INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Data Analysis

Bureau d'étude 2020-2021

Stations of Velib' in Paris
Instructor: Olivier Roustant

Nguyen Hai Vy
Hoang Van Hao
Bertin Alexandre
Benzitouni Fethi

4 GMM A Promo 55

- 5 mai 2021 -

Sommaire

Introduction

Bike sharing systems have become more and more popular because they can help solve problems such as excess CO2 emissions and traffic congestion. The analysis work is very necessary to control the system to obtain better performance as well as to develop appropriate strategies to meet user needs.

In this project, we are working with the 'Vélib' data set, related to the bike sharing system of Paris, which is available on R. The data are loading profiles, collected every hour, of various bike stations over one week (from Monday 12 am to Sunday 11 pm). The loading of a station corresponds to the number of available bikes divided by the maximum number of bikes that the station can accommodate. A loading of 1 means that all bikes are available. A loading of 0 means that the station is empty.

First,we are going to apply some descriptive statistics to better know and understand the data set. Then, we will use the principal component analysis (PCA) to reduce significantly the number of variables (Here the variables correspond to the hours, so we have initially 168 variables). We also compare different results obtained by studying the initial full data set and by using the PCA. Finally, we will implement various clustering methods to detect the structure of the data, in particular, finding different groups of stations corresponding to different features. From there, it is possible to predict in a relative way the user group corresponding to each station.

I Descriptive statistics

To get a global overview of this data, we calculate the service level of all stations on weekdays and group the results by different intervals of values.

Loading≥	100%	90%	80%	70%	60%	50%	40%	30%
Number of station	0	3	40	114	210	343	516	709
Ratio	0	0.0025	0.0336	0.0958	0.1766	0.2884	0.4339	0.5962

TABLE 1 – Available bike level

1-Loading ≥	100%	90%	80%	70%	60%	50%	40%	30%
Number of station	0	82	290	480	673	846	979	1075
Ratio	0	0.0689	0.2439	0.4037	0.5660	0.7115	0.8233	0.9041

TABLE 2 – Global service level

The table 1 shows that there are only 3 stations that few people use. The table 2 shows that no station keeps the status on-service at all times, there are 82 stations whose service level is greater than 90%. For those 82 stations here, it is important to increase the number of bikes to improve the system.

