Statistical Analysis of a Bike Sharing Transportation System

JUSTE RAIMBAULT[1,2]
[1]Erasmus Mundus Master in Complex Systems Science, Ecole Polytechnique ParisTech
[2]LVMT, Ecole Nationale des Ponts et Chaussées

December 16, 2013

⋆ ⋆
⋆

Erasmus Mundus Master in Complex Systems Science

Fall 2013 Project

Class: *Therapeutic Evaluation and Complex Systems*

⋆ ⋆
⋆

1 Introduction

1.1 Context

Bike Sharing system have recently been the object of common but also scientific interest. The multiplication of implementation across many countries of the world, following the example of European Cities, has shown his potential as a new flexible and ecological transportation system ([1]). In [2] this new transportation mode is presented as being totally complementary to the overall transportation part of an urban system.

It however raises particular issues in the conception and in the exploitation of the system, because of the spatial and temporal heterogeneity in use patterns. Therefore a consequent number of studies have been lead in operationnal research in order to optimise the initial design of the system ([3, 4] for example) or the bike redistribution process which is essential for maintaining a good level of service ([5, 6]).

The understanding of the mechanisms of such a system is essential for a good management of it. Therefore one can use statistical models, such as the work done on bike-sharing system of Lyon, for which statistical cyclic models were used ([7, 8]), including elaborated statistical estimators ([9]) and spatial analysis were done ([10, 11]). An other approach is closer to datamining techniques, such as basic visualisation of many systems ([12]) or application of clustering datamining techniques ([13, 14]). We will place ourselves between the two approaches in the following.

1.2 Presentation of the project

Our aim is to apprehend both generally and specifically a large set of data representing exhaustively the working of a bike sharing system during a given time period, that is Paris' bike-sharing system during approximatively 3 months. That system, the largest in the world, has been studied in [15]. We want to extend this analysis with other methods inspired from ones used on other systems. A part of our work will be aimed at providing a statistical parametrisation for an agent-based model that we won't detail here since that statistical approach will have in itself self-consistence.

2 Data collection

Type of data Data for the statistical analysis are public available data (open data) from the bike-sharing system of Paris ("V'Lib"), provided by the operating company in direct time on a dedicated website (url `http://api.jcdecaux.com`, for which the format of request is precised on `developer.jcdecaux.com`). It provides only the status of docking stations at the request time so we had to automatize the data collection process on a large time period in order to have significant time-series. Process is detailed in the following.

We chose to proceed that way for the data collection first because the obtention of more precise data from the company can raise several problems such as confidentiality issues or more constraining for our research, lead to an lack of independance in the design of the modeling process, since most of the time delivering of data had its price that is at least answering to some question asked by the company. Secondly, we argue that our experience will be one way of testing the possibilities and limits of open data: if the public provided data can lead to relatively good results compared to what can be obtained with a larger set. However, if our research process becomes quicky limitated by the lack of precision or diversity in the data, that will bring one essential question on front, that is that open data does not necessarly means freedom not exhaustivity, and that the control of the provided data can implicitely be highly dangerous for the global opening process. On that point, we follow Banos in [16] when he argues that a necessary condition of an open scientific cumulative process is a total transparency in the methods and an exhausting sharing of implementations of models of simulation and of data. Furthermore we wanted to avoid any risk of implicit reporting spin since it stays a major issue today for the quality of research as it is claimed by Ravaud & *al.* in [17].

The purpose of data sharing by the company in our case was surely, because of the nature of the available data, i. e. only current time stations status, nothing more than current time information and mapping. However, we will see that we can use them for statistical analysis and obtain quite good results.

Data collection process A script requesting current data to the API and saving it into a file have been written and scheduled each 5 minutes on a remote server (we did not choose finer temporal granularity for a material reason, because the size of data becomes quickly huge and storage becomes then an issue). Data on remote server is then zipped everyday for storage purpose. When needed we download the files and process them with R using [18] in order to store them locally on a reduced form (csv) that can be called directly by our data processing algorithms. Note that it would have been more logical to process the data remotely and store them under the reduced form but technical reasons were an obstacle (in particular the installation of R on the remote machine). We also extracted from extensive files static information such as numerotation and coordinates of docking stations, what have been useful after for example to create a geographical file for map drawing with [19]. Fig. 1 shows a flowchart of the data collection and primary processing process. We collected data for all Paris during around 3 month, following statistical analysis are done on these data.

