# Shared Bicycles in a City: A Signal Processing and Data Analysis Perspective

Pierre Borgnat, Céline Robardet, Jean-Baptiste Rouquier, Patrice Abry, Patrick Flandrin, Eric Fleury

Advances in Complex Systems


# SHARED BICYCLES IN A CITY: A SIGNAL PROCESSING AND DATA ANALYSIS PERSPECTIVE

Pierre BORGNAT, Céline ROBARDET, Jean-Baptiste ROUQUIER,
Patrice ABRY, Eric FLEURY, and Patrick FLANDRIN

*P. Borgnat, P. Abry and P. Flandrin are with CNRS, Laboratoire de Physique (UMR 5672 CNRS) of École Normale Supérieure de Lyon; 46 allée d'Italie, 69364 Lyon Cedex 07 France. E-mail: {pborgnat;pabry;flandrin}@ens-lyon.fr*

*C. Robardet is with Université de Lyon, INSA-Lyon, CNRS, LIRIS (UMR CNRS 5205); Batiment Blaise Pascal 69621 Villeurbanne cedex, France. Email: celine.robardet@insa-lyon.fr*

*J.-B. Rouquier and E. Fleury are with LIP (UMR CNRS INRIA 5668) and IXXI (Institut des Systèmes Complexes) of École Normale Supérieure de Lyon; 46 allée d'Italie, 69364 Lyon Cedex 07 France. E. Fleury is member of INRIA/D-NET; Email: jrouquie@gmail.com and Eric.Fleury@inria.fr.*

Received (received date)
Revised (revised date)

Community shared bicycle systems, such as the Vélo'v program launched in Lyon in May 2005, are public transportation programs that can be studied as a complex system composed of interconnected stations that exchange bicycles. They generate digital footprints that reveal the activity in the city over time and space, making possible a quantitative analysis of movements using bicycles in the city. A careful study relying on nonstationary statistical modeling and data mining allows us to first model the time evolution of the dynamics of movements with Vélo'v, that is mostly cyclostationary over the week with nonstationary evolutions over larger time-scales, and second to disentangle the spatial patterns to understand and visualize the flows of Vélo'v bicycles in the city. This study gives insights on the social behaviors of the users of this intermodal transportation system, the objective being to help in designing and planning policy in urban transportation.

*Keywords*: Community bicycle sharing program; Vélo'v; Cyclostationarity; Nonstationarity; Dynamic network; Network community

## 1. Introduction

Community shared bicycle programs have been under development in the recent years all over Europe, as an answer to an increasing need of green and versatile public transportation in cities. Lyon's shared bicycle program, called Vélo'v and operated by the JCDecaux agency [1], is a major one of its kind, having started in May 2005. Besides their evident interest as a new means to think about public

transportation, such community shared programs offer a new way to look into the dynamics of movements inside a city, and more generally into its activity. In a sense, the Vélo'v system provides digital footprints that reveal the activity of people in the city over time and space, and makes possible their analysis.

Different issues motivate the study of such a system. Some questions are about the usage patterns of this kind of transport, with reference to social or economical studies of transportation, while others are about the system itself: does the service work correctly? Can it be optimized? Can one regulate the availability of bicycles? An objective in this paper is to make first steps in such directions by proposing relevant tools for the study of the space and time patterns of activity from all the trips made with Vélo'v, going from an empirical point of view that can be compared to previous studies of equivalent systems in Paris (the *Vélib'* program studied in [2]) or in Barcelona (*Bicing*; studied in [3, 4]), to a more quantitative point of view on the activity of the stations, and their properties.

A contribution of the paper is to use of methods from signal processing and data analysis to study the Vélo'v system, so as to exhibit some features of the system and to begin to answer some economical questions linked to such community system. Many social questions can be addressed using this dataset, and some specific ones are chosen in this study. How many trips are made using the rented bicycles, and is there an evolution in time of the use of the system? Is it then possible to forecast the use of the bicycles, as a help toward better regulation of the service. We will turn to statistical signal processing to address these questions. A second set of questions pertains to the spatial distribution of the system. The service is deployed in the whole city which is not uniform. The objective here is to learn, from the moves of rented bicycles, what is the dynamics of movements in the city at various hours of the day: Where do people go? What are the main flows between different parts of the city? As the dataset is large, data mining methods are needed to work on this topic. Finally, if compared to what social surveys and enquiries provide, the use of digital footprint to study the movements of bicycles gives new insights on properties of trips with bicycles in a city (length of trips, frequency of use, influence of external factors such as weather,...). On this aspect, this work shares a perspective similar to the one in [5], using digital footprints of a given means of urban transportation, first to understand how this method of transportation is used, and more globally to reveal some features of the moves in a city.

The paper is organized as follows. In Section 2, a general presentation of the Vélo'v program is given, highlighting its key features. Section 3 is concerned with a description of the data, in both time and space, that can be accessed for studying the system. Section 4 is then devoted more specifically to the global activity in time for which a predictive model is developed using signal processing tools, whereas Section 5 is concerned with spatial patterns of activity, with results in terms of clustering and communities obtained using data mining methods.

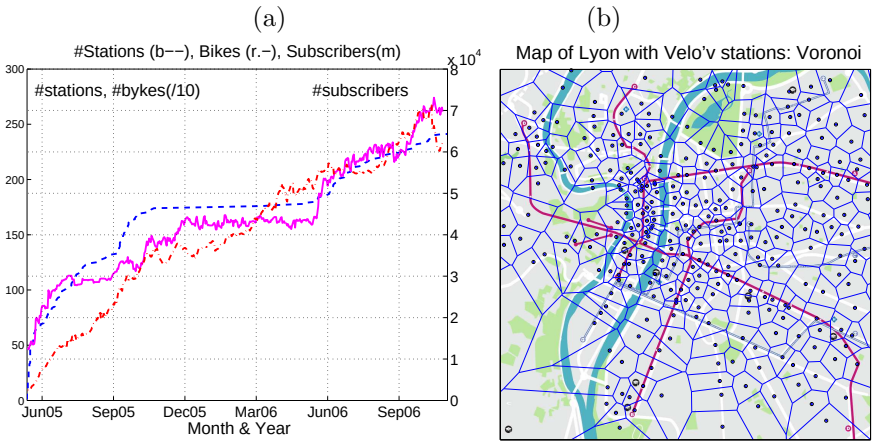


Fig. 1. **General features of the Vélo'v system.** (a) Time evolutions of the numbers of stations (dashed line, in blue), of available bicycles $N_v$ (dot-dashed line, in red), and of year-long subscribers $N_s$ (solid line, in magenta). (b) Map of Lyon with Vélo'v stations (dots), their Voronoi diagram (blue lines), subway lines (thick red lines), rivers (in blue), and parks (in green).

## 2. Vélo'v: A community bicycle system

The Vélo'v program is deployed in the city of Lyon[a], in France, since May 2005. It now consists of 4000 bicycles (also called Vélo'v) that can be hired at any of the 340 stations, spread all over the two cities and returned back later at any other station. In contrast to old-fashioned rental systems, the rental operations are fully automated: the stations are in the street and can be accessed at anytime (24h a day, 7 days a week), and the rentals are made through a digital terminal at the station using a credit card to obtain a short-term registration card, or using a year-long subscription system. First, this makes possible the collection of the complete data of rentals, and so of movements made with Vélo'v—a dataset not readily available for other means of transportation. Second, a global and fine management of the program can be envisioned since a real-time survey of the system is done. Currently, automated station reports are collected into a central database and mostly used a posteriori, if one excepts online reports about the availability of bicycle or free stand to return one at stations [6]. Yet, there is a strong incentive to evolve toward less empirical management of the system, for instance by being able to increase or redeploy in real-time the available bicycles to answer the demand.

Anonymized data from May 2005 to the end of 2007 were made available to us by JCDecaux and the "Grand Lyon" City Hall. The dataset consists of the records of

[a]Most of the stations are in downtown Lyon, in the southern and northern campuses of Lyon and in the town of Villeurbanne in the North, all part of the "Grand Lyon" Urban Community. The rest of the article uses simply the name "Lyon" to name the area of deployment of the program, and Grand Lyon City Hall to name the administrative service of the "Grand Lyon" Urban Community.

