoons. The red curve depicts a different pattern with higher activity early morning and late afternoon. In addition, considering the volume of displacements during Saturdays, Sundays and Mondays, the typical weekend pattern for the one-day users lasts until Tuesday. It is reasonable to assume that these trips occurring at weekends are more leisure and recreational oriented, and the ones occurring at weekdays are more utilitarian oriented. These temporal trends are not a peculiarly French feature. Froehlich et al. [9,10], who studied Bicing, the Bike Sharing System of Barcelona, similarly identified a distribution of trips marked by peaks at key moments of the day such as lunch time, which occurs at 2 pm in Spain.

These general trends are also affected by external factors such as urban environment, city infrastructure [24,8,7], sociological, economical context and weather conditions. We focus on the latter in this paper with in mind the strong assumption that the use of soft modes of transport decreases under bad weather conditions (extreme cold or heat, rain). Fig. 3(a) (resp. Fig. 3(b)) displays the total number of recorded trips during weekends (resp. weekdays) of September 2011, observed over a 24-hour day, with respect to specific days and rainfall intensities. The upper part of each figure displays the number of trips for three rainy days (in red, green and blue), which rainfall intensities are shown in the lower part of the figure. Lastly, the observed patterns on rainy days are compared with the average number of trips recorded during sunny weekends or weekdays (in purple colour). Such temperature and rainfall intensities were extracted every five minutes from a weather station located at the centre of Paris (source: http://www.wunderground.com/, weather-station-id: *I*75003*PA*1).

After looking at the two figures, a first important conclusion is that rain has greater impact on recreational demands occurring during the weekends, whereas it has not on primarily utilitarian demands of the weekday commutes. A study carried out by [29], who analysed the temporal variations of bicycle demands in the Netherlands, similarly underlined that recreational/utilitarian distinction in such conditions. Let us now focus on the weekends (Fig. 3(a)). The effect of bad weather on biking is first immediate: the number of trips observed on the afternoons of September 11th, 2011 (red curve) and September 18th, 2011 (blue curve) inevitably drops when it rains – even if it is a little rain (0.3 mm/h). On the

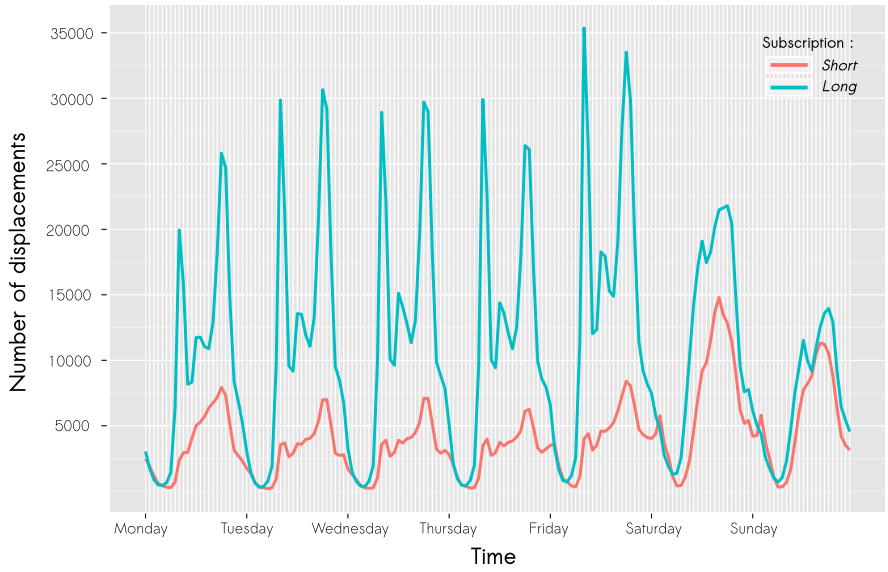


**Fig. 2.** Total number of trips during April 2011, summed over each hour of each day of the week, from Monday to Sunday. The blue line (resp. red line) corresponds to the displacements carried out by one-year (resp. one-day) subscribers. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

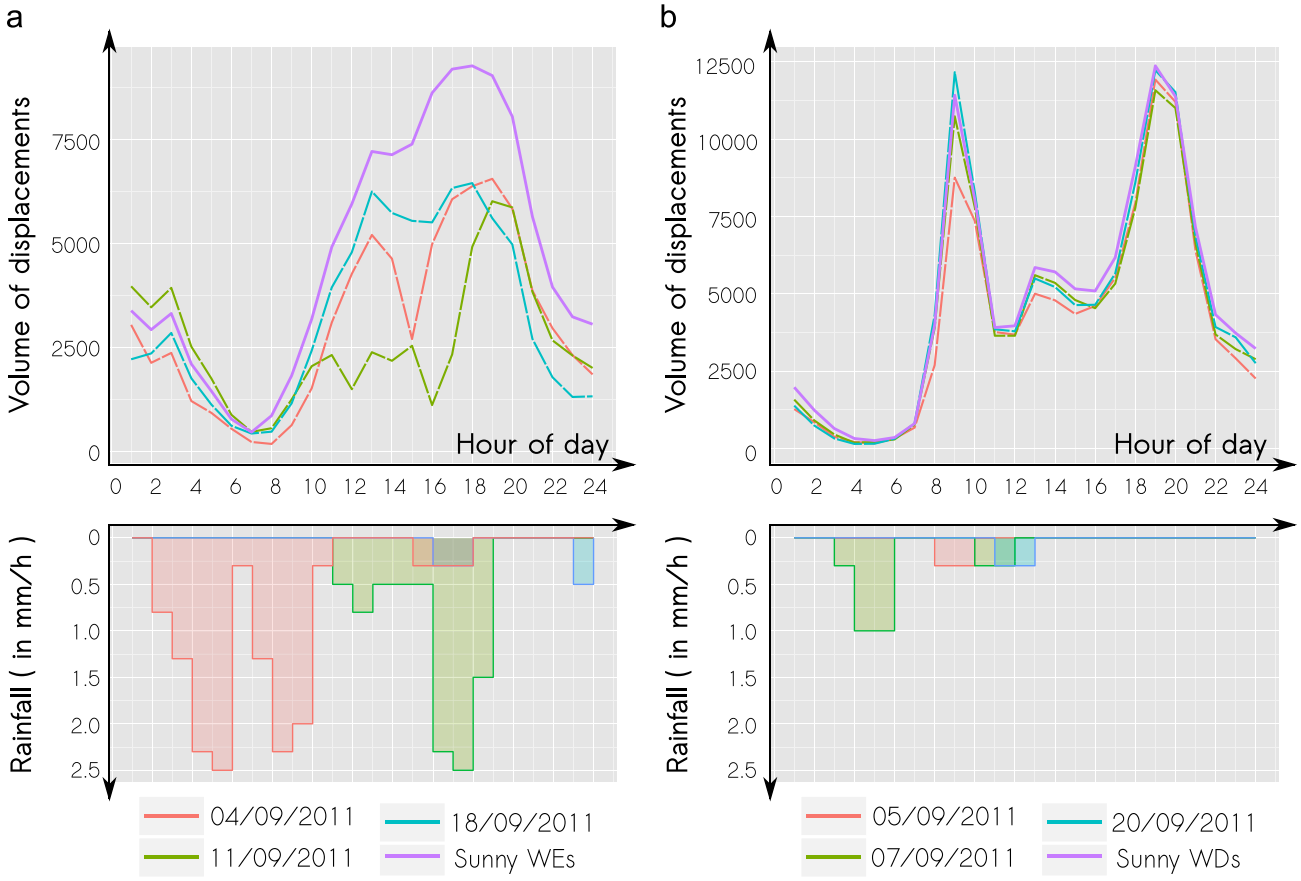


**Fig. 3.** Total number of trips (upper part of the figures) recorded during weekends (a) and weekdays (b) of September 2011, observed over a day, with respect to specific days and rainfall intensities (lower part of the figures). The red, green and blue curves depict the total number of trips for three rainy days. The purple curve displays the average number of trips recorded on sunny weekends (b) or weekdays (a). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

other hand, high rainfall intensities in the morning of September 04th, 2011 (red curve) may have demotivated the bikers, inducing lower activity in the afternoon. These observations concur with those made by [21], who investigated more thoroughly the relationship between weather conditions and cycling in Canada.

Let us now focus on the individual trips which are often affected by the current system configuration. The trips are mostly short, one in two being less than 1.6 km and four out of five being less than 3 km, as shown in Fig. 4(a), where the histogram represents the trip distance as the crow flies for the observed

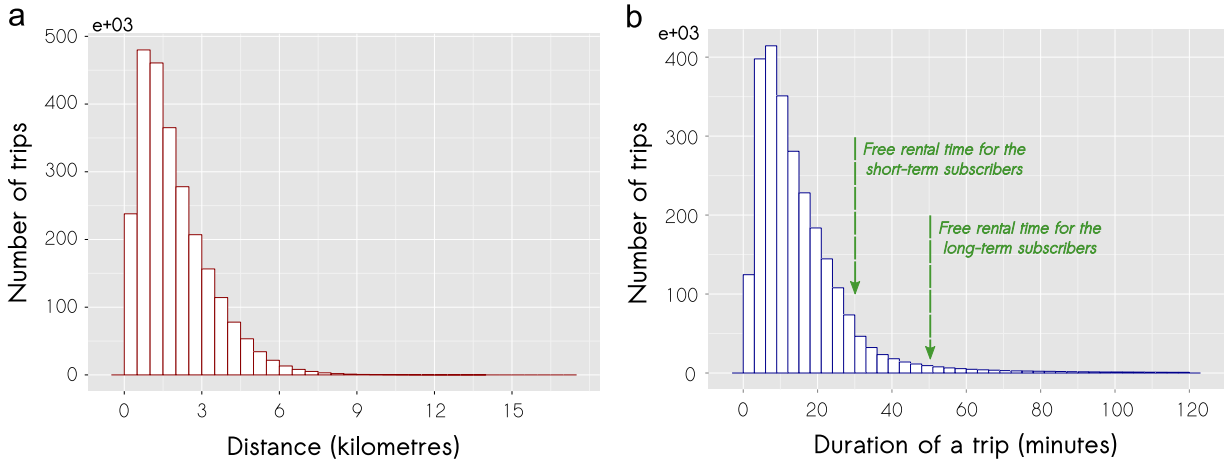


**Fig. 4.** Histograms of trip distance as the crow flies (a) and trip duration (b) computed on the Vélib' April 2011 data. (a) Trip distance (crow ies) histogram and (b) Trip duration histogram. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

