Analysis of bike sharing system by clustering: the Vélib' case

Yunlong Feng, Roberta Costa Affonso, Zolghadri Marc

Analysis of bike sharing system by clustering: the Vélib' case

Yunlong Feng, Roberta Costa Affonso, Marc Zolghadri

Quartz-Supmeca, 3 rue Fernand Hainaut, 93407 Saint-Ouen, France

(e-mail: yunlong.feng@supmeca.fr, roberta.costa@supmeca.fr, marc.zolghadri@supmeca.fr)

Abstract: Bicycle sharing system has become more and more popular as it can help partly solve the problems such as CO_2 over-emission and traffic congestion. Some systems have been operated for several years and the analysis work is very necessary for controlling and redesigning the system in purpose of getting better performances. In this paper we analyze the bike-sharing stations by clustering algorithms in order to mine the inner-station patterns, and these clustering results are essential for the system control and redesign. In this study, we take Vélib' bike sharing system in Paris as the study case.

Keywords: bike-sharing; service level; k-means; hierarchical clustering techniques; occupation rate; station behavior pattern

1. INTRODUCTION

With highlighted concerns about the global warming and urban traffic congestion, decision makers and policy experts look for sustainable transportation alternatives such as bike sharing. The shared bicycle fleets have been put into reality to encourage more citizens to use public transportations instead of private cars. Usually combined with bus and subway systems, the bike sharing service has become an approach to facilitate the short-distance trips within the city. The trip refers to the distance between home and the subway station or between a bus stop and the workplace, which maybe too far to walk. Thus, the bike sharing service is necessary to fill the unpleasant gap in the transportation system.

As one of the most successful large-scale bike sharing systems in the world, Vélib', was launched by JCDecaux in 15 July 2007 which encompassed around 18000 bicycles and 1230 docking stations covering Paris and its close suburbs. By July 2014, there have been more than 200 million trips and more than 274000 annual subscriptions in Vélib'. It offers non-stop service (24/7) and each station is equipped with an automatic rental terminal. An open data system about the stations' status is available on-line.

The Vélib' network will be extended in the coming years to cover new areas and will probably integrate electrical bicycles. The extension of this network raises some issues which could cause the change in several aspects: management of the current network, the definition of extended architecture (dimensioning and locations) and the enabling services such as maintenance, electrical batteries supply, etc. To solve these problems, we need to understand the existed network regarding to the static locations and dynamic behaviors.

Our first goal is to analyze the behaviors of bike sharing stations by following various performance indicators (e.g. availability rate). The user's satisfactory level can be seen as a global indicator related to the system performance, but it can not reflect the dynamics among the stations; the availability rate can show the status of available bicycles in the stations, and different stations behave differently because of their geographical locations. For instance, those stations which locate at the central business area must not have the same usage patterns compared with those which situate in the tourist places.

An extensive analysis of the system allows us therefore to determine their dynamic behaviors, which means a necessary understanding on station dynamics is required for further research works in terms of system control or redesign. In this process, it is too complicated to analyze the stations one by one; and it is also too general if we consider the station set as a whole object. Therefore, we need to separate the stations into several groups and make sure that inner-group stations are as similar as possible. As we did not have any knowledge on the stations, we choose clustering methods to help us decide the station divisions. In the literature there are numerous works which analyze bike sharing systems by clustering techniques. However, they do not give a common understanding about the number of classes that should be considered and they do not treat this problem focusing on rush hours of the day.

In this paper, we will present clustering analyses of Vélib' system by using hierarchical and partitioning approaches. This paper has been organized as follows. Section 2 discusses the related research work and gives the motivation of our research. Sections 3 and 4 describe the analysis works by descriptive method and unsupervised learning. We present and discuss the experimental results in Section 5. At last, we close the paper by some conclusions and the definition of future works.

2. RELATED WORKS

Almost all bike sharing systems open their station-usage database to public. This offers great convenience to researchers. In this research field, since several years ago, the works which were related to the system analysis, control and design or redesign flourished. Next we will only introduce those works regarding the system analysis.

2.1 System Analysis

Focusing on the system analysis, two major school of thoughts can be distinguished.