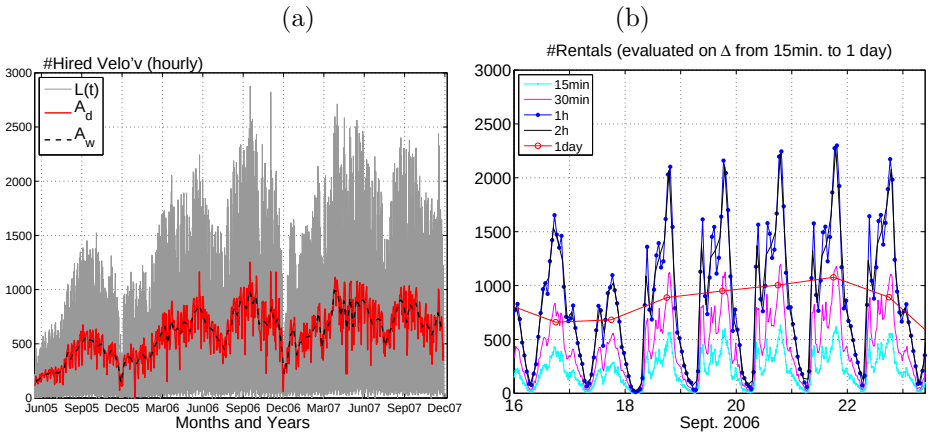


Fig. 2. **Time evolution of the number of hirings.** (a) Number of bicycles hired per hour $L(t)$, and its average per day $A_{\mathrm{d}}$ and per week $A_{\mathrm{w}}$. (b) Image of a typical week (16 Sept. 2007 is a Saturday), at different aggregation times $\Delta$ (the different $\Delta$ are given in the legend inside the graph); for the clarity, when data is aggregated at 2 hours and one day, we divided the amplitude to renormalize it as a number of rental per hour (yet estimated on aggregation over $\Delta$).

all bicycle trips, over more than two years of exploitation. During this period, there were more than 13 million bicycle trips. Each trip is documented with its starting time and station location, its ending time and destination station, the duration and length of the travel (as recorded on the bicycle), and specific tags when the movement is not a rental but a maintenance operation (first deployment of a Vélo'v bicycle, or movement to a repair workshop).

An important characteristic is that this bicycle program was expanded while already open. The Vélo'v system opened on May 19th, 2005 and stations and bicycles have been introduced continuously during the take off and lifetime of the system (no more stations are currently added, but this phase is not in the studied dataset). Fig. 1 (a) depicts the capacity of the system (station and bicycles being open/equipped regularly between May 2005 and October 2005). After this period, deployment reaches a plateau (October 2005 to May 2006) before a new phase of expansion that ends in January 2008 where the final number of installed stations was reached (340 stations). It relates to the increase along time of the number of year-long subscribed users (displayed also in Fig. 1 (a)). Note that bicycles can also be used without subscription, with short-term registration cards bought on spot.

Before turning to a more detailed analysis of the data, let us comment on a spatial property of the system. Fig 1 (b) displays a map of Lyon, showing the current deployment of the Vélo'v stations in the city, and a Voronoi diagram [7] around the stations. It gives an idea of the variation of the density, higher near city center and major axis of transportations, yet putting almost any point of downtown no further than 500 m from a station. However, the stations differ both in neighborhood

and number of stands, so that some inhomogeneity is expected in their use. Vélo'v movements can then be seen as a dynamic process over the transportation network that connects all stations. An analysis of the flows of bicycles on this network will be useful to find spatial patterns of the Vélo'v activity.

## 3. Descriptive statistics of Vélo'v data

Let us first derive basic facts on the Vélo'v using empirical features from the data.

### 3.1. *Temporal Patterns*

As depicted in Fig. 1 (a), the increase in the number of available bicycles and stations parallels the increase of the number of subscribers. The progressive deployment and the increase in popularity of the program generate a nonstationary behavior of the whole system. Fig. 2 (a) shows the number of rentals per hour, aggregated by hours, days and weeks, for the whole network. A main characteristic is the nonstationary evolution of the use of Vélo'v (its increase), combined with a cyclostationary pattern over the week. This will be studied and modeled in Section 4.

A first question when one is confronted to data based on a large number of individual events is to choose a proper scale of representation in time (a question reminiscent of studies on Internet packets [10]). Let us call $\Delta$ the time scale over which to aggregate the number of new rentals. The trade-off is usual: the smaller $\Delta$ is, the larger the fluctuations are, whereas a larger $\Delta$ may smooth the signal with the risk of losing relevant temporal features. Fig. 3 (a) displays the distribution of rental durations, and in (b) the same histogram is given in log-log axis. This distribution of durations is large, yet there is a mode at 9 min and the median equal to 11 min is representative of its core. Let us note in Fig. 3 (b) that, for duration between 26 and 34 min (the 2 dashed lines), a subtle drop is seen, reflecting the fact that the first 30 minutes are free and the bicycles beep after 25 minutes of use.

We varied $\Delta$, typically from 15 minutes to 2 hours, so as to remain within the scales that are sufficient to smooth out the effect of individual rentals, while keeping the global evolutions of their collection, most importantly the one over the day. As an example, Fig. 2 (b) shows, for a typical week, the number of rentals made aggregated on a time scale of 15 min, 30 min, 1 h, 2 h and one day. The aggregation at 1 hour gives a good trade-off between resolution of details and fluctuations. On this specific week for instance, one sees clearly a repetition of modes each working day. Using smaller $\Delta$, it is less clear due to fluctuations. For $\Delta = 2$h, it is smoothed out (especially the peak around noon). The aggregation scale will thus be 1h.

### 3.2. *Spatial Patterns*

In Fig. 4, spatial patterns of the traffic at each station are displayed: For a given hour, the amount of incoming and outgoing traffic is proportional to the area of the semi-circles at each station, incoming traffic on bottom, outgoing one on top. Then

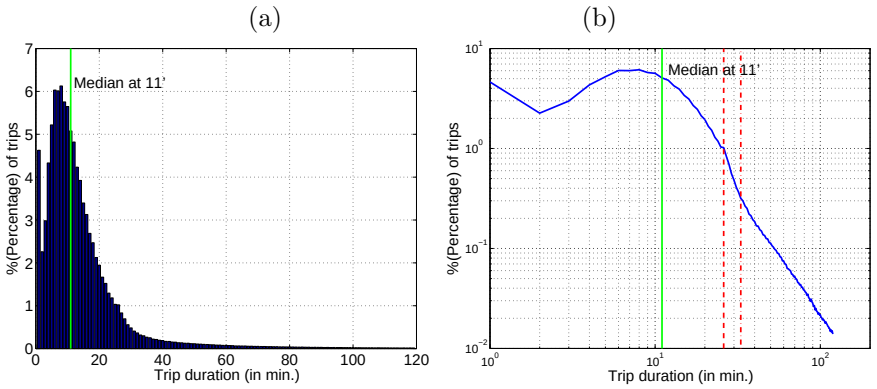


Fig. 3. **Temporal features of Vélo'v.** (a) Rental duration distribution (in %). (b) The same distribution in log-log axis (dashed red lines point on the interval [26, 34] min.)

the average of the directions of incoming trips at each station is represented with a light (green) vector whose direction and length represents the anisotropy of the set of trips arriving at this station. Let $\Omega_{\text{in}}(m) = \{\text{trips into station } m\}$; the complex representation of this vector is computed as the average $\sum_{k\in\Omega_{\text{in}}(m)} e^{i\theta_k}/|\Omega_{in}(m)|$, where $\theta_k$ is the angle coordinate in the plane of the origin-destination vector (destination being station $m$). Dark (blue) arrows represent the same average direction computed for leaving bicycles, with $\Omega_{\text{out}}(n) = \{\text{trips from station } n\}$. Zooms on specific parts of the city are shown in Figs. 4 (2), (3) and (4).

