Model-based count series clustering for Bike-sharing system usage mining, a case study with the Vélib' system of Paris.

CÔME ETIENNE and OUKHELLOU LATIFA

November 21, 2012

Abstract:
The bicycle sharing systems are increasingly numerous nowadays. These transportation systems generate sizable transportation data the mining of which can reveal the underlying urban phenomenons linked to city dynamics. This paper introduces a statistical model to automatically analyze bike sharing system trips data. This model will introduce a latent variable to partition the stations in terms of their temporal dynamics over the day with respect to the number of rented and returned bikes. This generative model is based on Poisson mixtures and introduces a station scaling factor that handles the discrepancy between the stations activities. Eventually, the difference of dynamics between week days and week-end will also be taken into account. This model will find the latent factors that shape the geography of trips. The results produced by such an approach give insights on the relationships between stations neighborhoods type (the amenities it offers, its sociology, ...) and the generated mobility pattern. In other words, the proposed method enables the discovery of regions of different functions, that induce specific usage patterns in BSS data. These potentials are demonstrated through an in-depth analysis of the results obtained on the Vélib' large-scale bike sharing system of Paris.

Key words: Bike-sharing systems, count data, clustering, generative model, model-based clustering

1 Introduction

The growth of population and of urban concentrations, as well as the increase of nuisance factors such as pollution, noise, congestion, greenhouse gas emissions, have urged the development of new sustainable mobility strategies in urban areas. Public authorities need to deploy urban mobility policies to organize differently passenger mobility, thus lessening the negative impact of mobility demands. One possible way adopted by cities and regions to face these problems

is the promotion of soft modes of transport such as walking and cycling, which are economical, healthy, less pollutant and more equitable [26, 6, 11].

The implementation of Bike Sharing Systems (BSSs) is one of the urban mobility measures proposed in many cities all over the world as an additional mean of sustainable intermodal transport. These last years, different BSSs have been implemented in European cities. The main motivation behind this concept is to provide users with free or rental bicycles especially suited for short distance trips in urban areas, thus reducing traffic congestion, air pollution and noise. In Europe BSSs are most popular in southern European countries, where a cycling tradition does not exist. BSSs become a modern mode of urban mobility. Due to their incontestable success [8, 6], more and more cities want to supply this mode of mobility in order to present themselves as sustainable and modern. In France, since the implementation of the BSS in Lyon in 2005 (called Vélo'v), bicycle sharing schemes have been launched in twenty French cities, among which one of the most large-scale bicycle share scheme, implemented in Paris (called Vélib').

A good knowledge of BSS usage and performance is the key of its success. This knowledge can be transferred afterwards to cities aiming to incorporate BSSs. To do this, the analysis of the data collected by BSSs operators and cities is instructive as shown in several studies like [14, 4, 19]. A statistical analysis of the collected data on such a scheme contributes to leverage the development of new and innovative approaches for a better understanding of urban mobility, as well as of the use and performance of BSSs. The design of BSSs, the adjustment of pricing policies, the improvement of service level of the system (redistribution of bikes over stations) can benefit from this kind of analysis [9, 20, 1]. It also helps sociologists and planners to apprehend the users mobility patterns within the cities.

However, the data collected on such systems are frequently sizable. It is therefore difficult to gain knowledge from them without the help of automatic algorithms that extract spatio-temporal patterns and give a synthetic view of the information. Many data sets collected on human mobility can indeed help to recover underlying urban phenomenons linked to city dynamics. Human mobility can be captured through GPS trajectories of vehicles or pedestrians [31], cell phone usage [27], as well as data related to bicycle sharing systems as it is the case here. This paper deals with a statistical model that will automatically cluster BSS stations according to their usage profile. The performed analysis will help in understanding the BSS stations attractiveness, in relation with city geography and sociology. In fact, the proposed method enables the discovery of regions of different functions, that induce specific usage pattern in BSS data. The model proposed here shares therefore some objectives with those highlighted in [30], *i.e.* finding functional regions in a city through the mining of mobility data (taxi trips in this application). However, the specific nature of the transport mode analyzed here (which is mainly used for short distance trips) requires the development of a particular model more fitted to these data. The clustering of the BSSs stations is closely related to the city functionalities (transport, leisure, employments) and can benefit a variety of applications, including urban planning and location choosing for a business as cited by the previous authors.

But, the analysis of the results provided by the model will furthermore give insights into the relations between the kind of neighborhood of the stations (the type of amenities it offers, its sociology, ...) and their associated usage profiles. The crossing of the model results with social and economical data is to this end carried out, and will show the close links between these two aspects and the use of bike sharing transport mode. This may at the end help for bikes redistribution planning and for designing new BSSs.

In an attempt to handle the challenges above, this paper has the following main contributions. A dedicated model based on count series clustering is developed in order to highlight spatio-temporal patterns in the BSS usage data. The model uses trips data to describe the station usage. A generative mixture model is proposed and an EM algorithm is derived to learn the model parameters and to perform the station clustering. The formalization of the model is general enough to take into account specific hypotheses related to the BSS case study. The proposed approach is validated through extensive investigations carried out on data collected on the Paris large-scale bicycle sharing scheme (Vélib').

This paper is thus organized as follows, in Section 2 we present a survey of previous works in relevant literature. The Vélib' case is detailed in Section 3. Section 4 is devoted to the proposed statistical model based upon count series clustering. Results are then given and discussed in Section 5, prior to a conclusion in Section 6.

2 Related work

Mobility patterns are traditionally analyzed through human and social sciences frameworks. The data used for such studies are collected either from sensing devices or through observational mechanisms, e.g. surveys. But, the emergence of information and communication technologies, as well as the advent of new observations and tracking capabilities, have boosted the availability of sizable spatio-temporal data. The availability of this kind of datasets contributes to emphasize the importance of the development of novel approaches based upon engineering and computer sciences. Indeed, tools for processing spatio-temporal data are needed for a better understanding of mobility patterns of travelers and goods, as well as of the use and performances of transportation systems.

Several requirements have motivated previous studies dealing with BSSs: improvement of existing systems, growth of knowledge on urban mobility, and more generally developing the BSSs of tomorrow. The design of new BSSs can benefit from the experience gathered on existing systems the analysis of which can help to better understand their usage. The long-term goal is to be able, before and after BSS implementation, to optimize station planning in terms of both urban planning, mobility needs and redistribution capacities of the system.

One of the main issues raised by users in recent surveys concerns the availability of bikes: users are confronted to empty stations when renting bikes, and full stations when returning them back. Redistribution of bikes is indeed necessary in most bicycle sharing schemes to compensate the uneven demand of

users by relocating the bikes among the stations, thus ensuring a good quality of service of the system. This is generally performed by redistribution trucks driving around the city moving bikes between stations. Several research works address the issues related to the optimization of bikes redistribution policies. The reader can refer to [2, 7, 24] for more information on this topic.

Other works from computer science or signal processing domains have been proposed to study the existing BSSs. The proposed approaches differ according to the kind of data they use and the goal they aim to solve. The collected data on existing systems could correspond to station state statistics such as station occupancy over the day, or over several time frames. They could also be of origin and destination (OD) matrix form, assuming available records of BSS trips every day, i.e. for each trip, the location of station and the starting time, the destination station and the stopping time are recorded. Two main topics have been investigated, namely clustering and prediction. Whereas the clustering topic aims to uncover spatio-temporal patterns in the BSS usage by partitioning the stations into different clusters having a similar usage, the prediction topic focuses on developing models able to predict the state of the stations (the number of bikes per station) or more globally the state of the network over time.