By Machine Learning Algorithms. Most of the researchers focus on the system analysis using algorithms and data structures in machine learning. Chabchoub and Fricker (2014) applied k-means in one-day trip dataset from Vélib' system by abstracting each station as a data vector and got six clusters (railway station, mixture, employment, periphery, habitation and entertainment). Sarkar et al. (2015) analyzed 996 stations included in 4.5-months data from ten cities by applying agglomerative (bottom-up) hierarchical clustering method considering station occupancy and activity level. They found four and six clusters regarding these two indicators. Vogel et al. (2014) have built the user profile and developed an analysis by k-means over an one-year database of Lyon's Vélo'v system and found nine clusters. Borgnat et al. (2011) discovered the temporal bike rental regularity and spatial traffic pattern for Vélo'v system by descriptive statistical methods jointly with k-means. Wong and Cheng (2015) found three clusters on weekdays and four clusters on weekends by analyzing station availability in Taipei's bike-sharing System. Xu et al. (2013) combined k-means and simulated annealing algorithm then applied it to the station segmentation in bike sharing system in Hangzhou. Vogel et al. (2011) implemented the clustering on about 760 thousand trip data of Vienna's Citybike Wien by k-means, expectation maximization algorithm and sequential information-bottleneck method. Their study yielded five clusters by locating the elbow point in cluster validation chart.

By Probabilistic Methods. Based on Poisson mixture models and the origin-destination flow of Vélib', the one-month data led to nine clusters (Randriamanamihaga et al., 2013) while two-month data yielded eight clusters (Randriamanamihaga et al., 2014). Fricker et al. (2012) performed mean field analysis and measured the system performance by calculating the stationary probability that a station is either empty or full. Montoliu (2012) used Latent Dirichlet Allocation to discover station behavior patterns in a Spanish bike sharing system. Chen et al. (2015) examined the Washington, D.C's system and detected the significant bike usage by sliding-window based method and selected unusual bike usage from its probabilistic distribution. They found that the most bicycle usages are located at the downtown areas, public parks, sports stadiums, and community centers. Corcoran et al. (2014) modeled the daily trip number by Poisson distribution and measured the effect of weather conditions and calendar events on bike usage data in Australia's CityCycle by multivariate regression method.

It can be concluded that among all the analysis techniques, the most widely used one is clustering. However, looking at those works that focus on Vélib' system, it might be seen that six, eight or nine clusters identified by different researchers. Therefore, there is no consensus about the number of classes.

2.2 Motivation and Contributions

Understanding such complex systems which combine random variables with a big number of stations and bikes is a hard task. To do so, as other researchers we looked for finding out a set of classes of stations by focusing on their dynamic behaviors. We performed then clusterings with more precise time scales while looking at the clustering quality indexes in order to find out the most appropriate number of classes that could help designers or controllers to do their job.

The goal was to identify the bike station service level and extract the temporal-spatial patterns of the bike usage. Service level is measured by station availability for both bike and dock. This indicator used to show the basic statistical information. We collect the station availability record from JCDecaux Open Data transferred to a visualization interface in connection with Google map. We applied k-means and hierarchical clustering algorithms to stations' dataset and compare the results through a set of quality indexes.

Our contributions can then be summarized as:

(1) Construction of station behavior database, started from March 2015.
(2) Analysis of the stations behaviors by clustering through unsupervised learning methods.

3. DESCRIPTIVE ANALYSIS

We defined the time duration and extracted data from that specific time period from Sep. 07, 2015 to Oct. 18, 2015 (six weeks), because neither public holiday nor big event was presented, so we can eliminate these exterior factors and focus on the bike-usage on weekdays. The single station data can be considered as data point, and the whole six weeks' dataset is formed by numerous discrete data points. At each point it contains the information about the station capacity, available bike number and parking place, current update time, etc.

We defined the service level as the binary measurement of availability. We consider a station as *on-service* if it has at least one available bicycle and at least one free parking place, otherwise we reckon it as *off-service*. For a station, we compute the ratio of on-service status during the six weeks' data as its service level.

