



**HAL**  
open science

# 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

## ► To cite this version:

Etienne Come, Njato Andry Randriamanamihaga, Latifa Oukhellou, Patrice Aknin. Spatio-temporal Analysis of Dynamic Origin-Destination Data Using Latent Dirichlet Allocation: Application to Vélib' Bike Sharing System of Paris. TRB 93rd Annual meeting, Jan 2014, France. 19p. hal-01052951

**HAL Id: hal-01052951**

**<https://hal.science/hal-01052951>**

Submitted on 29 Jul 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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

1 **ABSTRACT**

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

## 1 INTRODUCTION

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

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

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

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

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

## 6 RELATED WORK

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

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

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

40 Other researches using OD-trips data are proposed by Vogel et al., Vogel and Mattfeld  
41 (9, 15) and aim to identify a reduced set of clusters of stations to get a better understanding of  
42 the spatial and temporal causes of imbalances between BSS stations. The proposed methodology,  
43 based on Geographical Business Intelligence process, was successfully applied to data collected  
44 from the Vienna's BSS Citybike Wien. It used feature vectors, *i.e* the per-hour and per-station

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

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

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

## 24 LATENT DIRICHLET ALLOCATION APPLIED ON DYNAMIC OD-MATRICES

### 25 Background on Latent Dirichlet Allocation

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

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

- 35 – **Corpus:** the corpus is a collection of  $M$  documents, denoted  $\mathbf{C} = \{\mathbf{d}^1, \dots, \mathbf{d}^M\}$ .
- 36 – **Documents:** Each of the  $M$  documents is a bag of words  $\mathbf{d} = \{\mathbf{w}^1, \dots, \mathbf{w}^L\}$ . Since the  
 37 size of each bag of words is not fixed, the number of words in the  $i^{th}$  document of the  
 38 corpus is neither fixed and is denoted  $L(i)$ .
- 39 – **Words:** the words are taken within a fixed vocabulary indexed by  $\{1, \dots, N\}$ . A word  
 40  $\mathbf{w}$  is represented by an indicator such that the  $N$ -vector  $\mathbf{w}$ , with  $w_j = 1$  and all other index  
 41 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  $\mathcal{D}(\cdot)$  of parameter  $\beta$ :

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

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  $\mathcal{M}(\cdot, \cdot)$
  - (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 maximisation of the log probability  $\log(p(\mathbf{C}|\Lambda, \alpha))$  of the corpus  $\mathbf{C}$  with respect to the parameters  $\alpha$  and  $\Lambda$ . 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).

## 1 LDA adapted to Dynamical OD-data mining

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

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

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

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

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

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

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

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



1           Such a model is promising and is likely to find interesting structure in the BSS data, known  
2 to be affected by cyclic regularities. The observed behavior of users during the Mondays, between  
3 8a.m and 9a.m, would for example be quite similar to the behavior of users observed during the  
4 same time frame, for other weekdays. To keep the analogy, two documents dealing with the same  
5 topics certainly share common lexical fields. Consequently, the cyclo-stationarity introduced by  
6 daily activities such as *Home* → *Work* → *Leisure* → *Home* may therefore be recovered by  
7 the model, which can give insight of the whole system behavior with only few OD-templates. The  
8 patterns specific to weekdays and weekends, expected to be hidden in the data, may play a role  
9 similar to those of the hidden topics behind the observed similarities in documents collections.  
10 These patterns may be caught by the OD templates: for example, we can think of a *Home* →  
11 *Work* OD-template which has OD-couples leaving from places with high population densities and  
12 going to places with high employment densities. Eventually, LDA will help to assess the timespan  
13 over which the system can be considered as stationary and to discern the associated change points.

