Clustering the Vélib' origin-destinations flows by means of Poisson mixture models

Andry Randriamanamihaga[1], Etienne Côme[1], Latifa Oukhellou [1], Gérard Govaert[2]

1 - Université Paris-Est, IFSTTAR, GRETTIA, F-93166 Noisy-Le-Grand, France

2 - Université de Technologie de Compiègne, UMR CNRS 6599 Heudiasyc,
Centre de Recherches de Royallieu, F-60205 Compiègne Cedex, France

Abstract. Studies based on human mobility, including Bicycle Sharing System analysis, has expanded over the past few years. They aim to give insight of the underlying urban phenomena linked to city dynamics. This paper presents a generative count-series model using adapted Poisson mixtures to automatically analyse and find temporal-based clusters over the Vélib' origin-destination flow-data. Such an approach may provide latent factors that reveal how regions of different usage interact over the time. More generally, the proposed methodology can be used to cluster edges of temporal valued-graph with respect to their temporal profiles.

1 Introduction

The soaring population, coupled with massive urban densification has urged public authorities to deploy urban mobility policies that organize differently how people move. One way adopted by major cities to face these problems is to promote more economical and less pollutant *soft* modes of transport. Within this context, the Vélib' Bike Sharing System (BSS) of Paris has been launched in July 2007. It has quickly become the second largest BSS in the world: 20 000 bikes distributed on 1208 fixed stations now generate 224 000 annual subscribers and an average number of 110 000 travels per day. The availability of such sizeable dataset contributes to leverage the development of new data mining based approaches that aim to recover the underlying urban phenomena linked to city dynamics. The BSSs can obviously benefit from this kind of analysis [1], which would surely help sociologists and planners in their task of apprehending the bikers' mobility patterns withing large megacities.

A statistical model that shall automatically cluster the observed O/D flows between BSS stations is proposed in this paper. The undertaken approach, based on count-series clustering, follows and specifies the work initiated in [2] and [3] that relies on temporal-profiles' analysis to cluster BSS stations and find spatio-temporal patterns describing the pulse of the city. The proposed methodology slighly differs from the latters, as we use a generative dedicated model to count time-series analysis that describe the O/D flow dynamics between couple of stations. Such O/D records enable the study of the interactions and the tradings established among couples of stations over the time. The crossing of the model results with social and economical data is carried out to this end, and will show the close links between these two aspects and the use of the Parisian

Bike Sharing transport mode. To begin with, the next section is devoted to the proposed statistical model based upon count series clustering.

2 Count-series based model

In order to derive the cluster of O/D flows that share similar temporal profiles, an adapted generative mixture model is presented. We first detail the structure of the data that is going to be processed.

2.1 Representation of the origin-destination flows

The displacement flow of bikes used in this paper corresponds to the whole trips (roughly 2 500 000 rides) of the Vélib' April 2011 data. Each displacement 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). We therefore introduce $X_{d,h}^{(u,v)}$ as the number of bikes going from origin station $u \in \{1, \dots, S\}$ to destination station $v \in \{1, \dots, S\}$, during day $d \in \{1, \dots, D\}$ and hour $t \in \{1, \dots, 24\}$. Similarly to [4], an one-hour aggregation was selected to generate the counts since it seems to give a good trade-off between detail resolution and fluctuations . Let $\mathbf{X}_d^{(u,v)}$ be :

$$\mathbf{X}_d^{(u,v)} = \left(X_{d,1}^{(u,v)}, \dots, X_{d,24}^{(u,v)} \right)$$

the description vector of hourly traffic between station u and station v in day d. All of the description vectors can be arranged in a tensor $\mathbf{X}$ of size S^2 x D x T , where S is the total number of stations (1185 in the Vélib' case), D the number of available days in the dataset (30 days for April 2011) and T the length of the description vector (24 hours).