Let us now underline the main trends among the use of bicycles. The first comment is the non-uniformity of use of the stations: the order of magnitude of the number of trips at less frequented station is very low as compared to the most frequented stations in the center of the city (less than 1/100 of their use). Zones A and C in Fig. 4 (1) and in zooms (2) and (3) correspond to university campuses. On Monday 8 am, these stations receive many bicycles whereas on Tuesday 4pm-5pm (see maps (2) and (3)), there are more leaving trips than incoming ones (and this usually lets the stations be in deficit of Vélo'v for the evening). In Fig. 4 (1), zone B corresponds to stations that are on the top of a hill (Croix-Rousse) and mostly have leaving trips (at all hours of the day). All these zones illustrate the unbalanced character of many stations. Related to that, many stations show an anisotropic activity: stations around the center of the city have usually incoming trips coming from the center and leaving ones going to the center (hence the appearance of a field of vector pointing toward the center of the city in Fig.4 (1)). In Fig.4 (4), mostly the center of the city is displayed: Zones D and F correspond to railway stations, and zone E is an active area with both shops and residential parts. All these three zones serve also as connection hubs with major subways and buses. These zones experience a rush of activity at almost anytime during the day. For instance, many people seem to return or take a Vélo'v near one of the train stations on Thursday 4pm-5pm, validating the idea that Vélo'v are used as one part of an intermodal transportation

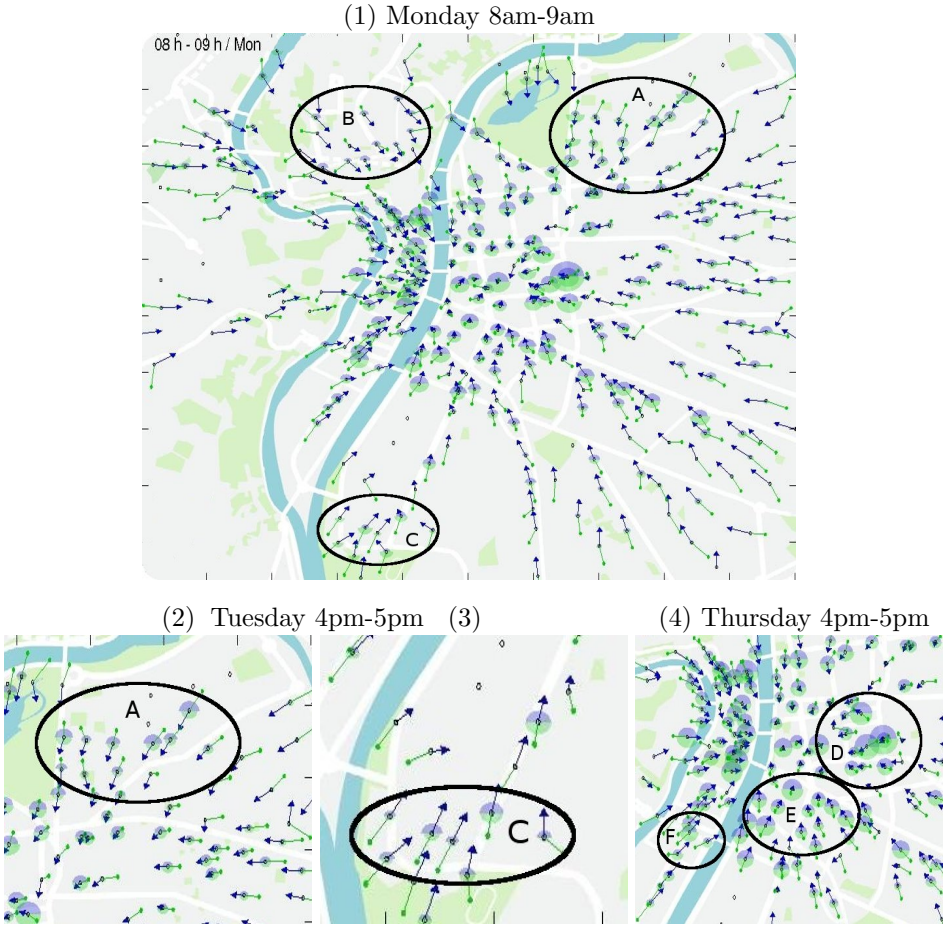


Fig. 4. **Visualization of the traffic at all stations.** For a given hour, the amount of incoming and outgoing traffic is proportional to the area of the semi-circles at each station, incoming in light grey (green) on bottom; outgoing in dark grey (blue) on top. The arrow gives the average direction (as defined in Sec. 3.2) of these trips: incoming in light green; outgoing in dark blue arrow.

system (with trains, buses or subways). These simple diagrams based on temporal patterns visualized at each stations allow us to differentiate their behaviors. Some stations (zones D and F in Fig. 4 (3)) act like hubs for Vélo'v. At several other stations, mostly one-way flows (reversing direction depending on the time of the day) are found, that leave the stations unbalanced during the day. This indicates a use of Vélo'v by people nearby the stations, using it to commute to or from works.

### 3.3. *Individual characteristics of trips*

Before aggregating the trips in space and/or time, studies can also be conducted on individual trip level. Basic features are displayed here. Fig. 5 (a) and (b) reports

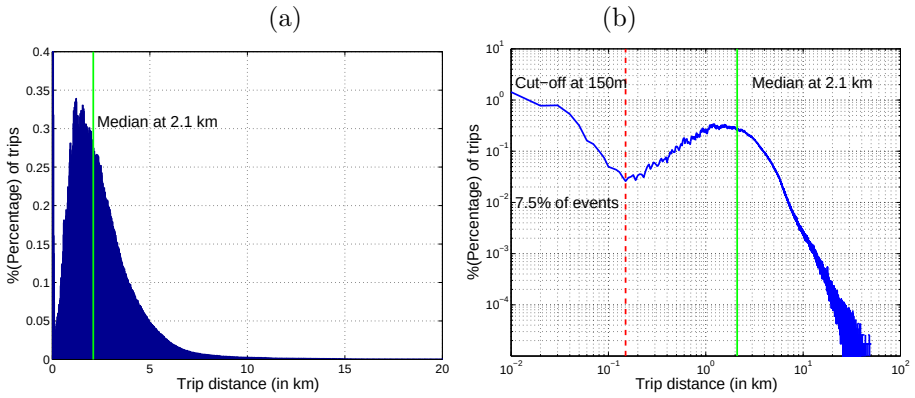


Fig. 5. **Individual characteristics of trips with Vélo'v.** (a) Distribution of lengths of each trip; (b) the same distribution in log-log axis.

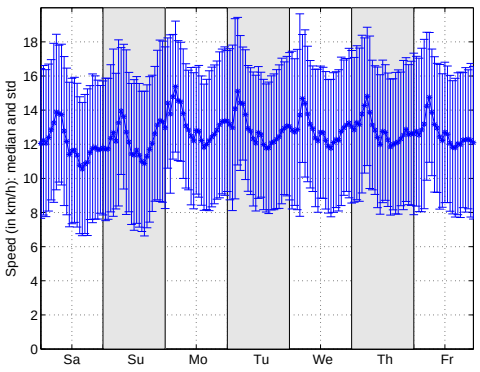


Fig. 6. **Median speed of individual trips with Vélo'v**, as a function of the day of week and hour. The standard deviation is represented around the median.

the distribution of lengths of each journey. Like the duration distribution, 3 parts can be distinguished: a sharp peak near 0 and up to 150m, which amounts to 7.5% of the traffic, that is associated to rentals of bicycles coming back to the departure point or that are out of order due to mechanical reasons; a mode with median near 2.1 km corresponding to normal use; a long tail up to more than 20 km. The tail accounts for just a small fraction of the rentals (around 5%), yet it exists and, if one would like to use some agent-based modeling, at least 3 different classes of rentals should be made. In the present study, aggregated analysis is favored rather than agent-based ones, especially because (due to privacy issue) we also lack any identification of users or bicycles.

A complementary characteristics, in Fig. 6, is the median velocity of the user

(computed from the data reported by the Vélo'v bicycles), averaged over all the trips that begun at the same time (with an aggregation scale of $\Delta = 1$h) during the week. Here again, there is a signature of the natural cyclostationarity of the week, people moving faster in the morning than later in the day, or faster during week-days than during week-ends (this has been studied in more details in [9].) Also, an interesting point is that the average velocity is between 12 and 14 km/h. As a comparison, the mean velocity in cities is 18 km/h for buses, 25 km/h for (regular) subways and only 17 km/h in the center for cars [8]. This proves that bicycles are actually a competitive means of transportation as compared specifically to cars.

All the properties discussed so far are interesting in that they would not be easily obtained using classical social surveys (usually with population sampling); the digital nature of the information is here precious. It provides a full characterization of the trips made with rented bicycles, and this is an important asset to models transportation and moves in a city.

## 4. Time dynamics

This Section deals with a statistical study of the time series of the number of bicycles hired along time, expanding upon first results reported in [11–13]. The goal is not only to identify its temporal patterns but, going way further in the modeling than previous studies such as [3], to propose a statistical model for the series, encompassing their cyclostationarity and their nonstationarity. Then, this model is used to predict the number of bicycle rentals on a daily or hourly basis.