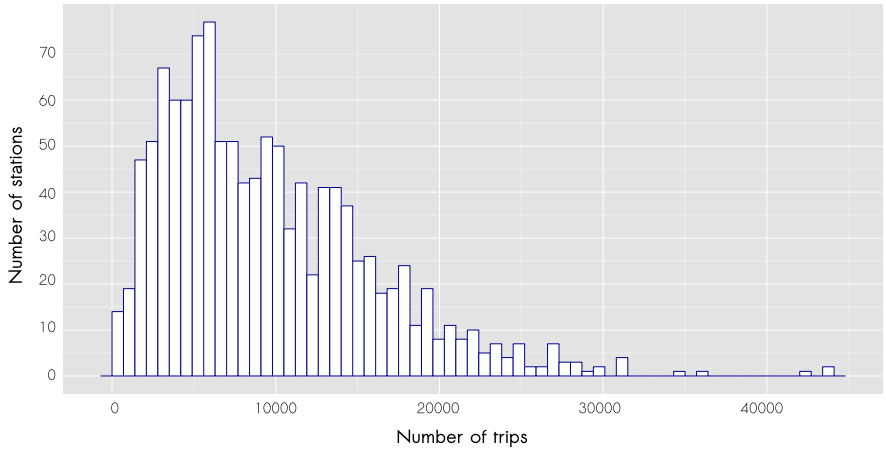


**Fig. 5.** Histogram of the total number of trip per station observed during April and September 2011.

period of April 2011. In order to maximize the number of trips per cycle per day, pricing policies that encourage short trips have been introduced. In Paris, the first half hour of usage is free and longer journeys are charged at an increasing rate. Such pricing policy explains in part Fig. 4(b), where the histogram represents the trip duration: 91% of the trips last less than 30 min, which is the free usage for one-day subscriptions and 96% of them last less than 45 min, which is the free usage for one-year subscriptions.

Let us now focus on the *spatial* distribution of trips. Fig. 6 represents a naive approximation of the *trip map* of Paris, using the total number of incoming plus outgoing trips logged during April and September 2011, for each individual Vélib' station. The mean total number of observed trips over a hexagon of area of approximatively 15 ha is computed in each coloured cell. The closer to red the colour is, the higher the global activity over the period in the area. The deepest red colour, which concerns the most significant displacements, is to be found in the centre of Paris and around three big metro stations (République, Châtelet-Les Halles and Nation). The next most attractive places are the parks (Vincennes and La Villette) and the river surroundings. In contrast, fewer trips are recorded in the south-west and north-west of Paris, although they are equipped with stations (see Fig. 1). Another simpler viewpoint is to look at the distribution of the trips among the stations as presented in Fig. 5, in which we are told that there are few highly popular stations (probably those located in the centre of Paris) while others seem to be deserted.

To explore such complex data, which varies in function of temporal, spatial and sociological factors, it is crucial to propose models that take some of these specificities into account. The following paragraphs detail the family of models proposed in this paper.

## 4. Model-based clustering of count-series

To derive the cluster of O/D pairs that share similar temporal profiles, a family of adapted generative Poisson mixture models based on count-series clustering is introduced in this section. We begin with the structure of the count-data that is going to be processed, and go on to introduce the *Expectation–Maximization* algorithm used to derive the clustering and to estimate the parameters of each model.

### 4.1. Representation of the Origin/Destination flows

The mobility data used in this paper corresponds to trips recorded on the Vélib' BSS of Paris. Each trip is described by a vector of the *origin station*, its *time of departure*, the *destination station*, its *time of arrival* and the corresponding *type of user-subscription* (long versus short). To keep the model general to all timestamped Origin/Destination data, this last variable is not used.

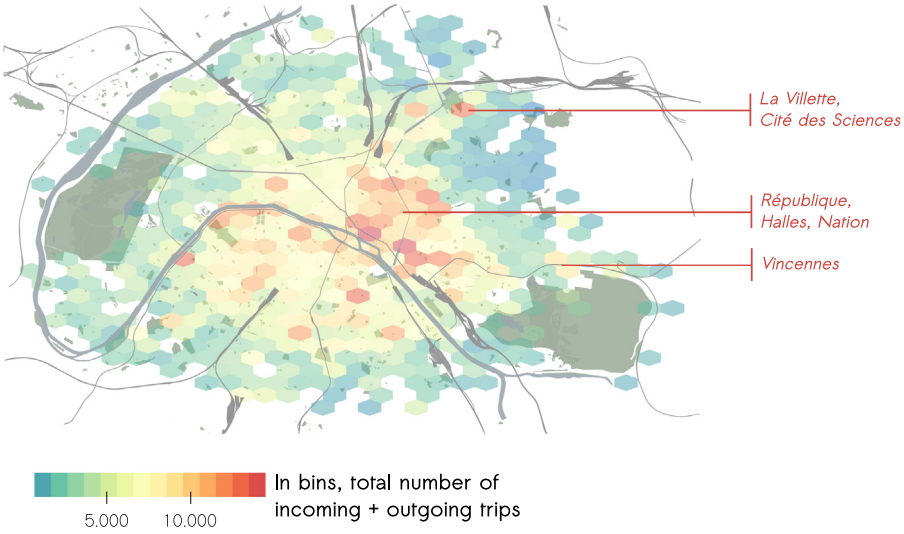


**Fig. 6.** Map of the total number of incoming plus outgoing trips, logged during April and September 2011. This map, which was computed using the mean total number of observed trips over a hexagon of area of approximatively 15 ha, gives a first overview of the busiest spots of Paris. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

We first introduce $X_{dt}^{uv}$ as the number of bikes going from origin station $u \in \{1,\ldots,S\}$ to destination station $v \in \{1,\ldots,S\}$, during day $d \in \{1,\ldots,D\}$ and hour $t \in \{1,\ldots,24\}$. Similar to [5], a one-hour aggregation, which gives a good trade-off between detail resolution and fluctuations, has been considered to generate the counts. Let $\mathbf{X}_d^{uv}$ be

$$\mathbf{X}_d^{uv} = (X_{d1}^{uv},\ldots,X_{d24}^{uv}),$$

the description vector of hourly traffic between station $u$ and station $v$ in day $d$. All of these description vectors can be arranged in a tensor $\mathbf{X}$ of size $S^2 \times D \times T$, where $S$ is the total number of stations (1204 in the Vélib' case), $D$ the number of available days in the dataset and $T$ the length of the description vector, fixed to 24 hours in this paper.

### 4.2. Clustering with Poisson mixture models

Since we are dealing with count data, conditional Poisson mixtures are well suited. They are used to construct a family of generative models which identify $K$ clusters of O/D pairs behaving in the same way over time. The most complicated model of this family, and therefore the most finely-tuned of them, is entirely described by the *Graphical Model* illustrated in Fig. 7, and formally defined as

$$Z^{uv} \sim Multinomial(1,\pi), \tag{1}$$

$$X_{d1}^{uv} \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_{dT}^{uv} | \{Z_k^{uv}=1, W_{dl}=1\}, \tag{2}$$

$$X_{dt}^{uv} | \{Z_k^{uv}=1, W_{dl}=1\} \sim Poisson(\alpha^{uv}\lambda_{klt}), \tag{3}$$

where $Z^{(u,v)} \in \{0,1\}^K$ is the latent indicator variable that encodes the cluster membership of the O/D pair $(u,v)$ and $W_d \in \{0,1\}^L$ is an observed indicator variable that gives the *category* of day $d$. Note that there is a unique correspondence between one O/D pair $(u,v)$ and its cluster membership $k$. Similarly, one day $d$ is member of one and only one day-category $l$, *i.e.*,

$$\forall (u,v),\ \sum_k Z_k^{uv} = 1 \quad \text{and} \quad \forall d,\ \sum_l W_{dl} = 1$$

This first model, denoted *Model*-3, is the third most complex model in this proposed family of Poisson Mixture Models. *Model*-3 assumes different usage patterns from day to day. For instance, it may encode the difference of usage between weekdays and weekends, or the fact that day $d$ is sunny or not, and so on.

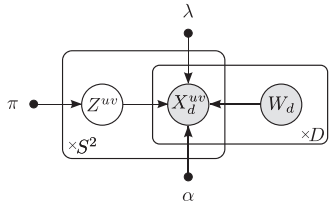


**Fig. 7.** Graphical model representation of our proposed *Model*-3 for the Vélib' count-data clustering. A box is a plate representing replicates. The main plate corresponds to the repeated conditional allocation of a cluster $k$ to each flow $X_{dt}^{uv}$.