2.2 Clustering with Poissonian generative mixture model

Since we are dealing with count data, conditionnal Poisson mixtures are used to construct the following generative model that discerns K clusters of O/D flows:

$$\begin{aligned} Z^{(u,v)} &\sim& Multinomial(1, \pi), \\ X_{d,1}^{(u,v)} \perp\!\!\!\perp \dots \perp\!\!\!\perp X_{d,T}^{(u,v)} &|& \{Z_k^{(u,v)} = 1, W_{dl} = 1\}, \\ X_{d,t}^{(u,v)} \,|\, \{Z_k^{(u,v)} = 1, W_{dl} = 1\} &\sim& Poisson(\alpha^{(u,v)} \lambda_{klt}), \end{aligned}$$

where **(i)** $Z^{(u,v)} \in \{0,1\}^K$ is the indicator variable that encodes the cluster membership of the O/D flow between (u, v) and **(ii)** $W_d \in \{(0,1),(1,0)\}$ encodes whether the day d belongs to the cluster l of the day-type (either weekday when $(0,1)$ or $(1,0)$ for the weekend). This model thus assumes two different usage patterns over the week. The former variable of cluster membership is drawn from a Multinomial distribution of parameter π, which is the prior proportions of each cluster of the mixture. They are observed conversely to the W_d . Regarding

$X_{d,t}^{(u,v)}$, they are drawn from a Poisson distribution of parameter $\alpha^{(u,v)}\lambda_{klt}$ from the moment that both the cluster $Z^{(u,v)}$ of the displacement and the cluster W_d of the day-type are known. Actually the scaling parameter $\alpha^{(u,v)}$ captures the flow magnitude on the edge (u, v). Regarding the parameters $\lambda_{kl.}$, they sketch the temporal variations of the O/D flows that are specific to the cluster k at day-type l, over the aggregated 24-hours of a day. To achieve identifiability, the λ_{klt} 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 the cluster l of day-type. Under the preceding assumptions, the likelihood of the parameters $\mathbf{\Theta}$ expresses as :

$$L(\mathbf{\Theta};\, \mathbf{X}|\mathbf{W}) = \sum_{(u,v)} \log \left(\sum_k \pi_k \prod_{d,t,l} p\left(X_{d,t}^{(u,v)};\, \alpha^{(u,v)}\lambda_{klt}\right)^{W_{dl}} \right)$$

The estimate $\widehat{\mathbf{\Theta}} = (\widehat{\alpha}, \widehat{\lambda}, \widehat{\pi})$ that locally maximises the log-likelihood is computed with an EM type algorithm [5], a well-fitted two-step iterative approach for maximum likelihood estimation in statistical problems involving latent variables. Fed with the O/D tensor $\mathbf{X}$, EM also gives the partitions of O/D flows by returning the *a posteriori* probabilities $t_k^{(u,v)}$ of the flow (u, v) to be in cluster k. From now on we will explicitly address the rationale for model selection.

2.3 Comparing and specifying the model

The presented model, called *Model 1* for convenience of reference, is now compared to other generative Poisson mixture models that cluster as well the edges of a temporal valued-graph. *Model 2* is the most basic: given $Z_k^{(u,v)}$, the $X_{d,t}^{(u,v)}$ are simply independent and indentically distributed Poisson random variables with parameter λ_{kt}. Here, all of the observed days d are of the same and unique day-type. *Model 3* fine-tunes the latter by capturing the specific patterns related to weekdays and weekends, similarly to our model. *Model 4* on the opposite relaxes the assumption related to the two day-types but makes the difference between the most and the least significant O/D flows through the scaling parameters $\alpha^{(u,v)}$, which capture the amplitude of the edges (u, v).

Their predictive performance were evaluated using the perplexity criterion [6] on a held-out test-data that corresponds to the whole Vélib' displacements recorded during September 2011. Figure 1(left) presents the test-set perplexity for each of the four mentioned models with respect to the number of cluster K: our model clearly outperforms all others.

Then, the number of cluster K is fixed. The BIC value on April data, highlighted in Figure 1(right), certainly decreases until $K = 25$. This is far too large to describe the primary usage trends of the system. A lower value of K that rather-well minimizes this criterion and eases the interpretation process is prefered, and leads us to choose $K = 9$. The next section provides analyses of few clusters of O/D flows.