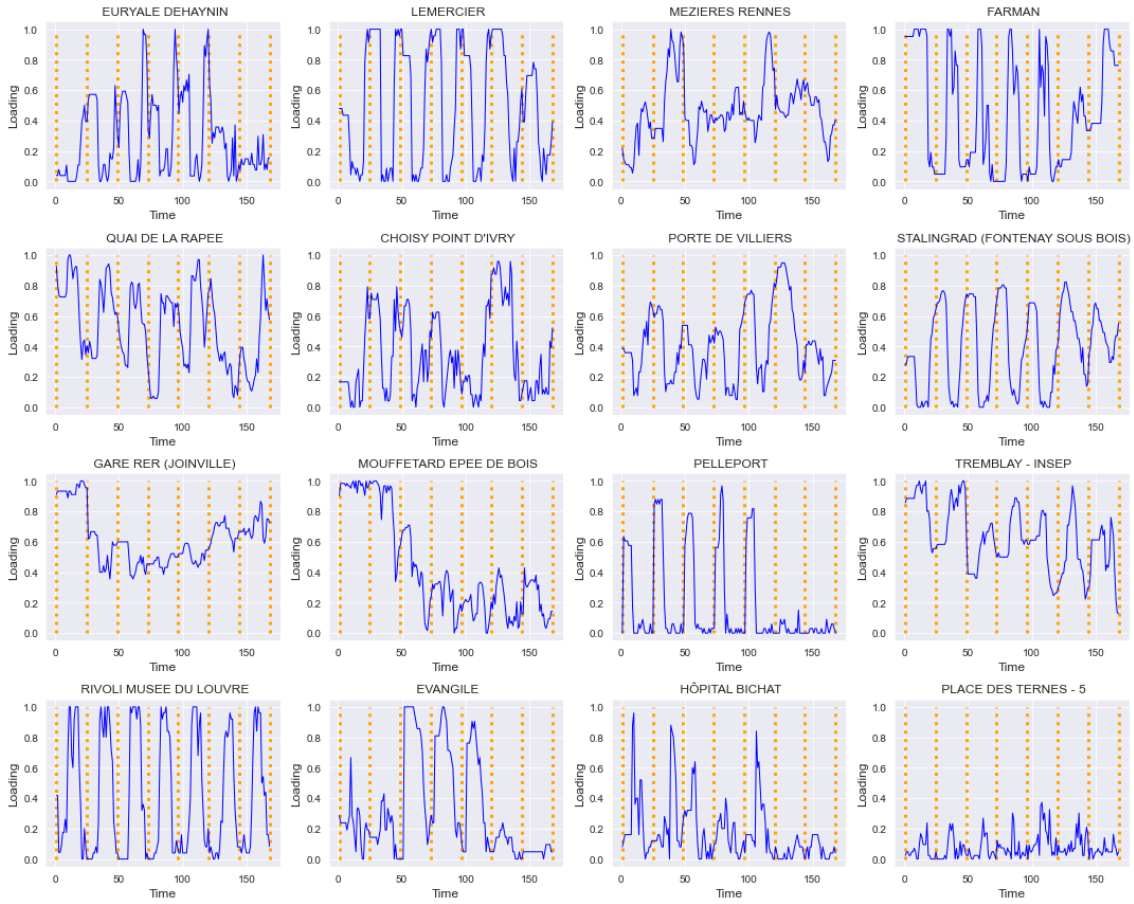


FIGURE 1 – First 16 stations analysis

The Figure 1 displays the service level of 16 first stations over 7 days. This helps us to familiarize ourselves with the data. Also, it gives us some basic information about the data. In particular, we

observe a form of oscillation over time of loading level in the stations. So, the time has to be an important detail to interpret the behavior of different groups that we will precise in the following parts.

The Figure 2 shows that there is a general trend that is a daily cycle.(although the amplitude of the oscillation changes, the vibrations follow the same shape). Furthermore, we also find that the variability is quite stable over time. The loading is lowest at around 19 or 20h, which means that there are many people using bikes around this time. This is quite logical because that is the moment people goes home after work. Moreover, we can clearly see that there is a strong discontinuity between 9h and 10h for the weekdays.This indicates that there is a sudden decrease in bicycle loading in the stations, or in other words, this is the time of day when the stations start to be "active". However, we do not observe this discontinuity during weekends, therefore we can deduce that the starting time of office hours is the reason for this discontinuity.

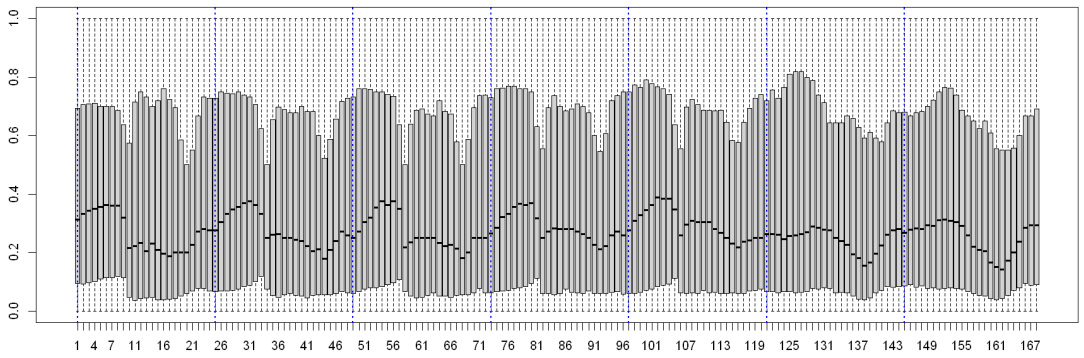


FIGURE 2 – boxplot all station

The Figure 3 displays the boxplots of all stations on different hour of Monday.