The reader interested by the prediction problem can refer to [14], [5], [17] and [23]. In the first study, the problem of forecasting near-term station usage is addressed by using Bayesian Networks the performance of which are analyzed with respect to factors such as time of day and station activity. The same problem is addressed in [17], using a time series analysis based on an ARMA process. Borgnat et al. predict the global rental volume using the cyclostationarity of the temporal series and finally Michau et al. attempt to relate, through a parsimonious statistical regression model, social, demographic and economical data of the various neighborhoods of the city with the actual number of trips made from and to the different parts of the city.

In all the clustering studies carried out until now, the bicycle sharing stations are grouped according to their usage patterns, thus highlighting the relationships between time of day, location and usage. The proposed approaches differ according to how they describe the stations usages and the clustering techniques they use. The first attempt in this line of work is due to Froehlich et al. They have analyzed in two studies a dataset from Barcelona Bicing system by means of clustering techniques. In [15], a mixture of Gaussian is used to cluster the stations according to a feature vector build from station state statistics, whereas in [14], two clusterings are compared both being performed by hierarchical aggregation. The first one use activity statistics derived from the evolution of station state while the second use directly the number of available bicycles along the day.

Other studies like [19] use similar clustering techniques to study the effect of changing the user-access policy in the London Barclays cycle hire scheme. The authors investigate how the change affected the system usage throughout the city via both spatial and temporal analysis of station occupancy data. Another approach proposed by [4] analyzes both temporal and spatial usage trends of the

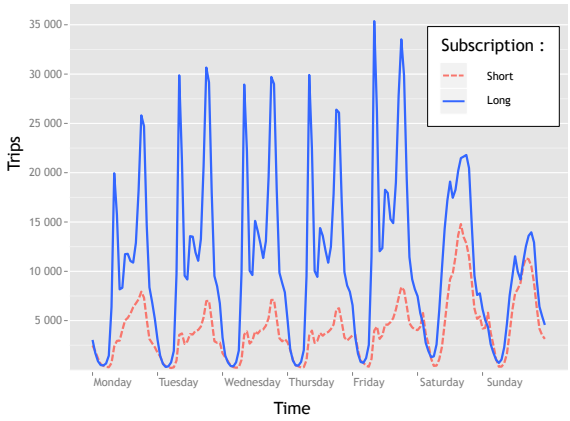

Figure 1: Number of recorded trips per hour and day of the week during April 2011 with respect to the type of subscription (long: one year, short: one day).

Lyon Vélo'v BSS. To this end two solutions are investigated, the first one is based on a graph clustering algorithm (Modularity optimization, [25]) that uncover communities of stations that exchange bikes in a preferential way. The second proposal aims to cluster flows of activity between stations that exhibit similar exchange dynamics using a K-means algorithm. In this paper, we investigate the analysis of the Vélib' system through the daily recorded trips during a month. Before detailing the proposed approach, a description of the Vélib'program is presented in the next section.

3 The Vélib' case

3.1 Historical

Since 2001, the city of Paris deploys urban policies aiming to favor public transportation and soft modes of transport such as bicycle, walking ... . Within this context, the Vélib' bike sharing system, has been launched in July 2007. Vélib' is operated as a concession by Cyclocity, a subsidiary company of the French advertising corporation JCDecaux. 7000 bikes were initially distributed on 750 fixed stations. Five years ago, the Vélib' system has been extended to reach 20 000 bikes spread out over 1 208 fixed stations and 224 000 annual subscribers with an averaged number of 110 000 travels each day. Vélib' is a large-scale scheme, the second largest BSS in the world after the BSS launched in China. Vélib' is available mainly in Paris *intramuros* but some stations are located in the suburbs of Paris. Vélib' offers a non stop service (24/7). Each Vélib' station is equipped with an automatic rental terminal. The whole network includes 40 000 docking points (between 8 and 70 per station). The bikes are locked to the electronically controlled docking points. Users can purchase a short-term subscription, over a day or a week, or a long-term subscription over a year. The

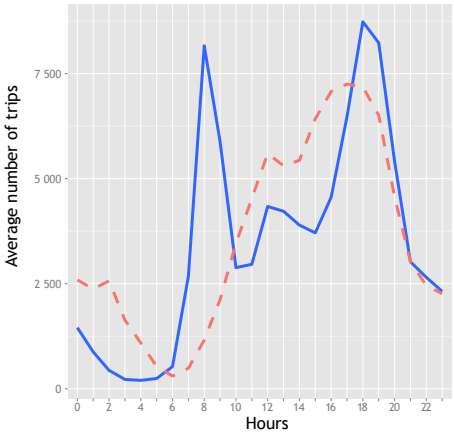

Figure 2: Average number of trips per hour during a week day (plain blue line) and a week-end day (dashed red line).

subscription allows an unlimited number of rentals, the first half hour (or the first 45 minutes for long-term subscription) of every individual trip being free. Registration of users is required. The bicycles can be hired at any of the stations and at any time and returned back at any other station and at any time.

Despite the boost in bike use in Paris that have followed the implantation of the BSS the cycling modal share is still very low compared to other cities in Europe. Analyzing modal splits [6] in Paris can give hints about the local cycling culture. Cycling share is still very low in Paris (3%) but has been on the increase in the last years. The modal part of BSS is about 2%. Public transport has an estimated modal share of 40% while car share is estimated to 21% [1]. Even if France has not a strong cycling culture (the primary purpose of cycling is for leisure), people seem to be very enthusiastic with bike public plans. Bike is viewed as environmentally friendly by 62% of people in France [1].

3.2 General view of the system

The aim here is to perform some general statistics to highlight global trends in the usage of Vélib'. Figure 1 shows the whole number of recorded Vélib' trips per hour and day of the week during a month with respect to the type of subscription (day or year). This figure reveals that the Vélib' usage is closely linked not only to the hour and the day it occurs but also to the kind of day (weekday or weekend) and to the type of subscription.

A first significant difference in Vélib' usage between short-term and long-term subscribers can be noticed. This difference is reflected in terms of the usage volume: most of the Vélib' trips are generated by long-term subscribers even if the difference between the two subscriptions is less important during the weekend. This remark can be linked to the fact that short-term subscriptions are mainly associated to leisure while Vélib long-term subscriptions tend to

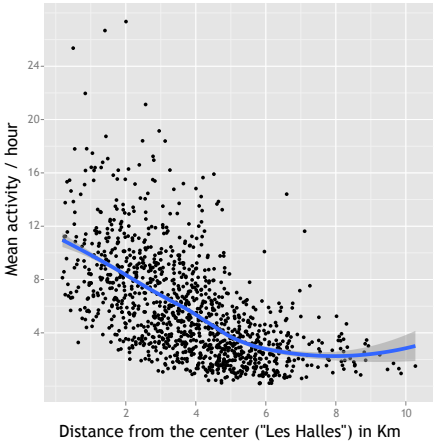

Figure 3: Average activity of stations (number of actions: departure or arrival) per hour with respect to the distance of the stations from the center of Paris ("Les Halles").

cover daily mobility routines of users.

