Using spatial outliers detection to assess balancing mechanisms in bike sharing systems

Rayane El Sibai, Yousra Chabchoub, Christine Fricker

Using spatial outliers detection to assess balancing mechanisms in bike sharing systems

Rayane EL SIBAI[1*], Yousra CHABCHOUB[2*] and Christine FRICKER[3†]

[*]LISITE Laboratory, ISEP, Issy-les-Moulineaux 92130, France
[1]rayane.el_sibai@etu.upmc.fr
[2]yousra.chabchoub@isep.fr
[†]INRIA, RAP Project, 2 Rue Simone IFF, Paris 75012, France
[3]christine.fricker@inria.fr

***Abstract*—Spatial outliers are objects having a behavior significantly different from their spatial neighbors, in a context where neighbors are heavily correlated. Moran scatterplot is a well-known method that exploits similarity between neighbors in order to detect spatial outliers. In this paper, we proposed first an improved version of Moran scatterplot, using a robust distance metric called Gower's similarity. We used the new version of Moran scatterplot to study the homogeneity of the Parisian bike sharing system (Velib). We carried out different experiments on a real dataset issued from the Velib system. We identified many spatial outliers stations, very different from their neighboring stations (often with much more available bikes or with much more empty docks during the day). Then, we designed and tested a new method that globally improves the distribution of the resources (bikes and docks) among bike stations. This method is motivated by the existence of spatial outliers stations. It relies on a local small change in users behaviors, by adapting their trips to resources' availability around their departure and arrival stations. Results show that, even with a partial users collaboration, the proposed method enhances significantly the global homogeneity of the bike sharing system and therefore the users' satisfaction.**

***Index Terms*—Outliers detection, spatial data mining, Moran scatterplot, Gower's coefficient, robust distance.**

I. Introduction

Outliers are defined as dataset observations that are inconsistent with the remainder of the dataset. The identification of outliers has practical applications in many areas, such as intrusion detection, fraud detection, fault detection and medical informatics (see [1] for a survey). Outliers detection is also an important feature in the data analysis process. It aims to detect abnormal patterns and leads to the identification of unusual phenomena, and to new knowledge concerning the monitored environment. Data quality can also be improved using outliers detection. As an example, in [2] authors enhance the robustness of the data analysis by detecting and replacing of erroneous values so that the results are no longer affected by the defective data.

To isolate outliers is it necessary to first characterize the normal observations, which can be provided by the past values of the same object or by the current values issued from other objects in the neighborhood. In this latter case, the outlier is said spatial. In a spatial context, each data is defined with two categories of attributes: spatial attributes and non-spatial attributes. Spatial attributes include the shape, position, and other topological characteristics of the sensor, and they are used to define the neighborhood of the spatial object. Non-spatial attributes include the ID, manufacturer, age, and sensor measure (called behavioral attribute). A spatial outlier represents a local instability and is only compared to the surrounding dataset [3]. This is based on the rule: 'Everything is related to everything else, but nearby things are more related than distant things' [4]. Spatial outliers detection is used in many applications, such as the detection of abnormal highway traffic patterns [5], the identification of disease outbreaks [6], the detection of tornadoes and hurricanes [7] and the identification of urban soils pollution [8].

Several algorithms have been developed to detect the outliers in a spatial context. They can be classified into two categories: graphical-based algorithms, and quantitative-based algorithms. Graphical-based algorithms are based on the visualization. They present for each spatial point the distribution of its neighbors and identify outliers as points in specific regions. This category includes variogram cloud, pocket plot, scatterplot, and Moran scatterplot methods. Quantitative-based algorithms perform statistical tests to distinguish the outliers from the rest of the data. These methods include z algorithm, iterative r, iterative z and median algorithm.

Scatterplot represents the data in a two-dimensional space where the X-axis represents the values of the non-spatial attribute (the observable) of each object and the Y-axis represents the mean value of the neighbors of this object. A regression line is used to identify outliers points [9]. Variogram Cloud [10] compares the distance between the spatial attributes to the distance between non-spatial attributes. It displays a scatterplot between the spatial distance (X-axis) and the difference of the observable values (Y-axis) for each pair of points in the dataset. Outliers are identified as pairs of points having a small spatial distance and a big difference for the observables measurements.

The Z statistic approach [11] is one of the most known quantitative-based algorithms for spatial outliers detection. For

each spatial object x, S_x denotes the difference between the attribute value of x and the average attribute value of its spatial neighbors. Spatial outliers are simply identified using a threshold based on μ_s and σ_s which respectively represent the mean and the standard deviation of the attribute value of S over all the spatial objects.

