Spatio-temporal Analysis of Dynamic Origin-Destination Data Using Latent Dirichlet Allocation: Application to Vélib' Bike Sharing System of Paris

Etienne COME, Njato Andry RANDRIAMANAMIHAGA, Latifa Oukhellou, Patrice Aknin

Spatio-temporal analysis of Dynamic Origin-Destination data using Latent Dirichlet Allocation. Application to the Vélib' Bike Sharing System of Paris.

Etienne Côme
Université Paris-Est, IFSTTAR, COSYS-GRETTIA, F-77447 Marne-la-Vallée, France

Andry Randriamanamihaga (Corresponding author)
Université Paris-Est, IFSTTAR, COSYS-GRETTIA, F-77447 Marne-la-Vallée, France

Latifa Oukhellou
Université Paris-Est, IFSTTAR, COSYS-GRETTIA, F-77447 Marne-la-Vallée, France

Patrice Aknin
Université Paris-Est, IFSTTAR, COSYS-GRETTIA, F-77447 Marne-la-Vallée, France

5700 words + 7 figures + 0 tables

September 16, 2013

ABSTRACT

This paper deals with a data mining approach applied on Bike Sharing System Origin-Destination data, but part of the proposed methodology can be used to analyze other modes of transport that similarly generate Dynamic Origin-Destination (OD) matrices. The transportation network investigated in this paper is the Vélib' Bike Sharing System (BSS) system deployed in Paris since 2007. An approach based on Latent Dirichlet Allocation (LDA), that extracts the main features of the spatio-temporal behavior of the BSS is introduced in this paper. Such approach aims to summarize the behavior of the system by extracting few OD-templates, interpreted as typical and temporally localized demand profiles. The spatial analysis of the obtained templates can be used to give insights into the system behavior and the underlying urban phenomena linked to city dynamics.

INTRODUCTION

The widespread use of smart card automated fare collection systems by transport operators can help in innovative studies on human mobility. In fact, these fare collection systems collect a large amount of data related to travels on the whole public transit networks in which they are deployed. In this way, they can be viewed as passive sensors for human mobility. Advanced analysis of the streams of trips produced by these systems can be used to give insights on human mobility, allowing transport operators to provide best quality service. It may also help sociologists and urban planners to apprehend the mobility patterns of users within the city. However, the volume collected is often large, which raises challenges for their exploitation. Automatic algorithms able to extract useful information from these sources has consequently become of great interest.

This paper deals with a data mining approach applied on Bike Sharing System Origin-Destination data, but part of the proposed methodology can be used to analyze other modes of transport that similarly generate Dynamic Origin-Destination (OD) matrices. The transportation network investigated in this paper is the Vélib' Bike Sharing System of Paris, deployed since 2007. Its access system generates streams of detailed travel information, recorded as Origin-Destination data. This work investigates the analysis of sizeable OD-matrices using an advanced statistical model called Latent Dirichlet Allocation (LDA). This model, initially developed to process document collections, was adapted to mine such OD-data in order to extract the main features behind the spatio-temporal behavior of the BSS. The results provided by this model address the following issues:

- Identify a reduced set of demand profiles, specific to such soft modes of transport. The spatial analysis of the resulting patterns can be used to get a better understanding of the underlying urban phenomena linked to city dynamics.
- Build links between the sociological, economical and geographical context of a city and the usage of its BSS. BSS operators can both benefit from this kind of analysis to better understand the system usage and learn how to improve the service quality of the existing system. In the future, such knowledge can be transferred to cities aiming to incorporate new BSSs.
- Get a better understanding of the problem of balancing load of bikes. One of the main issues raised by BSS users in recent surveys is the availability of bikes: users are confronted to empty stations when they want to rent bikes, and full stations when they return them back. Redistribution of bikes, which consists of relocating them among the stations, is then necessary in most BSSs to compensate the uneven demand of users. This issue is traditionally addressed within the field of Operation Research, in which optimization policies of bikes redistribution are developed. In this paper, we will focus on a data mining approach aiming to give indicators on the imbalances of stations, which may be used as inputs of advanced Operation Research algorithms.

The paper is organized as follows: Section 3 is devoted to related work conducted on BSS data analysis. In section 4, we detail the data mining approach based on Latent Dirichlet Allocation, which was used to achieve the BSS data analysis. The obtained usage patterns as well as bike stations unbalances are also analyzed. In Section 5, contextual elements of the Bike Sharing

System of Paris are given. Then the results of the proposed methodology applied to the Vélib OD-data are presented and discussed in Section 6, as well as new operational indicators, the obtained usage patterns and the per-station bike imbalance of the BSS. Conclusion and perspectives are finally presented to show how data mining approaches applied on new available data sources can lead to innovative modeling and better understanding of urban mobility.

RELATED WORK

Several research studies have been conducted on BSS data over the past few years. They generally arise from two main fields of research: Operation Research and Data Mining. The works from the former field mainly concerns the optimization of the load balancing of bikes, often necessary to compensate the uneven demand of bikes. This is usually performed with trucks that move some bikes between the stations. The reader interested by this topic can refer to Benchimol et al., Chemla et al., Nair et al., Lin and Yang (*1, 2, 3, 4*).

Data Mining approaches have been applied in various ways to BSS data. Two main topics have been investigated: Clustering and Prediction. The Prediction topic focuses on developing models able to forecast the usage of stations or, more generally, the behavior of the transportation network in either the short term or the long term (see Froehlich et al., Borgnat et al., Kaltenbrunner et al., Michau et al., Vogel et al. (*5, 6, 7, 8, 9*)). The Clustering topic aims to uncover spatio-temporal patterns in the BSS usage, thus highlighting the relationships between time of day, location and usage. This is classically done by partionning the set of stations into clusters of similar patterns. However, one of the key differences among the researches concerns how the usage of the BSS is described. A major part of the researches on BSSs use public data sampled from the operator's website which consist of station-occupancy statistics, such as the number of available bicycles and free slots per-station along a day. The remaining part of the studies directly focuses on the mining of anonymized and individual dynamic OD-trips provided by the BSS operators.