We follow the time-line to implement the service level calculation: first we start from the first non-zero availability data point in our dataset, and we stop at the first-appeared zero availability point, then we calculate the time duration in the unit of second by using the update time-stamp of ending points to subtract the one from starting point, we continue the computation until the whole time line is covered. We set the accumulated time duration as

dividend and the total time duration as divisor, and finally the station service level is yielded by the quotient.

From the global view, we calculate the service level of all stations on the six weeks' data on weekdays and group the results by setting different value intervals.

Table 1 displays that no station keeps on-service status for all the time and there are only 14% of stations whose service level is higher than 95%. This indicates that the system improvement is quite necessary. To achieve this objective we need to acquire more comprehension of the system, which means more analysis work should be involved.

Table 1. Global Service Level

	100%	≥95%	≥90%	≥85%	≥80%
Stations	0	172	417	660	831
Ratio	0	14.0%	34.0%	53,9%	67,8%

4. CLUSTERING

Clustering is the process of grouping elements such that the inner-group elements as similar as possible while the elements of different groups are as dissimilar as possible. Clustering analysis methods can be classified into supervised or unsupervised learning. Unsupervised learning studies the patterns inside the input data without any previous understanding or expert knowledge about the awaited groups while supervised learning methods are dedicated to discovering input data patterns by a well-labeled training set. In our current research, as we look for the classes without previous knowledge about them, we have applied two unsupervised learning methods. Among all cluster analysis methods we have chosen two very largely used techniques: k-means and hierarchical clustering method.

4.1 K-means and Hierarchical Clustering Algorithms

K-means This algorithm groups the data points into a certain number (assume as K) of clusters. First we define k initial cluster centroids by assigning them to k data points randomly. Next we calculate the distances between every data point and its nearest centroid among the k newly defined centroids. After each point is passed through, we get k data clusters and now we need to calculate the new centroids by averaging the points in their cluster. We re-execute the data assigning phase and get newer centroids, this iteration continues until no more change is generated. Readers may refer to Hartigan (1975) for more details and examples of this technique.

Hierarchical Clustering Algorithm It expresses a series of operations which create data partitions based on the partition generated in the former process. An agglomerative type of algorithm first considers each of the n data point as an individual cluster, and then a pair of clusters is found and merged into a new one according to the global distance measurement, so the size of dataset changes to $n-1$. This process advances until a single cluster containing n data points appears. A divisive algorithm runs in the reverse way, see Johnson (1967) for more details.

The Fig. 1 shows a typical process of agglomerative (bottom-up) hierarchical clustering represented by a dendrogram. The horizontal axis represents the elements or items to classify while the vertical axis shows the distance between elements or groups of elements. For instance the distance between a and b is less than 0.1. One advantage of hierarchical method lies on its flexibility to determine how many clusters are considered. In fact, we can consider three classes if we cut the dendrogram at the distance level 0.75. In this case, the groups are: (a, b, c, d, e), (f, g) and (h, i, j). If more relevant clusters are required, the cutting level could be at 0.5 which gives 5 clusters: (a, b), (c, d, e), (f, g), (h) and (i, j).

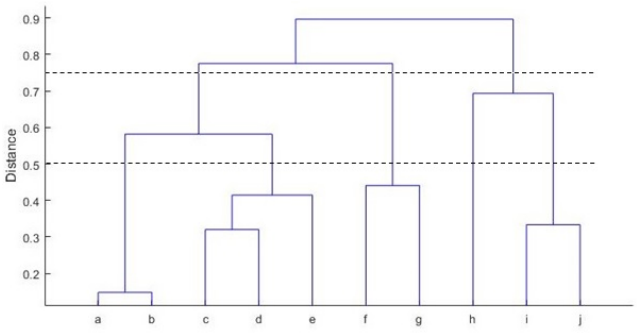

Fig. 1. Hierarchical Clustering Example

5. EXPERIMENTS AND RESULTS

Thanks to the literature review and discussions with experts of the Vélib' operator, we have defined two critical times of the day: 7:00 to 10:00 and 17:00 to 20:00 referred in the following as rush-hours.

5.1 Data Preprocessing

The raw data are always noisy (unusable values or has a long update time gap due to data server issues). So we filtered the dataset and re-constructed it before applying the clustering algorithms.