In [12] authors propose two iterative algorithms (iterative r and iterative z) for the detection of spatial outliers. These algorithms detect the outliers on several iterations. Each iteration detects a single outlier and modifies its value in order to reduce its negative impact on its neighbors in the next iteration.

We apply in this paper the spatial outliers problem to a particular case study: the evaluation of a balancing mechanism in Bike Sharing Systems (BSS). Nowadays, public authorities are more and more encouraging this ecological mean of transport by expanding the BSS to the suburbs and building new bike paths. Since its launch in 2007, Velib (the Bike Sharing System -BSS- in Paris) has emerged in the Parisian landscape and has been a model for similar systems in many international cities. Velib provides a significant proportion of people travels as it daily ensures about 110,000 trips. It involves about 1800 stations with an average distance of 300 meters.

A major problem in the Velib system and in BSS in general, is the problem of empty stations and full stations caused by the asymmetric attendance to the stations. According to the annual satisfaction survey of Velib, only 50% of users are satisfied with the availability of bikes and docks in the stations [13]. Despite the performed regulation (moved bikes using trucks), users often find themselves in front of stations that are totally full or empty.

In most cities, operators provide open access to real-time status reports on their bike stations. Several studies show the interest of using these data (Froelich *et al.* [14] and Borgnat *et al.* [15], Vogel and Mattfeld [16]). Their main objective is to understand and characterize the behavior of the users in order to help in designing and planning policy in urban transportation. Among these studies, one can cite the partition of the BSS stations into several classes using different clustering algorithms (see [17] and [18] for more details). Other studies, performed a classification of the flows of trips as analysis of the trips in the Velo'v system in Lyon proposed by Borgnat *et al.* in [15].

We focus, in this work, on the problematic Velib stations which are often almost empty or almost full. First, we use an adapted version of Moran scatterplot to explore and characterize the neighborhood of such stations. For this purpose, we introduce Gower's similarity to evaluate the similarity between Velib stations using their capacities and the geographical distance. Results show a local heterogeneity in Velib station: in a small area, bikes availability is often very variable, depending on the station. This local heterogeneity motivates the second part of this paper, where we propose a new method that naturally enhances resources' distribution among the Velib stations. It is an incitative mechanism based on a local small change in users trips. In this natural regulation, users are redirected to another station in the neighborhood to locally reduce stations heterogeneity. Experiments, using real data trips, show the effectiveness of the proposed solution. Even with a partial users collaboration, it reduces significantly the number of problematic stations and decreases globally the duration of stations invalidity.

The rest of the paper is organized as follows. We describe in Section II the data used in this work, and we highlight the main problem of Velib system. In Section III, we describe the so-called Moran scatterplot technique and the proposed adaption to the Velib context. We also detail in this section the experiments carried out to illustrate Velib system heterogeneity. In Section IV, we present and validate our new solution to balance the Velib system and to improve bikes distribution among the stations.

II. Dataset description and problem definition

In order to promote innovation and collaboration with scientists, different kinds of data relative to the Velib system are "Open Data", available for the research community. We performed all the experiments presented in this paper on these datasets.

First, we have the static data describing the Velib stations. They consist of spatial attributes: the geo-coordinates of the station (latitude and longitude), and non-spatial attributes: ID of the station and its capacity (total number of docks).

Then, we have the dynamic data which are of two kinds: occupancy data and trip data. Occupancy data depict the number of bikes present in each station for each timestamp t and they are provided in real time. This parameter is varying during the day and is closely dependent on users activity. Trip data represent the data corresponding to the trips of Velib' users. A trip is characterized by a departure and arrival timestamp, and a departure and arrival station. The analysis of several months of trips showed a very strong periodicity, so the trips can be divided into two main categories: the working days and the weekends. Two days of the same category are very similar. That is why we used a single day trips in our experiments. We focus in this paper on the working days and we choose to analyze 24 hours trips: trips that took place on Thursday, October, the 31^{st}, 2013. This duration includes: $121,709$ trips, involving 1226 Velib stations. 1.03% of the trips are related to maintenance (bikes taken for repair) and 1.48% are trips of regulation (bikes moved by trucks).

According to many research studies ([19], [20] and [18]), the Velib system has some weaknesses caused by the strong attractiveness of some stations that can be explained by their location near a railway station or a monument or business area. Such stations are very often completely empty (no available bike) or completely full (no available dock to put a bike).