Using station occupancy data collected from the Bicing BSS of Barcelona, Froehlich et al. (*10*) and Froehlich et al. (*5*) proposed methodologies that identify its main usage patterns and performed a prediction of station usage within a prediction window ranging from 10 to 120 minutes. Lathia et al. (*11*) investigated how a new user access policy in the London Barclays Cycle Hire Scheme affected the system usage across the city, using both spatial and temporal analysis of station occupancy data. Other approaches use trips data to analyze BSS usage, such as the recent study of Borgnat et al. (12) on the Lyon Vélo'v BSS data in which different graphs are used to extract similar profiles of usage (in terms of arrivals/departure count correlations) between pairs of stations during weekdays and weekends, which lead to cluster the stations. Carried out on the same BSS, another approach similarly based on a dynamical view of the transportation network and proposed by Borgnat et al. (*13*) aims to uncover communities of stations that exchange bikes in a preferential way: the activity between the stations was clustered using graph clustering algorithm, and exhibited similar exchange dynamics. A statistical approach based on Poisson Mixture model has been proposed by Randriamanamihaga et al. (*14*) in order to discover usage patterns of the Vélib' BSS data on the basis of clustering of flows.

Other researches using OD-trips data are proposed by Vogel et al., Vogel and Mattfeld (*9, 15*) and aim to identify a reduced set of clusters of stations to get a better understanding of the spatial and temporal causes of imbalances between BSS stations. The proposed methodology, based on Geographical Business Intelligence process, was successfully applied to data collected from the Vienna's BSS Citybike Wien. It used feature vectors, *i.e* the per-hour and per-station

normalized number of incoming and outgoing trips recorded during weekdays and weekends, to describe the stations. Three clustering algorithms (K-Means, Gaussian Mixture Model estimated through the EM algorithm and sequential Information-Bottleneck (sIB)) are then compared.

The approach undertaken in this paper is based on Latent Dirichlet Allocation, a text-categorization algorithm initially introduced in the seminal paper of Blei et al. (*16*). Conversely to Montoliu (*17*), who uses LDA to analyze a BSS using occupancy data, this work deals with OD-trips data. The second key differences concerns the formulation of the approach. In Montoliu (*17*), as in most of the previous studies, the clustering-step partitions a set of stations whereas in this paper, we aim to extract few global and recurrent demand profiles that describe the behavior of the BSS. In order to give a clear overview of the system dynamic, post-processing tools are furthermore introduced to analyze the results provided by LDA.

From a methodological point of view, topic models such as LDA are Probabilistic Generative Models that aim to recover the latent structure of a document collection. Although initially developed to analyse text documents, Probabilistic Topic Models have been applied to other issues: Farrahi and Gatica-Perez (*18*) aims to discover some location-driven routines using mobile phone data , Huynh et al. (*19*) extracts daily human routines from wearable sensors and Niebles et al. (*20*) analyzed trajectory and modeling semantic region on video scenes. These topic models are used here to uncover the underlying mobility patterns, assuming the key idea that the usage of a mode of transport can be summarized by a finite set of demand profiles, or routines, encoded within typical OD-templates. The topic model involved in this paper is the LDA model, which background is recalled in the next Section, as well as its re-interpretation in the context of mining Dynamic Origin-Destination matrices.

LATENT DIRICHLET ALLOCATION APPLIED ON DYNAMIC OD-MATRICES

Background on Latent Dirichlet Allocation

LDA is a three-level Hierarchical Bayesian Model for discrete data. Originally developed to process document collections, it was first introduced in the seminal paper of Blei et al. (*16*). The basic idea is that each document of a given document collection can be efficiently represented as mixture of latent topics, since each of them deals with a relative small number of topics that induces the use of a specific vocabulary (a semantic field).

This intuition is formalized into a Statistical Generative Model that involves latent variables. One of the simplest way to describe LDA is to detail this generative model *i.e.* the random process by which the model assumes the documents arise. This generative model involves the following elements:

- **Corpus:** the corpus is a collection of M documents, denoted $\mathbf{C} = \{\mathbf{d}^1, \ldots, \mathbf{d}^M\}$.
- **Documents:** Each of the M documents is a bag of words $\mathbf{d} = \{\mathbf{w}^1, \ldots, \mathbf{w}^L\}$. Since the size of each bag of words is not fixed, the number of words in the i^{th} document of the corpus is neither fixed and is denoted $L(i)$.
- **Words:** the words are taken within a fixed vocabulary indexed by $\{1, \ldots, N\}$. A word $\mathbf{w}$ is represented by an indicator such that the N-vector $\mathbf{w}$, with $\mathbf{w}_j = 1$ and all other index set to zero, is the j^{th} word in the vocabulary.

In this view, topics are first defined as distributions over a fixed vocabulary. For example the Statistic topic would have words about transports such as estimation, likelihood and variance with high probability. These distributions of topics, denoted by $\Lambda^{(k)}$ for a topic k, are supposed to be known in advance before any documents has been generated using a Dirichlet distribution $\mathcal{D}(.)$ of parameter β:

$$\Lambda^{(k)} \sim \mathcal{D}(\beta),\ \forall k \in \{1, \ldots, K\}. \tag{1}$$

Thereafter, a document $\mathbf{d}^i$ is generated according to the two-stage process described below:

1. Choose the proportions of the different topics $\pi^{(i)}$ for the document, using $\pi^{(i)} \sim \mathcal{D}(\alpha)$
2. For each of the $L(i)$ word of document i:
   (a) Draw a topic T from the distribution over topics $\pi^{(i)}$ from step (1), using $T \sim \mathcal{M}(1, \pi^{(i)})$. The Multinomial distribution is denoted $\mathcal{M}(.,.)$
   (b) Choose a word W from the corresponding distribution over the vocabulary using $W \sim \mathcal{M}(1, \Lambda^{(T)})$.

This statistical model reflects the intuition that documents exhibit multiple topics. Each word of each document is drawn from one of the topics, which one is chosen from the per-document distribution over topics preliminary drawn in step (1). For example, this paper may have been generated using such a scheme with a distribution over topics mainly concentrated on two topics: Transport and Data Mining, which may have induced our specific distributions over words. The formal description of this model involves intensively the Dirichlet distribution and its conjugate prior, the Multinomial distribution. To sum up, the main quantities of interest produced by LDA are:

- the $\Lambda^{(k)}$, which characterize each latent topic by a discrete distribution over words and encodes the keywords of the topics.
- the $\pi^{(i)}$ which summarize each document by a K-vector of topic proportion and encodes to a short description of text content.