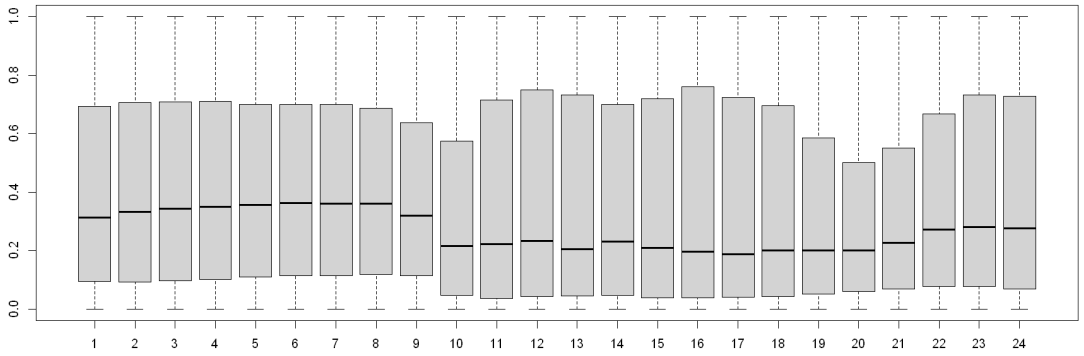


FIGURE 3 – boxplot all station on Monday

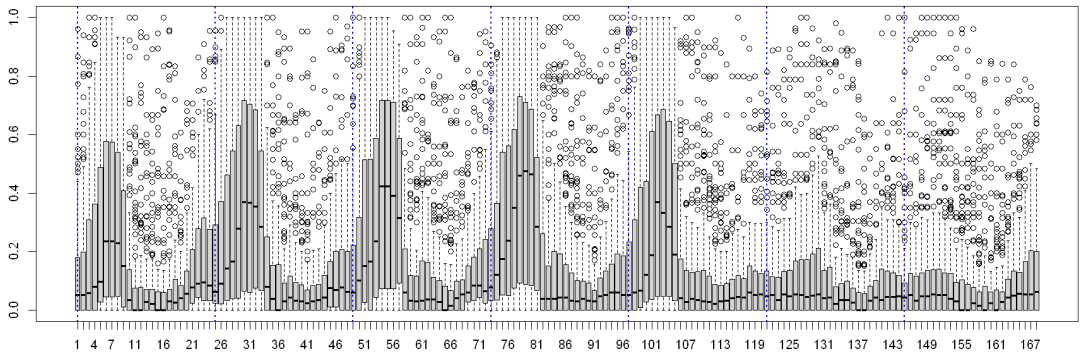


FIGURE 4 – The boxplots of stations on hills

From the figure 4, we notice that there are many outliers. This implies that behavior of bike loading of each station is quite different from each other, i.e using of bikes between stations is not homogeneous. Furthermore, from 9 a.m to 4 a.m the following day for weekdays and all day of weekends, the ratio of the available bike is very low, very close to 0. This shows us 2 possibilities :

— The bike is very popular for people who lives on hill and maybe people tend to hang out on the hills on weekends.

— The number of bicycle stations is too few to serve the people on hills.

Moreover, we notice that the distance between the "active moment" and "non-active moment" is quite high. So, we can deduce that on the weekdays, there is a strong dependence of loading on time slot.

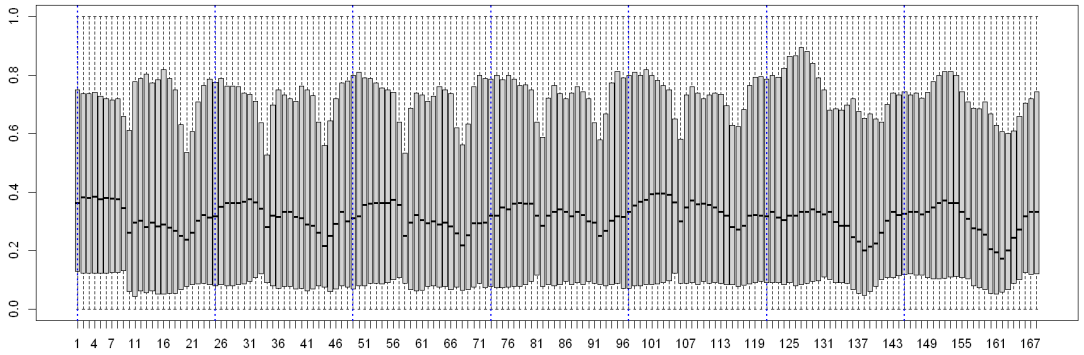
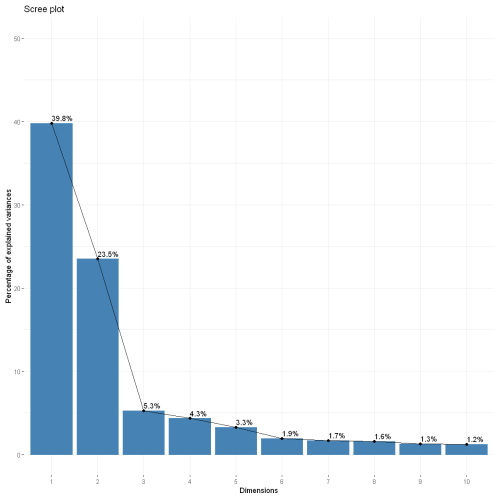


FIGURE 5 – The boxplots of stations not on hills

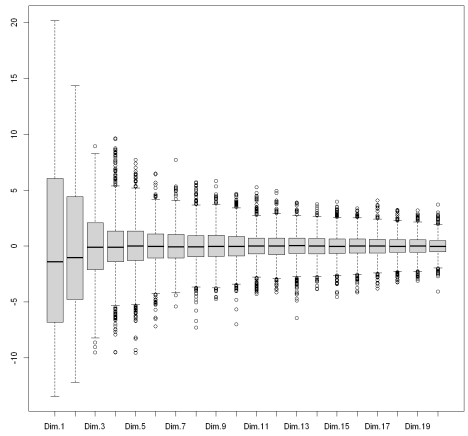
Observing the Table 5, we notice that the curve of the median is rather stable over time and the phenomenon of discontinuity is very small. We can deduce that bicycle using is very stable on regions that are not on hills. This is reasonable because in non-hill areas, there are many production and service activities and the population segment is also more diverse, so working hours and living activities of people are so diverse, so there is no sudden change on bicycle using. Therefore, contrary to hill-areas, the bicycle using in the non-hill-areas is much less dependent on the moment of the day.