Despite the performed regulation (bikes moved by trucks), the system is still unbalanced, causing users dissatisfaction.

The unbalanced stations are referred to as *problematic* stations. More precisely we introduce the following definition: a station is said problematic at a timestamp t if its *occupancy rate* is under 10% or more than 90%. The *occupancy rate* of the station, at a timestamp t, is defined as follows:

$$occupancy\,rate_{\,t} = \frac{Number\ of\ bikes\ present\ at\ t}{Capacity\ of\ the\ station} \times 100\%$$

Our objective is to improve resources' availability in the Velib system by reducing the number of problematic stations. For this purpose, we propose and test in Section IV a new incitative method, based on a natural and ecological regulation performed by Velib users. The main idea behind the proposed method is to balance the global system by performing small changes in the trips in small local areas. A preliminary study is provided in Section III to check the existence of several isolated problematic stations. In other words, the aim of this part is to show that around a given problematic station (in a distance smaller than 500 meters), there are many balanced stations (with an occupancy rate around 50%), which make it possible for Velib users to balance this problematic station by slightly changing their trips (with an award, extra-time for example).

III. Spatial outliers detection with an improved Moran scatterplot

The objective of this section is to estimate the number of isolated problematic stations at a given timestamp t, which motivates the incitative method detailed in the following Section. Such station satisfies both following conditions: First, it is almost empty or almost full at timestamp t. Second, its occupancy rate is significantly different from the average occupancy of the neighboring stations at the same timestamp t. Thus, the isolated problematic stations are among the spatial outliers.

In order to detect spatial outliers, we opted to use Moran scatterplot [21] that we adapted to the specificities of our context.

A. Moran scatterplot

Moran scatterplot [21] illustrates the similarity between an observed value and its neighboring observations. It measures the global spatial autocorrelation over a geographical area, the well-known *Moran's I*.

Let us denote by $Z = \{z_i : 1 \leq i \leq n\}$ the set of the different values of the considered observable at a fixed given time t, in n different locations. For each location, the neighborhood is defined based on the geographical distance. Moran scatterplot visualizes the relationship between the values z_i and their neighborhood average $W_i.Z$, where W is a weight matrix that defines a local neighborhood around each location. The observations Z (x-axis) and $W.Z$ (y-axis) are represented by their standardized values.

Moran scatterplot contains four quadrants, corresponding to four types of spatial correlation. The upper-right and lower-left quadrants consist of the locations with positive spatial correlation: association between similar values. In the upper-right quadrant, the high values are surrounded by high neighbors values, while in the lower-left quadrant, the low values are surrounded by low neighbors values.

The upper-left and lower-right quadrants incorporate the locations with negative spatial correlation: association between dissimilar values. The upper-left quadrant contains low values surrounded by high neighbors values, while the lower-right quadrant contains high values surrounded by low neighbors values.The objects located in these two quadrants are considered as spatial outliers and can be identified by the statistical test function:

$$Z_i \times \sum_j w_{ij} Z_j < 0$$

W is the contiguity matrix of weights. It indicates the spatial relationship between every couple of objects. W is also called the row-normalized neighborhood matrix. It is based on a threshold d of the geographical distance: i and j are considered as neighbors if and only if $0 \leq d_{ij} \leq d$, where d_{ij} is the distance between i and j. Moreover, all the neighbors of i are equivalent and have the same impact on the calculation of the neighborhood average $W_i.Z$.

Thus, the contiguity matrix W is given by:

$$w_{ij} = \begin{cases} \frac{1}{Number\ of\ neighbors\ of\ i}, & \text{if } \ 0 \leq d_{ij} \leq d \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

To apply Moran scatterplot to the context of Velib, one has to estimate the crucial parameter d, which represents the highest distance between two neighboring Velib stations. The choice of d has to achieve the following trade-off: On the one hand, this distance has to be small enough to let the users slightly change their trips at a local scale, and on the other hand, it has to be high to make sure that most stations have a reasonable number of neighboring stations. Velib stations are generally close to each other and concentrated in the center of Paris and near attractive locations whereas they are distant in the suburbs.

To address this problem, we plotted in Figure 1 the distribution of the number of neighbors for all the Velib stations. We tested different values for the threshold distance d (300, 400 and 500 meters). According to Figure 1, a distance of 400 meters is reasonable as, in this case, a given Velib station has on average about 5 neighboring stations. Moreover, with $d = 400$, Only 4.4% of the stations do not have any neighboring station.