This figure shows also a difference in Vélib' usage during the weekdays and the weekend. A cyclostationarity pattern can be seen in the Vélib' usage during the weekdays. Three peaks of weekday usage can indeed be distinguished in Figure 2: the two most significant correspond to the commutes (8am and 18am) while the third one which appears at 12am, can be associated to the lunch break. As can be expected, the morning peak usage disappears during the weekend, where the Vélib' usage gradually increases to reach a maximum in the afternoon. One can notice that Friday corresponds to the peak of usage among the weekdays. These temporal trends of BSS usage can be informative on sociological characteristics of the city. Considering the study carried out by [14] on the Barcelona Bicing system, some sociological differences between the two cities can indeed be highlighted. The lunch peak occurring at 2pm on Barcelona Bicing data occurs at 12am for Vélib' data, reflecting thus the late lunch culture of Spain (resp. the earlier lunch culture of France). Secondly, Friday is the least active day in Barcelona Bicing usage (resp. the most active one in Vélib usage).

Simultaneously with those temporal trends in the use of Vélib bicycles, spatial trends closely linked to geographical aspects of the city can also be identified. Figure 3 shows the average activity of stations per hour, quantified through the number of rented and returned bikes with respect to the distance from the center of Paris. It is clear that the mean activity of a station is more significant if the station is located near the center of Paris. Furthermore, the duration and distance of trips can also be used as indicators of the Vélib' usage. As shown in Figure 4, half of the Vélib' trips last twelve minutes. This can be linked to the pricing policy of the Vélib' (free for half an hour). It can be noticed that

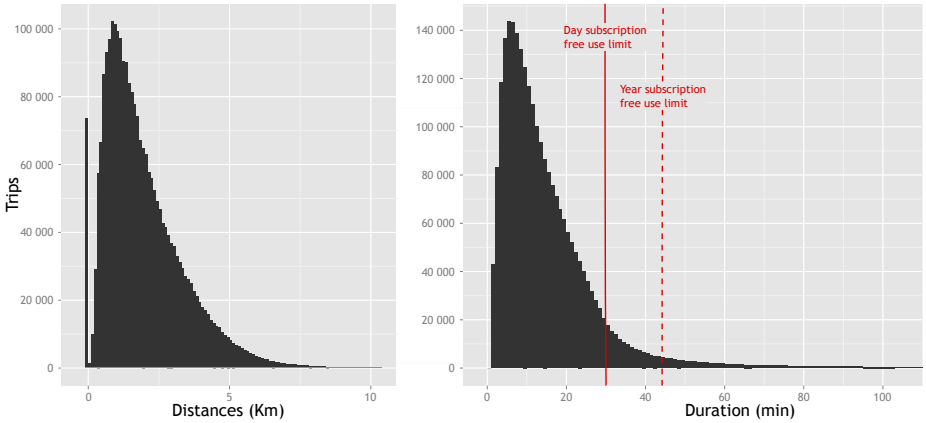

Figure 4: Histogram of trips length in kilometers (left) and of trips duration in minutes (right).

the trips recorded with null distances correspond to loop travels: users rent and return back bikes in the same station.

These first statistics carried out in this section show the global dynamic of the Vélib' system. The next section will present the statistical model proposed to automatically extract thiner details from BSS data .

4 Count series clustering

The approach undertaken here follows the line of work initiated by [15] and [19], but with a new tool tailored to fit the specificity of the data. These two previous studies propose to cluster the stations of a BSS with respect to their temporal dynamics over the day, in order to find spatio-temporal patterns which describe the activity of the city. The clusters produced by this kind of approaches group the stations based on their usage patterns. It is then possible to map the obtained clusters and see how usage relates to the city geography and sociology.

In [15] the clustering is illustrated on Bicing data from Barcelona. The data correspond to station state statistics in form of free slots, available bikes over several time frames and other station activity statistics derived from stations state raw data scraped every 5min. The clustering is then performed with a Gaussian Mixture model estimated by an EM algorithm. In [14] and [19] each station is described by a time series vector which corresponds to the normalized available bicycle (NAB) value of the station along the day. Each element of the feature vector is therefore equal to the number of available bicycles divided by the station size (the 95th percentile of the sums of free slots and available bikes). These time series are then smoothed using a moving average and clustered using a hierarchical agglomerative algorithm [12] p. 552, with a cosine distance. Such approaches have demonstrated that spatio-temporal patterns can be extracted

from usage data.

The approach proposed here aims also to find spatio-temporal pattern in the BSS usage data, but differs from these previous propositions with respect to several points. First, to describe the stations dynamics we do not use the stations states but count time series derived from trips data. Such a description of the stations dynamics is significantly ampler than the one used in these two previous studies since informations on arrivals and departures are available. In particular, it may differentiate situations where a lot of bikes come and leave from the station from the cases where there is no activity at the station, whereas the descriptions built from station state data cannot account for such a difference since in both cases the state of the station is unchanged. The proposed model will also deal with the differences of temporal dynamics observed during week days and week-end days (as observed in Figure 2), whereas the features used in the previous studies do not describe these differences. Eventually, the proposed model will handle the specific nature of the observations, *i.e.* that they are counts and therefore belong to $\mathbb{N}$.

To achieve these goals, we propose a generative mixture model and derive the associated EM algorithm to estimate the parameters of the model and the clustering. This work adopts therefore the model-based clustering framework [22, 13] with specific hypotheses related to the phenomena under analysis that we discuss in the next paragraphs. But we first describe more formally the stations features vectors construction and introduce the notations used in the rest of the paper.

4.1 Trips data and count time series

The dataset used here corresponds to one month of trips data (April 2011) and contains the following informations for each trip: station of departure, time of departure, station of arrival, time of arrival, type of user subscription (day / year). From roughly 2 500 000 trips recorded in April 2011, the following counts statistics are derived:

- X^{out}_{sdh} : Number of bikes leaving from station $s \in \{1, \dots, S\}$ during day $d \in \{1, \dots, D\}$ and hour $h \in \{1, \dots, 24\}$;
- X^{in}_{sdh} : Number of bikes coming to station $s \in \{1, \dots, S\}$ during day $d \in \{1, \dots, D\}$ and hour $h \in \{1, \dots, 24\}$.

The aggregation at 1 hour was used to produce the counts since it gives a good trade-off between resolution of details and fluctuations [4]. These two time series of counts are then concatenated in a vector $\mathbf{X}_{sd}$ describing the arrival and departure activity of station s during day d

$$\mathbf{X}_{sd} = (X^{in}_{sd1}, \dots, X^{in}_{sd24}, X^{out}_{sd1}, \dots, X^{out}_{sd24}). \tag{1}$$

These activity vectors can then be arranged in a tensor (or three-way array) of size $N \times D \times T$, with N the number of stations (1 185 in the Vélib' case), D the

number of available days in the dataset (30 for this study) and T the length of the description vector, which is 48 here since non-overlapping windows of one hour are used to compute the arrivals and departures counts.