Parameters estimation for this model is achieved by the maxisimation of the log probability $\log(p(\mathbf{C}|\mathbf{\Lambda}, \alpha))$ of the corpus $\mathbf{C}$ with respect to the parameters α and $\mathbf{\Lambda}$. This corpus probability may be written with respect to L, the number of words per documents. The corpus $\mathbf{C}$ is encoded as a sparse count matrices where $\mathbf{C}_{in}$ represent the number of occurrences of word n in document i. The maximization of this quantity can be performed using either Variational EM algorithm or Gibbs Sampling as in Grün and Hornik (*21*). A Variational EM algorithm is used instead of an ordinary EM algorithm since the expected complete likelihood in the E-step is computationally intractable. For an introduction into Variational Inference, the reader can refer to (22) and for some applications of these methods on huge corpus, see Hoffman et al. (*23*). LDA has also been extended in several ways to take into account additional aspects such as document co-variables (*i.e.* metadata) as proposed by Mimno and McCallum (*24*) or topic drift presented in Blei and Lafferty (*25*).

LDA adapted to Dynamical OD-data mining

In order to adapt the previous LDA for text-categorization to our OD-trips analysis, a first assumption is necessary: we assume that a finite and small set of timestamped OD-trips is sufficiently informative. Thus, one way to adapt LDA using the BSS transit data is to make the following analogy: (i) the M bags of words are replaced with M bags of successive OD-trips (denoted OD-bags) and (ii) the words are replaced with OD-couples. The hidden topics are then interpreted as OD-templates and the vector of OD-templates proportions of each OD-bag summarizes, for each OD-template, a specific temporal behavior of the system. Let us point that each of these bags is equivalent to an OD-matrice that simply counts the occurences of each OD-couple during the timespan of the bag. Since the OD-bag (*i.e* OD-matrices) are sorted temporally, the inputs of the LDA algorithm are Dynamical OD-matrices.

Using this analogy the whole generative process is rewritten in what follows. First, the latent demand profiles or OD templates are drawn using a Dirichlet distribution over the set of OD:

$$\Lambda^{(k)} \sim \mathcal{D}(\beta),\ \forall k \in \{1, \ldots, K\}. \tag{2}$$

Then, each OD-bag of successive trips i in the set of bags $\{1, \ldots, M\}$ is similarly generated according to the following two-stage process:

1. Draw the proportions of the templates in the bag using $\pi^{(i)} \sim \mathcal{D}(\alpha)$
2. For each trip of the bag i:
   (a) Draw its template T using $T \sim \mathcal{M}(1, \pi^{(i)})$
   (b) Draw an OD couple W using the OD template T using $W \sim \mathcal{M}(1, \Lambda^{(T)})$

The generative process did not change at all with this reinterpretation, this is still the classical LDA model. It assumes, however, that the system generating the OD is stationary during short time frames (one hour for example): in fact, the OD-couples of one particular OD-bag are generated by the one same distribution. More formally the random process generating the observed OD is supposed to be fixed during a certain period before it may switch into another regime. As previously the main output from LDA will be the $\Lambda^{(k)}$ and the $\pi^{(i)}$ which can be interpreted in the context of dynamical OD matrices analysis as follow:

- the $\Lambda^{(k)}$ are the discrete distribution over the OD couples. They can be interpreted as typical demand profiles and describe the typical geography of trips.
- the $\pi^{(i)}$ summarize each OD matrix by a K-vector of template proportion. Since the OD-matrices are temporally sorted, using the OD-templates may give a compact representation of the temporal behavior of the system.

Regarding the items, the vocabulary are made of OD-couples which can be seen as elements of the Cartesian product of two finite sets of stations. This has obviously no impact on the model, except that the Multinomial law used to draw one OD-couple is parametrized by a matrix of probabilities that sum to one. The OD-bags of trips are described by matrices of counts, denoted by $\mathbf{C}$, where $\mathbf{C}_{tij}$ is the number of occurrences of OD couple (i, j) in bag t. In other word, $\mathbf{C}_{t..}$ is a classical OD-matrix.

Such a model is promising and is likely to find interesting structure in the BSS data, known to be affected by cyclic regularities. The observed behavior of users during the Mondays, between 8a.m and 9a.m, would for example be quite similar to the behavior of users observed during the same time frame, for other weekdays. To keep the analogy, two documents dealing with the same topics certainly share common lexical fields. Consequently, the cyclo-stationarity introduced by daily activities such as $Home \rightarrow Work \rightarrow Leisure \rightarrow Home$ may therefore be recovered by the model, which can give insight of the whole system behavior with only few OD-templates. The patterns specific to weekdays and weekends, expected to be hidden in the data, may play a role similar to those of the hidden topics behind the observed similarities in documents collections. These patterns may be caught by the OD templates: for example, we can think of a $Home \rightarrow Work$ OD-template which has OD-couples leaving from places with high population densities and going to places with high employment densities. Eventually, LDA will help to assess the timespan over which the system can be considered as stationary and to discern the associated change points.

Further in this paper, these intuitions are validated through the application of this methodology on trips data recorded by the Vélib' Bike Sharing System of Paris. Prior to this analysis, some contextual elements on the Vélib' BSS are supplied.

THE VÉLIB' BIKE SHARING SYSTEM OF PARIS

The Bike Sharing System of Paris, called Vélib' has been deployed since July 2007 and is operated as a concession by Cyclocity, a subsidiary of the French outdoor advertising company JCDecaux. It offers a non stop service 24/7 and at its debut in 2007, 700 bicycles were spread across 750 fixed stations. In four years, it has expanded to more than 1200 stations which hire out around 18,000 bikes throughout the city. Considering the number of annual subscribers, 224,000 and still growing, 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. Vélib' is available mainly in Paris *intramuros*, some stations being located in the suburbs. At the stations, the bikes are locked to the electronically controlled docking points: the whole network includes 40,000 of them, inducing from 8 to 70 docking points per station. Regarding the policy, a user can purchase a short-term daily or weekly subscription, or a long-term annual subscription which allows an unlimited number of rentals. In both cases, the first half hour (or the first 45 minutes for a long-term subscription) of every individual trip is free of cost.