However, when detecting spatial outliers, the assumption that all the neighbors have the same impact on the neighborhood

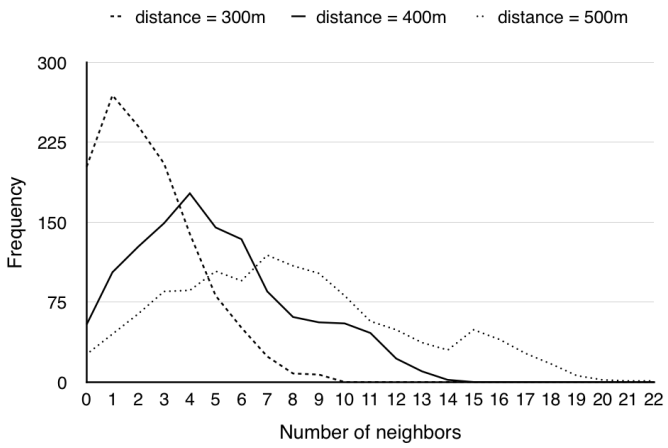

Fig. 1. Distribution of the number of neighbors

average may lead to missing some true spatial outliers. In the dataset described in section II, there are 1226 stations. As plotted in Figure 2, the capacity of the stations is highly variable between 8 and 114 bikes, with an average of about 31 bikes. As Velib stations have different capacities, we defined the occupancy rate in order to compare normalized bikes availability in these stations. The key idea is that two neighboring stations should have almost the same occupancy rate if they have similar capacities. That is why the capacity of the station has to be taken into account when calculating neighborhood occupancy average.

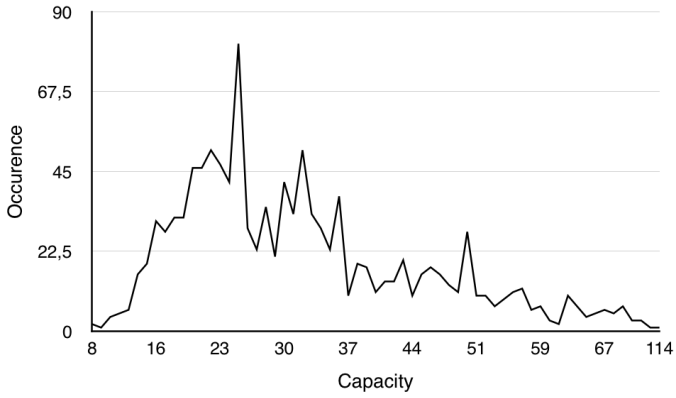

Fig. 2. Distribution of the capacity of stations

B. Improvement of Moran scatterplot using Gower's coefficient

We will replace W with a new weight matrix $\tilde{W}$ also based on the degree of similarity between the station i and the corresponding neighboring stations. This new matrix will take into account the distance and also the difference of capacities between a station and its neighbors. The set N_i of neighbors of station i is defined as previously by the stations with a maximal distance d from station i.

In order to measure the similarity degree between two spatial objects, the Euclidean distance is most often used. However, in our case, the use of this distance is inappropriate since the location and capacity attributes are measured on different scales.

Hence, we propose to use the Gower's coefficient [22] to calculate the similarity between two stations. Gower's coefficient is a similarity measure which computes the distance between two instances on each attribute k, and then aggregates all of them to finally calculate the similarity degree.

Gower's similarity degree $GOWER_{ij}$ between two stations i and j is defined by:

$$GOWER_{ij} = \frac{\sum_{k=1}^{n} W_{ijk} \times S_{ijk}}{\sum_{k=1}^{n} W_{ijk}} \tag{2}$$

where

- W_{ijk} is the weight associated to the attribute k,
- S_{ijk} is the similarity between two stations i and j for the k^{th} attribute, given by

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}.$$

where x_{ik} is the observable attribute k in station i and r_k is a standardization for the attribute k since each attribute is of different unit.

In the context of Velib stations, we calculate the similarity S_{ij} of the location SD_{ij} and capacity SC_{ij} between two stations i and j by:

$$SD_{ij} = 1 - \frac{d_{ij}}{d}$$

$$SC_{ij} = 1 - \frac{|Capacity_i - Capacity_j|}{Capacity_{max} - Capacity_{min}}$$

where

- d_{ij} is the distance between the two stations and d is the maximal distance.
- $Capacity_{max}$ and $Capacity_{min}$ are respectively the maximal and minimal stations capacities in the neighborhood of station i.