According to Eq. (1), the cluster membership variable $Z^{(u,v)}$ is drawn from a multinomial distribution of parameter $\pi = (\pi_1,\ldots,\pi_K)$, in which each $\pi_k$ is the prior proportions of one of the $K$ clusters of the mixture. They are latent, whereas $W_d$ is not. Next, Eq. (2) recalls the assumptions of independence of the O/D pairs from time to time, day to day and cluster to cluster. Finally, the counts $X_{d,t}^{(u,v)}$ are drawn from a Poisson distribution of parameter $\lambda = \alpha^{(u,v)}\lambda_{klt}$ from the moment that both the cluster $Z^{(u,v)}$ of the displacement and the cluster $W_d$ of the day-category are known (see Eq. (3)). The $\alpha^{(u,v)}$ are scaling parameters which capture the flow magnitude between station $u$ and station $v$. The bigger it is, the more people move from station $u$ to $v$. Regarding the $\lambda_{kl.}$ parameters, they represent the temporal variations of the O/D pairs in cluster $k$, at day-category $l$, over the aggregated 24 hours of a day. To achieve identifiability, they are constrained under

$$\sum_{l,t} D_l \lambda_{klt} = DT, \quad \forall k \in \{1,\ldots,K\},$$

where $D_l = \sum_d W_{dl}$ is the number of days in cluster $l$ of the day-category. Under the preceding assumptions, we can easily derive the conditional density $f(x_{dt}^{uv} | \{Z_k^{uv}=1, W_{dl}=1\})$ of each count $x_{dt}^{uv}$:

$$f(x_{dt}^{uv} | \{Z_k^{uv}=1, W_{dl}=1\}) = p(x_{dt}^{uv}; \alpha^{uv}\lambda_{klt}), \tag{4}$$

where $p(x;\lambda) = (\lambda^x/x!)e^{-\lambda}$ is the classical density of a Poisson distribution with mean $\lambda$. By stressing the equality condition $W_{dl}=1$ on the observed day-category variable, the latter density becomes

$$f(x_{dt}^{uv} | \{Z_k^{uv}=1, W_{dl}=1\}) = \prod_l p(x_{dt}^{uv}; \alpha^{uv}\lambda_{klt})^{W_{dl}}, \tag{5}$$

since $W_{bl}=0,\ \forall b\neq d$. Finally the log-likelihood of the parameters $\boldsymbol{\theta}$ expresses as

$$L(\boldsymbol{\theta};\mathbf{X})=\sum_{u,v}\log\left(\sum_k \pi_k \prod_{d,t,l} p(x_{dt}^{uv};\alpha^{uv}\lambda_{klt})^{W_{dl}}\right). \tag{6}$$

The estimate $\widehat{\boldsymbol{\theta}}=(\widehat{\alpha},\widehat{\lambda},\widehat{\pi})$ that locally maximizes the log-likelihood is computed with an *Expectation–Maximization* (EM) type algorithm [19,18], the main steps of which are presented in the following section.

### 4.3. The EM framework

The EM algorithm is a well-known and well-suited two-step iterative approach for maximum likelihood estimation in statistical problems involving latent variables and unknown parameters. The algorithm begins with the *E-step* which computes a lower bound of the log-likelihood; the latter is then maximized in an *M-step*. These two steps are iterated until they converge towards a local maximum of the log-likelihood. However, to build the lower bound it is in practice more convenient to introduce the completed log-likelihood, which incorporates the latent cluster-membership variable $Z$. In this paper, it is denoted $L_c$ and is given by

$$L_c(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})=\sum_{u,v,k} Z_k^{(u,v)}\log\left(\pi_k \prod_{d,t,l} p(x_{dt}^{uv};\alpha^{uv}\lambda_{klt})^{W_{dl}}\right). \tag{7}$$

#### 4.3.1. E-step

The lower bound of $L$, defined as

$$\mathbb{E}[L_c(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})|\mathbf{X},\boldsymbol{\theta}^q]=\sum_{u,v,k}\tau_k^{uv}\log\left(\pi_k \prod_{d,t,l} p(x_{dt}^{uv};\alpha^{uv}\lambda_{klt})^{W_{dl}}\right), \tag{8}$$

is equal to the conditional expectation of $L_c$ over $Z$, with respect to the current parameter values. Its computation, during the *E-step*, requires the values of $\tau_k^{uv}$, given by

$$\tau_k^{uv}=\frac{\pi_k\prod_{d,t,l}p(x_{dt}^{uv};\alpha^{uv}\lambda_{klt}^{ij})^{W_{dl}}}{\sum_k \pi_k\prod_{d,t,l}p(x_{dt}^{uv};\alpha^{uv}\lambda_{klt}^{ij})^{W_{dl}}}. \tag{9}$$

They correspond to the *a posteriori* probabilities of the pair $(u,v)$ to be in cluster $k$ at the $q$th iteration of the algorithm. There are $K\times S^2$ of them and from the moment they are known, the clustering that associates to each O/D pair a single cluster is simply derived using a *map* rule, *i.e.* picking the value of the $k$ that maximizes the $\tau_k^{uv}$.

#### 4.3.2. M-step

In order to increase the likelihood, the expectation (8) is maximized with respect to the parameter. This is performed in the *M-step* and leads to the following update rules:

$$\hat{\alpha}^{uv}=\frac{1}{DT}\sum_{d,t}x_{dt}^{uv} \tag{10}$$

$$\hat{\pi}_k=\frac{1}{S\times S}\sum_{u,v}\tau_k^{uv} \tag{11}$$

$$\hat{\lambda}_{klt}=\frac{1}{\sum_{u,v}\tau_k^{uv}\alpha^{uv}\sum_d W_{dl}}\sum_d W_{dl}\sum_{u,v}\tau_k^{uv}x_{dt}^{uv} \tag{12}$$

which have natural interpretations, since the flow magnitude $\alpha^{uv}$ between a O/D pair can simply be seen as the average number of trips that has moved from station $u$ to $v$ during the monitoring period $DT$. The expression of $\lambda_{klt}$ is similarly a ratio of counts per time: however, it is more specific since these are counts proper to both the cluster and the type of day. At last, since $\hat{\pi}_k$ is updated using the normalized sum of the *a posteriori* probabilities of all O/D pairs to be in cluster $k$, it expresses the weight of cluster $k$ in the mixture. Table 1 summarizes for convenience the main variables and parameters used in this EM algorithm. In what follows, alternative models that handle different specificities of the system-usage are introduced.

**Table 1**
In order of appearance in this paper, detailed list of the observed and latent variables (or parameters) used in the *EM* step of each of the proposed family of Poisson mixture models.

| Variable | Domain | Variable description and interpretation |
|---|---|---|
| $X_{dt}^{uv}$ | $\mathbb{N}$ | Observed number of bikes going from station $u$ to station $v$, during day $d$ and hour $t$ |
| $Z^{(u,v)}$ | $\{0,1\}^K$ | Latent indicator variable. Encodes the cluster of the O/D pair $(u,v)$ among $K$ clusters |
| $W_d$ | $\{0,1\}^L$ | Observed indicator variable. Gives the *category* of day $d$, among $L$ day-categories |
| $\pi$ | $[0;1]^K$ | Latent parameter vector of prior proportions of the $K$ clusters. The $\pi_k$ sum to one |
| $\alpha^{(u,v)}$ | $\mathbb{R}^{+*}$ | Scaling parameters that capture the flow magnitude between station $u$ and station $v$ |
| $\lambda_{kl.}$ | $\mathbb{R}^{K.+*}$ | Temporal variations of the O/D pairs in cluster $k$, at day-category $l$, over the 24-hours of day |
| $\tau_k^{uv}$ | $\mathbb{R}^{K.+*}$ | *A posteriori* probabilities of the pair $(u,v)$ to be in cluster $k$ |

### 4.4. Three alternative but simpler Poisson mixture models

The presented *Model*-3 will now be compared to other simpler Poisson mixture models. *Model*-0 is the most basic of all. Fig. 8 (b) depicts its graphical model, the mathematical formulation of which is

$$\begin{aligned}&Z^{uv}\sim Multinomial(1,\pi),\\ &X_{d1}^{uv}\perp\!\!\!\perp\ldots\perp\!\!\!\perp X_{dT}^{uv}|\{Z_k^{uv}=1\},\\ &X_{dt}^{uv}|\{Z_k^{uv}=1\}\sim Poisson(\lambda_{kt}),\end{aligned}$$

where, given $Z_k^{(u,v)}$, the $X_{d,t}^{(u,v)}$ are simply independent and identically distributed Poisson random variables with parameter $\lambda_{kt}$. Here, all of the observed days $d$ are of the same and unique day category: since the $\lambda$ are simply indexed by $k$, there will be a unique temporal profile for each cluster.

*Model*-2, illustrated in Fig. 8(c), fine-tunes the latter by capturing the specific patterns related to the categories of days, similar to *Model*-3. From now on the $\lambda$ are indexed by both $k$ and $l$, inducing as many temporal profiles as distinct categories of days. Formally, it is the simplest *Model*-0 upgraded with the $W_{dl}$ cluster-membership conditions:

$$\begin{aligned}&Z^{uv}\sim Multinomial(1,\pi),\\ &X_{d1}^{uv}\perp\!\!\!\perp\ldots\perp\!\!\!\perp X_{dT}^{uv}|\{Z_k^{uv}=1,W_{dl}=1\},\\ &X_{dt}^{uv}|\{Z_k^{uv}=1,W_{dl}=1\}\sim Poisson(\lambda_{klt}).\end{aligned}$$

Model *Model*-1, on the contrary, relaxes the assumption related to the day-category but differentiates between the most and the least favoured O/D pairs through the scaling parameters $\alpha^{(u,v)}$. Its graphical model is illustrated in Fig. 8(d) and it is formally defined as

$$\begin{aligned}&Z^{uv}\sim Multinomial(1,\pi),\\ &X_{d1}^{uv}\perp\!\!\!\perp\ldots\perp\!\!\!\perp X_{dT}^{uv}|\{Z_k^{uv}=1\},\\ &X_{d,t}^{uv}|\{Z_k^{uv}=1\}\sim Poisson(\alpha^{uv}\lambda_{kt}).\end{aligned}$$

This section concludes with a discussion of the parameter $W_d$.

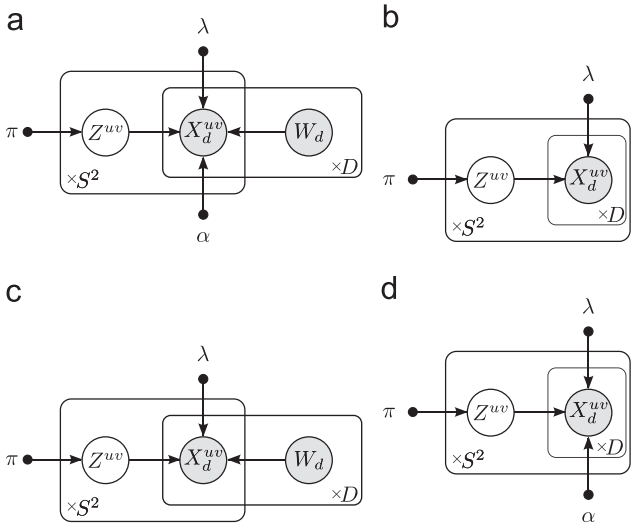


**Fig. 8.** Graphical models of *Model*-3, *Model*-0, *Model*-2 and *Model*-1. (a) Our model. (b) Model-0. (c) Model-2. (d) Model-1.