The aim here is to obtain some general statistics to highlight the global trends and usage of Vélib'. The dataset used to estimate these global statistics and analyze the results of the proposed methodology corresponds to one month of trips data recorded in April 2011. This corresponds to roughly 2,500,000 trips after data cleansing, which removed the trips with a duration of less than one minute and with the same station as point of departure and destination. These trips correspond to user misoperation and not to real trips.

Figure 1 (left) displays the total number of recorded trips, observed over a week, with respect to the type of subscription: annual (plotted in blue) or during 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 (8a.m and 6p.m) 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 totally 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 are more leisure and recreational oriented. These temporal trends of BSS usage can provide information on the sociological characteristics of the city. Considering the study carried out by Froehlich et al. (*5*) on the Barcelona Bicing BSS, some sociological differences between Barcelona and Paris can be highlighted: the lunch peak occurring at 2p.m in Barcelona Bicing data occurs at 12 noon in the Vélib' data, reflecting the late lunch culture of Spain (resp. the earlier lunch culture of France). Secondly, although Friday is the least active day in Barcelona, in France it is not.

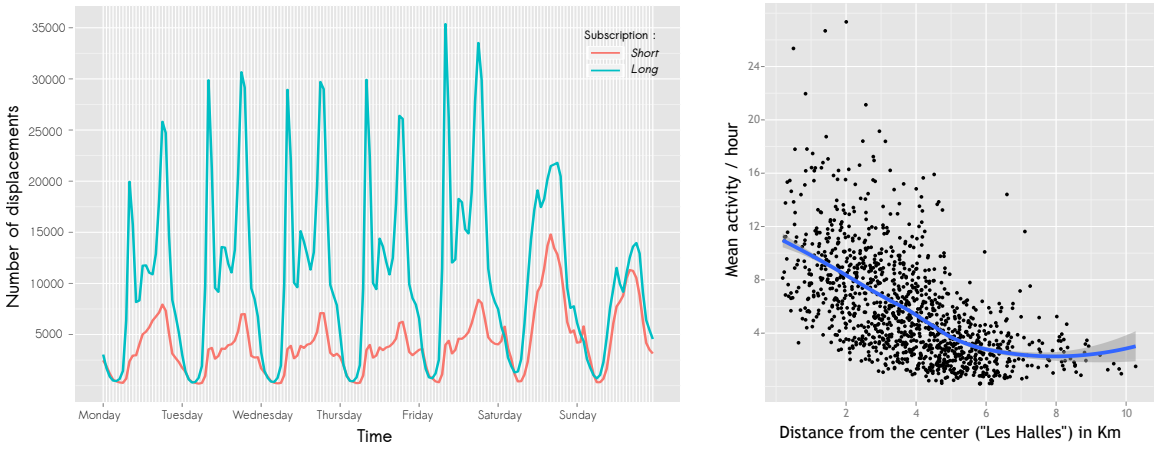

FIGURE 1 (Left) Total number of displacements, summed over each hour of each day of the week, from Monday to Sunday. The blue line (resp. red line) corresponds to the displacements carried out by one-year (resp. one-day) subscribers. (Right) Average activity of stations (number of actions: departure or arrival) per hour with respect to the distance of the stations from the center of Paris ("Les Halles").

In addition to these temporal trends, spatial trends closely linked to geographical aspects of the city can also be identified. Figure 1(right) shows the average number of observed departures and arrivals per hour with respect to the distance from the station to the center of Paris. It is clear that the closer the station to the center of Paris, the greater its mean activity. Furthermore, the duration and distance of trips can also be used as indicators of Vélib' usage. As shown in Figure 2, half of the trips last less than twelve minutes: this can be linked to the Vélib' pricing policy (free for half an hour).

These first statistics show the global dynamic of the Vélib' system. Let us now examine the clustering results obtained using the adapted LDA to automatically extract finer details from BSS data.

RESULTS AND DISCUSSION

Pre-processing and description of the Vélib' OD-data

The proposed methodology was applied on two months of trips data recorded by the Vélib' BSS in April and September 2011. These datasets contain the following informations for each trip: station of departure, time of departure, station of arrival, time of arrival, type of user subscription (day / year). Their are roughly 2 500 000 trips in the April dataset and 3 000 000 trips in the September

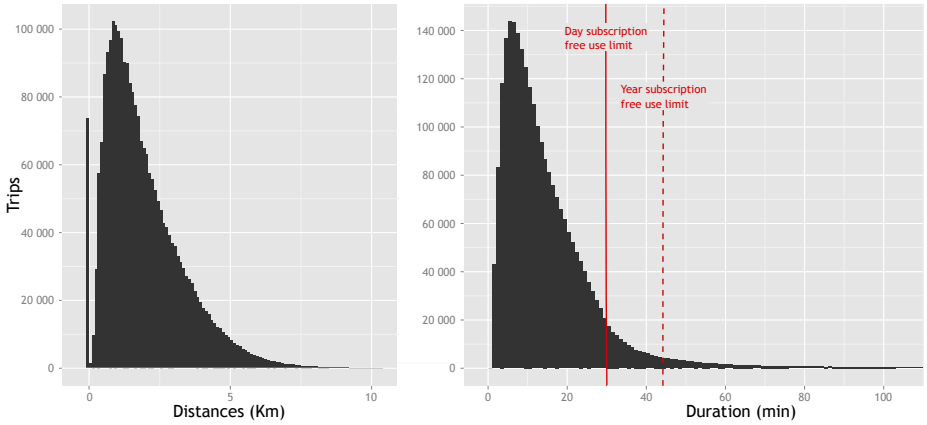

FIGURE 2 Histogram of trip length in kilometers (left) and of trip duration in minutes (right). The trips recorded with null distances correspond to round trips.

dataset. The April dataset was used for training (fitting of the model) and September for testing and model selection purpose (performance evaluation).

First, the trips were sorted according to their starting date and then cut in OD-bags of 5 000 successive trips each. For each OD-bag the number of occurrences of each possible OD-couple was computed. Since 1188 stations were operating during these two months, the number of possible OD-couples was $1188^2 \approx 1,500,000$: much less were in fact observed during these two months since less than $500\,000$ different OD-couples were observed during April 2011.

As often in text mining, rare words are removed from the analysis. The OD-couples observed in less than five OD-bags were similarly removed. This final preprocessing gives us $117\,838$ possible OD-couples describing the OD-bags of trips. The number of occurrences of each one of these OD-couples in each OD-bag were then the inputs of the Variational EM algorithm used to fit the model.