4.2 Generative Model

Since the observed data are counts, we propose to use Poisson mixtures to build the generative model. Poisson mixtures have already been used successfully in several applicative domains and can take different forms depending on specific assumptions [28, 29, 18, 16]. Like these previous works we will consider that conditionally on the clusters the observed variables are drawn from Poisson distributions, but we will adapt the model to our needs by making further hypotheses on the model parametrization. The generative model that we propose uses two additional sets of variables. The first a classic one corresponds to indicator variables (denoted by Z_s) which encode the cluster membership of the stations and take their values in $\mathcal{Z} = \left\{\{0,1\}^K : \sum_k Z_{sk} = 1\right\}$, these variables are not observed and must be recovered. The variables in the second set denoted by W_d are also indicator variables, but they are attached to the days and encode the differences between week and week-end days (which present very different usage profiles visible in Figure 1 and 2). These variables take their value in $\mathcal{W} = \left\{\{0,1\}^2 : \sum_l W_{dl} = 1\right\}$ and we consider that they are observed. Using these two sets of variables the following generative model is then assumed for the observed data:

$$\begin{aligned} Z_s &\sim \mathcal{M}(1,\pi) \\ X_{sd1} \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_{sdT} &\mid \{Z_{sk}=1, W_{dl}=1\} \\ X_{sdt}|\{Z_{sk}=1, W_{dl}=1\} &\sim \mathcal{P}(\alpha_s \lambda_{klt}), \end{aligned}$$

with $\mathcal{P}(\lambda)$ the Poisson distribution of parameter λ and $\mathcal{M}(1,\pi)$ the Multinomial distribution of parameter π. This generative model assumes therefore that knowing the cluster of the station and the cluster of the day the departure and arrival counts of each hour are independent and that they follow a Poisson distribution of parameter $\alpha_s \lambda_{klt}$. The parameter α_s is a scaling factor specific to station s and will capture the global activity of the station. The parameters λ_{klt} describe the temporal variations of departure / arrival and are specific to each station clusters and day type (week / week-end). For the parameters to be identifiable we must have constraints on the λ. The following constraints will ensure that the model is identifiable up to the permutation undetermination unavoidable in all mixture models.

$$s.t. \sum_{l,t} D_l \lambda_{klt} = DT, \forall k \in \{1,\ldots,K\}, \tag{2}$$

with $D_l = \sum_d W_{dl}$ the number of days in day cluster l. The conditional independence assumption relates this model to the naive Bayes model, and can be criticized; it is nonetheless a good first approximation. The Poisson hypothesis

is natural for count data and furthermore it enables the introduction of the station scaling factor α_s [28, 16]. These scaling factors are necessary to produce interesting results since we may observe a lot of activity variations between stations with a clear centrality effect (see Figure 3). Using these assumptions the conditional density of an activity vector $\mathbf{x}_{sd}$ can be derived as:

$$f(\mathbf{x}_{sd}|\{Z_{sk}=1, W_{dl}=1\}) = \prod_{t,l} p(\mathbf{x}_{sdt}; \alpha_s\lambda_{klt})^{W_{dl}} = \prod_{t,l} \left(\frac{(\alpha_s\lambda_{klt})^{\mathbf{x}_{sdt}}}{\mathbf{x}_{sdt}!} \exp^{-\alpha_s\lambda_{klt}} \right)^{W_{dl}},$$

with $p(., \lambda)$ the density of a Poisson distribution of mean λ. Therefore, the log-likelihood of such a model is given by:

$$L(\mathbf{\Theta}; \mathbf{X}|\mathbf{W}) = \sum_s \log \left(\sum_k \pi_k \prod_{d,t,l} p(X_{sdt}; \alpha_s\lambda_{klt})^{W_{dl}} \right) \tag{3}$$

The maximization of this quantity can be achieved by an EM type algorithm described in the next section.

4.3 EM Algorithm

The EM algorithm [10, 21] is a popular algorithm for maximum likelihood estimation in statistics when the problem involves missing values or latent variables. It is an iterative algorithm that alternates between maximizing a lower bound of the log-likelihood and updating the bound. This bound is classically obtained from the completed likelihood which introduces the latent variable Z:

$$Lc(\mathbf{\Theta}; \mathbf{X}, \mathbf{Z}) = \sum_{s,k} Z_{sk} \log \left(\pi_k \prod_{d,t,l} p(X_{sdt}; \alpha_s\lambda_{klt})^{W_{dl}} \right) \tag{4}$$

During the E step of the algorithm the conditional expectation of this function over Z with respect to the current parameter values is computed. This expectation will provide the lower bound of the log-likelihood that will be maximized during the M step. This expectation is given by:

$$\mathbb{E}[Lc(\mathbf{\Theta}; \mathbf{X}, \mathbf{Z})|\mathbf{X}, \mathbf{\Theta}^{(q)}] = \sum_{s,k} t_{sk} \log \left(\pi_k \prod_{d,t,l} p(X_{sdt}; \alpha_s\lambda_{klt})^{W_{dl}} \right), \tag{5}$$

where the t_{sk} are the *a posteriori* probabilities (given the current parameters estimate $\mathbf{\Theta}^{(q)}$) of each cluster given by:

$$t_{sk} = \frac{\pi_k^{(q)} \prod_{d,t,l} p(X_{sdt}; \alpha_s^{(q)}\lambda_{klt}^{(q)})^{W_{dl}}}{\sum_k \pi_k^{(q)} \prod_{d,t,l} p(X_{sdt}; \alpha_s^{(q)}\lambda_{klt}^{(q)})^{W_{dl}}}. \tag{6}$$

These quantities are computed during the E step of the algorithm. During the M step, this expectation is maximized with respect to the parameters in order

ALGORITHM 1: EM algorithm to estimate the models parameters and the clustering

Input: Data $\mathbf{X}$: tensor of size $(N \times D \times T)$, W indicators of day clusters: matrix of size $(D \times 2)$, desired number of cluster K
Output: Estimated parameters $\mathbf{\Theta} = (\alpha, \lambda, \pi)$, posterior probabilities t_{sk}
Initialization ;
for *each station* s *in* $\{1, \dots, N\}$ **do**
    compute the stations scaling factor ;
    $\alpha_s = \frac{1}{DT} \sum_{d,t} X_{sdt}$;
end
for *each cluster* $k \in \{1, \dots, K\}$ **do**
    initialize $\hat{\pi}_k^{(0)}$;
end
for *each station* $s \in \{1, \dots, N\}$, *cluster* $k \in \{1, \dots, K\}$ *and day cluster* $l \in \{1, 2\}$ **do**
    initialize $\hat{\lambda}_{klt}^{(q)}$;
end
repeat
    E step : compute the a posteriori probabilities;
    for *each station* $s \in \{1, \dots, N\}$ *and cluster* $k \in \{1, \dots, K\}$ **do**
        $t_{sk} = \frac{\pi_k^{(q)} \prod_{d,t,l} p(X_{sdt}; \alpha_s \lambda_{klt}^{(q)})^{W_{dl}}}{\sum_k \pi_k^{(q)} \prod_{d,t,l} p(X_{sdt}; \alpha_s \lambda_{klt}^{(q)})^{W_{dl}}}$;
    end
    M step : update the parameters ;
    for *each cluster* $k \in \{1, \dots, K\}$ **do**
        $\hat{\pi}_k^{(q)} = \frac{1}{N} \sum_s t_{sk}$;
    end
    for *each station* $s \in \{1, \dots, N\}$, *cluster* $k \in \{1, \dots, K\}$ *and day cluster* $l \in \{1, 2\}$ **do**
        $\hat{\lambda}_{klt}^{(q)} = \frac{1}{\sum_s t_{sk} \alpha_s \sum_d W_{dl}} \sum_{s,d} t_{sk} W_{dl} X_{sdt}$;
    end
until *convergence*;

to increase the likelihood. This maximization, detailed in Appendixes A and B, leads to the following update formulas:

$$\hat{\alpha}_s = \frac{1}{DT} \sum_{d,t} X_{sdt} \, , \, \hat{\pi}_k = \frac{1}{N} \sum_s t_{sk} \tag{7}$$

$$\hat{\lambda}_{klt} = \frac{1}{\sum_s t_{sk} \alpha_s \sum_d W_{dl}} \sum_{s,d} t_{sk} W_{dl} X_{sdt} \tag{8}$$