### 4.5. Several specifications of $W_d$ leading to versions of Model-3 and Model-2

*Model*-3 and *Model*-2, as introduced above, both handle an observed day-category parameter, which was deliberately not defined precisely, since it can be refined in several ways. In this paper we have investigated three different specifications of $W_d \in \{0,1\}^L$ that lead to three versions of *Model*-3 and one version of *Model*-2.

The first proposed day-category arises from the clear difference in usage between weekdays and weekends, as shown in Fig. 2. $L$ is in such a case set to two, and $W_d$ is $(0,1)$ when day $d$ is a weekday, and $(1,0)$ otherwise. Model-3 and Model-2 with such day-categorization will be called *Model*-3.*a* and *Model*-2.*a*. In a second trial we defined *Model*-3.*b*. With $L$ set to seven, $W_d$ now encodes one category per day of the week (Sunday, Monday, …, Saturday). Such a model, which assumes a specific usage of the Vélib' BSS each day, still makes sense: Fig. 2 subtly shows that the volume of displacements recorded on Mondays and Fridays are quite a bit different from the volume recorded on Tuesdays, Wednesdays and Thursdays. This remark holds for weekends, too. Finally, we introduced *Model*-3.*c*, which attempts to encode the weather conditions. There are eight categories of days in this setting, being the cross product of three binary variables: the first one handles the difference between weekdays and weekends; the second, two levels of precipitation intensities (*rain* versus *no rain*); and the third, two levels of temperature (either $>15^\circ$ or not). This last specification of *Model*–3 assumes that bikers are very much affected by bad weather conditions, as previously shown in Fig. 3.

Subsequently, all these clustering models which are summarized in Table 2 and which encode different assumptions – *i.e.* three versions of *Model*-3, *Model*-0, one version of *Model*-2 and *Model*-1 – were compared and applied on real Vélib' data.

## 5. Results

In this section, we first detail the real BSS data used to train and test the previously presented models, before going on to compare them. The *training set* used for parameter estimation is a mix of 20 randomly chosen days from April 2011, plus 20 others from September 2011. It is made up of 27 weekdays and 13 weekends and accounts for 3,700,000 trips. To keep focused on the global usage of the system, the casual trips – that are to be found in the pairs $(u,v)$ of stations that do not generate at least one trip per day – were removed in formatting and preprocessing the data, leaving 13,100 O/D flows. The corresponding O/D matrix is sparse, revealing that there are lots of pairs of stations that rarely exchange bikes, such as the pairs of stations far apart from each other which interactions are discouraged by a pricing policy that "taxes" longer journeys. Hence the *training set*, now denoted $\boldsymbol{X}^{\boldsymbol{t}train}$, is of size $S^2 \times D \times T = 13.100 \times 40 \times 24$. A *testing set* $\boldsymbol{X}^{\boldsymbol{t}est}$ to assess the performance of the different models is similarly defined and preprocessed using the 20 days of April and September that had not been chosen.

**Table 2**
In ascending order of complexity, the different *Model*-0, 1 and the model declinations *Model*-2.a, 3.a, 3.b and 3.c are compared with respect to two parameters: the first one is the scaling parameter which makes the difference between the pairs of stations $(u,v)$ that exchange – or not – a lot of bikes. Last parameter is the category of day $d$ which encodes a specific usage of the system per category of day. For *Model*-3.c the symbol $\times$ refers to the cross product. Thus a day $d$ is defined by the cross product of its day-type, precipitation and temperature.

| *Model* | *O/D scaling?* | *Day category?* | *How many?* | *Day-category levels l, for any day d* |
|---|---|---|---|---|
| Model-0 | No | No | | |
| Model-1 | Yes | No | | |
| Model-2.a | No | Yes | 2 | {*weekday*,*weekend* } |
| Model-3.a | Yes | Yes | 2 | {*weekday*,*weekend* } |
| Model-3.b | Yes | Yes | 7 | {*Monday*, .., *Sunday*} |
| Model-3.c | Yes | Yes | 8 | {*weekday*,*weekend*} $\times$ {*rainy*,*sunny*} $\times$ {*cool*,*mild*} |

### 5.1. Models comparison

The first results presented in this section concern the comparison of the previously introduced models. To this end, we computed the classical *Bayesian Information Criterion* (BIC) [27,15], for all the models, for a number of clusters ranging from three to twenty-five. To compute the predictive performances of the models, we used the perplexity measure: the six proposed models are trained with $\boldsymbol{X}^{\boldsymbol{t}train}$ and the corresponding value of $\hat{\boldsymbol{\theta}}^{\boldsymbol{t}train}$ is used to compute the model perplexity on the held-out testing set $\boldsymbol{X}^{\boldsymbol{t}est}$. By doing so on unseen data, we intended to assess to what extent the assumptions encoded within each model describe the global usage of the system. Given model $M$, it is formally defined as

$$perp(M;\boldsymbol{X}^{\boldsymbol{t}est}) = \exp\left(\frac{-\log\left(\mathbb{P}(\boldsymbol{X}^{\boldsymbol{t}est}|M,\hat{\boldsymbol{\theta}}^{\boldsymbol{t}train})\right)}{S^2 D}\right), \tag{13}$$

and is closely tied to the concept of likelihood. Its numerator can be seen as a decreasing function of the log-likelihood $L(M,\hat{\boldsymbol{\theta}}^{\boldsymbol{t}train};\boldsymbol{X}^{\boldsymbol{t}est})$ of the testing set, with respect to the estimated parameters of $M$, computed on the training set. The lower the perplexity score, the better the generalization performance of the tested model.

Fig. 9(b) presents the test perplexity for each model with respect to $K$, and Fig. 9(a) presents the evolution of the BIC under the same settings. The proposed *Model*-3.*a* gave interesting results: as expected, for all $K$ it outperforms the simpler *Model*-0, 1, 2.*a* in terms of perplexity and BIC, underlining that the difference in usage between weekdays and weekends must *at least* be taken into account. Furthermore, the scaling factors definitely appear to be essential since *Model*-0 and 1 rank last.

As for the three versions of *Model*-3, we will only discuss the design of $W_d$. The different day-categorization also favours *Model*-3.*a* since it outperforms *Model*-3.*b* and matches *Model*-3.*c* in terms of perplexity, the conclusions being less clear for the BIC. Actually, with respect to the BIC these three models are close to each other: *Model*-3.*b* and *Model*-3.*c* scored slightly better than *Model*-3.*a* for lower values of $K$, and slightly worse afterwards.

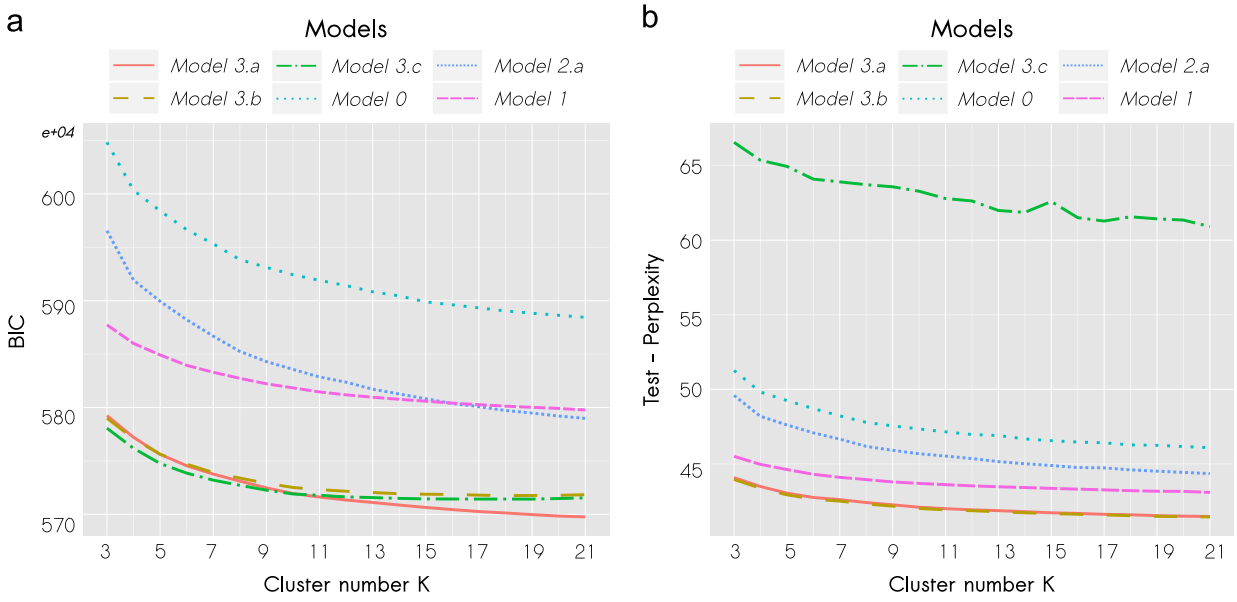


**Fig. 9.** With $K$ ranging from 3 to 21: (a) BIC and (b) perplexity evaluated on test-set data for *Models*-0, 1, 2, 3.a, 3.b, 3.c.

The test perplexity gives another picture since *Model*-3.*c* is clearly outperformed by *Model*-3.*a*, 3.*b* for all $K$. These observations may be explained by the relatively small size of our data sample (only two months) and may have led to overfit *Model*-3.*c*. More days may be needed to reach satisfactory performances. Taking these facts into account now leads us to detail the results obtained with *Model*-3.*a*. Although *Model*-3.*b* did as well as 1.*a*, $L=7$ in the former and is *just* two in the latter, which will considerably facilitate the interpretation process.