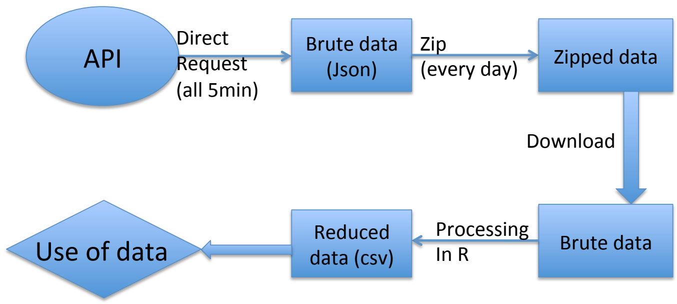


Figure 1: Flowchart of data collection process

3 Statistical analysis

3.1 Data visualisation

Many basic means for a global visualisation of data behavior are available such as the ones proposed in [12], so we won't go too much into detailed representation since it is not the first puspose of our study. Note that this step is however essential, especially during the elaboration of algorithm and the choice of methods for statistical treatment.

To have an idea of the cyclic character of daily mobility patterns, we can plot the total number of available bikes at docking stations against time. If we suppose the total number of bikes constant over the time duration of the plot, what seems reasonnable even on the all time period our data cover (even if there are surely variations because for example of bike reparations, they are surely neglectible regarding the total number of bikes, which is around 15000), this plot is exactly the complementary of the quantity of current travel as a function of time, what allows to visualize mobility tendancies. Fig. 2 shows the obtained curve that fits the expected results, showing in particular the distinction between week days and weekends.

We can also for example draw maps for the understanding of spatial patterns in system use. One can expect for example to see distinction in time between residential and activity areas for the quantity of available bikes in stations. This allows to visualise global and local heterogeneity patterns. Fig. 3 shows an example of such maps on a particular district.

3.2 Extraction of patterns

A first step in the treatment of data is to extract typical patterns in use of the system. In [13], datamining techniques, and especially clustering of activity profiles, are used to extract typical patterns in station use. We propose to use similar methodology in order to identify typical overall day profiles and classify them. We expect to be able to differentiate weekdays from week-ends for example, but also see the influence of climate on use patterns. The clustering of time-series offer an alternative for a predictive model, as the cyclic model proposed in [8, 7].

A day is exhaustively represented by the time-series, defined on all the stations of the system $s \in S$, and on a discrete time sample $T = \{0, \tau, ..., N\tau\}$ (with τ time step of the data, 5min in our case), $(b(s,t))_{s \in S, t \in T}$ of availables bikes at each stations. Each station has a maximal capacity $c(s)$ that allow to define the number