The update formulas have natural interpretations, the scale factor of station s, α_s is simply given by the average of its activity vectors along all the time frames and days. Since they do not depend on the t_{sk}, they can be computed only one time. The proportions π_k are classically updated by the mean of the *a posteriori* probabilities of each cluster. The λ_{klt} are given by a weighted mean of the activity of cluster k stations in day cluster l and time frame t. Eventually,

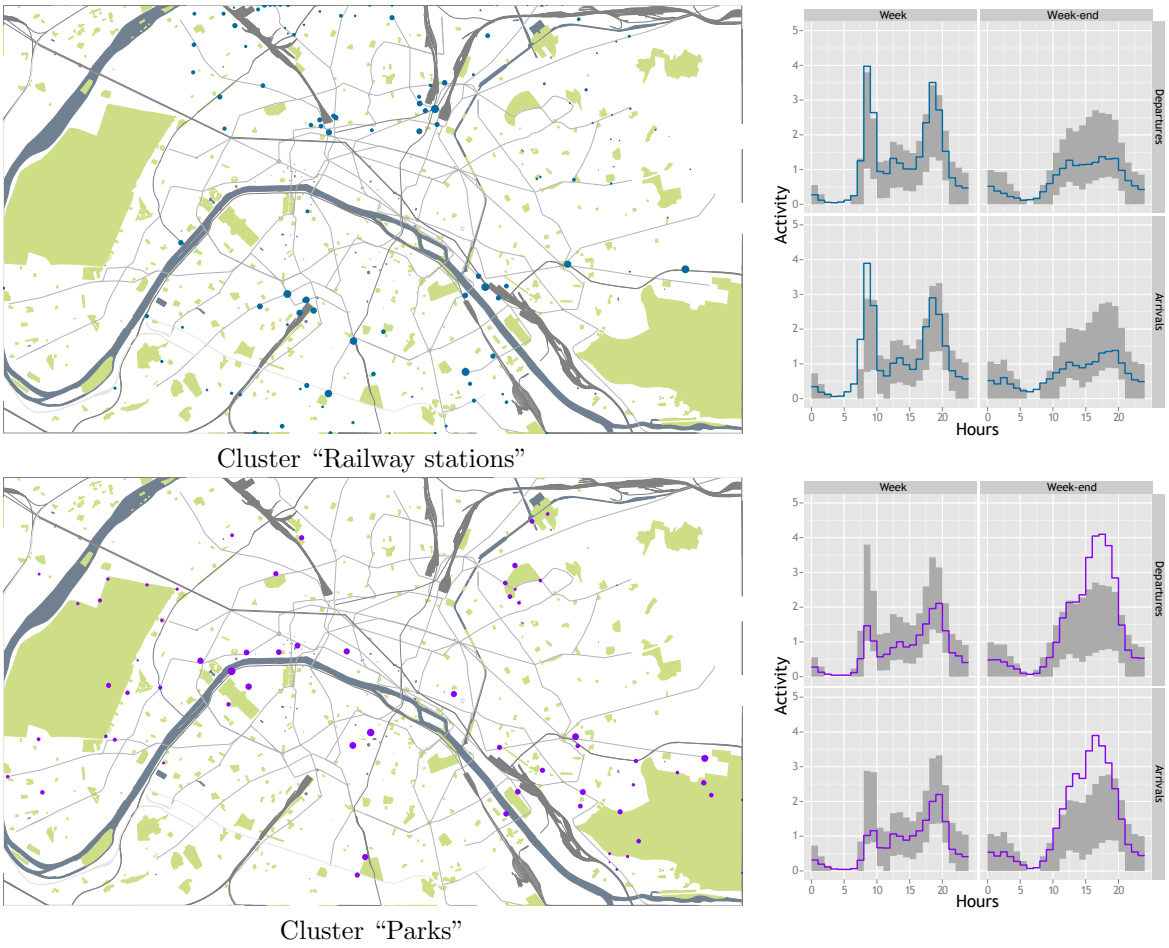

Figure 5: Maps of stations positions for clusters "Railway stations" and "Parks". The map background presents the subway and railway lines, the parks and the Seine. The areas of the dots representing the stations are proportional to the station scaling factor α_s. Each cluster map is completed with the temporal profile of the cluster, the parameters λ_{klt} are to this end arranged according to departure/arrival and week/week-end. The quantiles 0.05 and 0.95 of the total population of stations activity (scaled by their average activity) are also shown in order to highlight the temporal specificities of each cluster.

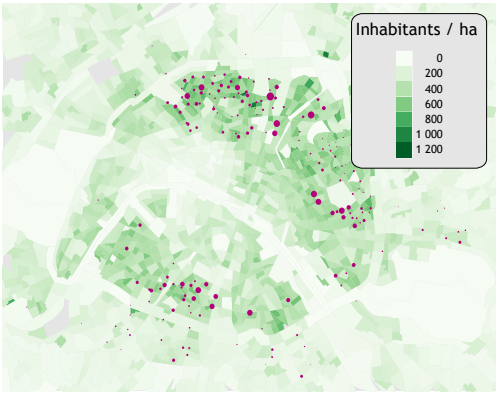

Figure 6: Map of stations positions for cluster " Housing". The map background presents the density of inhabitants per hectare. The areas of the dots representing the stations are proportional to the station scaling factor α_s (Sources "Recensement 2008", "Base permanente des équipements", Insee).

the E and M steps are iterated to build an EM algorithm (see Algorithm 1) that will converge towards a local maxmimum of the log-likelihood.

5 Results

The proposed algorithm was tested on the Vélib' April 2011 dataset, with a varying number of clusters. A good trade-off between complexity of the clustering and interpretability was found for $K = 8$. We therefore analyze in more details the clustering found for this value of K in this section. A first way to investigate the nature of the different clusters found is to look at their temporal profiles given by the parameters λ of the model. In order to give a clear overview of these profiles we organize them according to the nature of the count departures/arrivals, in row, and to the day type (week / week-end), in column. The results for two specific clusters are presented in Figure 5. We name the first cluster "Railway stations" and the second "Parks" because these two clusters correspond to stations that are close to these two kinds of amenities. The maps presented in Figure 5 clearly show the relationship between these two clusters and their corresponding amenity. The temporal profiles present also interesting points, the profile of the "Railway stations" cluster shows an important activity around peak hours for both departures and arrivals, the other time frame being in the average of the total population of stations. The parks profiles give a totally different picture with a rush of activity in the afternoon of the week-end days and a low activity during the peak hours of the week. The maps (see Figure 5) which depict the positions of the clusters stations confirm the interpretation of these two usage clusters. All the railways stations of Paris are clearly visi-

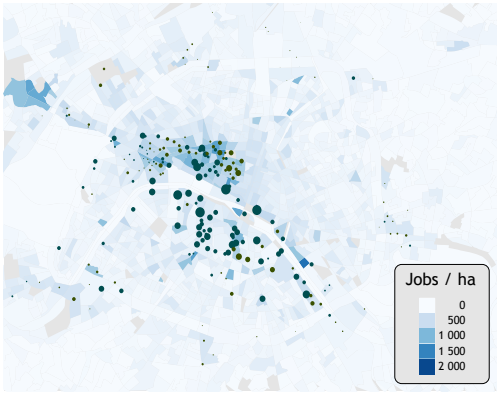

Figure 7: Map of stations positions for clusters "Employment (1)", with light green dots, and "Employment (2)", with light blue dots. The map background presents the density of jobs per hectare. The areas of the dots representing the stations are proportional to the station scaling factor α_s (Sources "Recensement 2008", "Base permanente des équipements", Insee).