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

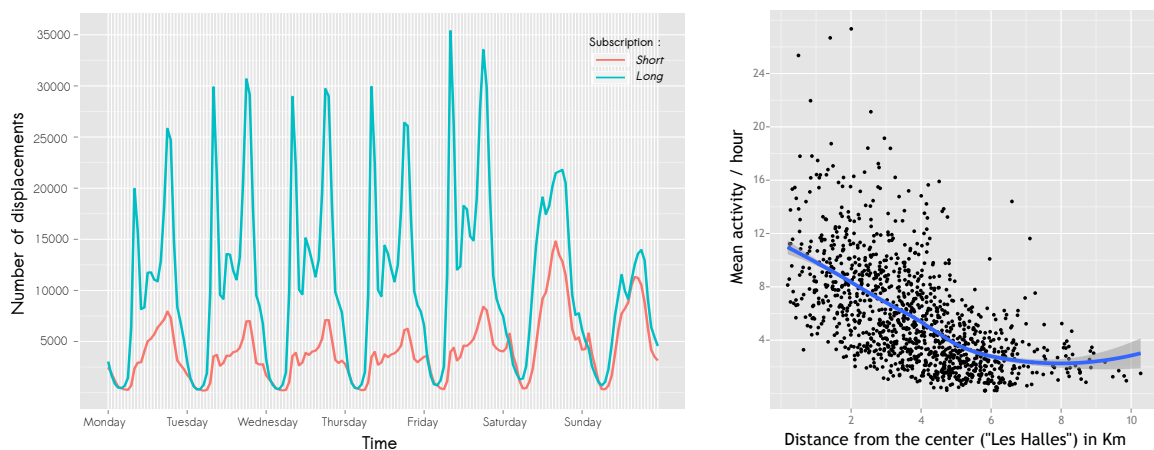
## 17 **THE VÉLIB' BIKE SHARING SYSTEM OF PARIS**

18 The Bike Sharing System of Paris, called Vélib' has been deployed since July 2007 and is operated  
19 as a concession by Cyclocity, a subsidiary of the French outdoor advertising company JCDecaux.  
20 It offers a non stop service 24/7 and at its debut in 2007, 700 bicycles were spread across 750  
21 fixed stations. In four years, it has expanded to more than 1200 stations which hire out around  
22 18,000 bikes throughout the city. Considering the number of annual subscribers, 224,000 and still  
23 growing, and the average number of 110,000 trips per day, Vélib' is large-scale and is now one  
24 of the largest Bike Sharing Systems in the world and the biggest Bike Sharing System in Europe.  
25 Vélib' is available mainly in Paris *intramuros*, some stations being located in the suburbs. At the  
26 stations, the bikes are locked to the electronically controlled docking points: the whole network  
27 includes 40,000 of them, inducing from 8 to 70 docking points per station. Regarding the policy,  
28 a user can purchase a short-term daily or weekly subscription, or a long-term annual subscription  
29 which allows an unlimited number of rentals. In both cases, the first half hour (or the first 45  
30 minutes for a long-term subscription) of every individual trip is free of cost.

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

37           Figure 1 (left) displays the total number of recorded trips, observed over a week, with  
38 respect to the type of subscription: annual (plotted in blue) or during one day (plotted in red).  
39 The blue curve shows a repetitive but distinct pattern depending on the type of day. Weekdays  
40 (Monday to Friday) are marked with peaks at the commutes (8a.m and 6p.m) and during the lunch  
41 break, whereas the highest volume usage at weekends (Saturday and Sunday) is evenly distributed  
42 throughout the afternoons. The red curve depicts a totally different pattern with higher activity  
43 early morning and late afternoon. In addition, considering the volume of displacements during  
44 Saturdays, Sundays and Mondays, the typical weekend pattern for the one-day users lasts until

1 Tuesday. It is reasonable to assume that these trips are more leisure and recreational oriented.  
 2 These temporal trends of BSS usage can provide information on the sociological characteristics of  
 3 the city. Considering the study carried out by Froehlich et al. (5) on the Barcelona Bicing BSS,  
 4 some sociological differences between Barcelona and Paris can be highlighted: the lunch peak  
 5 occurring at 2p.m in Barcelona Bicing data occurs at 12 noon in the Vélib' data, reflecting the late  
 6 lunch culture of Spain (resp. the earlier lunch culture of France). Secondly, although Friday is the  
 7 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").**