The station is represented by a data vector with specific number of dimensions. The data normalization need to be done first. For a station i, we set its capacity (the total number of bike parking places) as c_i; at time t, the available bike number is denoted as $n_{i,t}$ and we normalized the availability data by calculating the occupation rate $o_{i,t}$ in station i at specific time t, which is

$$o_{i,t} = \frac{n_{i,t}}{c_i} \tag{1}$$

Equation (1) is used to calculate the normalized availability at one single data point. We set up a time sampling period to allocate the data point to its right time interval. For instance, with a sampling period of 12 hours, we get two intervals; with one hour sampling period we get 24 intervals, see 2. Afterwards, the mean occupation rate of the intervals is calculated and the respective multi-dimensional vector is formed and is ready for clustering.

5.2 Data Representation

The sampling period affects both data vector precision and computational cost. For a large period provides the average values that can not describe the real occupation rate level (averaging bias). A very small period may contain insufficient data points because of the imbalance of data upgrades by the JCDecaux's servers and also it increases the data matrix dimension and cause too much computation. We have set six different sampling periods as shown in Fig. 2. They are: 12 hours, 6 hours, 4 hours, 1 hour, 30 minutes, 15 minutes; each point stands for the mean occupation rate of the time interval with the vertical line indicating the range of variation(mean value ± standard deviation).

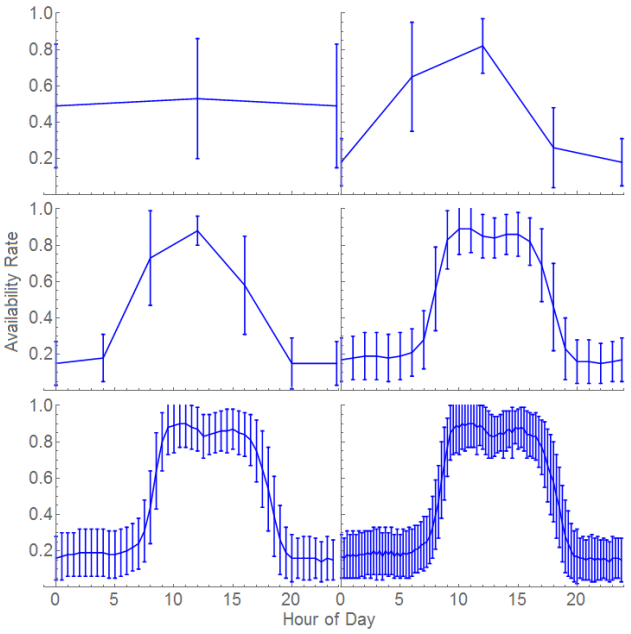

Fig. 2. Comparison of time slot sampling on station 8015

As we want to find out a suitable trade-off between the data precision and complexity, we used two different sampling periods: 30 minutes for normal hours and 2 minutes for rush hours. For clustering, we use the Euclidean Distance for both algorithms. For k-means process, we find the cluster numbers from two to twelve considering the interpretation difficulties. For hierarchical process, we apply average linkage as the criterion of the measurement between two merged clusters.

5.3 Discussion of Results

We use quality indexes explained in Rendón et al. (2011) to evaluate the clustering results as we only have the intrinsic information from the clusters themselves, so here are several cluster validation tools which are used to measure the quality of the clustering results:

- Davies-Bouldin index (Davies and Bouldin, 1979): the smaller the better.
- Dunn index (Bezdek and Pal, 1995): the bigger the better.
- Silhouette index (Rousseeuw, 1987): the bigger the better.
- Calinski-Harabaz index (Caliński and Harabasz, 1974): the bigger the better.
- Pakhira-Bandyopadhyay-Maulik (PBM) index (Pakhira et al., 2004): the bigger the better.

All these quality indexes concern about the cohesion and separation of clusters, for the calculation they use between-cluster distance, within-cluster distance, paired point distance, etc. Interested readers may refer to Rendón et al. (2011) for a complete survey of these basic concepts of the quality indexes.

Clustering comparison The general clustering results are computed on different sampling period and qualified by the aforementioned quality indexes for the different cluster number k. We pick up the 30-minute sampling slot, normalize the index results and plot them together as shown in Fig. 3.