Table 1: Mean of each cluster with respect to population density (number of inhabitants per hectare), employment density (number of jobs per hectare), services density (number of personal services such as restaurants, barber, ... per hectare) and shops density (number of shops by hectare). Sources "Recensement 2008", "Base permanente des équipements", Insee.

Cluster name	inhabitants/ha	jobs/ha	services/ha	shops/ha
All	162	237	4.2	3.7
"Spare-time (1)"	367	189	**6.3**	**4.4**
"Spare-time (2)"	261	322	**7.7**	**6.9**
"Parks"	172	90	2	1.7
"Railway Stations"	209	206	2.4	1.8
"Housing"	**375**	108	3.8	2.7
"Employment (1)"	138	**409**	4.5	2.8
"Employment (2)"	157	**456**	5.7	5.6
"Mixed"	301	163	3.8	2.8

ble in the first map, along with several important subway stations like Nation, Denfert-Rochereau, Porte d'Orléans and Vincennes. The map of the stations which belong to the "Parks" cluster gives also a clear view of the nature of this cluster, all the stations are close to parks like Vincennes, Buttes-Chaumont, Montsouris, La Villette, ...

The remaining clusters shown in Figures 8 and 9 also found their origins in geography and sociology. The clusters "Spare-time (1)" and "Spare-time(2)" present high activity values during the week-end nights. The difference be-

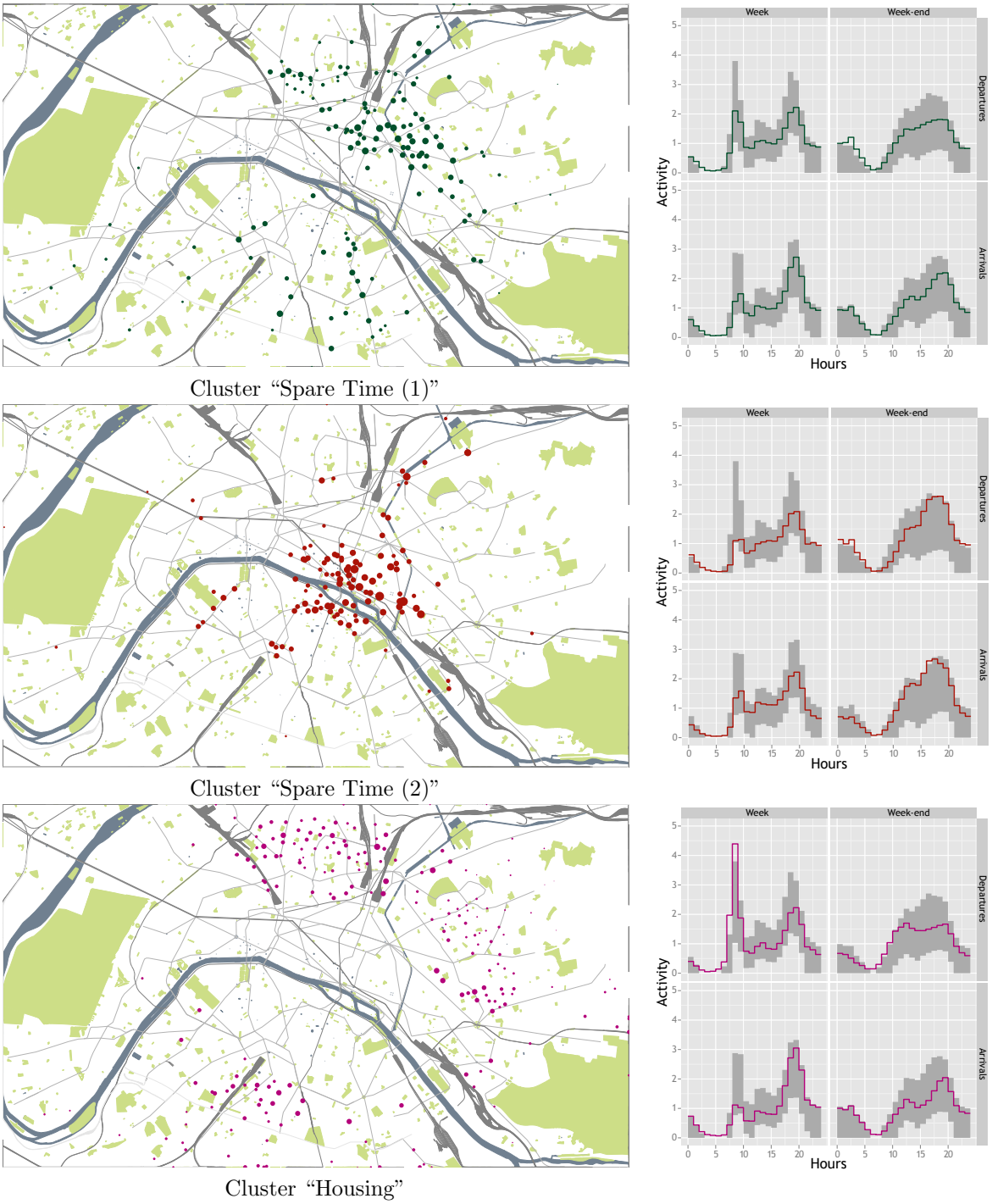

Figure 8: Maps of stations positions for clusters “Spare-time (1)”, “Spare-time (2)” and “ Housing”. The map background presents the subway and railway lines, the parks and the Seine. The areas of the dots representing the stations are proportional to the station scaling factor α_s. Each cluster map is completed with the temporal profile of the cluster, the parameters λ_{klt} are to this end arranged according to departure/arrival and week/week-end. The quantiles 0.05 and 0.95 of the total population of stations activity (scaled by their average activity) are also shown in order to highlight the temporal specificities of each cluster.