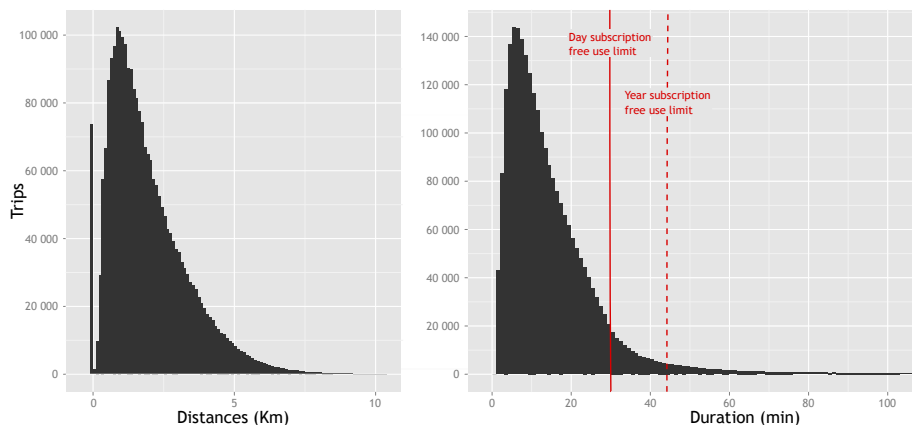
8 In addition to these temporal trends, spatial trends closely linked to geographical aspects  
 9 of the city can also be identified. Figure 1(right) shows the average number of observed departures  
 10 and arrivals per hour with respect to the distance from the station to the center of Paris. It is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

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

## 18 RESULTS AND DISCUSSION

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

20 The proposed methodology was applied on two months of trips data recorded by the Vélib' BSS in  
 21 April and September 2011. These datasets contain the following informations for each trip: station  
 22 of departure, time of departure, station of arrival, time of arrival, type of user subscription (day /  
 23 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.**