Model selection

As usual in clustering, one important element to fix is the number K of OD-templates: models with K ranging from two to thirteen OD-templates were therefore fitted and compared, using the model Perplexity on test data (September 2011). Such quantity (see Grün and Hornik (*21*) for an introduction) measures to what extend each tested-model is confused by new data. Figure 3 represents this quantity with respect to K, in which a significant drop of the Perplexity value is observed when $K = 5$. Since the Perplexity has to be be minimal to achieve a good description of the data, this value of five OD-templates is fixed and seems a good candidate to performs the analysis.

Taking these observations into account now leads us to detail the results obtained with five OD templates in the remaining paragraphs.

Temporal segmentation

A first way to look at the results obtained by LDA is to plot the template proportions $\pi^{(i)}$ of the OD-bags with respect to time. This is presented in Figure 4 where each OD-bag is depicted with a colored stacked bar charts. Each color is, in this Figure, associated with an OD-template: the bar height represents the estimated number of trips per hour expected to be drawn from the OD-

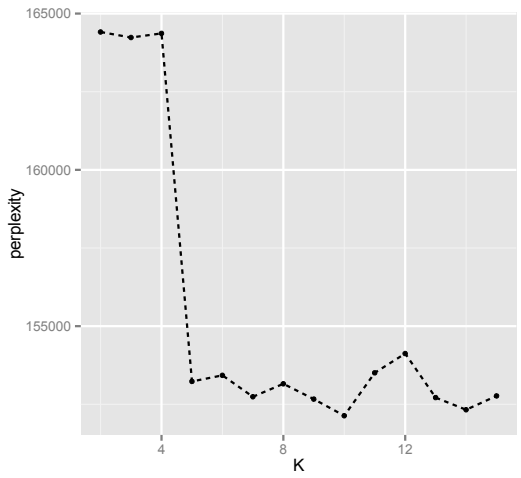

FIGURE 3 Perplexity on the September dataset with respect to the number of latent OD templates of the models.

template, and the bar width corresponds to the time span of the OD-bag.

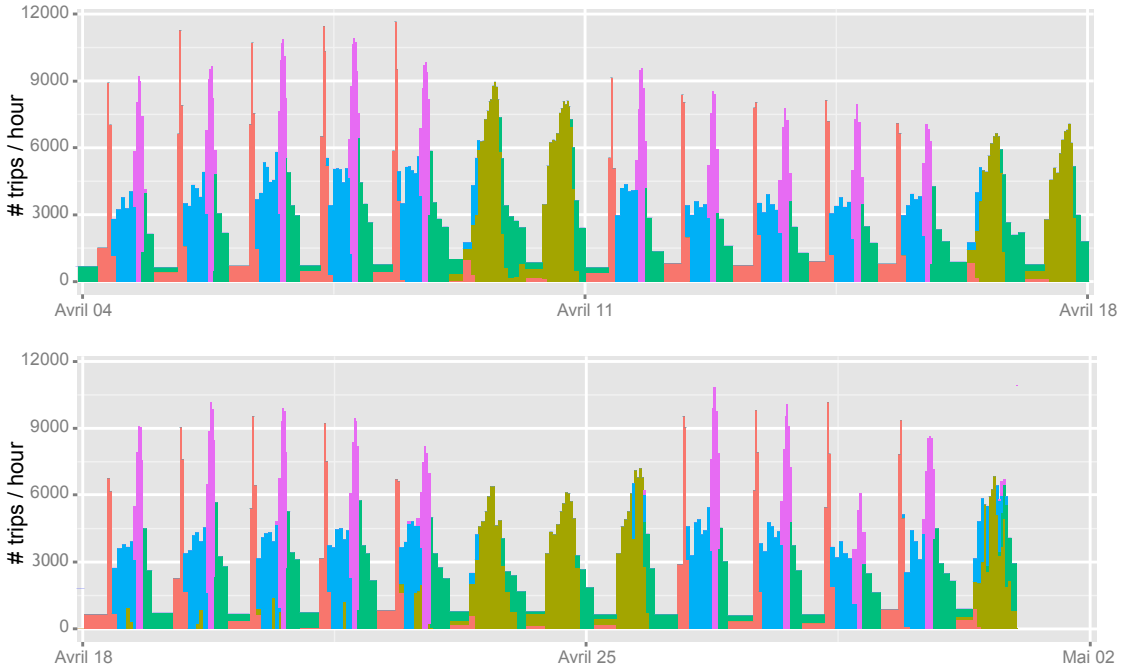

FIGURE 4 Temporal evolution of the OD-templates proportions $\pi^{(i)}$. Each bag is represented by stacked bars were the height of each bar represents the estimated number of trips per hour expected to be drawn from each OD-template (one colour per OD-template). The bars width encodes to the timespan of the OD-bags.

The expected cyclic patterns are clearly visible in this Figure. Five days with the same

shape are followed by two days with another shape and so forth, except on April 25 (Easter day) which presents a shape similar to a weekend. A typical weekday has the following form: lots of *red trips* occur during the morning pick, followed by a large majority of *blue trips*. Then *magenta trips* are observed during the second pick of the day and eventually *green trips* (or a mix of *green and red trips*) are seen during the evening and at night.

The different colored OD-templates are thus clearly identifiable: the *red OD-template* corresponds to the $Home \rightarrow Work$ commute, the blue OD-template to the system behavior during Lunch time (denoted by $Lunch$), the *magenta OD-template* to the $Work \rightarrow Home$ commute and the *green OD-template* to the evening behavior (denoted $Evening$). The last OD-template (brown color) is mainly observed during weekends and little during Lunch time the week before Easter: it is denoted $Spare\,time$. Another important aspect about this segmentation is that the majority of the OD-bags can be quite easily associated to a unique OD-template. Although some bags are mixtures of OD-template, there are not that many. All the OD-templates being quite easily identifiable, we may then gain insight into the system behavior during each of these time period by looking at the distribution of each OD-template.

OD-templates analysis

Since each OD-template corresponds to a discrete distributions over OD-couples, these distributions can be used to describe the behavior of the system in the timespan were the OD-template is responsible for the large majority of trips. These distributions are quite big since we analyzed more than 1000×1000 OD-couples. It is therefore interesting to rely on indicators, that are summaries of each station, to analyze these distributions. We detail in the next paragraph such summaries.