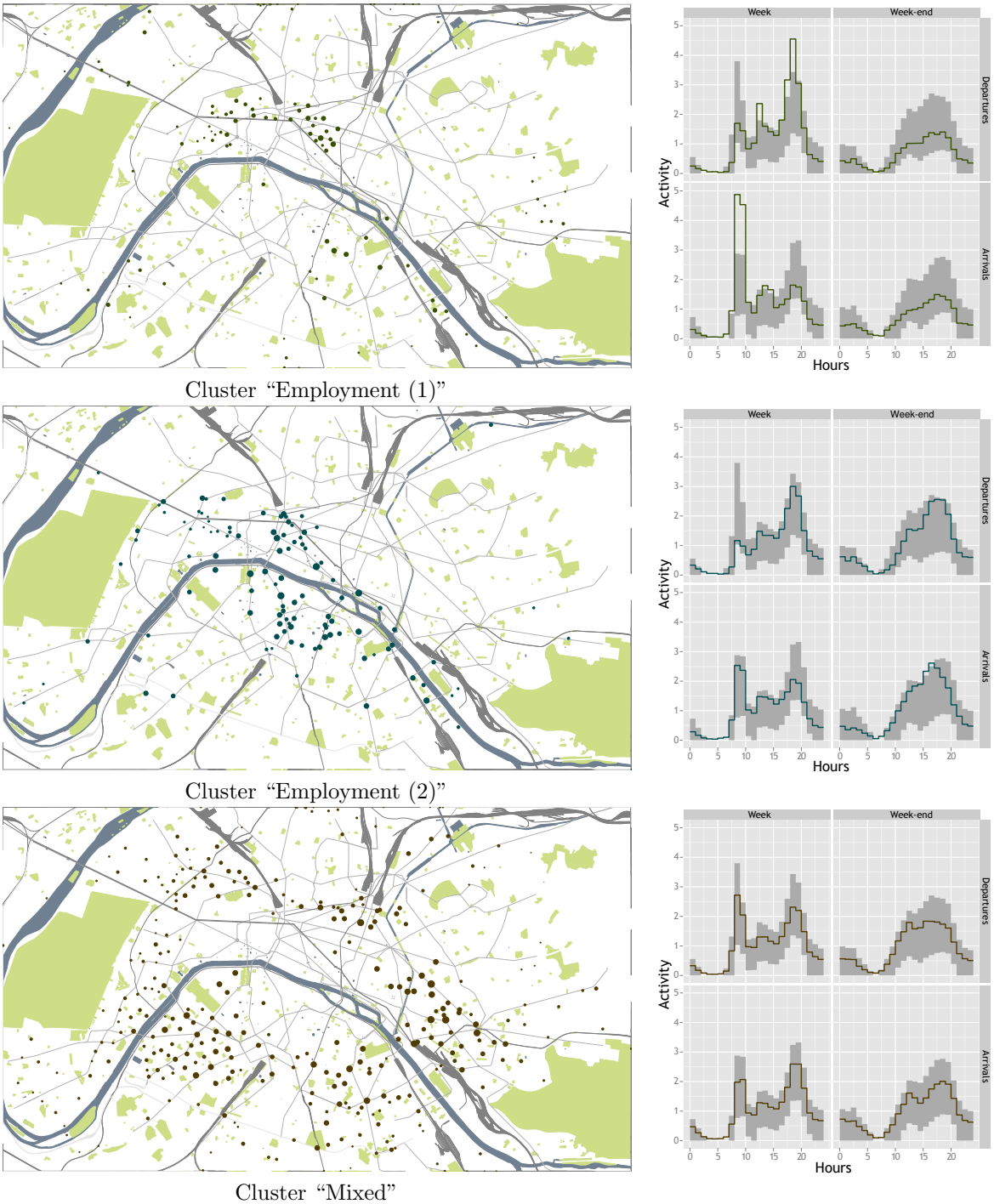

Figure 9: Maps of stations positions for clusters "Employment (1)", "Employment (2)" and " Mixed". The map background presents the subway and railway lines, the parks and the Seine. The areas of the dots representing the stations are proportional to the station scaling factor α_s. Each cluster map is completed with the temporal profile of the cluster, the parameters λ_{klt} are to this end arranged according to departure/arrival and week/week-end. The quantiles 0.05 and 0.95 of the total population of stations activity (scaled by their average activity) are also shown in order to highlight the temporal specificities of each cluster.

tween these two sets of stations appears during the week-end, when the cluster "Spare-time (2)" has a higher activity. This cluster is also more central with stations near "Les Halles" whereas the stations from "Spare-time (1)" come from neighborhoods with night activities like Pigalle, Mouffetard, ... The cluster "Housing" presents a dissymmetry in its profile with a lot of departures during the morning rush but few arrivals, and the reverse during the out of work peak. The stations belong to a belt surrounding the close center of Paris which presents a high population density visible on Figure 6. The next two clusters "Employment (1)" and "Employment (2)" present a dissymmetry, contrary to the one in "Housing": a lot of arrivals during the morning rush but few departures and the reverse during the out of work peak. During the week-end the two clusters present differences with more activity in stations from "Employments (2)". These two clusters correlate with the employments density as shown in Figure 7. Finally, the last cluster "Mixed" seems to be formed by stations with a mixed usage: its temporal profile is medium without specific features. The previous observations are confirmed by an analysis of the mean of each cluster with respect to the population density, employment density, service (restaurants, barber ...) and shops density which are presented on Table 1. An analysis of variance confirms that the clusters are significantly different with respect to these four variables. As expected, the local density of inhabitants is particularly high for the "Housing" cluster, the density of employment being at the opposite high for the "Employment (1)" and "Employment (2)" clusters. Finally, the shops and services densities are important for the "Spare-time" clusters.

6 Conclusion

This paper has presented a new model-based clustering methodology to explore the usage statistics generated by bike-sharing systems. This model introduces a latent variable to encode the stations cluster membership, and an observed variable which deals with the difference of usage between week days and week-end days. Conditionally on these variables the observed counts are supposed to be Poissonians and independent. Their intensities take into account a station scaling factor that handles the discrepancy between the global stations activities. An EM algorithm is then derived to estimate the parameters of the model. Eventually, the methodology is tested to mine one month of usage data from the Paris Vélib' system. The clustering found is rich with interpretable clusters which can be easily linked to the presence of certain type of amenities like parks and railway stations, and to sociological variables like population, jobs and services densities. The clusters are richer than the one obtained in other BSS usage mining studies, since the model uses information on arrivals / departures and not only on the stations states (free places, available bikes).

Their are nonetheless some room for possible improvements, the naive assumption of conditional independence between the time frames could perhaps be removed with benefit using approaches like [18] and [29]. The use of Zero

inflated Poisson or Negative Binomial laws to model the observed counts would also deserved to be tested and compared with the approach proposed here. Eventually, the use of a mix-membership mixture model like LDA [3] will be interesting to describe the mixed nature of the city neighborhoods.

APPENDIX

A Maximization of the lower bound with respect to λ_{klt}

The optimization must take into account the constraints $\sum_{l,t} D_l \lambda_{klt} = DT, \forall k \in \{1, \ldots, K\}$, with $D_l = \sum_d W_{dl}$ the number of days belonging to cluster l. The Lagrangian associated with these K equality constraints is given by:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{s,d,t} \sum_{k,l} t_{sk} W_{dl} \left(X_{sdt} \log(\alpha_s \lambda_{klt}) - \alpha_s \lambda_{klt} \right)) + \sum_k \gamma_k (DT - \sum_{l,t} D_l \lambda_{klt}), \tag{9}$$

with γ_k the Lagrange multiplier associated with the k^{th} constraints.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda})}{\partial \lambda_{klt}} &= \sum_{s,d} t_{sk} W_{dl} \left(\frac{X_{sdt}}{\lambda_{klt}} - \alpha_s \right) - \gamma_k D_l = 0 \qquad (10) \\
&\Rightarrow \sum_{s,d} t_{sk} W_{dl} X_{sdt} - \sum_s t_{sk} \alpha_s D_l \lambda_{klt} - \gamma_k D_l \lambda_{klt} = 0 \\
&\Rightarrow \sum_{s,d} t_{sk} W_{dl} X_{sdt} - D_l \lambda_{klt} \left(\sum_s t_{sk} \alpha_s + \gamma_k \right) = 0 \\
&\Rightarrow \sum_{l,t} \left(\sum_{s,d} t_{sk} W_{dl} X_{sdt} - D_l \lambda_{klt} \left(\sum_s t_{sk} \alpha_s + \gamma_k \right) \right) = 0 \\
&\Rightarrow \sum_{s,d,t} t_{sk} X_{sdt} - \sum_{l,t} D_l \lambda_{klt} \left(\sum_s t_{sk} \alpha_s + \gamma_k \right) = 0 \\
&\Rightarrow \sum_{s,d,t} t_{sk} X_{sdt} - DT \left(\sum_s t_{sk} \alpha_s + \gamma_k \right) = 0 \\
&\Rightarrow \gamma_k = \frac{1}{NT} \sum_{s,d,t} t_{sk} X_{sdt} - \sum_s t_{sk} \alpha_s \qquad (11)
\end{aligned}$$