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

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

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

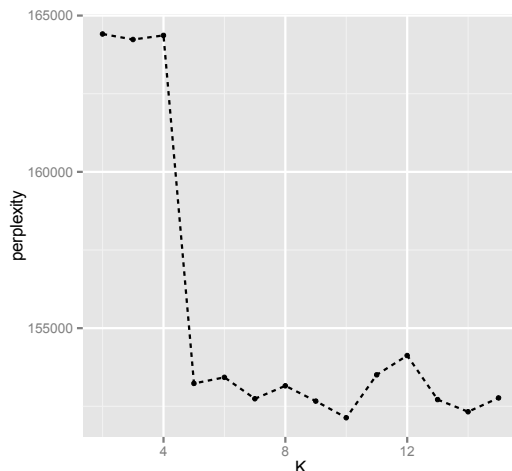
### 13 Model selection

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

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

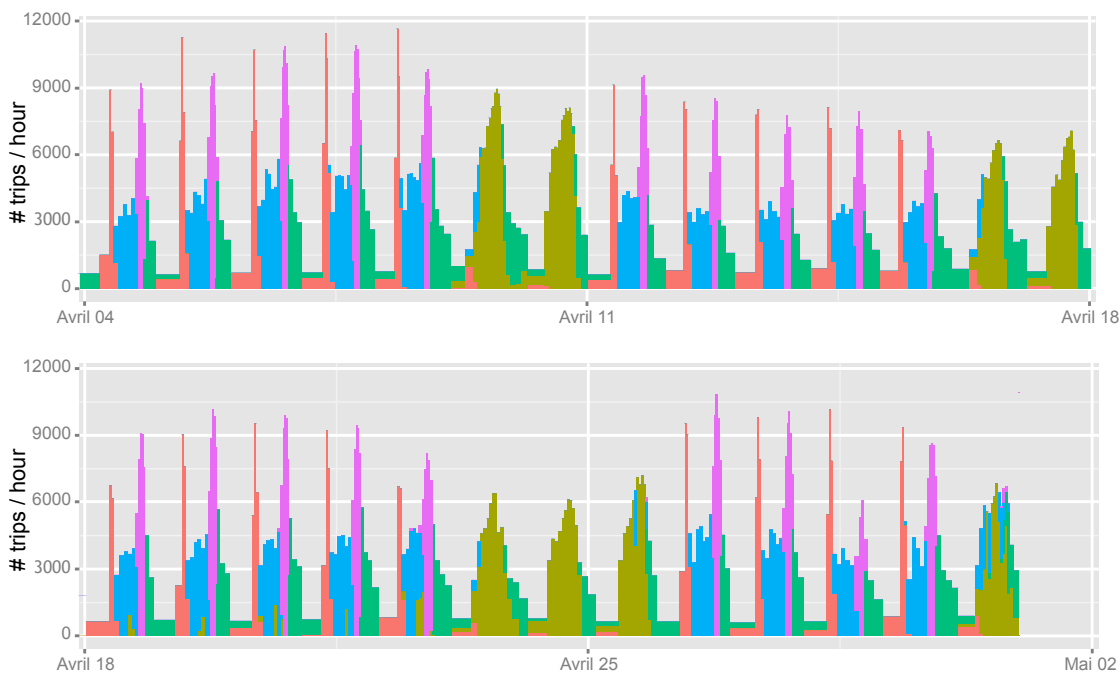
### 24 Temporal segmentation

25 A first way to look at the results obtained by LDA is to plot the template proportions  $\pi^{(i)}$  of the  
 26 OD-bags with respect to time. This is presented in Figure 4 where each OD-bag is depicted with  
 27 a colored stacked bar charts. Each color is, in this Figure, associated with an OD-template: the  
 28 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.

1 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 where 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.

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

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

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

## 16 OD-templates analysis

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

### 22 Arrival and Departure specificities

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

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

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

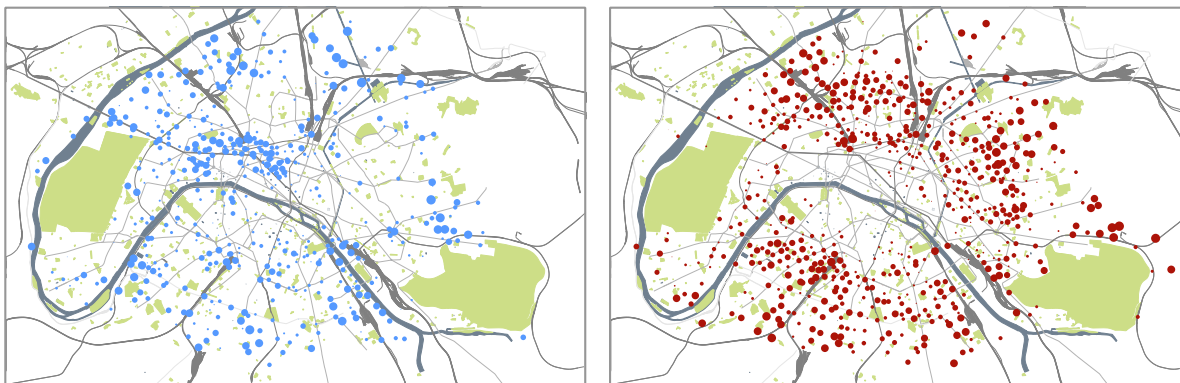
$$pa_s^{(k)} = \sum_j \Lambda_{js}^{(k)}, pd_s^{(k)} = \sum_j \Lambda_{sj}^{(k)}$$

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

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

1 where  $C_{tij}$  is the number of trips of OD-bag  $t$  moving from station  $i$  to  $j$ .

2 Under such settings, the stations with a departure (resp. arrival) specificity greater than  
 3 zero experience an increasing number of departures (resp. arrivals) for OD-template  $k$ . For each  
 4 OD-template, these indicators offer a natural way to highlight trips *generators* and *attractors*. They  
 5 can be easily mapped, as shown in Figure 5 (left) which depicts, through dots of different sizes,  
 6 the arrival specificities on stations, for the latent OD-template  $House \rightarrow Work$ . The departure  
 7 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.

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

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

1 inducing that the system seems to be partially re-balanced. The *Evening* maps present an impor-  
 2 tant number of station with a high arrival and departure specificities in the North, the North-East of  
 3 Paris and in a lesser extend in the South-West. Finally, the stations of the center have an important  
 4 departure specificities. The last two maps, which correspond to the *Spare time* OD-template, are  
 5 quite symmetric and the influence of parks, of the “Canal St Martin” and of important tourist places  
 6 such as the Eiffel Tower, the “Cit  des Sciences et de l’Industrie” (a science museum complex),  
 7 the old historical center of Paris are clear. In this template, all these places experience important  
 8 departure and arrival specificities.