Arrival and Departure specificities

One way to process these distributions is to focus on each OD-template and look at the stations that present an increasing number of incoming and outgoing trips, with respect to the mean behavior of the system. In other words, we are looking at station that gets or looses more bikes than average in OD-template k. To formalize this idea, the station arrival specificity $A_s^{(k)}$ and departure specificity $D_s^{(k)}$ is introduced, for each OD-template k . For an OD-template k, these quantities are defined for a station s as:

$$A_s^{(k)} = \log\left(\frac{pa_s^{(k)}}{pa_s^g}\right),\ D_s^{(k)} = \log\left(\frac{pd_s^{(k)}}{pd_s^g}\right), \tag{3}$$

with $pa_s^{(k)}$ (resp. $pd_s^{(k)}$) the probability that a trip ends (resp. starts) in station s according to an OD-template k and pa_s^g (resp. pd_s^g) the average probabilities that a trip ends (resp. starts) in station s. Each of the OD-template probability is computed using:

$$pa_s^{(k)} = \sum_j \Lambda_{js}^{(k)},\ pd_s^{(k)} = \sum_j \Lambda_{sj}^{(k)}$$

with $\Lambda_{ij}^{(k)}$ the probability of OD (i, j) to be in template k, estimated using to LDA. The global probability pa_s^g (resp. pd_s^g) is the empirical probabilities that a trip ends (resp. starts) in station s. It is estimated on the entire dataset using:

$$pa_s^g = \frac{\sum_{j,t} \mathbf{C}_{tjs}}{\sum_{i,j,t} \mathbf{C}_{tij}},\ pd_s^g = \frac{\sum_{j,t} \mathbf{C}_{tsj}}{\sum_{i,j,t} \mathbf{C}_{tij}}$$

where $\mathbf{C}_{tij}$ is the number of trips of OD-bag t moving from station i to j.

Under such settings, the stations with a departure (resp. arrival) specificity greater than zero experience an increasing number of departures (resp. arrivals) for OD-template k. For each OD-template, these indicators offer a natural way to highlight trips *generators* and *attractors*. They can be easily mapped, as shown in Figure 5 (left) which depicts, through dots of different sizes, the arrival specificities on stations, for the latent OD-template $House \rightarrow Work$. The departure specificities on stations, for the latent OD-template $House \rightarrow Work$, are shown in Figure 5 (right).

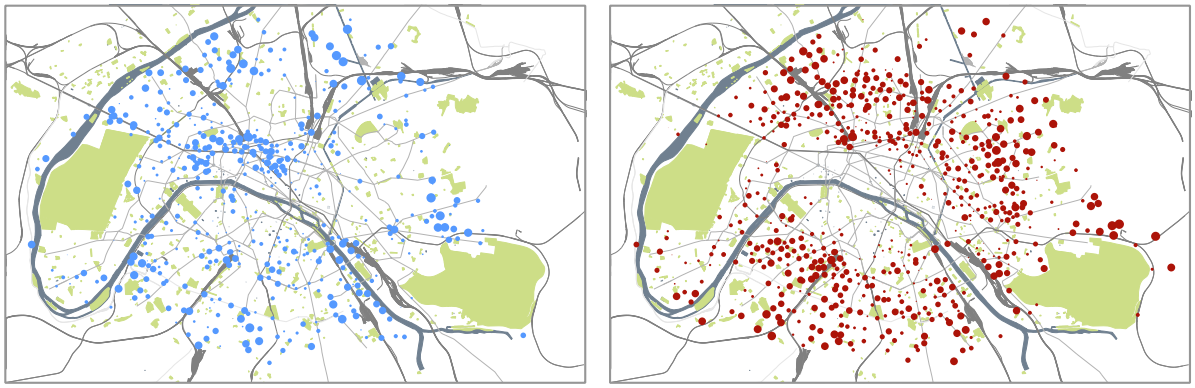

FIGURE 5 Arrival (left) and departure (right) specificities on stations, observed for the latent OD-template $House \rightarrow Work$ commute. The dots of different sizes encode the stations specificities, using a linear scale starting from zero. Stations with negative specificities are not shown on these maps.

These two maps give a clear overview of the system behavior during the $House \rightarrow Work$ commute. Peripheral stations have an important departure specificity, whereas stations close to "Les grands boulevards" and "Saint-Lazarre" railway station reach important arrival specificities. Other BSS stations close from the big railway stations (such as "Montparnasse", "Austerlitz", "Gare de Lyon","Gare du Nord" and "Gare de l'Est") present high values for both the departure and arrival specificities. Since these two maps strongly differ, the imbalance introduced in the system during this pattern is important: some stations experience more incoming than outgoing trips, and vice versa.

Similar maps are drawn for the four other latent OD-templates and are presented in Figure 6. For the $Lunch$ template the two maps are quite similar with high specificities values located in the Paris center. Since the departure and arrival maps are similar, this pattern do not introduce a big imbalance in the system. The $Work \rightarrow House$ commute present however a strong asymmetry at the opposite from the $House \rightarrow Work$ commute: the bikers clearly leave the "Grands Boulevards" and move to peripheral stations. The observed value are however not as important as during the morning rush, and some differences are clearly visible. For instance, stations close to the Seine, between the railway stations "Gare de Lyon" and "Austerlitz", have important departure specificities in this template: they were not visible in the maps of the $House \rightarrow Work$ commute,

inducing that the system seems to be partially re-balanced. The $Evening$ maps present an important number of station with a high arrival and departure specificities in the North, the North-East of Paris and in a lesser extend in the South-West. Finally, the stations of the center have an important departure specificities. The last two maps, which correspond to the $Spare\,time$ OD-template, are quite symmetric and the influence of parks, of the "Canal St Martin" and of important tourist places such as the Eiffel Tower, the "Cité des Sciences et de l'Industrie" (a science museum complex), the old historical center of Paris are clear. In this template, all these places experience important departure and arrival specificities.