II PCA Method

The PCA is a process which aims to change the coordinate space. All news variables are linearly uncorrelated. The process to construct these news variables allows for the user to choose those which explains the most the inertia of the system studied.



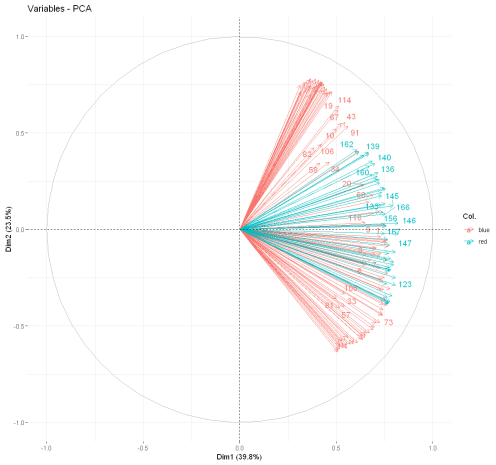
(a) Percentage of variance explained by the first 10 components



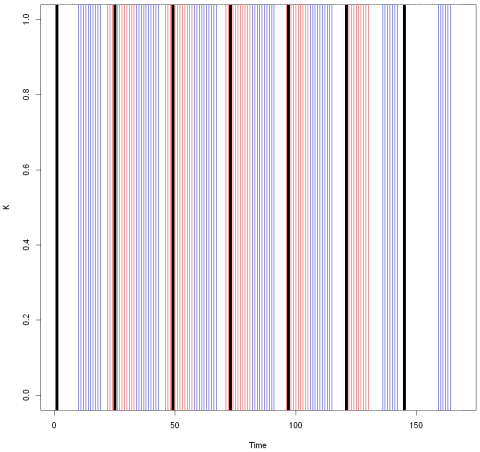
(b) Boxplots of first 20 principal components

FIGURE 6 – PCA graph 1

The Figure 6a displays us the percentage of variance explained. The first two components carry over 60% of the information contained in primary variables. From Figure 6b, we observe that the size and the median of these boxplots are do not change so much after 5 components. Hence, we may keep also components 3, 4, 5. Besides, we also see a slight jump in the Figure 6a after the 5-th components. Eventually, keeping 5 components is justifiable here.



(a) Variables factor map



(b) Contrast effect on the second principal component

FIGURE 7 – PCA graph 2

In the Figure 7a, we color weekdays in red and weekends in blue.

— First, the length of almost all the arrows is approximately close to one, thus most of the variables are close to their projections on the first two components. We can notice that on the first principal component, the abscissas of all the variables are strongly positive, and between 0.4 and 0.6, hence the first component is approximately 0.5 times the sum of service level during the week. In other words, the first component represents the average loading number.

— Second, all the vectors are located equally in 1st quadrant and 4th quadrant. It may represent a contrast between variables with a positive ordinate, and variables with negative ordinate.

— Third, we can clearly see that the weekdays lies around the first component in positive sense and their ordinate (i.e 2^{nd} component) is rather small, so information for weekdays is strongly included in the first component.

Next, to check the contrast effect on the second principal component, we show the 7b, we color the time steps in blue if their value on Dim 2 is greater than 0.25 or in red if their value on Dim 2 is lower than - 0.25. From the Figure 7b, we clearly see that there are two variables are associated with night and day during the week (except for the week-end). So we may conclude that the second component highlights the effect of contrast between night loading and day loading during the 'active' days.

Finally we could say that the loading profiles of a station could be mainly explained by these situations :

— Time :
  — During weekends or not.
  — During the day or in the evening.
  — During working hours or not.
— Location : Being on a hill or not.