The raw data here is the number $L(t)$ of hired Vélo'v between $t$ and $t+\Delta$ (thus, aggregated over the time scale $\Delta$). In the following, time instants will therefore be discrete and understood as integer multiples of the aggregation scale, i.e., of the form $t = k\Delta$ with $k \in \mathbb{N}$. As seen in Fig. 2, two features are dominant: The mean is nonstationary and evolves with time, and there is a periodic repetition over the week. The first feature is related to the increase in size and popularity of the program (commented above); a complementary reason is that the use of Vélo'v also depends on the season (with less users during winter, or during holidays). The second feature of cyclic evolution over the week, more properly referred to as cyclostationarity, comes from the obvious fact that from a social point of view, days and hours are not equivalent for people. Those two features, nonstationarity and cyclostationarity, are precisely the ones that the model proposed in this Section aims at accounting for.

### 4.1. *Model for the cyclic temporal patterns*

Let us first study nonstationary patterns on time scales larger than the day. An estimation is obtained by computing, from the rentals $L(t)$ aggregated on $\Delta$, the number of rentals $A_{\mathrm{d}}(d)$ at a given day ($d$ is the variable of day):

$$A_{\mathrm{d}}(d) = \sum_{t \in (d)} L(t). \tag{1}$$

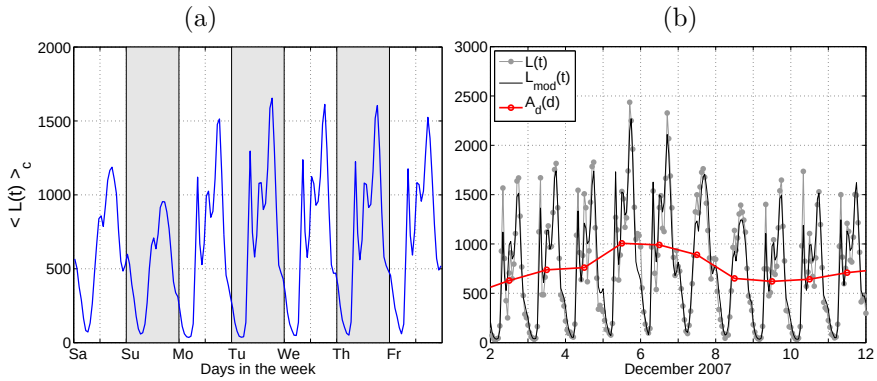


Fig. 7. **Cyclic models and comparison to data.** (a) Model $\langle L(t)\rangle_c$ giving the typical expected evolution over the week. (b) Examples of $L(t)$ for some chosen days, compared to the model $L_{\mathrm{mod}}(t)+\widehat{F(t)}$. Here, we choose to zoom on the days around the 8th of December, a Lyon festivity (which was a Saturday in 2007), showing qualitatively that the model holds well.

Then, inspired from cyclostationary methodologies [14, 15], an estimate of the cyclic mean for $L$ over the week is the periodic average:

$$\langle L(t)\rangle_c = \frac{1}{N_w}\sum_{k=0}^{N_w-1} L(t+k\,w_\Delta), \tag{2}$$

for $t$ expressed in multiples of $\Delta$, from 0 to one week, and $w_\Delta$ is the duration of the week in unit of $\Delta$ ($w_\Delta = 167$h if $\Delta = 1$h); $N_w$ is the number of weeks of data used. This equation describes the periodic average of the data, evaluated over a period of one week. The result is displayed in Fig. 7 (a). It shares similarities with observations made on the Barcelona program [3, 4]. During week-days, three peaks are seen: in the morning (8am-9am), at noon (12am-1pm) and by the end of afternoon (5pm-7pm, this one being the highest and broadest). During week-ends, the pattern changes, with mostly a large peak spread during the afternoon, having a maximum around 5pm (with only a small increase on its top at noon). These features match intuitive interpretations about the fact that people use bicycle transportations mostly during the day to commute, or during lunch break, whereas during the week-end, the major trend is to take an afternoon pleasure ride or go to recreational area in the city.

Let us write $A_{\mathrm{mod}}(d_7) = \sum_{t\in(d_7)}\langle L(t)\rangle_c$ the average number of rentals per day $d_7$, where $d_7$ simply marks the day of week, from Monday to Sunday. Mathematically, $d_7$ is equal to $d$ (the variable of day) modulo 7 (hence the choice of notation). As a quantitative approach of the time activity, the model is the following:

$$L(t) = L_{\mathrm{mod}}(t) + F(t) = A_{\mathrm{d}}(d)\,\frac{\langle L(t)\rangle_c}{A_{\mathrm{mod}}(d_7)} + F(t), \tag{3}$$

where $F(t)$ is the part of the data not accounted by the cyclic model. In Fig. 7 (b), we illustrate the model for a specific range of days, to show that it usually

holds well, even when specific occasions change the flow of days, such as holidays or festivities (here we illustrate that on the 8th of December, which is a specific festivity day in Lyon). Quantitatively, when using the value of $A_{\rm d}$ estimated from the data, the error (in variance) for the model is 16% (i.e, 130 bicycles per hour). For an operational use, prediction of $A_{\rm d}$ and $F$ is necessary; this the purpose of the next paragraph.

### 4.2. *Forecasting of the number of rentals, and anomalies*

Let us now turn to the prediction of the evolution of the hourly number of rented bicycles, taking into account factors that are external to the cyclic pattern. Using the model, Eq. (3), prediction is split into two subparts: First, the prediction of the non-stationary amplitude $A_{\rm d}(d)$ for a given day; Second, the prediction of the fluctuations $F(t)$ at a specific hour. The corresponding time scales being different, it is appropriate to predict them separately.

**Prediction of $A_{\rm d}(d)$** It seems fair to look for factors explaining $A_{\rm d}(d)$ among the following ones:

(i) the weather and seasons summarized by the average temperature $T(d)$ over one day (in $^{\rm o}$C and centered according to $\delta T(d) = T(d) - \langle T(d) \rangle$) and the volume of rain $R(d)$ (in mm) during day $d$ (for which the reference value is 0); we used weather data collected at the weather station of Lyon;
(ii) the development and popularity of the program: The number of subscribed users $N_s(d)$, the number of bicycles available $N_v(d)$; here again, we take deviations $\delta N_s(d)$ and $\delta N_v(d)$ between the real value and the value at the end of the data (December 2007) where the system is supposed to have reached its final state;
(iii) specific conditions such as holidays, with a marker $J_h(d)$ taking value 0 usually and 1 for those specific days, or strikes with marker $J_s(d)$.

A linear regression model is written as:

$$\begin{aligned}\widehat{A_{\rm d}(d)} = \alpha_0(d_7) + \alpha_1 \delta N_s(d) + \alpha_2 \delta N_v(d) + \alpha_3 \delta T(d) \\ + \alpha_4 R(d) + \alpha_5 J_h(d) + \alpha_6 J_s(d),\end{aligned} \tag{4}$$

where features $\delta N_s$, $\delta N_v(d)$, $\delta T(d)$ and $R$ have been normalized to variance 1, and where the term $\alpha_0(d_7)$ describes the mean of the number of rentals. Because, as seen in Fig. 7 (a), the expected number of rentals each day varies from Monday to Sunday, the term $\alpha_0(d_7)$ has to depend on position of day $d_7$ during the week. The rentals are, for instance, less numerous during the week-ends. A term linear with $A_{\rm mod}(d_7)$ is thus added in $\alpha_0(d_7)$, with a coefficient $c_1$, to describe this dependence:

$$\alpha_0(d_7) = A_0 + c_1 \left( A_{\rm mod}(d_7) - \langle A_{\rm mod}(d_7) \rangle_{d_7} \right). \tag{5}$$

The constant $A_0$ is finally the constant in the linear regression. Solving this problem of linear regression using standard least square minimization, we obtain the results

Table 1. **Statistical model for $A_{\mathrm{d}}(d)$ as per eq. (4).** For the different linear coefficients associated to the factors in play, we report the estimated value (est.) and its Confidence Interval (under Gaussian assumption), given by [CI$_-$, CI$_+$].

| Variable | $\delta N_s(d)$ | $\delta N_v(d)$ | $\delta T(d)$ | $R(d)$ | $J_h(d)$ | $J_s(d)$ |
|---|---|---|---|---|---|---|
| Unit | Subscr. | Bicycles | °C | mm | | |
| ref. | 62 250 | 3 000 | 13.0 | | | |
| std. | 8 030 | 400 | 7.7 | 0.37 | | |
| coeff. | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ |
| est. | 1 860 | -120 | 2270 | -1280 | -2900 | 20 |
| CI$_-$ | 1 210 | -720 | 1980 | -1520 | -3700 | -2900 |
| CI$_+$ | 2 560 | +490 | 2560 | -1030 | -2100 | +2900 |

reported in Table 1. Confidence intervals are reported along with the estimated values of the coefficients because, even though computed under Gaussian hypothesis, which does not hold for many factors, it assists us in the interpretation of the relevance and importance of each factor. Note that errors are found to be sub-Gaussian, i.e., the distribution is sharper and more concentrated toward 0 than the Gaussian one with the same variance. The confidence intervals are thus over-estimated. Results call for the following comments.