### 9 *Stations expected balances*

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

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

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

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

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

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

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

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

## 30 CONCLUSION

31 In this paper, the problem of spatio-temporal analysis of Origin-Destination (OD) data is investi-  
 32 gated within a data mining framework based on an advanced statistical model called Latent Dirich-  
 33 let Allocation (LDA). The proposed methodology, applied to the mining of OD-matrices collected  
 34 on the V lib’ Bike Sharing System of Paris, has the advantage to identify a reduced set of demand

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

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

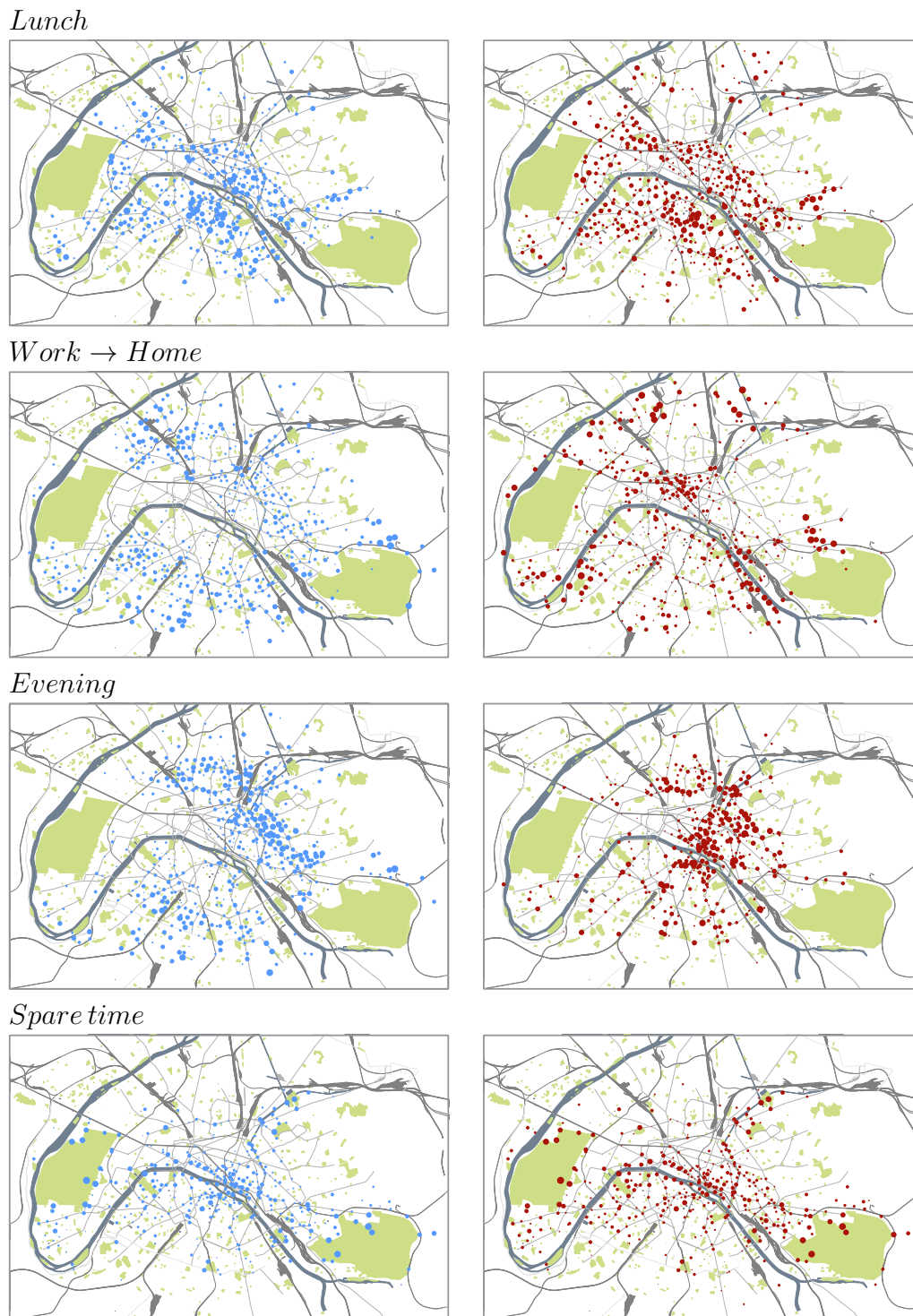
## 12 REFERENCES

- 13 [1] Benchimol, M., P. Benchimol, B. Chappert, A. De La Taille, F. Laroche, F. Meunier, and  
14 L. Robinet, Balancing the stations of a self-service bike hire system. *RAIRO-Operations Re-*  
15 *search*, Vol. 45, No. 1, 2011, pp. 37–61.
- 16 [2] Chemla, D., F. Meunier, and R. Wolfler Calvo, Balancing a bike-sharing system with multiple  
17 vehicles. In *Congrès annuel de la société Française de recherche opérationnelle et d'aide à la*  
18 *décision, ROADEF2011*, Saint-Etienne, France, 2011.