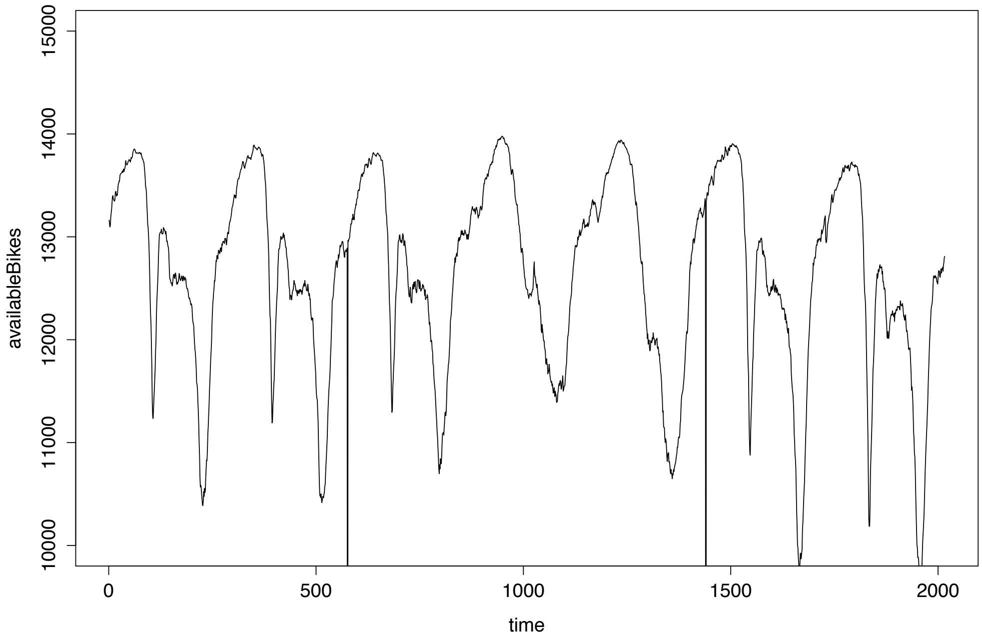
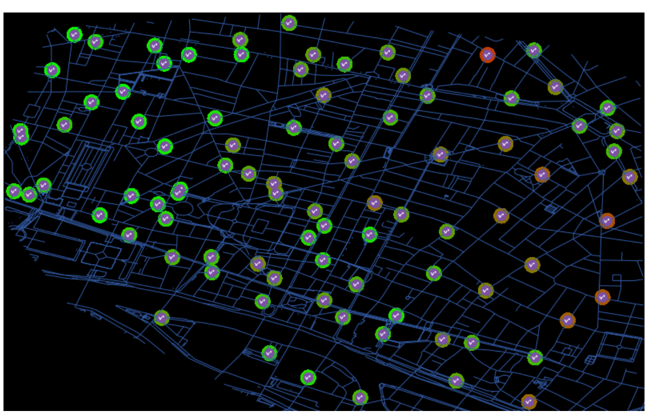
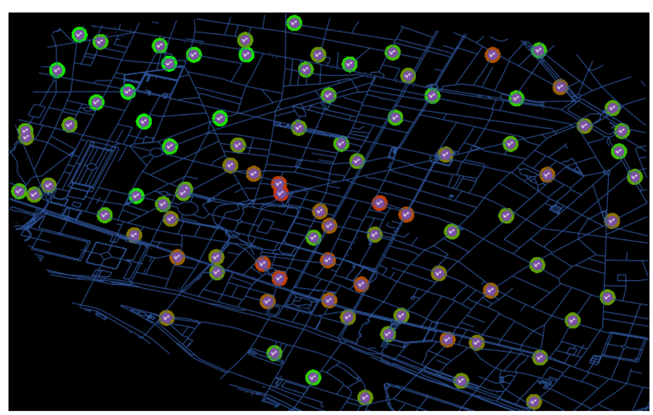


Figure 2: Quantity of total available bikes over a week. We observe the typical patterns of the daily mobility, with two minima corresponding to morning and evening affluence. The two day in the middle correspond to saturday and sunday since the serie begins on a wednesday. These weekend days present only one minimum, what is logical (no affluence in the morning) and confirms the results of other studies.