(1) The term depending on the day $\alpha_0(d_7)$ is simple enough: it consists of a constant $A_0$ whose value is close to the average number of hired bicycles per day during the last months in the data set (17 500 during the last 4 months of 2007), with a linear correction (with factor close to 1) that takes into account the dependence with the day of week.
(2) A larger number of subscribers increases $A_{\mathrm{d}}(d)$;
(3) Weather factors act in an expected manner: the warmer, the larger the number of bicycles used (and conversely) whereas, under heavy rain, $A_{\mathrm{d}}(d)$ decreases. For the rain, the effect seems to be relatively small because of averaging over the day: it is often the case that the rain lasts only for a part of the day. When turning to hourly analysis in the next paragraph, rain will have a deeper and more immediate impact.
(4) The factor pertaining to holidays $J_h$ also impacts $A_{\mathrm{d}}(d)$: There is a decrease (whose relevance is assessed by the confidence interval) during holidays — a feature that appears qualitatively in Fig. 2 (a) and is explained by the fact that people are out of city during holidays.
(5) The number of available bicycles does not impact much $A_{\mathrm{d}}(d)$ and this can interpreted by looking again at Fig. 1 (a): the numbers of subscribers and the number of bicycles follow roughly the same time evolution. This lack of influence hence results from the fact that a part of the evolution is already accounted for by the evolution of $N_s$, and by the fact that there seems to be no major

depletion of bicycles as confronted to subscribers.
(6) Strikes are a non conclusive factor, mostly because of the scarce number of such events in the current dataset.

Using this linear regression model, it becomes possible to predict the amplitude of the number of bicycles rented per day, depending on all the external factors proposed here. If one would use only the average number of hirings $A_{\mathrm{mod}}(d_7)$ adjusted only for the day of week $d_7$, without any other non-stationary factors, the root-mean-square error between the observed data $A_{\mathrm{d}}(d)$ and this number, as normalized by the mean value of this amplitude, would conduct to 30% of mean relative error. Using the model $\widehat{A_{\mathrm{d}}(d)}$, it decreases to 12%. Clearly there is still room for improvement, yet the quantitative gain is not negligible and, more importantly, the interpretation of the dependence with the various factors shows their relevance. Turning to $L(t)$, a zoom is shown in Fig. 7 (b) comparing the resulting model with actual data. The agreement is already good.

**Prediction of hourly fluctuations** Let us now turn to the fluctuation term $F(t)$, whose standard deviation is 210 (in bicycles hired per hour; it can be compared to the mean of $L(t)$ that is equal to 655 hired bicycles per hour). A standard empirical spectrum analysis shows that it is well modeled by an auto-regressive process of order 1 with exogenous input (ARX(1)) [16, 17]:

$$F(t) = a_1 F(t-\Delta) + \beta_1 R(t) + I(t), \tag{6}$$

where $a_1$ is the coefficient of the AR(1) part, and $\beta_1$ is the linear regression coefficient for the rain $R(t)$ (in mm) and $I(t)$ is a white innovation. Using a quadratic error minimization, the estimates are $a_1 = 0.59 \pm 0.02$ and $\beta_1 = -40 \pm 4$ (Vélo'v/$\Delta$/mm of rain). The coefficient $a_1$ and the order of the model were estimated using a classical algorithm on correlations [17].

This leads to a general prediction scheme for the number of hourly rentals that follows eq. (3) with $\widehat{A_{\mathrm{d}}(d)}$ obtained from Eq. (4) and

$$\widehat{F(t)} = a_1 (L(t-\Delta) - L_{\mathrm{mod}}(t-\Delta)) + \beta_1 \widehat{R}(t), \tag{7}$$

where $\widehat{R}(t)$ is the weather forecast for the hour (available from a weather station). In Fig. 8, the displayed model is built using these estimates $\widehat{A_{\mathrm{d}}(d)}$ and $\widehat{F(t)}$ and eq. (3). It works satisfactorily to follow the observed variations of $L(t)$ along time. Using this improved scheme including prediction of the fluctuations, the standard deviation of the error of the global prediction decreases from 210 bicycles to 120 bicycles per hour, i.e., the standard deviation of the innovation $I$, which, by nature of the approach, cannot be predicted. However, as a perspective, the model formulated here can be used to detect unusual changes in the number of rentals, when the measured remaining innovation is different from what is obtained here; it would be an indication of unusual anomalies in the functioning of the system.

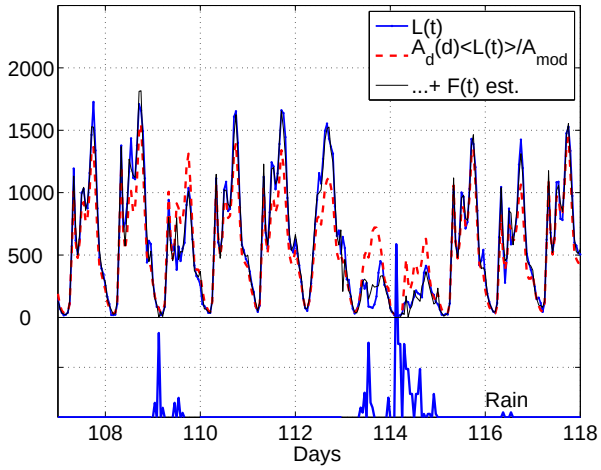


Fig. 8. **Hourly fluctuations of rental numbers: model and data.** The actual data (thin solid line), the model without the ARX(1) part (dashed line) and the full prediction with the ARX(1) part for $\widehat{F(t)}$ are superimposed. On the bottom, the rain for these days is drawn (on arbitrary scale), showing that a major correction obtained by the ARX(1) is actually due to the rain.

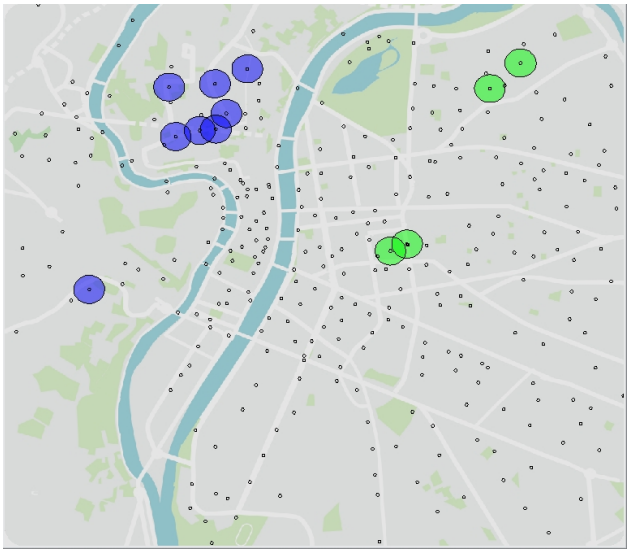

Fig. 9. **Unbalanced stations** in Lyon: the stations where the number of incoming trips is much larger (resp. smaller) than outgoing one are in light (green) circles (resp. dark blue circles); see text, Sec. 5.1 for more details.

## 5. Spatial patterns

### 5.1. *The Vélo'v system as a dynamical network*

Following [13], the Vélo'v system can be interpreted as a dynamical network, where bicycles move from a station to another. Stations are seen as nodes. A central question is to understand how the flows are distributed spatially along the network. Keeping in mind the dominant cyclic and nonstationary features, analyses in time should be combined with analyses in space.

For that, data is studied in the form of the matrix of the flows between stations, also standing for a directed graph or a complex network, where the dimension of time is added: $T[n, m](t)$ denotes the number of trips from station $n$ to station $m$, at time $t$ (aggregated over a time duration $\Delta$). Let $N$ stand for the number of stations, there are $N^2$ directed edges in the full network (including trips back to departure station), whose weights at time $t$ are the number of trips $T[n, m]$. Edges have different weights for each direction.

In order to study the evolution of this network of stations along time, stations will first be arranged in groups that exchange a large number of bicycles at coarse time scales ($\Delta \geq 1$ week), using a classification based on the trips, to and from each of them (cf. Section 5.2). Second, we will turn to the flows between stations and assess, on a finer time scale ($\Delta = 1$h), which pairs are most active, depending on the time in the week (cf. Section 5.3).