### 5.2. In-depth analysis of the clustering results

Dealing with *Model*-3.*a*, the number of clusters $K$ is now fixed. The BIC value on April data shown in Fig. 9(a) obviously decreases until $K=21$. However, since this is far too large to describe the primary usage trends of the system, a lower value of $K$ that better minimizes this criterion in terms of slope variation and facilitates the interpretation process is preferred. To clarify the choice of the number of clusters $K$, the BIC decrease for *Model*-3.*a* is plotted in Fig. 10. It shows the difference of two consecutive BIC values, with $K$ ranging from 3 to 21: the 74% drop in BIC between $K=3$ and $K=8$ is as significant as the drop in BIC between $K=8$ and $K=21$. Taking these facts into account now leads us to detail the results obtained with *Model*-3.*a* and $K=8$, which seems to be a good agreement in terms of model selection criteria and interpretation. Below are the results under these settings.

The clustering results can be analysed thanks to the different information supplied by the model parameters. As outlined in Section 4 describing the proposed models, the scaling parameter $\alpha^{uv}$ captures the flow magnitude between any station $u$ and station $v$. The parameter $\lambda_{kl.}$ shows the temporal variations of the O/D pairs specific to the cluster $k$ at day-type $l$, over the day. The values of $\lambda_{kl.}$ are plotted in Fig. 11 which shows the temporal profiles over weekdays and weekends for each cluster $k$. For any given $k$, the red (resp. green) curve represents the estimated values of $\lambda_{klt}$, where $l$ encodes the weekdays (resp. weekends) and $t$ ranges from 1 to 24. They are compared with the grey-shaded range, upper and lower-bounded by the empirical quantiles (0.75 and 0.25) estimated on the whole population of O/D pairs. As presented before, the *a posteriori* probabilities computed by the EM algorithm are also used to cluster the different O/D pairs by using the *map* rule. The O/D

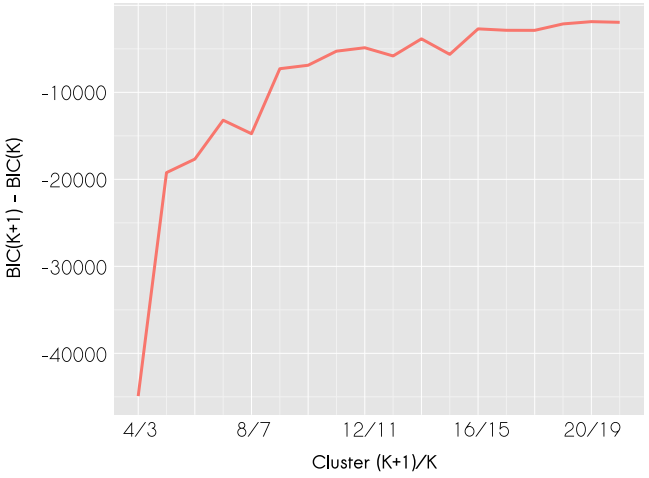


**Fig. 10.** With respect to $K$, difference of consecutive BIC values for *Model*-3.a.

pairs of each cluster may then be mapped to explore their spatial distribution. For example, Fig. 12(a) maps the O/D pairs of clusters 1. Each of its O/D pair is represented as an arrow leaving station $u$ and going to station $v$. Each arrow is also described by its colour intensity: darker coloured arrows indicate higher $\alpha^{uv}$.

Let us now explore these different aspects further. First we focus on the value of the scaling parameters $\alpha^{uv}$. The top ten pairs of stations in terms of $\alpha^{uv}$ are given in Table 3. The pair (15005,15016), which reaches the top position, links a residential area known to be poorly served by public transport and a Vélib' station *Place Trefouel* close to the *Pasteur metro station*. An important point is that *Pasteur* is the only metro station in the 14th and 15th arrondissements that gives access to both *line* 6 and *line* 12 of the Paris metro. The same remark applies for the pair (13053,13120) in third position. Station 13053 is one minute's walk from *Bibliothèque Francois Mitterand*, a big station in Paris served by both the *RER C*[2] and the metro *line* 14. Such findings show the

[2] RER *Regional Express Network* serves both Paris and the suburbs.

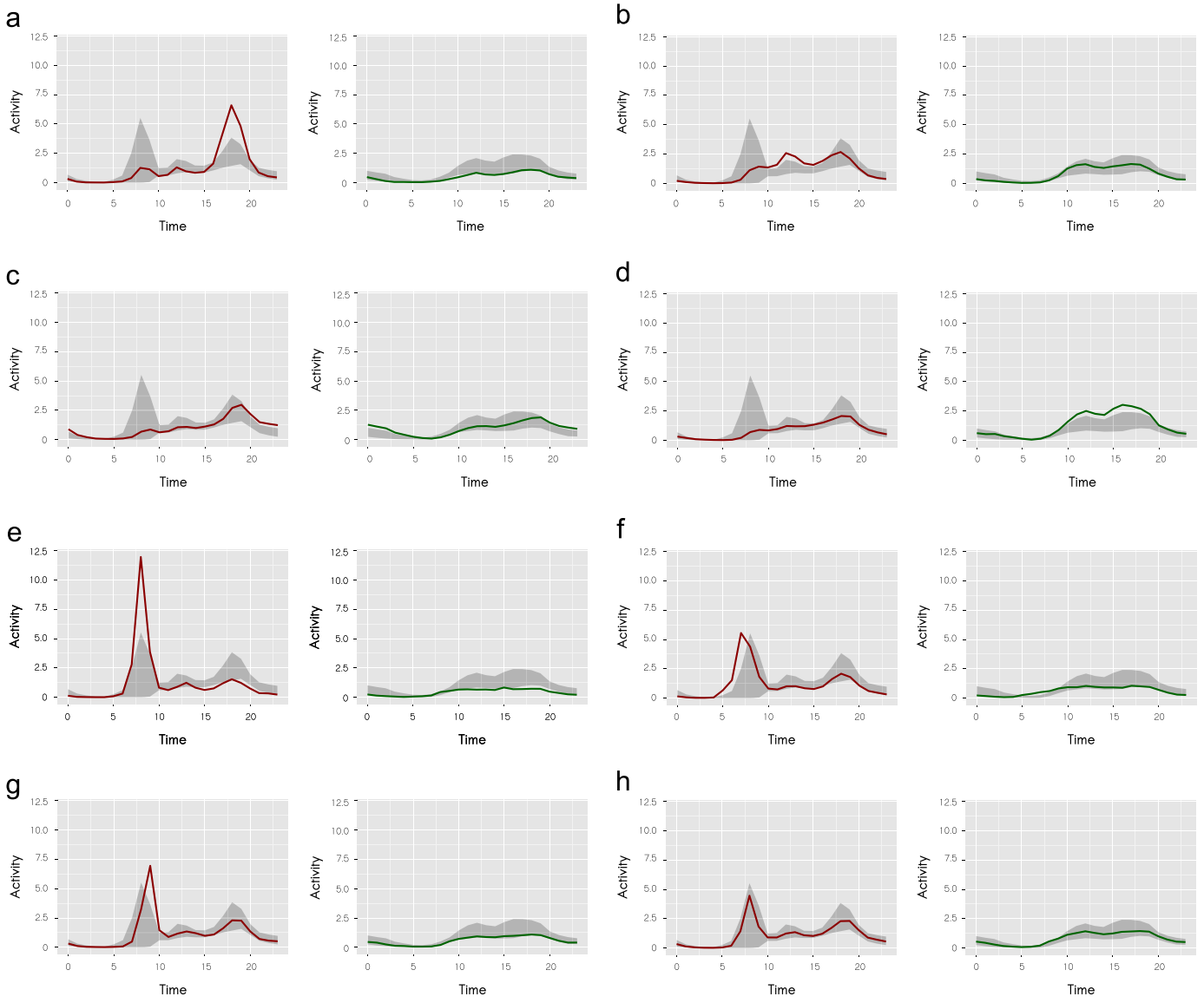


**Fig. 11.** Estimated values of $\lambda_{klt}$ for each category of day for clusters 1–8. They shape the temporal profiles of usage for weekdays (red curve on the left) and weekends (green curve on the right). They are compared with the grey-shaded range given by the empirical quantiles (0.75 and 0.25) estimated on the whole population of O/D pairs. (a) Cluster 2 – Temporal profile. (b) Cluster 3 – Temporal profile. (c) Cluster 4 – Temporal profile. (d) Cluster 5 – Temporal profile. (e) Cluster 1 – Temporal profile. (f) Cluster 6 – Temporal profile. (g) Cluster 7 – Temporal profile. (h) Cluster 8 – Temporal profile. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

strong links between Vélib' and other forms of public transport and underline the great interest of a BSS for multi-modal trips and its use as a *last-mile* connection with the metro system.

Turning now to the temporal profiles in Fig. 11, we see in Fig. 11(e) the temporal profile of cluster 1, which represents high activity during weekday mornings and relatively low activity the rest of the time. This may correspond to O/D pairs leaving from residential areas to transit spots or business centres, since the peak of activity occurs during the morning commutes. Cluster 2 (see Fig. 11(a)), on the other hand, represents high activity during weekday evenings and relatively low activity the rest of the time and may therefore correspond to the evening commutes, *i.e.* riders leaving from the transit spots or business centres and going back home. Another behaviour is depicted in the temporal profile of cluster 3 (Fig. 11(b)), which shows quite high activity during weekday lunchtimes and weekend afternoons. This may correspond to short journeys, transit trips occurring during the day, such as those during the one-hour lunch break in which people usually stay close to their starting point to save time. The temporal profile for cluster 4, Fig. 11(c), similarly depicts movements occurring during weekdays and at weekends, although they now take place at the night. The pattern shown in Fig. 11(d) for cluster 5 is probably more specific to weekend outings, long bike rides and recreational activities, given the large and significant peak at weekends. The profiles for clusters 6 and 7 (Fig. 11(f), (g)) are similar to that of cluster 1, the slight difference being in the time of the peak: earlier for cluster 6, later for cluster 7. Lastly, cluster 8 represents an average behaviour.