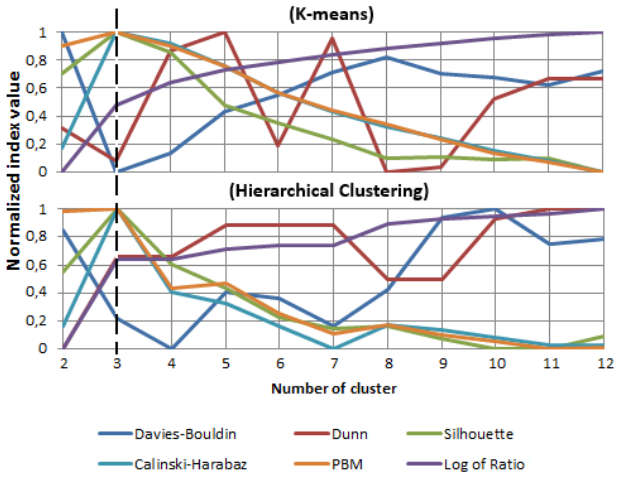

Fig. 3. Index values for k-means and hierarchical clustering

Apart from the five indexes we have also introduced another quality indicator called logarithm of ratio of summed squares (Hartigan, 1975), by which an *elbow point* is observed to help determine the cluster number k. An elbow point on a curve representing a function y of x means that the ratio between the variations of y regarding the variations of x becomes too small to be worthwhile. Therefore, often it corresponds to the best trade-off between the knowledge we get about y from varying x and the efforts we need to use to do so.

Table 2 summarizes the clustering results from the k-means and hierarchical algorithms.

Table 2. Overview of the clustering results on different time-slot

	K-means			Hierarchical method		
	1-hour	30-min	15-min	1-hour	30-min	15-min
Davies-Bouldin	3	3	3	5	4	7
Dunn index	12	5	11	12	11	12
Silhouette	3	3	3	3	3	2
Calinski-Harabaz	3	3	3	3	3	3
PBM index	3	3	3	3	3	2
Log of Ratio	3-6	3-6	3-6	3-8	3-8	4-8

As we diminish the time-slot length, the evaluation results do not change too much. Dunn index tends to get bigger

cluster number than others. It gets too many clusters that would be hardly usable for any interpretation. Moreover, looking at this index, it might be seen that it does not have a stable behavior and does not show a real trend. Silhouette, PBM and Calinski-Harabaz and Davis-Bouldin yield clearly three clusters whereas four and five clusters are also supported. Using these suggested values of the cluster numbers, we plot the dynamic behavior of the cluster centroid in k-means case to examine the temporal patterns for three, four and five clusters at the same time, shown as Fig. 4.