In this definition, $W_{ijk} = W_{ij}$ previously defined by equation (1).

We propose in the following to modify the construction of the contiguity matrix of weights by incorporating the spatial and non-spatial attributes and in a weighted manner in the calculation of the weights associated with neighbors. For each neighboring station j, its new weight $GOWER_{ij}$ regarding the station i is given by equation (2).

The normalization of the contiguity matrix of weights is done per line, so for each station i, the weight of each neighboring

station j is divided by the sum of the weights of all the neighboring stations of i.

Thus, the new contiguity matrix $\tilde{W}$ is given by:

$$\tilde{w}_{ij} = \begin{cases} GOWER_{ij}, & \text{if } 0 \leq d_{ij} \leq d \\ 0, & \text{otherwise.} \end{cases}$$

We applied the improved version of Moran scatterplot to detect the isolated problematic stations. Recall that these stations are defined as spatial outliers with a critical occupancy rate. We used the same dataset described in Section II.

Moran scatterplot representation for the occupancy data of the stations at a fixed timestamp: $10:00$ *am* is given in Figure 3. At this time of day, we can expect that the system is highly unbalanced, as in general in a working day a lot of trips take place in the morning around $8:00$ *am*. The spatial outliers stations (almost 300 stations) are located in the upper-left and lower-right quadrants. One can notice that there are less points in these quadrants compared to the locations with positive correlation.

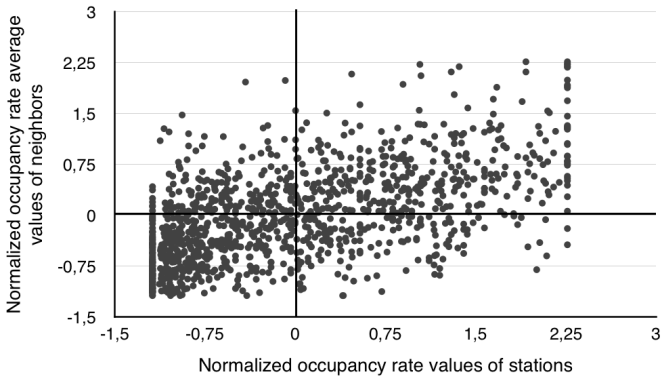

Fig. 3. Improved Moran scatterplot based on occupancy data of Velib on Thursday 10/31/2013 $10:00$ *am*

The number of detected isolated problematic stations at $10:00$ *am*, depending on the allowed distance, is given in Table I. Recall that isolated problematic stations are defined as spatial outliers with critical occupancy rate. According to this table, there are about 50 isolated problematic stations at $10:00$ *am*. The allowed distance does not have a considerable impact on the number of outliers and the isolated problematic stations. Moreover, with a local change of their trips, Velib users can enhance the occupancy rate of about 300 stations (spatial outliers), which represents 24.48% of Velib stations.

IV. Improving resource distribution in the Velib system

According to a recent annual survey on the Parisian bike sharing system, only 50% of users are satisfied with the availability of bikes and docks (free terminals) in the stations [13]. Many stations are often unusable for some users during some amount of time because of the lack of bikes or docks due to their attractive location.

Using the dataset described in section II, we plotted in Figure 4 the evolution of the number of current trips during the day (on Thursday 10/31/2013), in order to understand the usage of the Velib system. One can easily identify two peaks at about 8:00 am and 6:00 pm. They clearly correspond to the trips to the offices and the return home after work, as it is a working day.

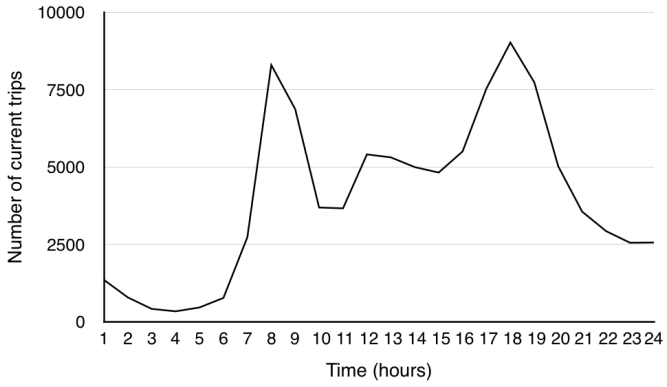

Fig. 4. Number of trips over time

Users trips unbalance the Velib system by making some stations problematic (almost empty or almost full). Based on the thresholds of station occupancy introduced before (10% and 90%), the current number of problematic stations is given in Figure 5. Despite the performed bike regulation using tracks, the number of problematic stations during the day remains high. The problematic stations are mainly composed of almost empty stations.