These results, in which a few system usages can already be deduced, are now completed by analysing the spatial distribution of the O/D pairs of the eight clusters and crossing results with the sociological information summarized in Table 4. Basic statistics on each cluster are given in Fig. 14, such as the violin plot of the trip distance as the crow flies per cluster (see Fig. 14(a)) and of the mean trip duration per cluster (see Fig. 14(b)).

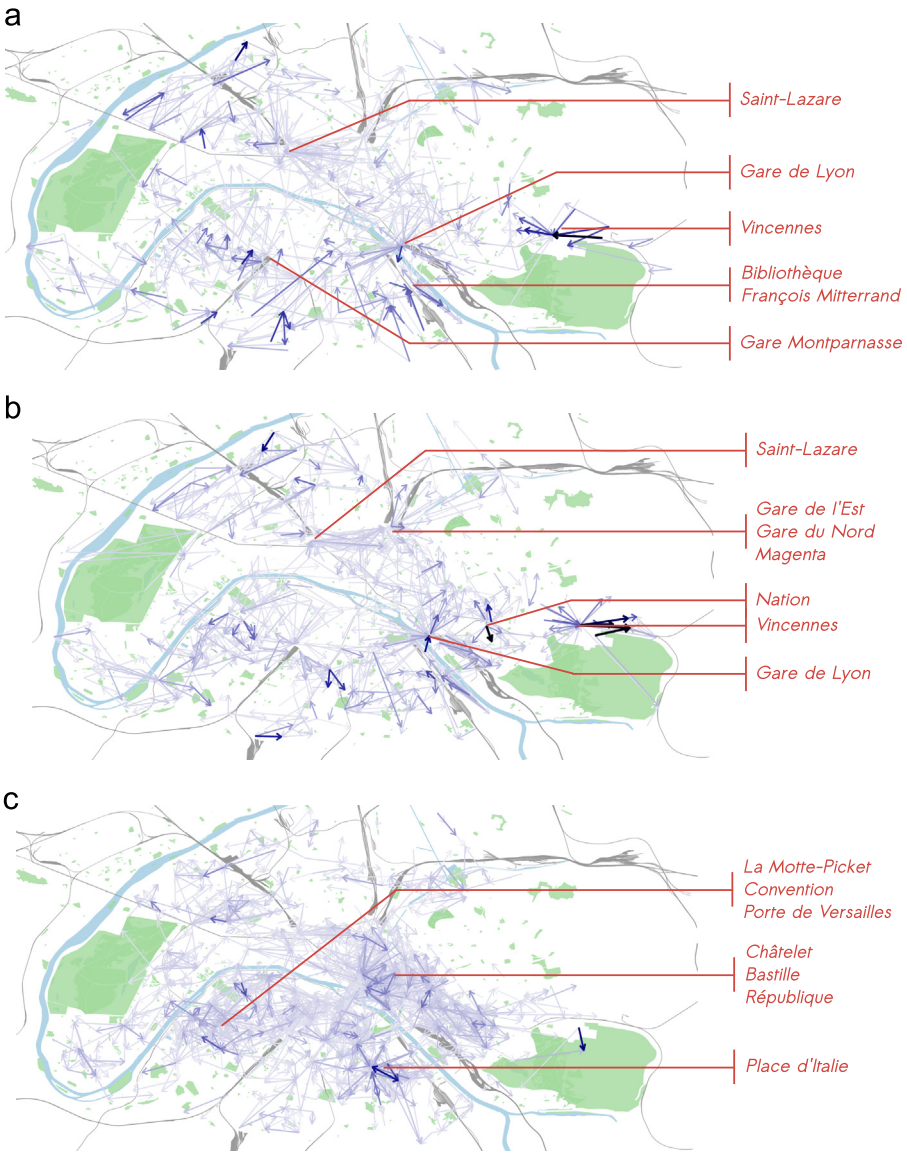


**Fig. 12.** Spatial distribution of the O/D pairs in clusters 1, 2 and 3. Each O/D pair is represented as an arrow, described by its colour intensity. Darker coloured arrows indicate higher $\alpha^{uv}$. POIs are noted in red. (a) Cluster 1 – O/D map. (b) Cluster 2 – O/D map. (c) Cluster 3 – O/D map. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

**Table 3**
Ten O/D pairs with the highest scaling parameter $\alpha^{uv}$ in terms of global usage. The scaling parameter makes the difference between the pairs of stations $(u, v)$ that exchange or not a lot of bikes.

| *Departure* | *Arrival* |
|---|---|
| Place Trefouel (15005) | Dutot (15016) |
| Chateau de Vincennes (12123) | Murs du Parc (43009) |
| Chevaleret Tolbiac (13053) | Vitry Desault (13120) |
| Nation (12014) | Gare de Reuilly (12019) |
| Aubert (43005) | Murs du Parc (43009) |
| Murs du Parc (43009) | Aubert (43005) |
| Murs du Parc (43009) | Chateau de Vincennes (12123) |
| Aubert (43005) | Fontenay (43008) |
| Fontenay (43008) | Aubert (43005) |
| Place Trefouel (15005) | Alleray (15018) |

To begin with, we assumed cluster 1 to be specific to morning commutes, and this is clearly confirmed in Table 4. The cluster concentrates geolocalized flows of users leaving their homes to go to work, since there is a 59% drop in population density and a jump of 77% and 67% respectively in office worker and shop worker density between the places of departures and arrivals. Furthermore, the geography of cluster 1 is strongly shaped by the railway stations which concentrate many O/D pairs, as can be seen in Fig. 12(a). The O/D maps of cluster 6 and cluster 7 are not shown here since they do not provide anything new: alike cluster 1 they depict high activity around the big railways stations. Such O/D pairs may be part of a "one mile" multi-modal transit journeys (see Fig. 14(a)) where the Vélib' is used to reach a railway station. The same remarks apply to the O/D map of cluster 2, presented in Fig. 12(b): the railway stations are similarly clearly visible even if some arrows seem to be reversed – though not always – since the end-of-day journeys seems to be shorter in distance and duration (Fig. 14). Nevertheless this seems to be logical since this cluster corresponds mainly to the *Work to Home* commute, which fact is also confirmed in Table 4 that shows higher population density near the arrival stations than at the departure stations. The map of cluster 4 is quite different since it mainly presents O/D pairs located in Paris, with many of them around big metro station such as "Place d'Italie", "Châtelet-Les Halles", "Nation", in other words

**Table 4**
For each of the clusters is presented the mean of the densities of population, office workers and shop workers around the departure (origin) stations and arrival (destination) stations of the flows that belong to the cluster, with *origin density mean→destination density mean*.

| *Cluster* | *Inhab./ha* | *Office-workers/ha* | *Shop-workers/ha* |
|---|---|---|---|
| 1 | 289→230 | 234→**311** | 170→**237** |
| 2 | 231→**296** | 286→222 | 211→163 |
| 3 | 284→298 | 265→258 | 182→186 |
| 4 | 312→338 | 234→197 | 169→137 |
| 5 | 289→292 | 223→225 | 161→163 |
| 6 | 347→**264** | 165→**252** | 119→170 |
| 7 | 285→287 | 255→257 | 185→193 |
| 8 | 327→300 | 192→229 | 135→164 |

strategically located places close to many restaurants, cafés and bars. Converse to the *Work to Home* end-of-day journeys, these leisure journeys occurring late at night are longer as shown in Fig. 14(a). For instance, such longer trips might be useful to reach further night service railways or bus stations, such as the Noctilien of Paris. Cluster 3, which shows a high level of activity during the evenings and at night, has (see Fig. 12(c)) O/D pairs spatially located in the city-centre and around neighbourhoods with bars, nightclubs and restaurants (such as Pigalle, Bastille and Mouffetard). Finally, in cluster 5 (Fig. 13(b)), which represents the weekend afternoon profile, the pairs have quite long trip distance (Fig. 14(a)) and are mostly concentrated around parks (Bois de Vincennes, La Villette), canals (Canal Saint-Martin) and nearby world-famous tourist attractions (Eiffel Tower). As for clusters 4 and 5, the spatial distribution is

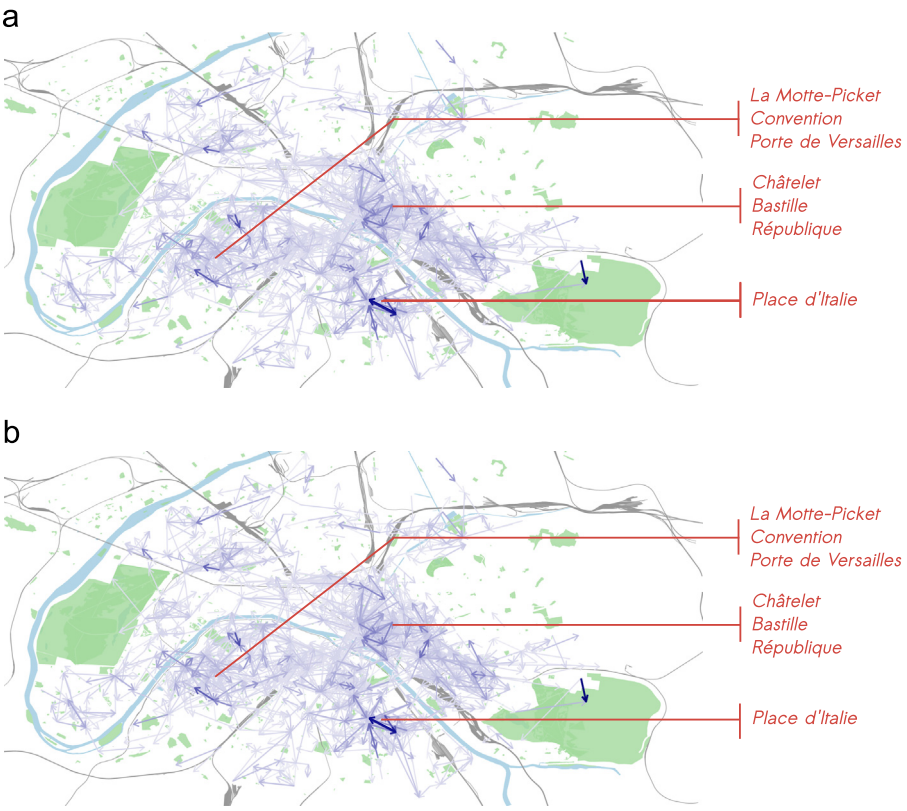


**Fig. 13.** Spatial distribution of the O/D pairs in clusters 4 and 5. (a) Cluster 4 – O/D map. (b) Cluster 5 – O/D map.