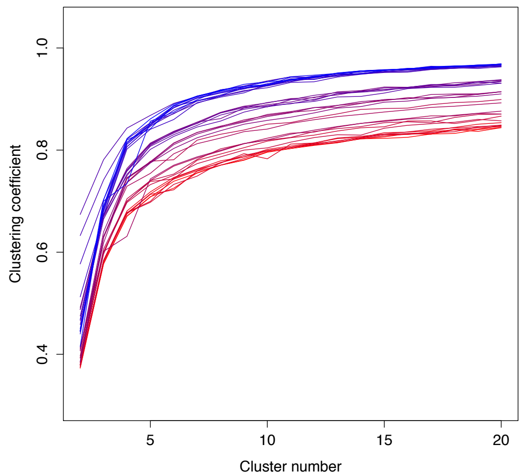


(a) Midnight

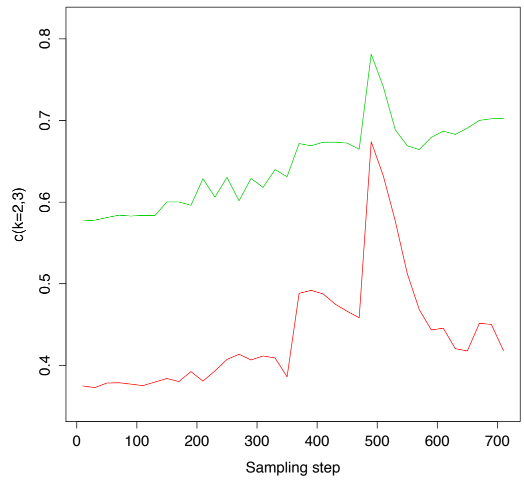


(b) Midday

Figure 3: Examples of heatmaps at two different moments of the day for the district of Chatelet. The color indicates, from green that corresponds to an empty station, to red to a full one, number of available bikes. Since it is a working district and not residential, stations in the center are overloaded during the day but empty during the night as expected.



(a) Clustering coefficient as a function of cluster number for different values of sampling step. The more blue the curve is, the more sampling step is large. If the curve goes quickier to 1, that means that points are less distinct and that statistical distribution contains less information. We observe a jump that is quantified in (b).



(b) Plot of the value of the clustering coefficient for k=2 (red) and k=3 (green), as a function of sampling step. We see the significative loss of information around a step of 400 minutes, which should correspond to the disapearance of pics in the curve, since they contribute significantly to the quantity of information.

Figure 4: Influence of sampling interval on quantity of conserved information in the clustering process.

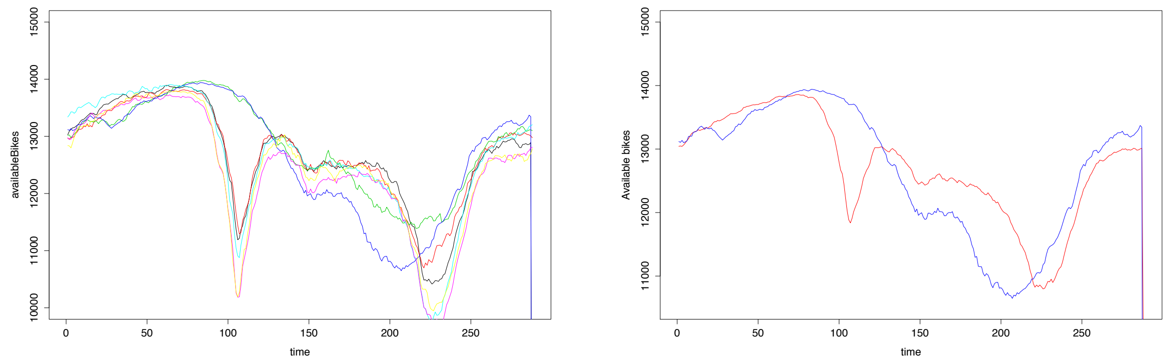
of free parking places $p(s) = c(s) - b(s)$ and the load factor wich can be more convenient to work with since it is normalised $lf(s) = \frac{b(s)}{c(s)}$. The overall clustering process first aims to reduce the dimension of the representation of a day without loosing majority of information, and then to be able then to classify days and make predictions on the day characteristics from its data.

First the dimension is reduced through a sampling process that can be seen as a projection from the space of complete time-series to a space of smaller dimension. If $\varphi \in \mathbb{N}^{\mathbb{N}}$ is a extraction then the sampling is defined as the canonic projection $\mathcal{S} : \mathbb{R}^{|T||S|} \to \mathbb{R}^{|\varphi(T)||S|}$. The question of the value of the time step for sampling is important. We tried for many values and looked at the possible loss of information through the evolution of clustering coefficient regarding number of clusters. It appeared that we had still good precision for large time steps such as one hour. See fig. 4 for more precision on the influence of sampling step.