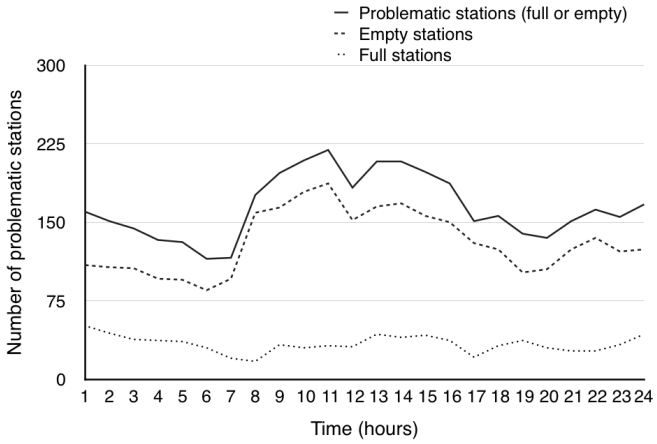

Fig. 5. Number of problematic stations

We propose in this section an incitative method that encourages Velib users to improve the homogeneity of the stations in terms of occupancy rate by slightly changing

TABLE I
NUMBER OF DETECTED OUTLIERS STATIONS WITH THE IMPROVED MORAN SCATTERPLOT

Allowed distance	Outliers	Outliers with critical occupancy rate
300	297	53
400	334	52
500	339	54

their trips. In the trips dataset, let us denote by A the station where the trip begins and by B where the trip ends. The neighborhood of the station is defined by a distance less than 400 meters. The key idea is to change the trips as follows:

For each trip, in terms of occupancy rate,

- station A will be replaced by the busiest station in the neighborhood of A,
- station B will be replaced by the emptiest station in the neighborhood of B.

The proposed method is inspired by Velib+ which consists of offering users of Velib an extra time (that can be cumulated) when they park their bike in a station having a high altitude. The main difference is that Velib+ regulation is static: Velib+ stations are well known and never change over time, whereas our preferred busiest and emptiest stations dynamically change. They vary during the time depending on their occupancy rate and the occupancy rate of their neighboring stations.

Figure 6 presents the impact of the proposed incitative method on the number of problematic stations. The results show a clear decrease in the number of problematic stations throughout the day. The average number of problematic stations drops from 164 in real trips to only 27 by slightly modifying each trip. Starting from a relatively high number of problematic stations (almost 150), users are able to balance almost all these stations within three hours. No new trips are either added or lost. The modification is done with exactly the same number of trips. The real trips are only locally modified. The obtained results confirm our intuition that resources global availability in the Velib system can be significantly improved by acting locally. This improvement would allow accepting new trips, where originally users are rejected due to a lack of bikes.

The performance of the proposed incitative method can also be measured by the number of spatial outliers in the Velib system. They consist of stations with an occupancy rate significantly different from the average occupancy rate in their neighborhood. These outliers stations are depicted in Section III using Moran Scatterplot. The comparison of the number of spatial outliers stations between the original and modified behaviors is given in Figure 7. With the improved user behavior, the number of spatial outliers drops significantly, which enhances stations homogeneity in the Velib system.

In Figures 6 and 7, all users trips are modified according to the proposed method. It is not a realistic scenario as in real life, many users will not accept to change their departure or arrival