A first remark is that the Vélo'v directed network is asymmetric and not self-regulated, because some stations have incoming and leaving traffics unbalanced. Indeed, to avoid saturation in some specific places that were evidenced in the analysis of Sec. 3.2, a small number of trucks are equipped with trailers to move bicycles from one station to another in order to balance the distribution on the network. From the data, we identify stations that reveal an unbalanced traffic. A station $n$ will be considered as unbalanced if the absolute value of the difference between their number of incoming and leaving trips, $|\sum_t \sum_m (T[m, n](t) - T[n, m](t))|$, is larger than 3 times the standard deviation of the distribution of these values over all the stations. This procedure finds 12 unbalanced stations, shown in Fig. 9. Among them, 8 have more leaving trips (dark blue circles) that are located on the top of the two hills that surround Lyon. The 4 remaining unbalanced stations (light green circles) have more incoming trips and are located near the central railway station (close also to the biggest shopping center), and on the university campus.

### 5.2. *Clustering stations in communities*

As seen in Sec. 4, the choice of $\Delta$ depends on whether one is interested in long trends (days or weeks) or short term details (intra-day). Section 5.2 focuses on long periods (typically on one month to one year), while Section 5.3 will concentrate on finer time scales.

To understand the impact of the inhomogeneities of the city on the long-term

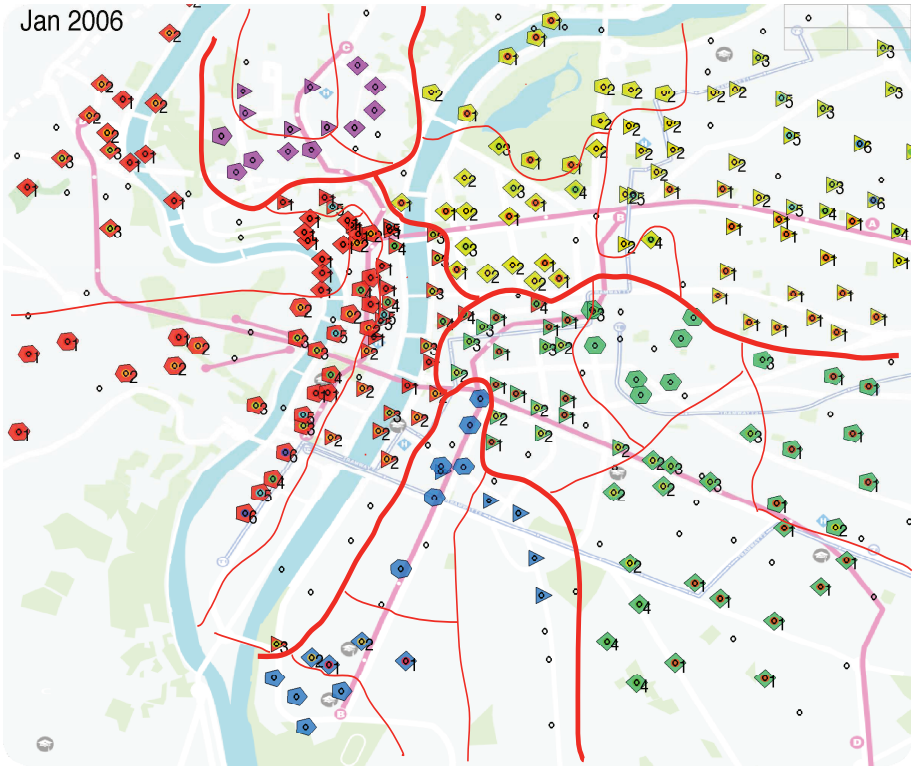


Fig. 10. **Hierarchical communities of stations** in Lyon for 2006, obtained by maximization of the modularity $Q$, as defined in eq. (8). The higher level communities are represented by colors and are separated by (red) thick lines; 5 communities are found at this level. At a second level, sub-comminities inside these 5 ones are distinguished by different marker shapes. Whenever they exist, third level communities are made explicit by a tag consisting of an integer value nearby the colored (1st level)-shaped (2nd level) station markers. (Note also that the colored dot inside the marker is an indication of the same third level in the hierarchy of communities). The communities are found to be mostly grouped by geographical proximity in the city. Stations not associated to a community were not yet in service in 2006.

activity of individual stations, let us look for groups of stations exchanging many bicycles. This amounts to detecting communities of stations in a network [18]. Communities are defined as dense subgraphs with few edges with other communities. They are found in many complex networks and they can correspond to groups with similar behaviors or interests (for people), with similar contents (for web pages), etc. Moreover studies shows that information (rumors for instance) spread more rapidly within communities than between communities. In the Vélo'v context, finding communities will help to aggregate spatially the individual stations on the basis of an objective criterion. Automated detection of communities in graph is a difficult problem that received recently considerable research effort, issues being both theoretical (conceptual definition of communities) and practical (definitions should

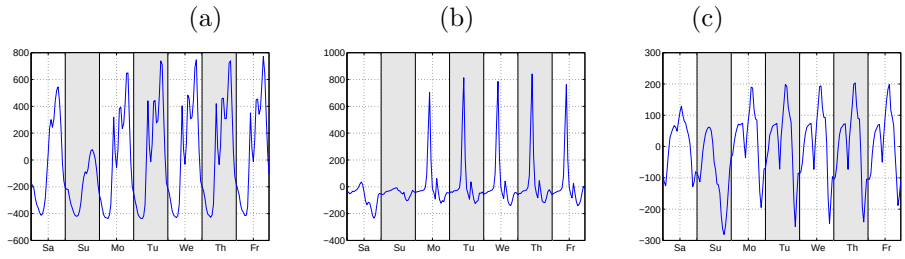


Fig. 11. **PCA analysis of** $T[m,n](t)$**.** The first 3 principal components of $T$ are represented along time. One recognizes in the first one the cyclic part of the model, $\langle L(t)\rangle_c$ (without its mean). The next components are mostly corrections on some of the peaks of this model: the second component changes mostly the mornings in the week-days, the third one brings corrections of opposite signs on noon and afternoon.

end up with quantities that can actually be computed at a reasonable load). Reviewing the literature reveals that proposed algorithms often suffer from either high computational costs, and hence cannot be used on an actual large database such as the Vélo'v one, or from a significant sensitivity to minor topology modifications, lacking robustness.

This review of the literature led us to resort to a definition of communities based on Newman's modularity [18, 19] The efficient algorithm proposed in [20] has been customized to the spatial analysis of the Vélo'v dataset. In this approach, graph modularity is defined as the average, over all pairs of nodes, of the difference between the actual $T[n,m]$ and that expected under the absence of community [18, 19]. For a directed and weighted network, modularity $Q$ takes the form:

$$Q = \frac{1}{N(N-1)} \sum_{n,m} \left[ T[n,m] - \frac{\sum_{j\neq n} T[j,n] \cdot \sum_{k\neq m} T[m,k]}{N(N-1)} \right] \delta_{c_n,c_m} \qquad (8)$$

where $c_n$ is the community where station $n$ is assigned, and $\delta$ the Kronecker delta symbol. Community definitions result from the maximization of the modularity over the set of possible partitions.

However, this optimization problem is NP-complete [21]. Approximations, such as the greedy approach of [22], tend to produce too large communities. The hierarchical algorithm proposed in [20] is used here because it builds a hierarchy of communities of increasing sizes (small community are grouped into larger ones). Also, it uses a local computation of the gain of modularity when increasing the size of the communities by merging some together, and hence shows a tractable computational efficiency. Finally, the method's output reads as a hierarchy of embedded communities: the first level of the hierarchy is the less detailed grouping into communities (the one that is found last in the unfolding of the algorithm); then each community of this first level can be split into several sub-communities on the second level. Then, in some cases, a third level breaks second-level communities

in several parts. The number of communities at each level of the hierarchy cannot be decided a priori by the practitioner and per se constitutes an important result of the analysis. The practical use of this output consists of first plotting the higher level communities. Then, information can be refined by considering the second level of the hierarchy that split the higher level communities into sub-communities, and so on and so forth. Therefore, the only choice left to practitioners is that of the level in the hierarchy at which the refinement superimposition should be stopped (often guided by readability of the result).