III Clustering

III.1 Hierarchical clustering

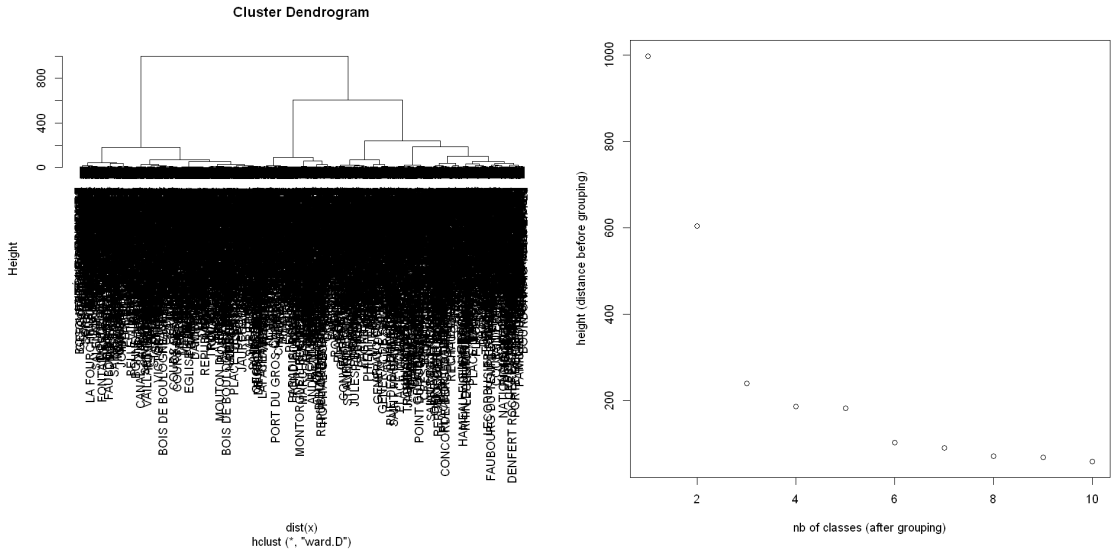


FIGURE 8 – Cluster Dendrogram and Distance before grouping vs number of class

Here, we use the hierarchical clustering method with the Ward criterion. The Figure 8 shows the dendrogram and the heights versus the number of classes. Based on these plots, choosing 6 classes seems logical. By observing the right panel of Figure 8 from right to left, we search for sudden changes in height, called 'jumps' : the first jump is between 6 and 5 classes. The second jump is between 4 and 3 classes. Thus we may choose 6 classes or 4 classes. We do not consider the next jumps, because if we do so, we would gather classes that are far from each other. Hence, we have to choose between 6 and 4 classes. Here 6 classes may be more suitable than 4, to be able to recover the diversity of the customer behaviors. When cutting the dendrogram with 6 classes (height around 100), the size of the classes seems large enough. In conclusion, the choice of 6 is most reasonable. Next, we plot 6 groups on the plane of the two first components of PCA. From the Figure 9, we can see that the 6 groups are quite well separated.

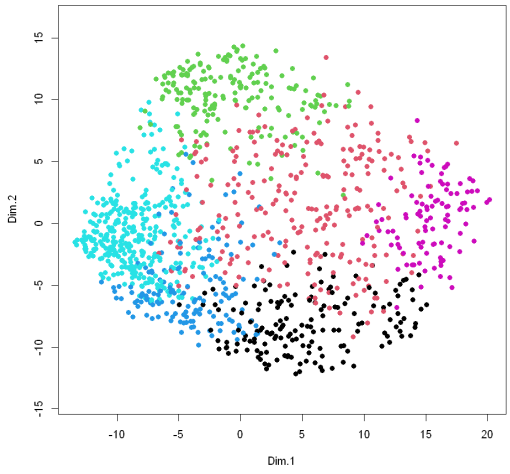


FIGURE 9 – Graph of each group projected on the two first components of PCA

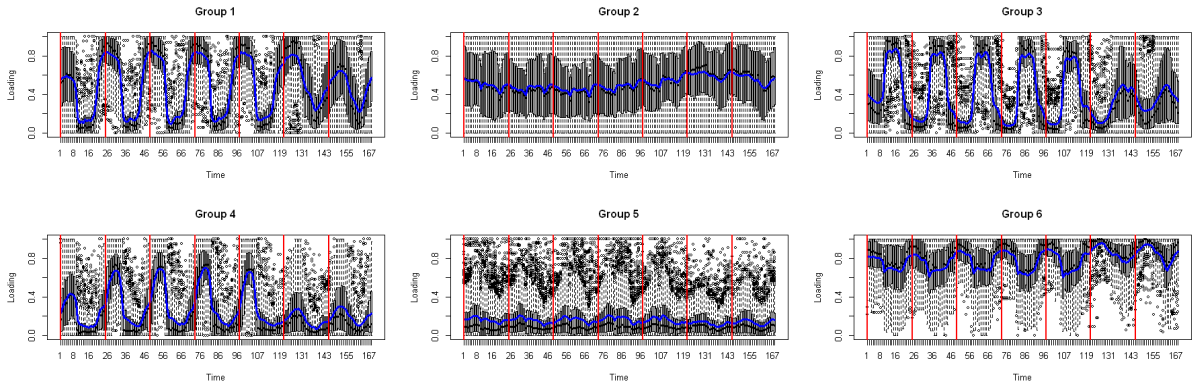


FIGURE 10 – The center of each class

The Figure 10 shows us the mean of each class. From The Figure 10, we can characterize each class :

— The center of class 1 corresponds to a high daily usage in the middle of the day and a low daily usage during night time. This difference in usage between day and night time is smaller during weekend. So,it may be reasonable to deduce that a big part of usage during weekday is for working purposes. Moreover, during weekends, the usage in nighttime if higher than during weekdays (corresponding to smaller loading). To explain this phenomenon, there are many reasons, we suppose here 2 hypothesis : (1) at night, people go out on weekends more than during the week, or (2) there are more night work on weekends than during the week.

— The center of class 2 corresponds to a relatively equal usage at every hours of the week. We suppose here a hypothesis that there are different production and service activities near this area with different time zones of operation all day and hence, users are more diverse and use bicycles in different moments of the day. This explains the reason why there is not much difference in usage between daytime and nighttime.