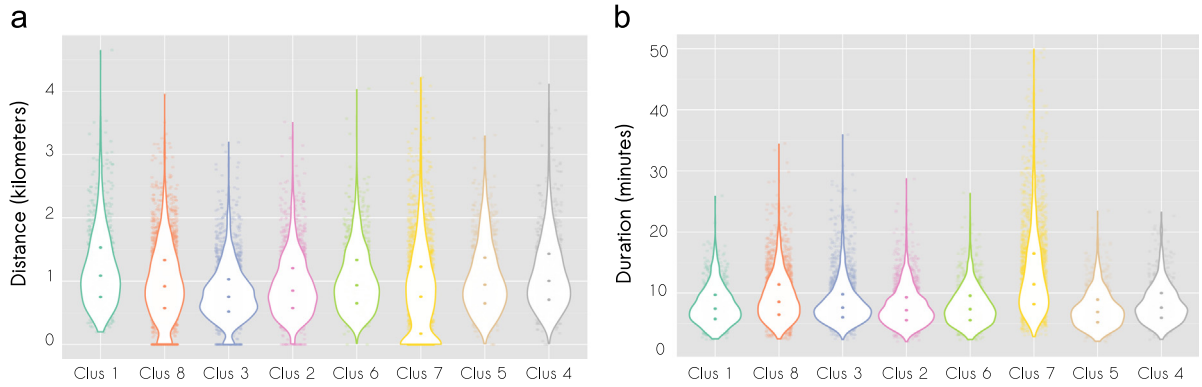


**Fig. 14.** Violin plot of trip distance as the crow flies per cluster (a) and trip duration per cluster (b). From top to bottom the dots represent the quantile 0.75, the mean and the quantile 0.25. (a) Trip distance (crow ies) per cluster and (b) Trip duration per cluster.

more informative than the sociological information provided in Table 4, since it does not contain information related to leisure activities, amenities or services. Finally, the O/D map of cluster 8, which is not presented here, is as average as its temporal profile and shows uniformly distributed trips throughout Paris.

## 6. Conclusions and perspectives

Using a Poisson mixture model and an EM algorithm for parameter estimation, this paper has introduced a family of generative Poisson mixture models for count-time series clustering, which was then applied on usage statistics generated by the Vélib' Bike Sharing System of Paris. It introduces a latent variable that encodes the cluster membership of each Origin/Destination pair. In addition, the most complex model we have propose handles a scaling parameter on station pairs encoding the difference of global observed volume between them, and an observed variable categorizing each day. This latter was specified in several ways and has led to three versions of the models. The first one deals with the difference in usage between weekdays and weekends, the second assumes one specific usage per day of the week and the last one assumes that riders are very much affected by the weather conditions. Model selection criteria were thus applied to compute the predictive power of this approach and to assess to what extent the assumptions encoded within each model describes the global usage of the system. It appears that the scaling factors were mandatory to obtain interesting results and that the usage of the system is mostly governed by the weekday/weekend distinction, which clearly differentiates between utilitarian and recreational usage.

This methodology was then tested to mine trips data from the Paris Vélib' Bike Sharing System, and the results were crossed with sociological information. Relevant answers to the common *why* and *when* of urban mobility analysis can be deduced from the resulting clusters, which are rich and interpretable. They are easily linked to the presence of parks, working places, residential areas and railways stations which underline the strong links between Vélib and other forms of public transport. More generally, such methodology could help to mine other dynamic Origin/Destination data in the field of transport such as in the road transport domain or in the particular case of car sharing-system.

There is nonetheless room for possible methodological improvements. For example, Zero-Inflated Poisson or Negative Binomial laws to model the observed counts may be worth investigating, since they handle the variability of counts and the excesses of zero (on rainy days for example) produced by such a complex sharing-system. Mixture of Generalized Linear Poisson Models that handle continuous variables, such as hourly precipitation intensities and temperature values, may also be an interesting alternative to encode the effects of weather conditions.

## Acknowledgements

The authors wish to thank Cyclocity – JCDecaux and Ville de Paris for providing the Vélib' Origin/Destination data.

## Appendix A. Maximization of the lower bound with respect to the parameters $\lambda_{klt}$

The maximization of the lower bound is detailed in this Appendix:

$$\mathbb{E}[L_c(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})|\mathbf{X},\boldsymbol{\theta}^q] = \sum_{u,v,k} \tau_k^{uv} \log\left(\pi_k \prod_{d,t,l} p(x_{dt}^{uv};\alpha^{uv}\lambda_{klt})^{W_{dl}}\right),$$

with respect to the model parameters $\lambda_{klt}$, under the $K$ equality constraints $\sum_{l,t} D_l \lambda_{klt} = DT, \forall k \in \{1,\ldots,K\}$, where $D_l$ is the number of days in category $l$. Such extrema is computed using the method of Lagrange multipliers and we consequently introduce the Lagrangian $\mathcal{L}(\alpha,\lambda)$ given by

$$\mathcal{L}(\alpha,\lambda) = \sum_{u,v,d,t} \sum_{k,l} \tau_k^{uv} W_{dl} (X_{dt}^{uv} \log(\alpha^{uv}\lambda_{klt}) - \alpha^{uv}\lambda_{klt}) + \sum_k \gamma_k \left(DT - \sum_{l,t} D_l \lambda_{klt}\right), \tag{A.1}$$

with $\gamma_k$ the Lagrange multipliers associated with the $k$th constraints. In what follows are classical computations of the first partial derivative of $\mathcal{L}(\alpha,\lambda)$: first with respect to $\lambda_{klt}$ and then with respect to $\alpha^{uv}$.

### A.1. Maximization of $\mathcal{L}(\alpha,\lambda)$ with respect to $\lambda_{klt}$

The first partial derivative of $\mathcal{L}(\alpha,\lambda)$ with respect to $\lambda_{klt}$ is given by

$$\begin{aligned}\frac{\partial \mathcal{L}(\alpha,\lambda)}{\partial \lambda_{klt}} &= \sum_{u,v,d} \tau_k^{uv} W_{dl} \left(\frac{X_{dt}^{uv}}{\lambda_{klt}} - \alpha^{uv}\right) - \gamma_k D_l \\ &= \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - \sum_d W_{dl}\lambda_{klt} \sum_{u,v} \tau_k^{uv}\alpha^{uv} - \gamma_k D_l \lambda_{klt} \\ &= \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \sum_{u,v} \tau_k^{uv}\alpha^{uv} - D_l \lambda_{klt}\gamma_k \\ &= \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \left(\sum_{u,v} \tau_k^{uv}\alpha^{uv} - \gamma_k\right)\end{aligned} \tag{A.2}$$

The Lagrange multipliers are obtained by setting expression (A.2) to zero:

$$\begin{aligned}&\frac{\partial \mathcal{L}(\alpha,\lambda)}{\partial \lambda_{klt}} = 0 \\ &\Leftrightarrow \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \left(\sum_{u,v} \tau_k^{uv}\alpha^{uv} - \gamma_k\right) = 0 \\ &\Leftrightarrow \sum_{l,t} \left(\sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \left(\sum_{u,v} \tau_k^{uv}\alpha^{uv} - \gamma_k\right)\right) = 0 \\ &\Leftrightarrow \sum_{u,v,d,t} \tau_k^{uv} X_{dt}^{uv} - \sum_{l,t} D_l \lambda_{klt} \left(\sum_{u,v} \tau_k^{uv}\alpha^{uv} - \gamma_k\right) = 0 \\ &\Leftrightarrow \sum_{u,v,d,t} \tau_k^{uv} X_{dt}^{uv} - DT \left(\sum_{u,v} \tau_k^{uv}\alpha^{uv} - \gamma_k\right) = 0 \\ &\Leftrightarrow \gamma_k = \frac{1}{DT} \sum_{u,v,d,t} \tau_k^{uv} X_{dt}^{uv} - \sum_{u,v} \tau_k^{uv}\alpha^{uv}\end{aligned} \tag{A.3}$$

Finally, the expression (A.3) of the Lagrange multiplier is used in (A.2):

$$\begin{aligned}&\frac{\partial \mathcal{L}(\alpha,\lambda)}{\partial \lambda_{klt}} = 0 \\ &\Leftrightarrow \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \frac{1}{DT} \sum_{u,v,d,t} \tau_k^{uv} X_{dt}^{uv} = 0 \\ &\Leftrightarrow \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \sum_{u,v,d,t} \tau_k^{uv} \frac{1}{DT} X_{dt}^{uv} = 0 \\ &\Leftrightarrow \sum_{u,v,d} \tau_k^{uv} W_{dl} X_{dt}^{uv} - D_l \lambda_{klt} \sum_{u,v,d,t} \tau_k^{uv}\alpha^{uv} = 0 \\ &\Leftrightarrow \hat{\lambda}_{klt} = \frac{1}{\sum_{u,v} \tau_k^{uv}\alpha^{uv} \sum_d W_{dl}} \sum_{u,v,d} \tau_k^{uv} W_{dl} x_{dt}^{uv}\end{aligned} \tag{A.4}$$

and results the update rules for the parameter $\lambda_{klt}$.