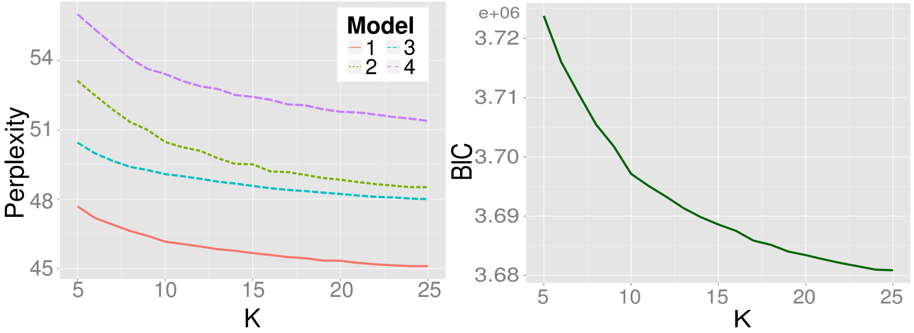

Fig. 1: With respect to the number of cluster K:(left) Predictive perplexity evaluated on September data for *Model 1, 2, 3, and 4* (right) BIC on *Model 1*

3 Empirical results on the Vélib' BSS data

The previous model is applied on the Vélib' April 2011 O/D data with $K = 9$. In preprocessing the data, the couples of stations that don't generate at least one trip per day were removed, leaving approximately 1210 O/D flows per cluster k. In this context, a cluster k corresponds to a group of geolocated edges of the Vélib' network that share the same $\lambda_{k..}$, i.e similar temporal profiles over the weekdays and weekends.

Figure 2A depicts the O/D flow behavior in cluster $k = 2$ (blue) and cluster $k = 6$ (red). Observing their curves of temporal profiles, it appears that cluster 2 is mainly about recreation displacements realized during the weekend afternoons (huge peak from 10 a.m to 19 p.m). Cluster 6 is more concerned about the evenings and the night periods (until 4 a.m) of the weekends. Indeed, the edges of cluster 2 associated with the weekend afternoon profile are condensed around parks (Vincennes, La Vilette), streams (Canal Saint-Martin) and nearby tourist attractions (Tour Eiffel). Thus, we label this cluster *Weekend Joyridings*. An analogous spatio-temporal analysis can be performed on cluster 6, which is therefore labelled *Night Life* as it concentrates leisure displacements moving towards the city bars, nightclubs and restaurants (Bastille, Mouffetard) between 11 p.m to 05 a.m. In these cases, spatial effects are quite self-explanatory, conversely to the next detailed clusters. Figure 2B presents the spatio-temporal results for cluster 7 (blue) and cluster 8 (red). Their temporal profiles reveal colossal peaks during the weekday mornings (from 6 to 11 a.m) that can be associated with bikers leaving home for work.

To extend the rationale behind the maps and temporal profiles, these flows are crossed with the socio-economical information summarized in Table 1. The cluster 8, that we call *Early Bird Works*, clearly concerns geolocalized flows of users leaving their housings to go to work: there is a 93% drop of the density of population and a jump of 94% of the employement density between the places