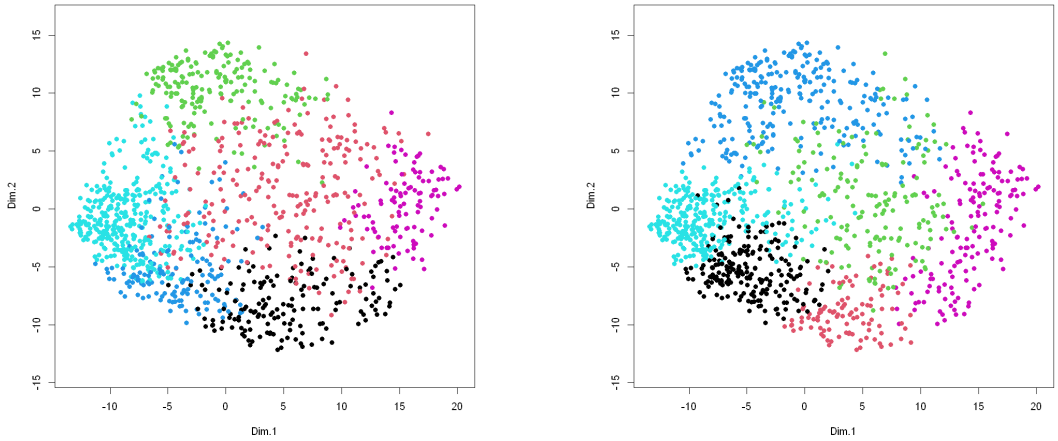
— The center of class 3 corresponds to a relatively higher daily usage in the morning than in the afternoon. The center of class 3 is in phase opposition to center of class 1 with almost the same amplitude.This phenomenon can be explained by a complementary of class 1 and class 3. We could think the bikes from the class's stations 1 are used to load the class's stations 3 and vice-versa. Besides, class 3 has the high loading rate during office time, we can say that class 3 is located near workplaces and class 1 represents a pattern of "Residence".

— To characterize the center of class 4, we will consider weekdays end weekends. During weekdays, the center corresponds to a relatively higher daily usage in the afternoon until night than in the early morning. This may be related to a particular group of users, for example, people working in this time frame. During weekends, the usage stays high all day, even in the early morning. So, it is quite reasonable to predict that this difference is due to tourists or people working during weekends. If we compare class 4 vs class 1, we can observe a net difference. That can be seen during the weekend. The class 4 is less loaded than the class 1. That can be explained by the nature of area. Class 1 is more linked to housing areas. While class 4 is linked to semi-housing areas. In class 4 we could observe residences but we could also observe pub, nightclub, bars, all places of entertainment during the night.

— The center of class 5 corresponds to a high daily usage at every hours of the week. The loading stays always in a very low level. This shows us that usage demand of customers in this group are very high.

— The center of class 6 corresponds to a low daily usage at every hours of the week. Hence it is not necessary to have too many bicycles in this group of stations. As a balance, the service provider might consider bringing a portion of bicycles in this group of stations to a group of stations with high user volumes, for example, group 5.

Now we perform the hierarchical clustering method on 5 first principal components instead of full data. The Figure 11 shows that we obtain the groups which are almost similar. This proves once again the suitability of the selection of the 5-component of PCA that we mentioned above.