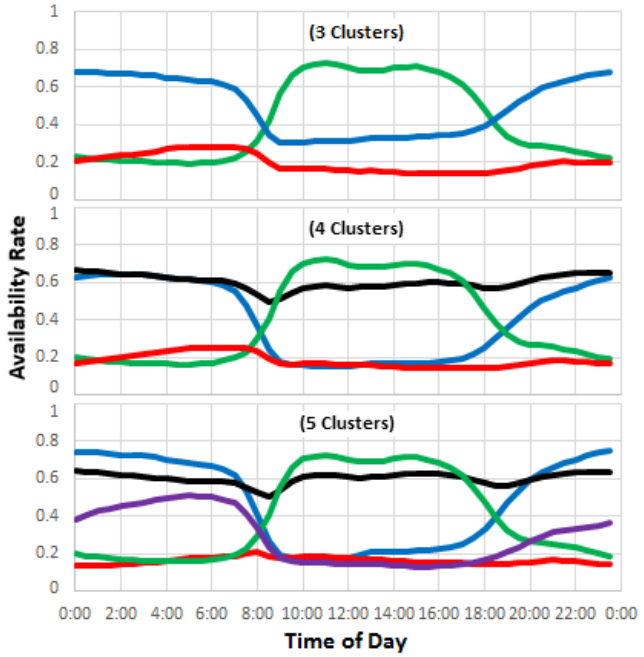

Fig. 4. Temporal patterns of general clusters on weekdays

The green curve is displayed by all three cases and it could be explained as the cluster "Employment" of which two morning and evening peaks (i.e. the low availability rate) are identified, as users arrive at the stations which are located around the working places and put the bikes off in the morning and take them away when they get off the work in the evening. Compared with the green one, the blue line shows the inverse behavior of which the occupation rate decreases in the morning and increases in the evening, which represents a pattern of "Residence". Users ride bikes away when they leave the stations in the morning and put the bicycles back when they get off work. The red and black lines, in the case of four and five clusters, both keeps relatively stable availability all the time corresponding to low and high occupation rate which can be thought as one line with different shifts. Both can cause problems. The red stations correspond to the "starving" stations while the black ones represent the "overfed" stations. The purple curve in the five-cluster figure is very similar to the blue one and it can be considered as the "semi-residential" stations combining those stations which cover mix industrial/economical and residential spaces. Therefore, from the temporal point of view, which focuses on the daily behaviors, it can be concluded that three clusters seems to be the most interesting clustering result. It defines the clusters which represent properly three types of behavior. It is quite intuitive. From spatial point of view, the stations within the same cluster tend to locate at neighbouring areas (no figure presented in the paper). Most of the stations in cluster-employment locate along the Paris's river and focus on central part of Paris, which is considered as the major working place region. Stations in cluster-residence can be found around the cluster employment with the inverse behavior. These ones are more located in residential places outside central business districts. This cluster contains almost half stations of the entire Vélib' system, most of which occupy the peripheral part of Paris.

Rush-hour Clustering We apply the same form of data vector with different time-cut to rush-hour dataset. The results are presented in Table 4. (K.M. as k-means, H.C. as hierarchical clustering)

Table 3. Overview of rush-hour clustering

			Davies-Bouldin	Dunn	Silhou-ette	Calinski-Harabaz	PBM	Log of Ratio
Morning Peak	K.M.	30-min	7	7	2	4	2	≤7
		2-min	4	9	2	2	2	≤7
	H.C.	30-min	12	12	3	6	2	≤6
		2-min	5	10	2	2	2	≤8
Evening Peak	K.M.	30-min	2	9	2	2	3	≤6
		2-min	2	10	2	2	3	≤5
	H.C.	30-min	2	7	2	2	3	≤6
		2-min	2	12	2	2	3	≤8

From Table 3 we can note that quality index results converge towards two or three clusters for evening rush hour, but the results of morning rush-hour clustering are not conclusive. The inaccuracy of index assessment is related to data sampling which considers the selection of sampling periods and affects the precision of data vector. For instance, the stations which locate at the central business area have more bicycle pick-ups and returns than the ones situated in suburban places, thus the whole dataset of station holds unsynchronized update frequency in the peak period of the day and this has an impact on station availability sampling and the clustering results.

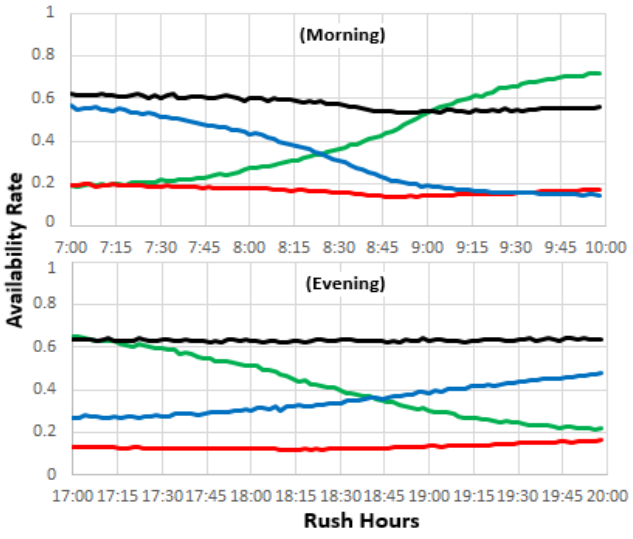

Fig. 5. Temporal patterns of rush-hour clusters on weekdays

Considering this problem, we have analyzed the behaviors of two, three and four clusters for both rush hours of the day according to the suggestion of the quality index; we exclude the Dunn index. Despite the difference between the morning and evening rush hours clustering, the behaviors of cluster are similar for these two time periods considering these three cases studied. Fig. 5 shows the 4-cluster result from which we can identify different behaviors of bicycle usages, and they are very similar to the behaviors presented in the 4-cluster clustering results as Fig 4, in condition of observing the same time range.