We proceed then for each day to a k-means algorithm on the sampled time-series (as described in [20]), in order to reduce more the dimension needed to represent a day. Intuitively, that corresponds to a classification of stations according to their "profile". We take in practice 20 clusters, what allows to divide by 100 the dimension. The final step is to cluster the representations of the days for establishing a classification of days. With two clusters, one expect to isolate weekdays from weekends, although kmeans can lead to bad results if cardinal of clusters appear to be imbalanced. In our case it worked quite well and we were able to reproduce that distinction. However, a finer distinction (e. g. between rainy and shiny days) was not possible and some work on a more specialized clustering algorithm (kmeans is very general) would be needed to obtain more precise results. Fig. 5 shows the comparison between real curves of available bikes and predicted curves by the clustering algorithm.

3.3 Inference of origin/destination fields

An other crucial point of the analysis is the estimation of real origin and destinations of users of the system. If the original purpose is in our case to obtain a parametrisation for an agent-based model as we already



(a) Curves of available bikes for all day of the week. Week days are superposed and correspond to the curves with two pics. the green and the blue curve are respectively saturday and sunday.

(b) Theoretical predicted curves for two clusters. As expected, we distinguish week days (red curve) from weekend (blue curve), according to the real curves.

Figure 5: Results of clustering process for classification of days: distinction between weekends and week days.

explained, this problem has its own internal value. Indeed a lot of research in economic geography and transportation geography aims at evaluating real Origin/Destination fields in order to integrate them into transportation/landuse models (see [21] for example).

Our statistical model for the inference of field is a non-parametric estimation with Gaussian kernels (described in [22]). Considering the real departures and arrivals in bike stations (that are easily calculated by discrete differentiation of data), we count each as a contribution to the global field at the current time step, smoothed with Gaussian kernel (that appeared to be enough in practice). At time t ,with a parameter σ fixing kernel sizes (each kernel has the same size, further work could be done to test the influence of multiple sizes, weighted by the maximum of the kernel distribution for example) and a set of effective arrivals $(d_i(t))$ at the corresponding coordinates $(\vec{x}_i(t))$, the spatial field of destinations is estimated as, with K normalisation factor,

$$[D(t)](\vec{x}) = \frac{1}{K}\sum_i d_i(t) \cdot exp(\frac{\|\vec{x}-\vec{x}_i\|}{2\sigma^2})$$

We do the same for the origin field. Kernel estimations are done with the ergonomic package `kernlab` ([23]). These extrapolated fields are then discretised and used as parametrisation for the agent-based models. We also use them to extrapolate bicycle flows on a day. Fig. 6 shows an example of an obtained field.

3.4 Mapping of bicycle flows

One interesting application of the preceeding point is the extrapolation and mapping of flows on a day. Especially, it allows to see if data we have are enough to understand the mechanisms of the system, i. e. infer reasonnably well possible travels, or if more would be needed.

Given mean time-series on a "standard day" (extracted through clustering and mean of the larger cluster), we extract on all the city the time-serie of effective departures and we calculate the origin field. Then for each step of time, we draw randomly the expected number of trips by the following process for each trip:

- Origin $\vec{o}$ is drawn from the origin probability field
- Trip distance r is drawn following a Gaussian law of mean 2km and standard deviation 800m according to the distribution of travel distances proposed in [15]

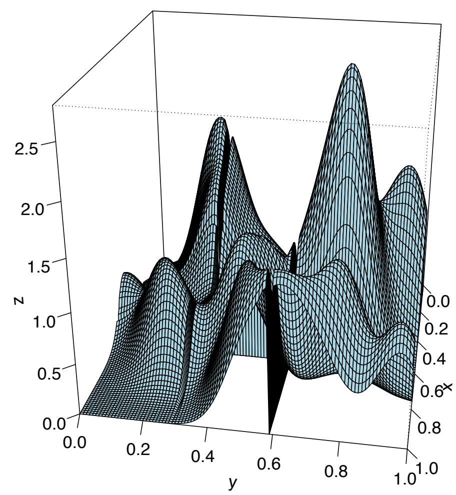


Figure 6: Destination field obtained through Gaussian kernel estimation on a square area.