The result of the unfolding of hierarchical communities, applied to one year of data, is displayed in Fig. 10, for the three higher levels of the hierarchy of communities. The most striking feature is that the communities are mostly organized as groups of stations close in space, even though the method uses no geographical information. This is in accordance with the short typical trip length (see Fig. 5): many trips are local. Inside the higher-level communities (found, by community clustering, to closely match the administrative districts of the city), finer-level communities reveal details such as groups of stations lined along major boulevards (and subway lines or bicycle paths often follow them). The stations on the Croix-Rousse hill (the zone on the north) are clearly grouped in a specific community, as are the ones near the northern campus of the Science University (La Doua).

Note that we have checked that removal of stations with an unbalanced behavior did not change the results reported in this sections about communities (nor would it change the results of the flow clustering in 5.3).

The conclusion is that grouping the stations by geographical proximity is a correct intuition. Indeed, close stations exchange more bicycles than distant stations: this is the meaning of the communities found by maximizing the modularity. These hierarchical communities provide guidelines to automatically group communities given a level of granularity in space that is wanted, instead of trying to do this task by hand and intuition only. The method provides us with a quantitative means of spatial aggregation.

### 5.3. *Clustering flows of activity between stations*

The second step of spatial analysis is the clustering of the flows between stations at finer time-scales. The objective is now to highlight the distribution in time along the week of the main spatial features of the Vélo'v use. Therefore, the $T[n, m](t)$ are now aggregated with $\Delta = 1$h, as in Sec. 4. Also, because of the nonstationary evolution at scales larger than the week, data used are either one specific week, or a mean over several weeks if one wants to study aggregated trends.

The high-dimensionality of the data involved ($N^2$ flows times 168 hours per week) calls for a dimensionality reduction. Thanks to the time analysis done, we know that the most important activities of the stations are characterized by 3 peaks every ordinary days (8am-9am, 12am-1pm and 5pm-7pm) and 2 peaks for weekends (1pm-2pm and 4pm-6pm). We select these 19 times stamps to be the features

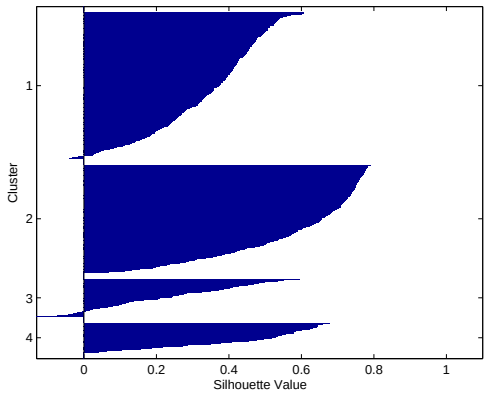


Fig. 12. **K-means silhouettes (Eq. (9)) of the clustering of Flows.** Computed for the 4 clusters of flows. Almost all values are positive: this is an indication of good and relevant clustering.

in time. Note that dimension reduction using Principal Component Analysis on $T[n,m](t)$ was performed in [13]. The PCA transforms the original attributes into a set of Principal Components (PC) that are non-correlated and obtained as linear combinations of the original variables. The first 3 obtained PCs along time are displayed in Fig. 11: one recognizes without surprise for the first PC the cyclic model (minus its mean) that was studied in Sec. 4; it accounts for 54.6% of the variance. The following components (accounting for 13.4%, then 5.6%) are corrections on this cyclic pattern, mostly on the various peaks of activity: the second component changes mostly the mornings in the week-days, the third one brings corrections of opposite signs on noon and afternoon. This leads us to retain these 19 peaks of activities during the week as the dominant time features. Then, we keep only 1046 pairs of stations where traffic is large enough between the pairs, meaning that the number of trips is at least 1 every 3 weeks.

In order to uncover the main properties of flows on the Vélo'v stations network, a K-means algorithm (see, e.g., [23]) is run on $T[n,m](t)$ for $t$ equal to the 19 selected time-features and $(n,m)$ being in the 1046 pairs of stations that are kept. The distances between every couples of pairs of stations were evaluated classically by the correlation between the temporal vectors of number of rentals [24]. Silhouette measures in a classical manner [25] the quality of a clustering by estimating how a pair of stations is similar to other pairs in its own cluster vs. pairs in other clusters, and ranges from -1 to 1. Silhouette is defined as

$$S(i) = \frac{\min_k(\overline{d_B}(i,k)) - \overline{d_W}(i)}{\max(\overline{d_W}(i), \min_k(\overline{d_B}(i,k)))}, \tag{9}$$

where $\overline{d_W}(i)$ is the average distance from the $i$-th point to the other points in its own cluster, and $\overline{d_B}(i,k)$ is the average distance from the $i$-th point to points in

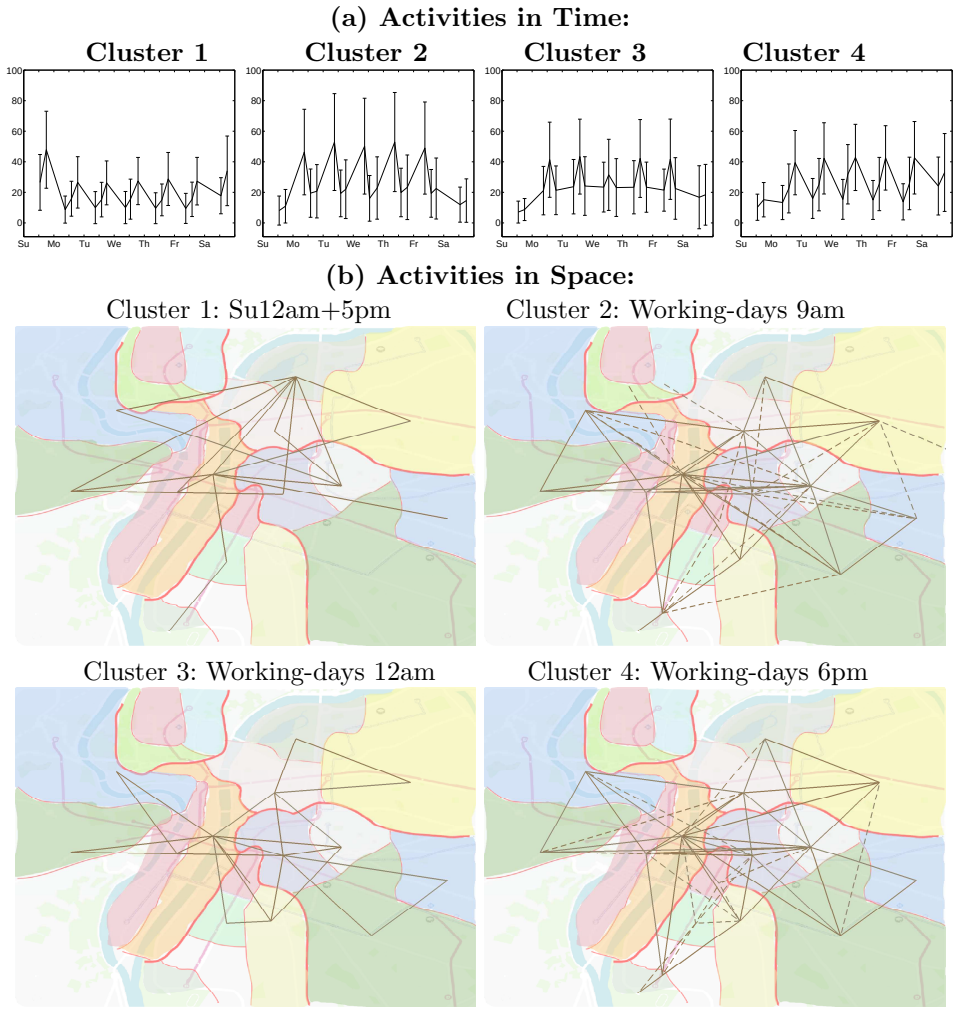


Fig. 13. **Clustering of the flows between stations.** (a) Activities in Time: At the 19 selected time-features, the mean and variance of the sum of the flows of the station are displayed for the 4 clusters. (b) Activities in Space: The map is coded in areas of different color background, each area being a community of the second level of the hierarchy displayed in Fig. 10. Then, in each sub-figure, a line is drawn between the centers of 2 communities if, in the displayed cluster, there exists a flow between stations of the 2 communities. For clusters 2 and 4, the solid lines are the ones appearing in both clusters 2 and 4; dashed lines are flows only in one of these clusters.

another cluster $k$. The procedure finds 4 well separated clusters whose silhouette values of pairwise stations are shown in Fig. 12. Pairs of stations are closer to the ones of the same cluster than to pairs of others clusters, except for 25 among 1046 pairs—this attests of the quality of the clustering.