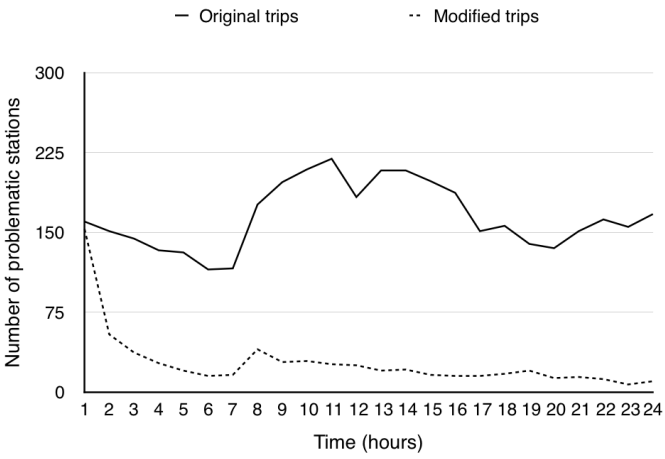

Fig. 6. Number of problematic stations

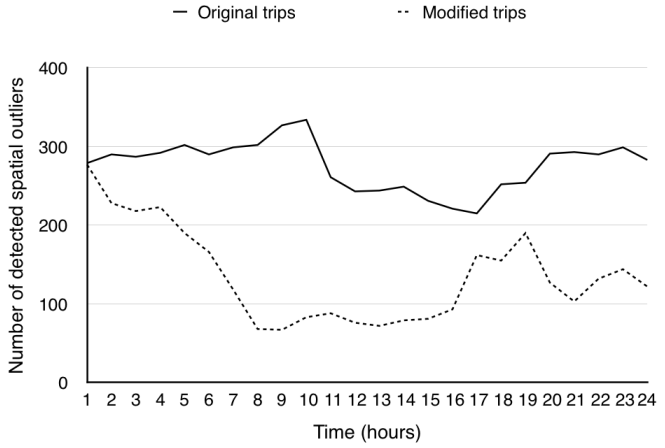

Fig. 7. Detected spatial outliers

station even if they are encouraged by a financial motivation or an extra offered time. To simulate a real-world situation, we plotted in Figure 8 the average number of problematic stations in the day under a variable collaboration rate of the users. One can see that, if only 20% of users accept to change their trips, the number of problematic stations will decrease by half. The decrease in the number of problematic stations is fast (faster than a linear decrease) which is an excellent result as we cannot expect that the majority of users will collaborate.

The number of problematic stations during the day is a good indicator to evaluate the quality of the service offered to Velib

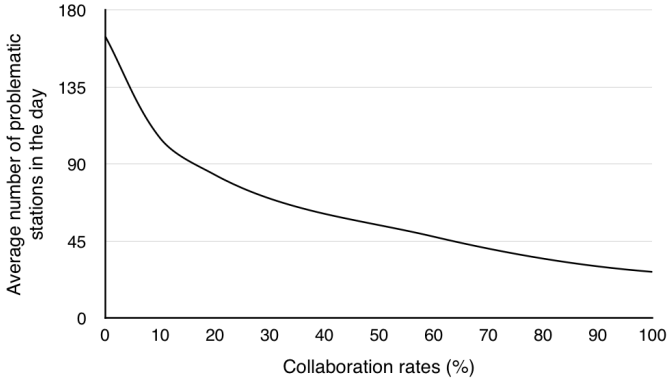

Fig. 8. Average number of problematic stations in the day

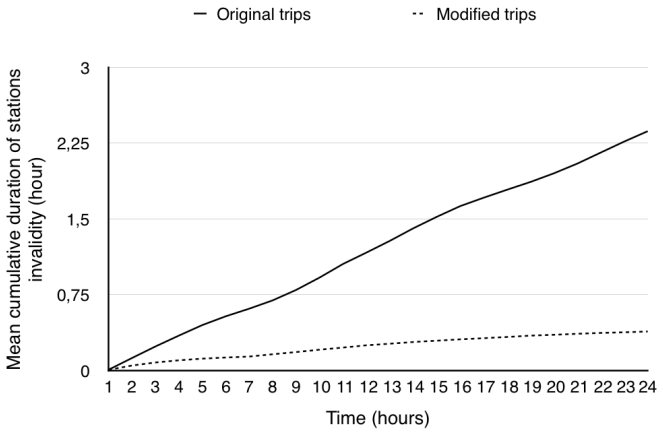

Fig. 10. Mean cumulative duration of stations invalidity

users. However, it cannot entirely qualify service availability. For a given station, the service is considered as interrupted if there is no bike or no dock in this station. In this case, the station is said invalid or out of service. Note that this concerns just one resource: bikes or free docks. To have a complete information, we plotted in Figure 9 the average duration of stations invalidity during each one-hour interval of the day, before and after the proposed improvement. One can notice that the mean duration of station invalidity has largely decreased, and likewise, the mean cumulative invalidity duration during the day has been widely improved (cf. Figure 10). According to Figure 10, at the end of the day, the mean invalidity duration of a Velib station drops from 141 minutes to only 22 minutes using our proposed improvement.