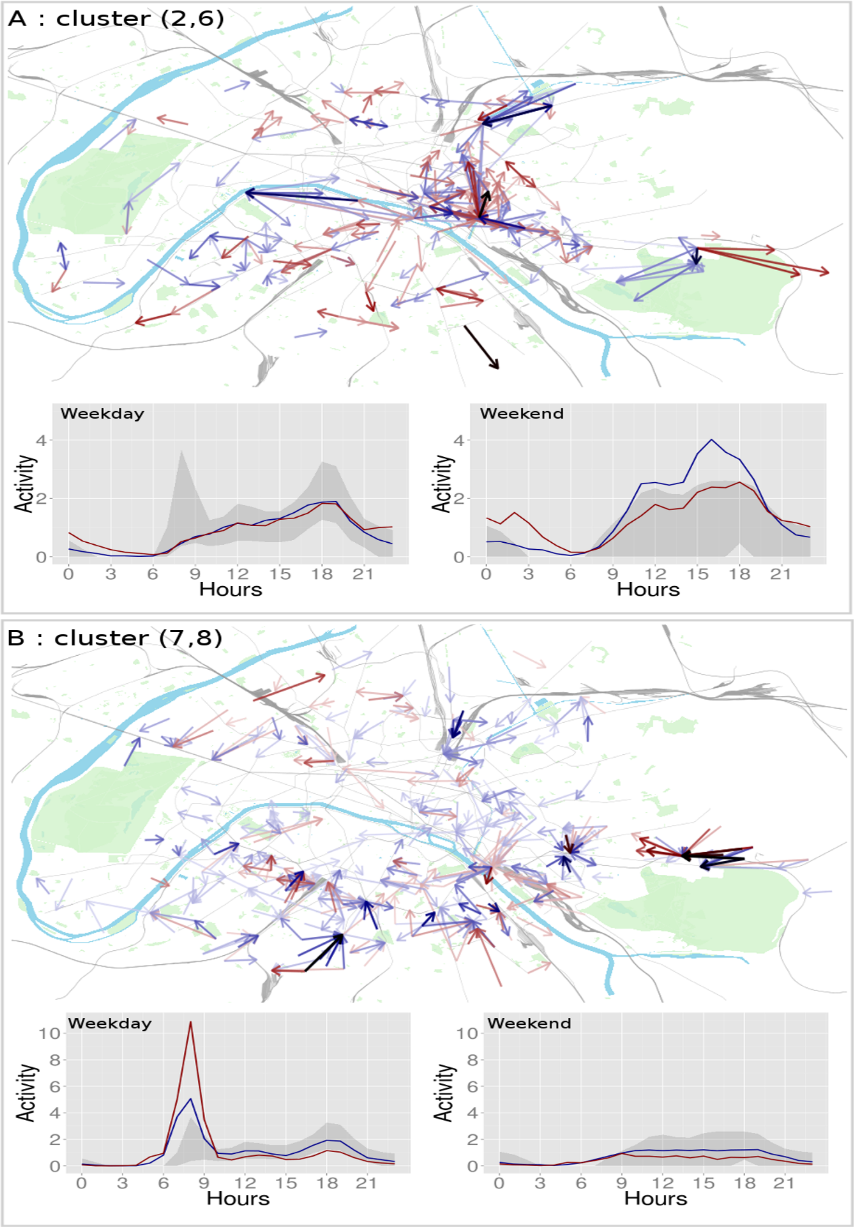

Fig. 2: Maps of the displacement flows and their respective temporal profiles of usage (weekday versus weekend, grey: 25-75 quantile) for the two groups of clusters $\{2,6\}$ (see block A) and $\{1,8\}$ (see block B). Each cluster of each group is respectively plotted in blue (cluster 2 and 7) and red colour (cluster 6 and 8)

of departures and arrivals. A similar behaviour is observed for cluster 7, which it is just labelled *Morning Works* to keep a more general interpretation.

n°	Label	inhab./ha	jobs./ha	comjobs./ha
2	Weekend joyridings	269 → 274	194 → 202	140 → 148
6	Night life	288 → 292	266 → 241	193 → 181
7	Morning works	**337** → 276	152 → **210**	102 → **148**
8	Early bird works	**329** → 236	175 → **269**	132 → **166**

Table 1: For each of the 4 clusters, this table presents the mean of the densities of population, employments and commercial employements arround the origins stations and destinations stations of the flows that belong to the cluster. Notation is: *origin density mean* → *destination density mean*

4 Conclusions and perspectives

By the use of a Poisonnian mixture model and EM algorithm for parameters estimation, this paper has introduced a novel model for temporal graph clustering, which was then used on usage statistics generated by a Bike Sharing System. It introduces a latent variable that encodes the cluster membership of each edge, and an observed variable which deals with the difference of usage between weekdays and weekend. Model selection critera were thus applied to assess the prediction power of this approach. This methodology is then tested to mine one month of trips data from the Paris Vélib' Bike Sharing System. Relevant answers to the common *why* and *when* of urban mobility analysis can be brought from the resulting clusters, which are rich and interpretable. There is nonetheless room for possible improvements. For example, Zero inflated Poisson or Negative Binomial laws to model the observed counts may deserve to be investigated. **Acknowledgments:** *The authors wish to thank Cyclocity - JCDecaux and Ville de Paris for providing the Vélib' origin-destination data.*

References

[1] Dell'Olio L., , Ibeas A., and J. L. Moura. Implementing bike-sharing systems. In *Proceedings of the ICE - Municipal Engineer*, volume 164, pages 89–101, 2011.

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

[3] Lathia N., Ahmed S., and L. Capra. Measuring the impact of opening the london shared bicycle scheme to casual users. *Transportation Research Part C, Emerging Technologies*, 22:88–102, 2012.

[4] Borgnat P., Abry P., Flandrin P., Robardet C., Rouquier J. B., and Fleury E. Shared bicycles in a city : A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(3):1–24, 2011.

[5] McLachlan Geoffrey J. and Krishnan T. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Wiley - Interscience, 2008.

[6] Brown Peter F., Della Pietra Vincent J., Mercer Robert L., Della Pietra Stephen A., and Jennifer C. Lai. An estimate of an upper bound for the entropy of english. *Comput. Linguist.*, 18(1):31–40, 1992.