Stations expected balances

A more direct way to assess the effect of the template on the bike distribution is to look at station balances, defined as the number of arrivals minus the number of departures. In fact, if all trips are drawn using a unique template k the OD matrix follows a multinomial law of parameters $\Lambda^{(k)}$. Let N_{dep} denote the number of trips, then:

$$OD \sim \mathcal{M}(N_{dep}, \Lambda^{(k)}).$$

The bike balance B_s (*Number of arrivals* minus *Number of departures*) for a station s is deduced from the previous equation and expresses as:

$$B_s = \overbrace{\sum_j OD_{js}}^{arrivals} - \overbrace{\sum_j OD_{sj}}^{departures},$$

The expectation $\mathbf{B}$ of the balance, for all the stations, can be computed easily using:

$$\mathbb{E}[\mathbf{B}] = N_{dep} \left((\Lambda^{(k)})^t - \Lambda^{(k)} \right) \mathbf{v} \tag{4}$$

with $\mathbf{v} = (1, \ldots, 1)^t$. For a given OD-template k, this vector describes how the station stocks evolve after N_{dep} trips are made: a negative expectation will characterize stations which lose bikes whereas a positive one will characterize stations which gets bikes. Such statistics can easily be mapped as shown in Figure 7, where the station balances for the $House \rightarrow Work$ commute template are represented.

Such a map summarizes the two view presented in Figure 5 into only one map. The trips generators and attractors already observed are clearly visible. From an operational point of view these statistics are also relevant since they point the stations which experience bike saturation or bike unavailability on a quantitative scale. Regarding this OD-template, one clearly see the flow from peripheral stations to more central stations, as shown in Figure 5: the stations of the neighborhood "Les grands Boulevards" receive bikes. For instance, one station gets more than 30 bikes per 10 000 trips made, which corresponds roughly to half the number of trips made during a typical week day during the morning rush.

CONCLUSION

In this paper, the problem of spatio-temporal analysis of Origin-Destination (OD) data is investigated within a data mining framework based on an advanced statistical model called Latent Dirichlet Allocation (LDA). The proposed methodology, applied to the mining of OD-matrices collected on the Vélib' Bike Sharing System of Paris, has the advantage to identify a reduced set of demand

profiles. The obtained results have shown that these demand profiles can be summarized by few OD-templates which are typical and temporally localized. Furthermore the proposed methodology may address the issue related to the balancing load of bikes since relevant indicators on the imbalances of stations was provided by such an approach to analyze the behavior of the system.

A perspective to this work is to use the proposed methodology to analyze OD-matrices generated by other modes of transport such as railway transport or road traffic data. Further works may concern the extension of the LDA model in order to take into account external factors such as sociological, economical and geographical context of a city. The long-term goal of this kind of research work is to build a dedicated tool able to automatically position and dimension the BSS stations considering a given city context which can be useful to extend an existing BSS or to implement a new BSS.

REFERENCES

[1] Benchimol, M., P. Benchimol, B. Chappert, A. De La Taille, F. Laroche, F. Meunier, and L. Robinet, Balancing the stations of a self-service bike hire system. *RAIRO-Operations Research*, Vol. 45, No. 1, 2011, pp. 37–61.

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

[3] Nair, R., E. Miller-Hooks, R. C. Hampshire, and A. Bušić, Large-Scale Vehicle Sharing Systems: Analysis of Vélib'. *International Journal of Sustainable Transportation*, Vol. 7, No. 1, 2012, pp. 85–106.

[4] Lin, J.-R. and T.-H. Yang, Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E-logistics and Transportation Review*, Vol. 47, 2011, pp. 284–294.

[5] Froehlich, J., 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*, 2009, pp. 1420–1426.

[6] Borgnat, P., E. Fleury, C. Robardet, and A. Scherrer, Spatial analysis of dynamic movements of VÃ©lo'v, Lyon's shared bicycle program. In *European Conference on Complex Systems, ECCS'09* (F. Kepes, ed.), 2009.

[7] Kaltenbrunner, A., R. Meza, J. Grivolla, J. Codina, and R. Banchs, Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, Vol. 6, No. 4, 2010, pp. 455–466.

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

[9] Vogel, P., T. Greiser, and D. Mattfeld, Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, Vol. 20, No. 0, 2011, pp. 514 – 523.

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

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

[12] Borgnat, P., C. Robardet, P. Abry, P. Flandrin, J. Rouquier, and N. Tremblay, *Dynamics On and Of Complex Networks, Volume 2*, Springer Berlin Heidelberg, chap. A Dynamical Network View of Lyon's Vélo'v Shared Bicycle System, 2013.

[13] Borgnat, P., P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury, Shared Bicycles in a City: A Signal processing and Data Analysis Perspective. *Advances in Complex Systems,*, Vol. 14, No. 3, 2011, pp. 1–24.

[14] Randriamanamihaga, A., E. Côme, L. Oukhellou, and G. Govaert, Clustering the Vélib' origin-destinations flows by means of Poisson mixture models. In *Proceedings of the 17th European Symposium on Artificial Neural Networks Computational Intelligence and Machine Learning, Bruges, Belgium, April 24-26*, 2013.

[15] Vogel, P. and 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.

[16] Blei, D. M., A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993–1022.

[17] Montoliu, R., Discovering Mobility Patterns on Bicycle-Based Public Transportation System by Using Probabilistic Topic Models. In *ISAmI*, 2012, pp. 145–153.

[18] Farrahi, K. and D. Gatica-Perez, Discovering routines from large-scale human locations using probabilistic topic models. *ACM TIST*, Vol. 2, No. 1, 2011, pp. 3–5.

[19] Huynh, T., M. F., and B. Schiele, Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*, ACM New York, NY, USA, ACM New York, NY, USA, 2008, pp. 10–19.

[20] Niebles, J. C., H. Wang, and L. Fei-Fei, Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *Int. J. Comput. Vision*, Vol. 79, No. 3, 2008, pp. 299–318.

[21] Grün, B. and K. Hornik, topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, Vol. 40, 2011, pp. 1–30.

[22] Wainwright, M. and M. Jordan, Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, Vol. 1, 2008, pp. 1–305.

[23] Hoffman, M., D. Blei, and F. Bach, Online learning for latent Dirichlet allocation. In *NIPS*, 2010, pp. 856–864.

[24] Mimno, D. and A. McCallum, Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of Uncertainty in Artificial Intelligence*, 2008, pp. 411–418.