(a) Group separated by hierarchical method on full data

(b) Group separated by hierarchical method on 5 first principal components

FIGURE 11 – Individual map of each group on first two principle component

III.2 K-means clustering

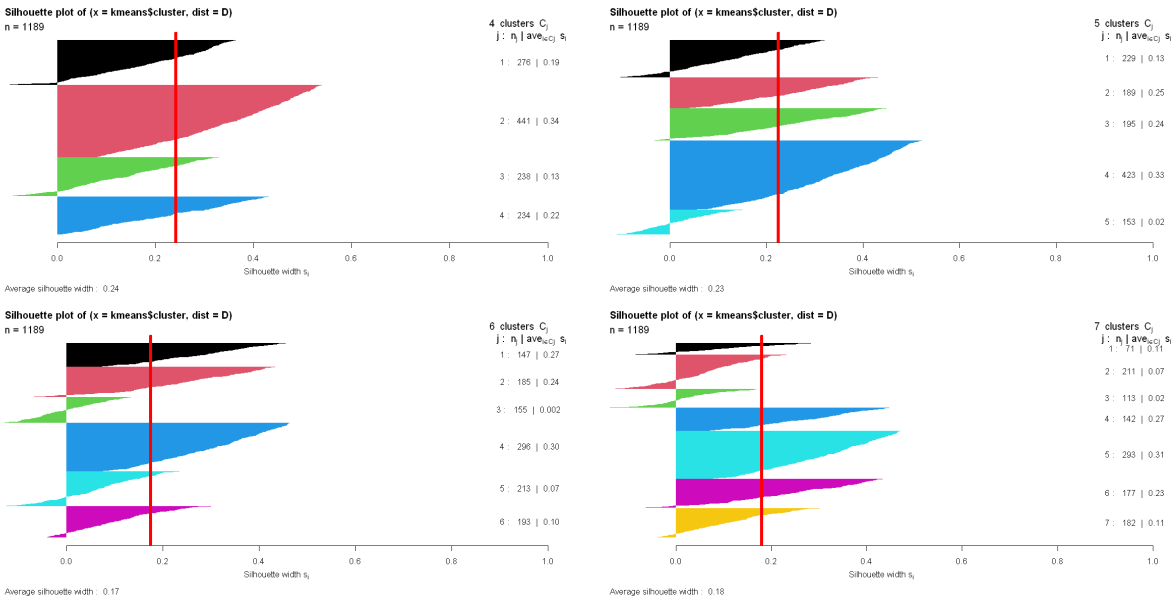


FIGURE 12 – The Silhouette plot by changing number of clusters

The Figure 12 shows that the number of clusters values of 5, 6, and 7 are a bad pick for the given data due to the presence of clusters with below-average silhouette scores and also due to wide fluctuations in the size of the silhouette plots. Also from the thickness of the silhouette plot the cluster size can be visualized. The non-homogeneity of the thickness of each group can be noticed for the number of clusters value of 5, 6, and 7. However, when the the number of clusters is equal to 4, all the plots are more or less of similar thickness and hence are of similar sizes. So we choose 4 as the number of clusters

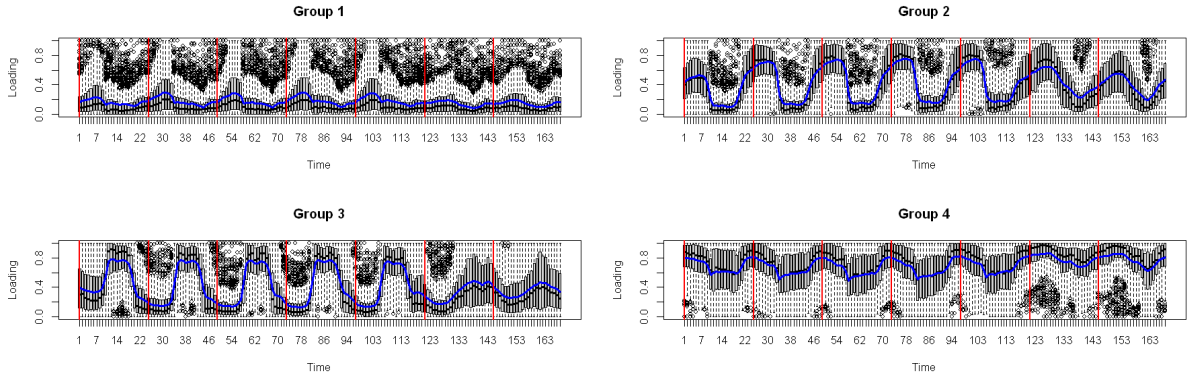


FIGURE 13 – The center of each class

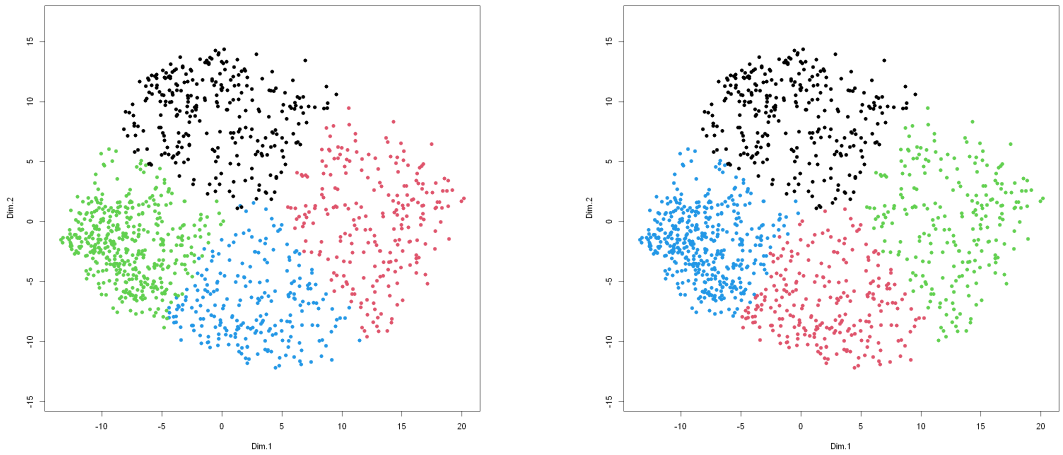
From Figure 13, we can notice that

— The center of class 1 corresponds to a high daily usage at every hours of the week. Normally, a particular group of users only use bicycles for a fixed time frame. In this case, the usage is high around the day so we can speculate that the users of this group of stations are very diverse : people working in different moments of the day, tourists... This group is similar to group 5 in CAH method.