Let us now comment on the identified clusters. Fig. 13 (a) shows the mean and standard deviation of the number of moves for the flows in each cluster. The peaks

of activity of each cluster are easily identified:

(1) Cluster 1 corresponds to rentals on Sundays at noon at 5pm;
(2) Cluster 2 corresponds to travels at 8am-9am, mostly on working days;
(3) Cluster 3 corresponds similarly to hirings around noon;
(4) Cluster 4 gathers afternoon travels at 5pm-7pm.

The clusters take a clear meaning as being a classification of the dynamics on the network in space and time.

Fig. 13 (b) locates on the map of Lyon the pairs of stations that are part of each cluster. To make the picture clearer, we grouped nearby stations according to their community (computed in Section 5.2) and plot a line between two communities if there exists at least one station in each forming a pair that belongs to the corresponding cluster. Communities are shown with the same code as in Fig. 10. The interpretations of the clusters follow:

(1) Trips in Cluster 1 are mainly along the two rivers and around the main parks of the city (in the north and the south of the map). We can also observe some travels between the university campus (or the periphery of the city) and the center of the city (North-east and the land between the two rivers).
(2) Clusters 2 and 4 share many similarities (the solid lines in Fig. 10 are the edges they have in common): they correspond to commuting to and from work (respectively Clusters 2 and 4). We identify the main network hubs (train stations, campus, business center, etc.) in the communities reached in these clusters.
(3) Cluster 3 is less dense; it includes short travels related to lunch break rides, and moves are often between close communities.

It is worth noting that these clustering results seem to be stable: similar results are obtained when applying the same methodology on a monthly basis.

## 6. Conclusion

The dataset made available to us by JCDecaux and the Grand Lyon City Hall is huge and unique in nature, consisting of the records of each and every Vélo'v trip over a two year long period. The exhaustive digital footprint kept by the system is unmatched by usual social enquiries, hence permitting real statistical and data analysis on issues pertaining to trips in bicycles.

Two kinds of analyses were performed. First, carefully combining standard statistical signal processing tools dedicated both to cyclostationarity or nonstationary trend analysis and to forecasting, enabled us to model the time evolution of hourly-aggregated bicycle rentals. It yielded a temporal pattern for the typical week mixing days and intra-days periodicities, most being naturally interpretable as related to professional activity rhythms (week days) or leisure (week-end) activities. This pattern closely resembles those observed in studies of different sharing programs in other cities. In addition, it enabled the forecasting of the number of bicycles rented

in the next hour, based on the knowledge of factors both internal to the deployment program (number of available bicycles or subscribers) and external (weather conditions), down to a ten per cent fluctuation accuracy. Second, computer science data mining tools were tailored to the analysis of the Vélo'v dataset to extract clusters of stations based either on an intra versus inter community preferred exchanges measure (modularity), yielding communities of stations exchanging regularly a large number of bicycles, or on a similarity measure in the time patterns of bicycle flows. Such analyses enabled us to gain a significant understanding on the social usage of the Vélo'v program in Lyon: Communities remain geographically concentrated (hence indicating a preferred short-range use of the bicycles) while time patterns of flows between stations display similarities so that they are grouped in clusters separating trips related to professional activities (week days and major communication hubs) from those used during leisure time (week end and parks). Finally, they showed that, depending on the time in the week, some stations are alternatively sinks or sources of Vélo'v.

Besides the usage conclusions they enabled us to yield, these contributions are also of methodological values: Notably, community mining for stations, and time pattern clustering for flows remain intricate issues both at the theoretical and practical levels. Also, the tools used here depend on a aggregation or resolution scale $\Delta$ at which analyses are conducted and that can be tuned to further address different questions and issues.

A large number of open questions remain, some of them being currently under investigations. Regular contacts with JCDecaux (the Vélo'v private operator) and the Grand Lyon City Hall (the political leader of the program), enabled the identification of various operational investigation objectives, ranging from the system optimization of bicycle removal/balancing operations to the evaluation and certification that the prescribed quality of service is actually achieved, most of them however not qualifying for public disclosure. Further developments will be oriented toward analyzing the Vélo'v system with respect to socio-economical information and quantitative data related to Lyon City, and collected by the French INSEE (Institut National de la Statistique et des Études Économiques).

## 7. Acknowledgments

JCDecaux and the Grand Lyon Urban Community are gratefully acknowledged for having made Vélo'v data available to us. The authors thank Antoine Scherrer and Pablo Jensen for interesting discussions. Work partially supported by IXXI-Lyon (Institut des Systèmes Complexes – Complex Systems Institute).

## References

[1] http://www.velov.grandlyon.com/
[2] Girardin, F. "Revealing Paris Through Velib' Data" http://liftlab.com/think/fabien/2008/02/27/revealing-paris-through-velib-data/ (2008).

[3] Froehlich, J., Neumann, J., and Oliver, N. "Measuring the Pulse of the City through Shared Bicycle Programs" *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems - UrbanSense08*, Raleigh, North Carolina, USA (November 4, 2008).
[4] Froehlich, J., Neumann, J., and Oliver, N. (2009) "Sensing and Predicting the Pulse of the City through Shared Bicycling" *Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, pp. 1420-1426 Pasadena, California, USA, (July 11 - 17, 2009).
[5] Roth, C., Soong Moon Kang, Batty, M., Barthelemy, M. "Commuting in a polycentric city", Publication: eprint arXiv:1001.4915 (January 31, 2010).
[6] http://www.velov.grandlyon.com/Plan-interactif.61.0.html
[7] Preparata, F. R. and Shamos, M. I. "Computational Geometry: An Introduction." New York: Springer-Verlag (1985).
[8] http://www.grandlyon.com/PDU.55.0.html
[9] Jensen, P., Rouquier, J.-B., Ovtracht, N. and Robardet, C. "Characterizing the speed and paths of shared bicycles in Lyon", *Transportation Research Part D: Transport and Environment*, vol. 15 (8), pp. 522–524 (2010).
[10] Abry, P., Baraniuk, R., Flandrin, P., Rieidi, R. and Veitch, D. "Wavelet and Multiscale Analysis of Network Traffic", *IEEE Signal Processing Magazine*, vol. 3 (19), pp. 28–46 (2002).
[11] Borgnat, P., Abry, A., Flandrin, F., and Rouquier, J.-B. "Studying Lyon's Velo'V: A Statistical Cyclic Model", *Proceedings of ECCS'09*, Warwick, UK (Sept., 2009).
[12] Borgnat, P., Abry, A., and Flandrin, F., "Modélisation statistique cyclique des locations de Vélo'v à Lyon", Symposium GRETSI-09, Dijon, FR (Sept., 2009).
[13] Borgnat, P., Fleury, E., Robardet, C., and Scherrer, A, "Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program", *Proceedings of ECCS'09*, Warwick, UK (Sept., 2009).
[14] Gardner, W., Napolitano, A., and Paura, L., "Cyclostationarity: Half a century of research", *Signal Processing*, vol. 86 (4), pp. 639–697 (2006).
[15] Serpedin, E., Panduru, F., Sarı, I., Giannakis, G. "Bibliography on Cyclostationarity", *Signal Processing*, vol. 85 (5), pp. 2233–2303 (2005).
[16] Priestley, M.B. *Spectral analysis and times series*. Academic Press, San Diego (1981).
[17] Ljung, L. "System identification: theory for the user (2nd edition)", Prentice-Hall, Englewood Cliffs, NJ (1999).
[18] Newman, M. and Girvan, M., "Finding and evaluating community structure in networks", *Phys. Rev. E*, vol. 69, 0206113 (2004).
[19] Leicht, E. and Newman M., "Community structure in directed networks", *Phys. Rev. Lett.*, vol. 100, 118703 (2008).
[20] Blondel, V. Guillaume, J.-L., Lambiotte, R. and Lefebvre E., "Fast unfolding of communities in large networks", *J. Stat. Mech.*, P10008 (2008)
[21] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. "On modularity clustering", *IEEE TKDE*, vol. 20 (2), pp.172–188 (2008).
[22] Clauset, A., Newman, M., and Moore, C., "Finding community structure in very large networks", *Phys. Rev. E*, vol. 70, 066111 (2007).
[23] MacKay, D."Information Theory, Inference and Learning Algorithms" Cambridge University Press (2003).
[24] Basseville, M. "Distance measures for signal processing and pattern recognition", *Signal Processing*, vol.18 (4), pp.349-369 (1989).
[25] Rousseeuw, P.J. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Computational and Applied Mathematics*, vol. 20, p. 53–65 (1987).