[25] Blei, D. M. and J. D. Lafferty, Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ACM, New York, NY, USA, 2006, ICML '06, pp. 113–120.

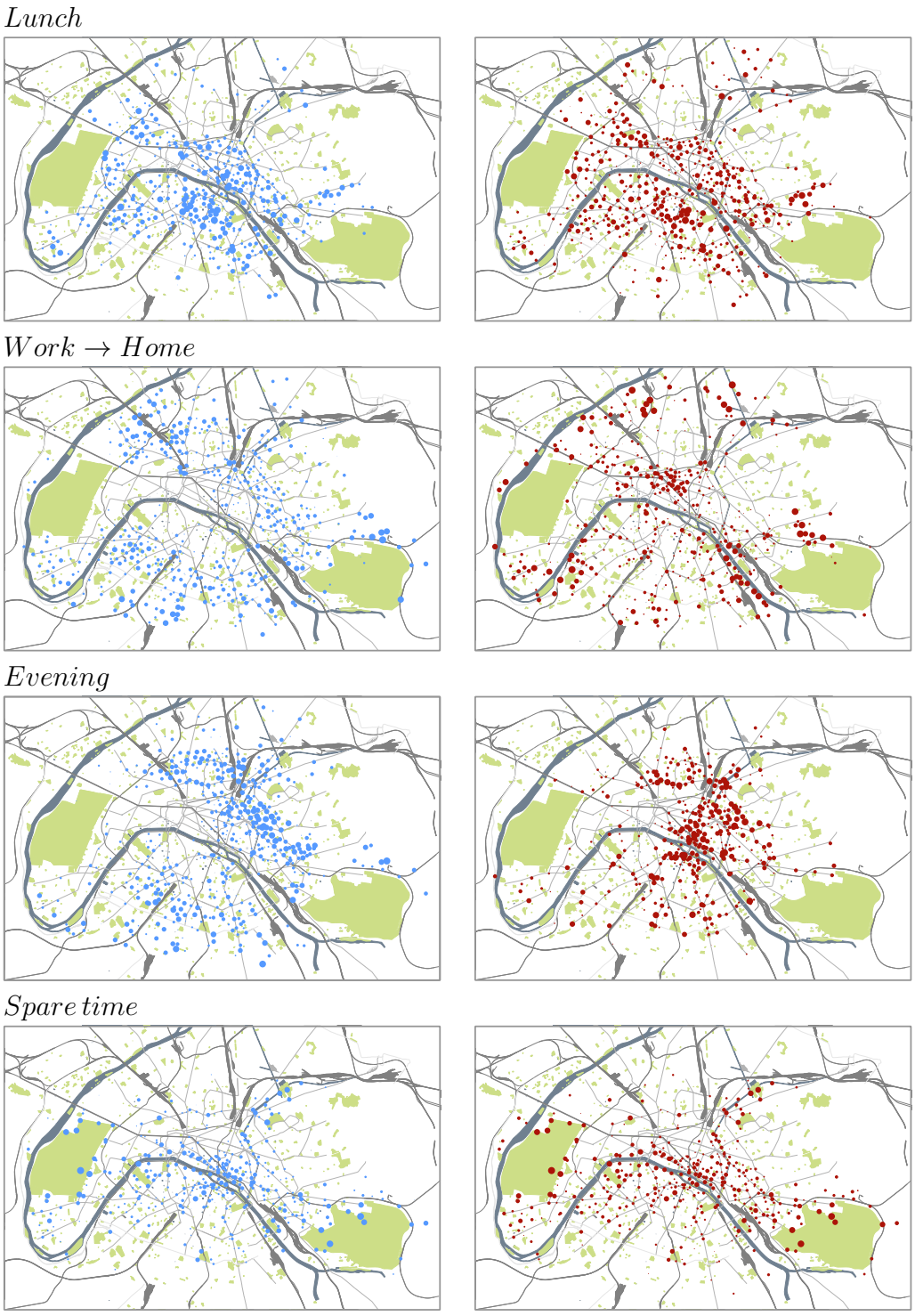

FIGURE 6 Arrival (left column) and departure (right column) specificities on stations, for four latent OD-templates. The dots of different sizes encode the stations specificities, using a linear scale starting from zero. Stations with negative specificities are not shown on these maps.

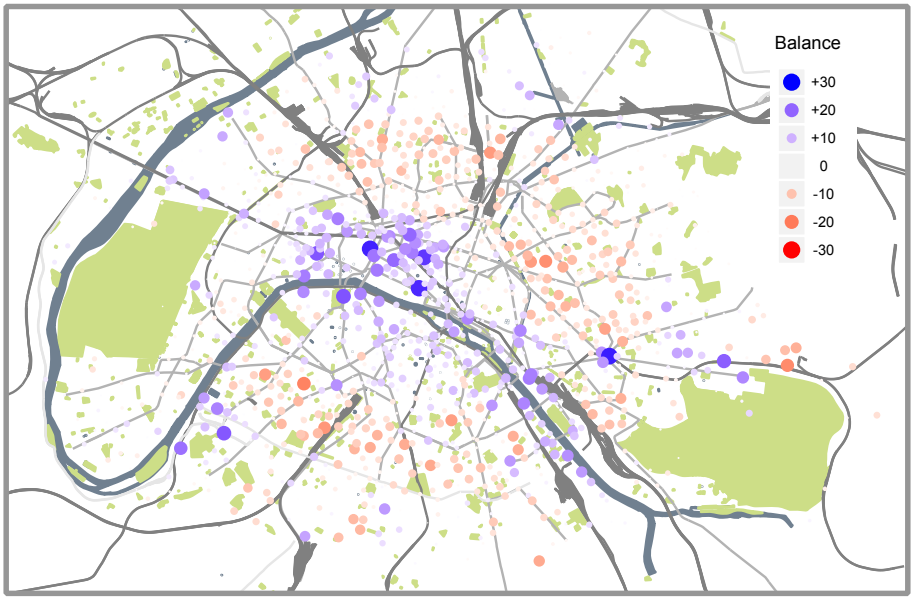

FIGURE 7 Latent activity for the OD-template $House \rightarrow Work$ commute. This Figure describes how the station stocks evolve after N_{dep} trips: a negative expectation will characterize stations which lose bikes whereas a positive one will characterize stations which gets bikes. Here, $N_{dep} = 10\,000$.