- We define the conditional destination field by $D_r(t) = K \cdot D(t) \cdot \mathbb{1}_{\|\vec{o}-\vec{x}\|=r\pm\varepsilon}$, where ε is a tolerance parameter and K a normalisation factor and draw a random destination according to this field.
- Following the shortest path between origin and destination in the network, we increase by one the cumulated flow for each link.

That gives the map of the cumulated flows on a all day. Implementation of GIS calculation and network paths was done in NetLogo ([24]) in the frame of the agent-based modeling. Fig. 7 shows the obtained map. We observe expected patterns in the center but irregularities for some links, especially in south-east. However, we claim having a good interpolation of the internal mechanisms of the system from only partial data, answering partially to the question raised in 2.

Conclusion

We were able to proceed to statistical analysis of large data for a better understanding of the working of a bike-sharing system. Although reults were sometimes quite limited, we showed that it was possible to create knowledge from Open Data that were not aimed at that, what is quite significant from an epistemological point of view. We argue that it is a proof of the need of such data opening process, in all fields, but also in a more extended way.

Further development of that project could be a stronger assessment and exploration of used methods, and also the insertion of methods from other fields such as statistical medicine: considering redistribution events as “treatment” on stations, and evaluating adverse events as load-factor over a threshold for example, it would be able to consider a district as a trial (treatment group: ay with redistribution, control group without) and proceed to a meta-analysis on all district, in order to quantify the effect of redistribution on the system. Such transposition may however remain discutable and need to be considered deeper.

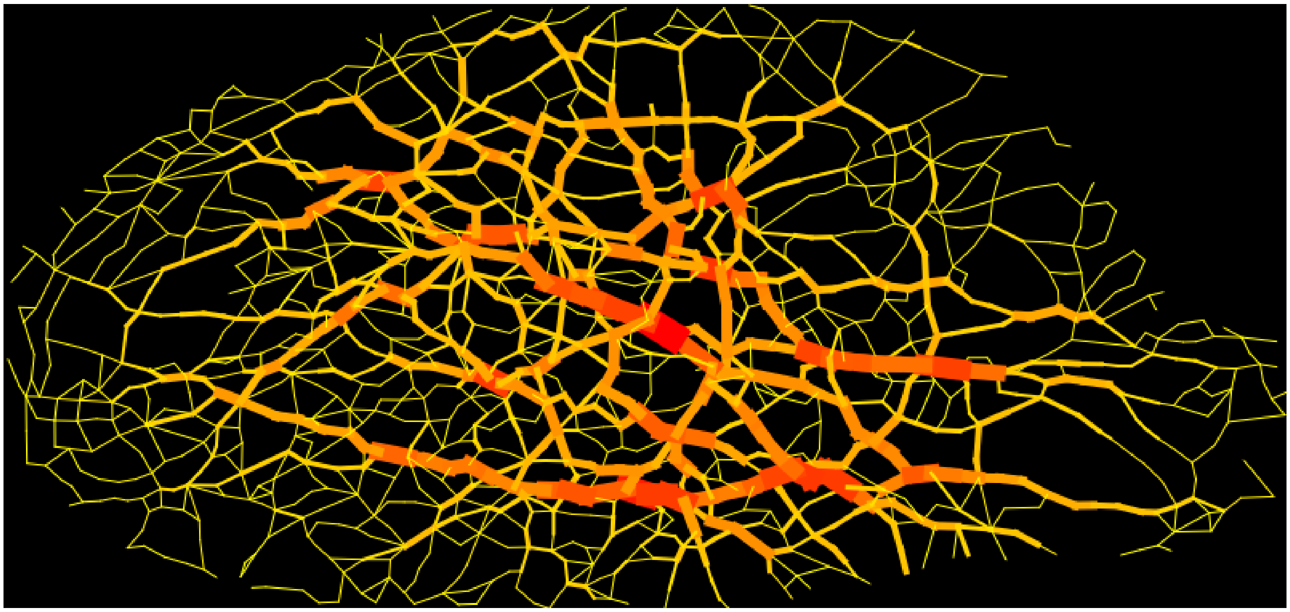


Figure 7: Map of extrapolated cumulated flows on a day for Paris.