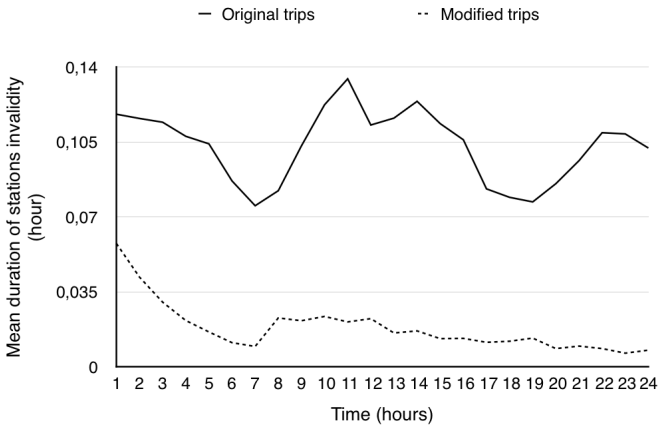

Fig. 9. Mean duration of stations invalidity

V. Conclusion

In order to improve the quality of service in the Velib system, we studied in this paper the distribution of bikes among Velib stations over time. First, we designed and applied a new version of Moran scatterplot, based on Gower's similarity, to evaluate the local heterogeneity of Velib stations in terms of occupancy rate. The obtained results showed the existence of a significant amount of spatial outliers stations. This first part was a preliminary study to motivate our proposed new method enhancing the homogeneity of the Velib system. The proposed method relies on users collaboration by locally changing their departure or arrival station. Results show the high performance of this scenario. Even if only a small proportion of users accept to apply this method, resources' availability in the Velib system can be significantly improved.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[2] Rayane El Sibai, Yousra Chabchoub, Raja Chiky, Jacques Demerjian, and Kablan Barbar. Assessing and improving sensors data quality in streaming context. In *Conference on Computational Collective Intelligence Technologies and Applications*, pages 590–599. Springer, 2017.

[3] Shashi Shekhar, Michael R Evans, James M Kang, and Pradeep Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.

[4] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46:234–240, 1970.

[5] Shashi Shekhar, Chang-Tien Lu, and Pusheng Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 371–376. ACM, 2001.

[6] Weng-Keen Wong, Andrew Moore, Gregory Cooper, and Michael Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. In *AAAI/IAAI*, pages 217–223, 2002.

[7] Chang-Tien Lu and Lily R Liang. Wavelet fuzzy classification for detecting and tracking region outliers in meteorological data. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pages 258–265. ACM, 2004.

[8] Chaosheng Zhang, Lin Luo, Weilin Xu, and Valerie Ledwith. Use of local moran's i and gis to identify pollution hotspots of pb in urban soils of galway, ireland. *Science of the total environment*, pages 212–221, 2008.

[9] Robert Haining. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, 1993.

[10] John Haslett, Ronan Bradley, Peter Craig, Antony Unwin, and Graham Wills. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, pages 234–242, 1991.

[11] Shashi Shekhar, Chang-Tien Lu, and Pusheng Zhang. A unified approach to detecting spatial outliers. *GeoInformatica*, pages 139–166, 2003.

[12] C-T Lu, Dechang Chen, and Yufeng Kou. Algorithms for spatial outlier detection. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 597–600. IEEE, 2003.

[13] Tout sur Vélib. http://blog.velib.paris.fr/blog/2014/07/15/7-ans-de-velib-des-records-de-frequentation-et-dabonnements/.

[14] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1420–1426, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[15] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Céline Robardet, Jean-Baptiste Rouquier, and Eric Fleury. Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03):415–438, 2011.

[16] 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.

[17] Y. Chabchoub and Fricker C. Analyse des trajets de vélib par clustering. *Extraction et gestion des connaissances, Clustering and Co-clustering*, 2014.

[18] E. Côme and L. Oukhellou. Model-based count series clustering for bike-sharing system usage mining, a case study with the velib' system of paris. *JACM-TIST Special Issue Urban computing*, 2012.

[19] Yousra Chabchoub and Christine Fricker. Classification of the vélib stations using kmeans, dynamic time wraping and dba averaging method. In *Computational Intelligence for Multimedia Understanding (IWCIM), 2014 International Workshop on*, pages 1–5. IEEE, 2014.

[20] C. Fricker and N. Gast. Incentives and regulations in bike-sharing systems with stations of finite capacity, special issue: Shared mobility systems. *EURO Journal on Transportation and Logistics*, 2014.

[21] Luc Anselin. *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*. Regional Research Institute, West Virginia University Morgantown, WV, 1993.

[22] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.