- 19 [3] Nair, R., E. Miller-Hooks, R. C. Hampshire, and A. Bušić, Large-Scale Vehicle Sharing  
20 Systems: Analysis of Vélib'. *International Journal of Sustainable Transportation*, Vol. 7,  
21 No. 1, 2012, pp. 85–106.
- 22 [4] Lin, J.-R. and T.-H. Yang, Strategic design of public bicycle sharing systems with ser-  
23 vice level constraints. *Transportation Research Part E-logistics and Transportation Review*,  
24 Vol. 47, 2011, pp. 284–294.
- 25 [5] Froehlich, J., J. Neumann, and N. Oliver, Sensing and Predicting the Pulse of the City through  
26 Shared Bicycling. In *Proceedings of the 21st International Joint Conference on Artificial*  
27 *Intelligence*, 2009, pp. 1420–1426.
- 28 [6] Borgnat, P., E. Fleury, C. Robardet, and A. Scherrer, Spatial analysis of dynamic movements  
29 of Vélo'v, Lyon's shared bicycle program. In *European Conference on Complex Systems,*  
30 *ECCS'09* (F. Kepes, ed.), 2009.
- 31 [7] Kaltenbrunner, A., R. Meza, J. Grivolla, J. Codina, and R. Banchs, Urban cycles and mo-  
32 bility patterns: Exploring and predicting trends in a bicycle-based public transport system.  
33 *Pervasive and Mobile Computing*, Vol. 6, No. 4, 2010, pp. 455–466.
- 34 [8] Michau, G., C. Robardet, L. Merchez, P. Jensen, P. Abry, P. Flandrin, and P. Borgnat, Peut-on  
35 attraper les utilisateurs de Vélo'v au Lasso ? In *Proceedings of the 23e Colloque sur le*  
36 *Traitement du Signal et des Images. GRETSI-2011*, 2011, pp. 46–50.
- 37 [9] Vogel, P., T. Greiser, and D. Mattfeld, Understanding Bike-Sharing Systems using Data Min-  
38 ing: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, Vol. 20, No. 0,  
39 2011, pp. 514 – 523.