Supplementary Material

Source code of all statistical processing in R, data collection script, and agent-based model in NetLogo (for the mapping part) in attached zipfile.

References

[1] Paul DeMaio. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):41–56, 2009.

[2] Peter Midgley. The role of smart bike-sharing systems in urban mobility. *JOURNEYS*, 2:23–31, 2009.

[3] Jenn-Rong Lin, Ta-Hui Yang, and Yu-Chung Chang. A hub location inventory model for bicycle sharing system design: Formulation and solution. *Computers & Industrial Engineering*, 2011.

[4] Jenn-Rong Lin and Ta-Hui Yang. Strategic design of public bicycle sharing systems with service level constraints. *Transportation research part E: logistics and transportation review*, 47(2):284–294, 2011.

[5] Rahul Nair and Elise Miller-Hooks. Fleet management for vehicle sharing operations. *Transportation Science*, 45(4):524–540, 2011.

[6] Alvina GH Kek, Ruey Long Cheu, and Miaw Ling Chor. Relocation simulation model for multiple-station shared-use vehicle systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1986(1):81–88, 2006.

[7] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Jean-Baptiste Rouquier, et al. Studying lyon's vélo'v: a statistical cyclic model. In *European Conference on Complex Systems 2009*, 2009.

[8] Pierre Borgnat, Patrice Abry, and Patrick Flandrin. Modélisation statistique cyclique des locations de vélo'v à lyon. In *XXIIe colloque GRETSI (traitement du signal et des images), Dijon (FRA), 8-11 septembre 2009*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2009.

[9] Gabriel Michau, Céline Robardet, Luc Merchez, Pablo Jensen, Patrice Abry, Patrick Flandrin, and Pierre Borgnat. Peut-on attraper les utilisateurs de vélo'v au lasso. In *Proceedings of the 23e Colloque sur le Traitement du Signal et des Images. GRETSI-201*, pages 46–50, 2011.

[10] Pierre Borgnat, Eric Fleury, Céline Robardet, Antoine Scherrer, et al. Spatial analysis of dynamic movements of vélo'v, lyon's shared bicycle program. In *European Conference on Complex Systems 2009*, 2009.

[11] Pablo Jensen, Jean-Baptiste Rouquier, Nicolas Ovtracht, and Céline Robardet. Characterizing the speed and paths of shared bicycle use in lyon. *Transportation research part D: transport and environment*, 15(8):522–524, 2010.

[12] Oliver O'Brien, James Cheshire, and Michael Batty. Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 2013.

[13] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523, 2011.

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

[15] Rahul Nair, Elise Miller-Hooks, Robert C Hampshire, and Ana Bušić. Large-scale vehicle sharing systems: Analysis of vélib'. *International Journal of Sustainable Transportation*, 7(1):85–106, 2013.

[16] Arnaud Banos. *Pour des pratiques de modélisation et de simulation libérées en Géographie et SHS*. PhD thesis, UMR CNRS Géographie-Cités, ISCPIF, Décembre 2013.

[17] Isabelle Boutron, Susan Dutton, Philippe Ravaud, and Douglas G Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA: the journal of the American Medical Association*, 303(20):2058–2064, 2010.

[18] Alex Couture-Beil. rjson: Json for r. *R package version 0.2*, 13, 2013.

[19] Timothy H Keitt, Roger Bivand, Edzer Pebesma, and Barry Rowlingson. rgdal: bindings for the geospatial data abstraction library. *R package version 0.7-1, URL http://CRAN. R-project. org/package= rgdal*, 2011.

[20] T Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[21] Fabien Leurent. *Modélisation du trafic, des déplacements sur un réseau et de l'accessibilité aux activités grâce au transport*. PhD thesis, Université Paris Dauphine-Paris IX, 2006.

[22] Alexandre B Tsybakov. Introduction to nonparametric estimation. (introduction à l'estimation non-paramétrique.). 2004.

[23] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab-an s4 package for kernel methods in r. 2004.

[24] U. Wilensky. Netlogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL., 1999.