6. CONCLUSIONS AND PERSPECTIVES

We develop our research based on exhaustive digital footprint of Vélib'. The computed availability data allows us to carry out the descriptive statistics analysis as well as the clustering process.

In descriptive analysis we defined two station status which are respectively on-service and off-service according to the availability, then we determined the global station service level by on-service status and it came to a conclusion that the system did not reach a fully available state and the improvement is necessary. Clustering process categorized all the stations into several groups, which revealed the bicycle usage patterns. By defining different station data models, two types of clustering were implemented by applying k-means and hierarchical method in terms of identifying the station typology. Based on observations made on the clustering yielded, we think that the four clusters (employment, residential, starving stations, and overfed) corresponds the best to the reality of the system control and re-design because much clearer strategies and recommendations may be determined for them (bike transferring from overfed to starving stations for instance).

This paper offers the insights on the entire system from the view of station availability, but more research work is still needed regarding to the clustering stability for a longer time period. Also, with the knowledge extracted from clustering, we could step forward to deploy supervised learning approaches into the research and gain more understanding about the system.

REFERENCES

Bezdek, J.C. and Pal, N.R. (1995). Cluster validation with generalized dunn's indices. In *IEEE Proceedings*, 190–193.

Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J.B., and Fleury, E. (2011). Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03), 415–438.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1–27.

Chabchoub, Y. and Fricker, C. (2014). Analyse des trajets de vélib par clustering. *EGC 2014*.

Chen, L., Yang, D., Jakubowicz, J., Pan, G., Zhang, D., and Li, S. (2015). Sensing the pulse of urban activity centers leveraging bike sharing open data. *Proc. UIC*, 2015.

Corcoran, J., Li, T., Rohde, D., Charles-Edwards, E., and Mateo-Babiano, D. (2014). Spatio-temporal patterns of a Public Bicycle Sharing Program: The effect of weather and calendar events. *Journal of Transport Geography*, 41, 292–305.

Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(PAMI-1), 224–227.

Fricker, C., Gast, N., and Mohamed, H. (2012). Mean field analysis for inhomogeneous bike sharing systems. *DMTCS Proceedings*, 365–376.

Hartigan, J.A. (1975). Clustering algorithms.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254.

Montoliu, R. (2012). Discovering mobility patterns on bicycle-based public transportation system by using probabilistic topic models. In *Ambient Intelligence-Software and Applications*, 145–153. Springer.

Pakhira, M.K., Bandyopadhyay, S., and Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3), 487–501.

Randriamanamihaga, A.N., Côme, E., Oukhellou, L., and Govaert, G. (2013). Clustering the vélib' origin-destinations flows by means of poisson mixture models. In *ESANN*.

Randriamanamihaga, A.N., Côme, E., Oukhellou, L., and Govaert, G. (2014). Clustering the vélib dynamic origin/destination flows using a family of poisson mixture models. *Neurocomputing*, 141, 124–138.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1), 27–34.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.

Sarkar, A., Lathia, N., and Mascolo, C. (2015). Comparing cities cycling patterns using online shared bicycle maps. *Transportation*, 42(4), 541–559.

Vogel, M., Hamon, R., Lozenguez, G., Merchez, L., Abry, P., Barnier, J., Borgnat, P., Flandrin, P., Mallon, I., and Robardet, C. (2014). From bicycle sharing system movements to users: a typology of vélov cyclists in lyon based on large-scale behavioural dataset. *Journal of Transport Geography*, 41, 280–291.

Vogel, P., Greiser, T., and Mattfeld, D.C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20, 514–523.

Wong, J.T. and Cheng, C.Y. (2015). Exploring activity patterns of the taipei public bikesharing system. *Journal of the Eastern Asia Society for Transportation Studies*, 11(0), 1012–1028.

Xu, H., Ying, J., Lin, F., and Yuan, Y. (2013). Station segmentation with an improved k-means algorithm for hangzhou public bicycle system. *Journal of Software*, 8(9), 2289–2296.