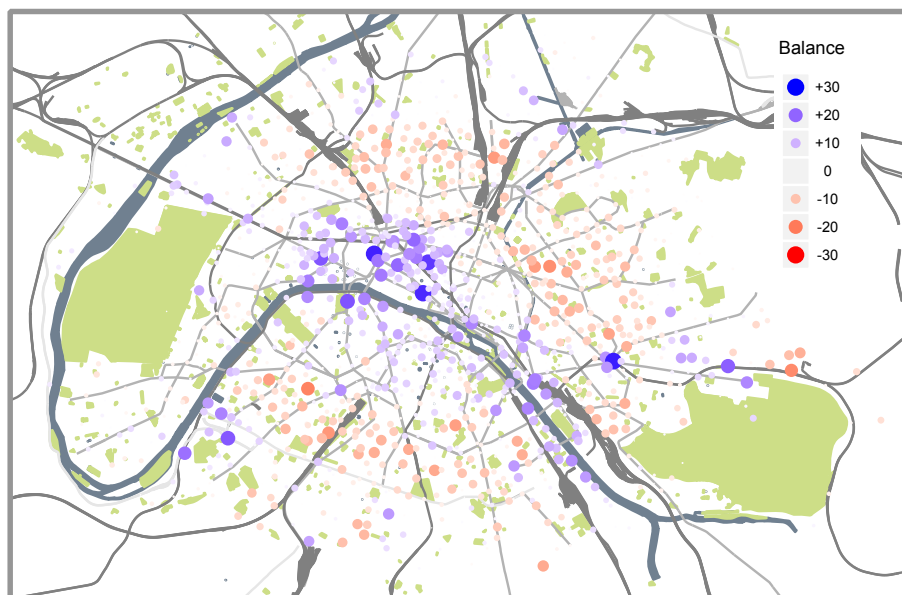


- 1 [10] Froehlich, J., J. Neumann, and N. Oliver, Measuring the pulse of the city through shared  
2 bicycle programs. In *Proc. of UrbanSense08*, 2008, pp. 16–20.
- 3 [11] Lathia, N., S. Ahmed, and L. Capra, Measuring the impact of opening the London shared  
4 bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*,  
5 Vol. 22, 2012, pp. 88–102.
- 6 [12] Borgnat, P., C. Robardet, P. Abry, P. Flandrin, J. Rouquier, and N. Tremblay, *Dynamics  
7 On and Of Complex Networks, Volume 2*, Springer Berlin Heidelberg, chap. A Dynamical  
8 Network View of Lyon’s Vélo’v Shared Bicycle System, 2013.
- 9 [13] Borgnat, P., P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury, Shared Bicycles  
10 in a City: A Signal processing and Data Analysis Perspective. *Advances in Complex Systems*,  
11 Vol. 14, No. 3, 2011, pp. 1–24.
- 12 [14] Randriamanamihaga, A., E. Côme, L. Oukhellou, and G. Govaert, Clustering the Vélip’  
13 origin-destinations flows by means of Poisson mixture models. In *Proceedings of the 17th Eu-  
14 ropean Symposium on Artificial Neural Networks Computational Intelligence and Machine  
15 Learning, Bruges, Belgium, April 24-26, 2013*.
- 16 [15] Vogel, P. and D. Mattfeld, Strategic and Operational Planning of Bike-Sharing Systems by  
17 Data Mining - A Case Study. In *ICCL*, Springer Berlin Heidelberg, 2011, pp. 127–141.
- 18 [16] Blei, D. M., A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation. *Journal of Machine  
19 Learning Research*, Vol. 3, 2003, pp. 993–1022.
- 20 [17] Montoliu, R., Discovering Mobility Patterns on Bicycle-Based Public Transportation System  
21 by Using Probabilistic Topic Models. In *ISAmI*, 2012, pp. 145–153.
- 22 [18] Farrahi, K. and D. Gatica-Perez, Discovering routines from large-scale human locations using  
23 probabilistic topic models. *ACM TIST*, Vol. 2, No. 1, 2011, pp. 3–5.
- 24 [19] Huynh, T., M. F., and B. Schiele, Discovery of activity patterns using topic models. In *Pro-  
25 ceedings of the 10th international conference on Ubiquitous computing*, ACM New York,  
26 NY, USA, ACM New York, NY, USA, 2008, pp. 10–19.
- 27 [20] Niebles, J. C., H. Wang, and L. Fei-Fei, Unsupervised Learning of Human Action Categories  
28 Using Spatial-Temporal Words. *Int. J. Comput. Vision*, Vol. 79, No. 3, 2008, pp. 299–318.
- 29 [21] Grün, B. and K. Hornik, topicmodels: An R Package for Fitting Topic Models. *Journal of  
30 Statistical Software*, Vol. 40, 2011, pp. 1–30.
- 31 [22] Wainwright, M. and M. Jordan, Graphical Models, Exponential Families, and Variational  
32 Inference. *Foundations and Trends in Machine Learning*, Vol. 1, 2008, pp. 1–305.
- 33 [23] Hoffman, M., D. Blei, and F. Bach, Online learning for latent Dirichlet allocation. In *NIPS*,  
34 2010, pp. 856–864.

- 1 [24] Mimno, D. and A. McCallum, Topic models conditioned on arbitrary features with dirichlet-  
2 multinomial regression. In *Proceedings of Uncertainty in Artificial Intelligence*, 2008, pp.  
3 411–418.
- 4 [25] Blei, D. M. and J. D. Lafferty, Dynamic topic models. In *Proceedings of the 23rd interna-*  
5 *tional conference on Machine learning*, ACM, New York, NY, USA, 2006, ICML '06, pp.  
6 113–120.



**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,  $N_{dep} = 10\,000$ .