$$
\begin{aligned}
&\Rightarrow \quad \sum_{s,d} t_{sk} W_{dl} X_{sdt} - D_l \lambda_{klt} \frac{1}{DT} \sum_{s,d,t} t_{sk} X_{sdt} = 0 \\
&\Rightarrow \quad \sum_{s,d} t_{sk} W_{dl} X_{sdt} - D_l \lambda_{klt} \frac{1}{DT} \sum_{s} t_{sk} \frac{1}{DT} \sum_{d,t} X_{sdt} = 0 \\
&\Rightarrow \quad \sum_{s,d} t_{sk} W_{dl} X_{sdt} - D_l \lambda_{klt} \sum_{s} t_{sk} \hat{\alpha}_s = 0 \\
&\Rightarrow \quad \hat{\lambda}_{klt} = \frac{1}{\sum_s t_{sk} \hat{\alpha}_s D_l} \sum_{s,d} t_{sk} W_{dl} X_{sdt} \qquad (12)
\end{aligned}
$$

B Maximization of the lower bound with respect to α_s

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda})}{\partial \alpha_s} &= \sum_{d,t} \sum_{k,l} t_{sk} W_{dl} \left(\frac{X_{sdt}}{\alpha_s} - \lambda_{klt} \right) = 0 \qquad (13) \\
&\Rightarrow \quad \sum_{d,t} \sum_{k,l} t_{sk} W_{dl} \left(X_{sdt} - \alpha_s \lambda_{klt} \right) = 0 \\
&\Rightarrow \quad \sum_{d,t} X_{sdt} - \alpha_s \sum_{k} t_{sk} \sum_{l,t} \sum_{d} W_{dl} \lambda_{klt} = 0 \\
&\Rightarrow \quad \sum_{d,t} X_{sdt} - \alpha_s \sum_{k} t_{sk} \sum_{l,t} D_l \sum_{t} \lambda_{klt} = 0 \\
&\Rightarrow \quad \sum_{d,t} X_{sdt} - \alpha_s DT = 0 \\
&\Rightarrow \quad \hat{\alpha}_s = \frac{1}{DT} \sum_{d,t} X_{sdt} \qquad (14)
\end{aligned}
$$

The authors wish to thank François Prochasson from la ville de Paris and Thomas Valeau from Cyclocity-JCDecaux for providing Vélib' data. We also thank Isabelle Saint-Saens form the Ifsttar for valuable discussions. We are grateful to Samuel Sellam from the Ifsttar for its help during the start of the study.

References

[1] APUR. Etude de localisation des stations de vélos en libre service. rapport. Technical report, Atelier Parisien d'Urbanisme, Paris, Décembre 2006.

[2] Mike Benchimol, Pascal Benchimol, Benoît Chappert, Arnaud De La Taille, Fabien Laroche, Frédéric Meunier, and Ludovic Robinet. Balancing the stations of a self-service bike hire system. *RAIRO-Operations Research*, 45(1):37–61, January 2011.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Céline Robardet, Jean-Baptiste Rouquier, and Eric Fleury. Shared Bicycles in a City: A Signal processing and Data Analysis Perspective. *Advances in Complex Systems,*, 14(3):1–24, June 2011.

[5] Pierre Borgnat, Eric Fleury, Céline Robardet, and Antoine Scherrer. Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program. In Francois Kepes, editor, *European Conference on Complex Systems, ECCS'09*, September 2009.

[6] H. Büttner, J. Mlasowky, T. Birkholz, D. Groper, a.C. Fernandez, Emberger G., and M. Banfi. Optmising bike sharing in european cities, a handbook. Technical report, Intelligent Energy Europe Program (IEE, OBIS projext), August 2011.

[7] D. Chemla, F. Meunier, and R. Wolfler Calvo. Balancing a bike-sharing system with multiple vehicles. In *In proceedings of Congrès annuel de la société Française de recherche opérationelle et d'aide à la décision, ROADEF2011*, Saint-Etienne, France, Mars 2011.

[8] P. De Maio. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):41–56, 2009.

[9] Luigi dell'Olio, Angel Ibeas, and Jose Luis Moura. Implementing bike-sharing systems. In *Proceedings of the ICE - Municipal Engineer*, volume 164, pages 89–101, 2011.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[11] J. Dill. Bicycling for transportation and health: The role of infrastructure. *Journal of Public Health Policy*, 30:S95–S110, 2009.

[12] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2. edition, 2001.

[13] C. Fraley and A. Raftery. Model based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[14] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1420–1426, 2009.

[15] Jon Froehlich, J. Neumann, and N. Oliver. Measuring the pulse of the city through shared bicycle programs. In *Proc. of UrbanSense08*, pages 16–20, 2008.

[16] Gérard Govaert and Mohamed Nadif. Latent Block Model for Contingency Table. *Communications in Statistics-Theory and Methods*, 39(3):416 – 425, January 2010.

[17] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[18] D. Karlis and L. Meligkosidou. Model based clustering for multivariate count data. In *Proceedings od the 18th International Workshop on Statistical Modelling*, pages 211–216. Katholieke Universiteit Leuven, July 2003.

[19] Neal Lathia, Saniul Ahmed, and Licia Capra. Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, 22:88–102, June 2012.

[20] J.R. Lin and T. Yang. Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E: Logistics and Transportation Review*, 47(2):284 – 294, 2011.

[21] G. J. Mclachlan and T. Krishnan. *The EM algorithm and Extension*. Wiley, 1996.

[22] G. J. Mclachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[23] G. Michau, C. Robardet, L. Merchez, P. Jensen, P. Abry, P. Flandrin, and P. Borgnat. Peut-on attraper les utilisateurs de vélo'v au lasso ? In *Proceedings of the 23e Colloque sur le Traitement du Signal et des Images. GRETSI-201*, pages 46–50, 2011.

[24] Rahul Nair, Elise Miller-Hooks, Robert C. Hampshire, and Ana Bušić. Large-Scale Vehicle Sharing Systems: Analysis of Vélib'. *International Journal of Sustainable Transportation*, 7(1):85–106, April 2012.

[25] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, June 2006.

[26] P. Pucher and R. Buehler. Making cycling irresistible: Lessons from the netherlands, denmark and germany. *Transport Reviews*, 28(4):495–528, 2008.

[27] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[28] Andrea Rau, Gilles Celeux, Marie-Laure Martin-Magniette, and Cathy Maugis-Rabusseau. Clustering high-throughput sequencing data with Poisson mixture models. Rapport de recherche RR-7786, INRIA, November 2011.

[29] Sarah Julia Thomas. *Model-based clustering for multivariate time series of counts*. PhD thesis, Rice University, 2010.

[30] Y. Yuan, J. Zheng and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *18th SIGKDD conference on Knowledge Discovery and Data Mining (KDD 2012)*, pages 186–194, 2012.

[31] Wangsheng Zhang, Shijian Li, and Gang Pan. Mining the semantics of origin-destination flows using taxi traces. In *Proceedings of UbiComp 2012*, 2012.