### A.2. Maximization of $\mathcal{L}(\alpha,\lambda)$ with respect to $\alpha^{uv}$

Similarly, the first partial derivative of $\mathcal{L}(\alpha,\lambda)$ with respect to $\alpha^{uv}$ expresses as

$$\begin{aligned}\frac{\partial \mathcal{L}(\alpha,\lambda)}{\partial \alpha^{uv}} &= \sum_{d,t}\sum_{k,l}\tau_k^{uv} W_{dl}\left(\frac{x_{dt}^{uv}}{\alpha^{uv}} - \lambda_{klt}\right)\\ &= \sum_{d,t}\sum_{k,l}\tau_k^{uv} W_{dl}(x_{dt}^{uv} - \alpha^{uv}\lambda_{klt})\\ &= \sum_{d,t} x_{dt}^{uv} - \alpha^{uv}\sum_{k}\tau_k^{uv}\sum_{l,t}\sum_{d} W_{dl}\lambda_{klt}\\ &= \sum_{d,t} x_{dt}^{uv} - \alpha^{uv}\sum_{k}\tau_k^{uv}\sum_{l,t} D_l\lambda_{klt}\\ &= \sum_{d,t} x_{dt}^{uv} - DT\alpha^{uv} \end{aligned} \tag{A.5}$$

Setting the derivatives to zero, we obtain the update rule for the $\alpha^{uv}$:

$$\begin{aligned}&\frac{\partial \mathcal{L}(\alpha,\lambda)}{\partial \alpha^{uv}} = 0\\ &\Leftrightarrow \sum_{d,t} x_{dt}^{uv} - DT\alpha^{uv} = 0\\ &\Leftrightarrow \hat{\alpha}^{uv} = \frac{1}{DT}\sum_{d,t} x_{dt}^{uv}\end{aligned} \tag{A.6}$$

## References


[1] M. Bagchi, P. White, The potential of public transport smart card data, Transp. Policy 12 (2005) 464–475.
[2] M. Benchimol, P. Benchimol, B. Chappert, A. De La Taille, F. Laroche, F. Meunier, L. Robinet, Balancing the stations of a self-service bike hire system, RAIRO-Oper. Res. 45 (2011) 37–61.
[3] P. Borgnat, E. Fleury, C. Robardet, A. Scherrer, Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program, in: F. Kepes (Ed.), European Conference on Complex Systems, ECCS'09, Complex Systems Society, Warwick, United Kingdom, 2009.
[4] P. Borgnat, C. Robardet, P. Abry, P. Flandrin, J. Rouquier, N. Tremblay, A dynamical network view of Lyon's Vlo'v shared bicycle system, Dynamics On and Of Complex Networks, vol. 2, Springer, Berlin, Heidelberg, 2013.
[5] P. Borgnat, C. Robardet, J.B. Rouquier, A. Parice, E. Fleury, P. Flandrin, Shared bicycles in a city: a signal processing and data analysis perspective, Adv. Complex Syst. 14 (2011) 1–24.
[6] D. Chemla, F. Meunier, R. Wolfler Calvo, Balancing a bike-sharing system with multiple vehicles, in: Congrs annuel de la socit Franaise de recherche oprationelle et d'aide la dcision, ROADEF2011, Socit Franaise de recherche oprationelle, Saint-Etienne, France, 2011.
[7] J. Dill, Bicycling for transportation and health: the role of infrastructure, J. Public Health Policy 30 (2009) 95–110.
[8] J. Dill, T. Carr, Bicycle commuting and facilities in major U.S. cities: if you build them, commuters will use them, Transp. Res. Rec.: J. Transp. Res. Board 1828 (2003) 116–123.
[9] J. Froehlich, J. Neumann, N. Oliver, Measuring the pulse of the city through shared bicycle programs, in: UrbanSense08, 2008, pp. 16–20.
[10] J. Froehlich, J. Neumann, N. Oliver, Sensing and predicting the pulse of the city through shared bicycling, in: 21st International Joint Conference on Artificial Intelligence, IJCAI'09, AAAI Press, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2009, pp. 1420–1426.
[11] A. Hofleitner, R. Herring, P. Abbeel, A.M. Bayen, Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network, IEEE Trans. Intell. Transp. Syst. 13 (2012) 1679–1693.
[12] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, R. Banchs, Urban cycles and mobility patterns: exploring and predicting trends in a bicycle-based public transport system, Pervasive Mobile Comput. 6 (2010) 455–466.
[13] D. Karlis, L. Meligkosidou, Model based clustering for multivariate count data, in: 18th International Workshop on Statistical Modelling, Katholieke Universiteit Leuven, 2003, pp. 211–216.
[14] N. Lathia, A. Saniul, L. Capra, Measuring the impact of opening the London shared bicycle scheme to casual users, Transp. Res. Part C: Emerg. Technol. 22 (2012) 88–102.
[15] E. Lebarbier, T. Mary-Huard, Le critère BIC : fondements théoriques et interprétation, Research Report, INRIA, 2004.
[16] J. Lin, T. Yang, Strategic design of public bicycle sharing systems with service level constraints, Transp. Res. Part E: Logist. Transp. Rev. 47 (2011) 284–294.
[17] L. Liu, A. Hou, A. Bidderman, C. Ratti, J. Chen, Understanding individual and collective mobility patterns from smart card records: a case study in Shenzhen, in: Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, October 3–7, 2009.
[18] G.J. Mclachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, NY, USA, 1996.
[19] G.J. Mclachlan, D. Peel, Finite Mixture Models, Wiley, New York, NY, USA, 2000.
[20] G. Michau, C. Robardet, L. Merchez, P. Jensen, P. Abry, P. Flandrin, P. Borgnat, Peut-on attraper les utilisateurs de vlo'v au lasso ?, in: XXIIIe Colloque GRETSI - Traitement du Signal et des Images, GRETSI, 2011, pp. 46–50.
[21] M. Miranda-Moreno, T. Nosal, L. Fernando, Weather or not to cycle; temporal trends and impact of weather on cycling in an urban environment, Transp. Res. Rec. 2247 (2011) 42–52.
[22] R. Nair, E. Miller-Hooks, R.C. Hampshire, A. Bušić, Large-scale vehicle sharing systems: analysis of Vélib', Int. J. Sustain. Transp. 7 (2012) 85–106.
[23] M. Pelletier, M. Trpanier, C. Morency, Smart card data use in public transit: a literature review, Transp. Res. Part C 19 (2011) 557–568.
[24] P. Pucher, R. Buehler, Making cycling irresistible: lessons from the Netherlands, Denmark and Germany, Transp. Rev. 28 (2008) 495–528.
[25] A. Rau, G. Celeux, M.L. Martin-Magniette, C. Maugis-Rabusseau, Clustering High-throughput Sequencing Data with Poisson Mixture Models, Technical Report RR-7786, INRIA, 2011.
[26] C. Roth, S.M. Kang, M. Batty, M. Barthélemy, Structure of urban movements: polycentric activity and entangled hierarchical flows, PLoS ONE 6 (2011).
[27] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[28] S.J. Thomas, Model-based clustering for multivariate time series of counts (Ph.D. thesis), Rice University, 2010.
[29] T. Thomas, R. Jaarsma, B. Tutert, Temporal variations of bicycle demand in the Netherlands: influence of weather on cycling, In: TBR Annual Meeting 2009 Compendium of Papers DVD, TBR, Washington D.C., January 11–15, 2009, pp. 1–15.
[30] P. Vogel, T. Greiser, D. Mattfeld, Understanding bike-sharing systems using data mining: exploring activity patterns, Procedia – Soc. Behav. Sci. 20 (2011) 514–523.
[31] P. Vogel, D. Mattfeld, Strategic and operational planning of bike-sharing systems by data mining – a case study, in: ICCL, Springer, Berlin, Heidelberg, 2011, pp. 127–141.
[32] Z. Wangsheng, L. Shijian, P. Gang, Mining the semantics of origin–destination flows using taxi traces, in: ACM Conference on Ubiquitous Computing, UbiComp'12, ACM, New York, NY, USA, 2012, pp. 943–949.
[33] J. Yuan, Y. Zheng, W. Xie, X. Xie, Y. Huang, T-drive: driving directions based on taxi trajectories, in: 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS'10, ACM, New York, USA, 2010, pp. 99–108.
[34] Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs, in: 13th ACM Conference on Ubiquitous Computing, UbiComp 2011, ACM, Beijing, China, 2011, pp. 89–98.


**Andry Njato Randriamanamihaga** received his Computer Science degree and his Master of Science degree in "Systèmes Intelligents pour les Transports" in 2012, from the *Université de Technologie de Compiègne*, France. He is currently a first-year Ph.D. student at the *French Institute of Science and Technology for Transport, Development and Networks* (IFSTTAR) and is interested in clustering Dynamic Origin/Destination matrices to mine large-scale data in the field of transport.

**Etienne Côme** is a researcher of the *French Institute of Science and Technology for Transport, Development and Networks* (IFSTTAR), Paris. He received his Master of Science degree in 2005, from the *Université de Technologie de Compiègne*, France, on "Technologie de l'information et des systèmes" and a Ph.D. from the same institution in 2009. His current research interests concern random graph models, mixture models and more generally probabilistic graphical models and their use to solve transportation problems.

**Latifa Oukhellou** received a Ph.D. from *Université Paris-Sud* (in Automatic and Signal Processing) in 1997 and an "Habilitation à diriger des Recherches" from *Université Paris-Est* in 2010. She is currently a senior researcher at the *French Institute of Science and Technology for Transport, Development and Networks* (IFSTTAR) and has been assistant professor at *Université Paris-Est Créteil* (UPEC). Her research interests concern pattern recognition, machine learning and information fusion applied to diagnosis problems as well as to spatio-temporal data analysis for supporting driving behaviour and mobility.

**Gérard Govaert** received his Computer Science degree in 1972 from *École Normale Supérieure, Cachan* (ENS) and obtained his Ph.D. on "Clustering with adaptive distance" in 1975 from *Université Paris-6*. He received his "Thse d'Etat" degree on bi-clustering in 1983 from *Université Paris-6*. Gérard Govaert is currently a professor at the *Université de Technologie de Compiègne*, researcher at the CNRS Laboratory Heudiasyc (Heuristic and Diagnostic of Complex Systems). His research interests include cluster analysis, statistical pattern recognition, model-based clustering and bi-clustering.