— The center of class 2 corresponds to a relatively higher daily usage in the middle of the day than in the beginning and the end of the day. The center of class 2 corresponds to stations placed in a housing area. Indeed, it's during the night time that the stations are the most loaded. That means that the bikes are less used. There is a reversal of the trend when the sun rises and the sun goes down. It's between (8 a.m to 17 .pm) that stations are the less loaded. This class is similar to groups 1 and 4 in CAH method.

— Contrary to class 2, the center of class 3 corresponds to a relatively higher daily usage in the beginning and the end of the day than in the middle of the day. The class 2 and the class 3, both keep relatively stable availability corresponding to low and high loading rate during office hours (8 a.m to 17 .pm) which can be located in different workplaces with different shifts. The class 2 and the class 3 are in opposite of phase. The two classes are linked to each other. This class is similar to group 3 in CAH method.

— Contrary to class 1, the center of class 4 corresponds to a low daily usage at every hours of the week. It is similar to group 6 in CAH method. So, please refer back to group 6 in CAH method for interpretation.



(a) Group separated by Kmeans method on full data

(b) Group separated by Kmeans method on 5 first principal components

FIGURE 14 – Individual map of each group on first two principle components

Next, we plot 4 groups on the plane of the two first components of PCA by performing two different methods : K-means on full data and K-means on 5 first principal components. We can see that the 4 groups are very well separated. All individuals of each group are almost not mixed with other groups. Besides, The Figure 14b shows that we obtain the groups almost similar. This proves once again the suitability of the selection of the 5-component of PCA that we mentioned above.

III.3 Gaussian Mixture Models

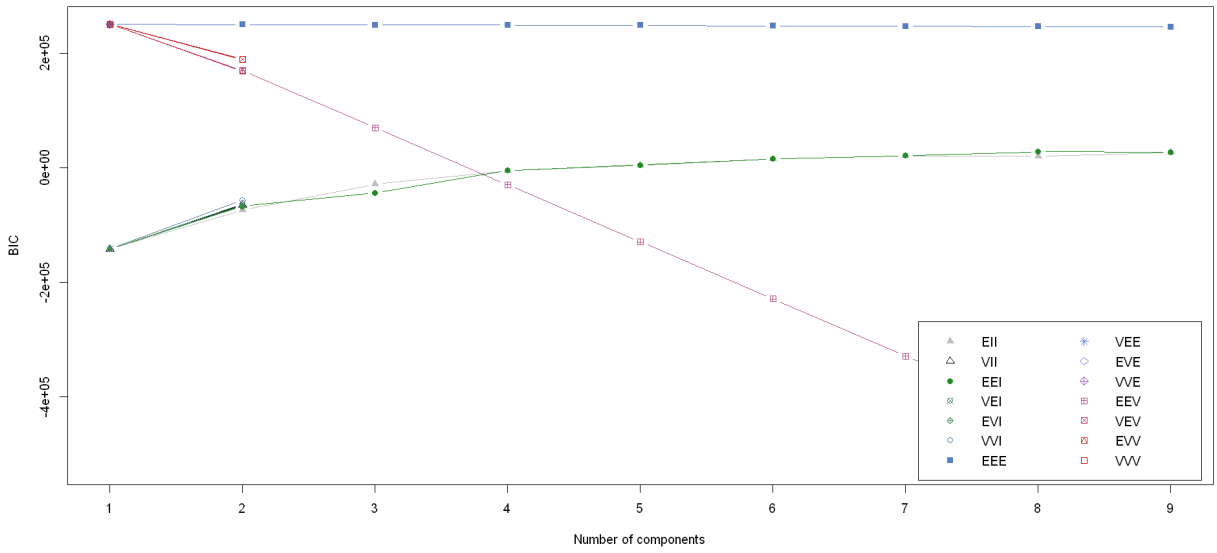


FIGURE 15 – The evolution of BIC on several in-build models

From Figure 15, we see that EEE model is markedly better than all the other ones. So we perform EEE model on the full data with different number of classes : 4,5,6 and 7. We get the following table with number of elements in each class.

№ of classes \ № of elements	1	2	3	4	5	6	7
4	639	548	1	1			
5	644	542	1	1	1		
6	644	541	1	1	1	1	
7	647	537	1	1	1	1	1

TABLE 3 – The performance of EEE model on full data by changing the number of classes

The Table 3 shows us the number of individuals in each group when performing the EEE model on full data. We can clearly see that the groups separated by this method are not balanced. In particular, almost all of the individuals are contained in the first two groups. So we would rather find another method to obtain better separation of groups.

Now, instead of taking all data, we perform the GMM method on just 5 first principal components. The Figure 16 displays the effectiveness of the GMM method with several in-build models.

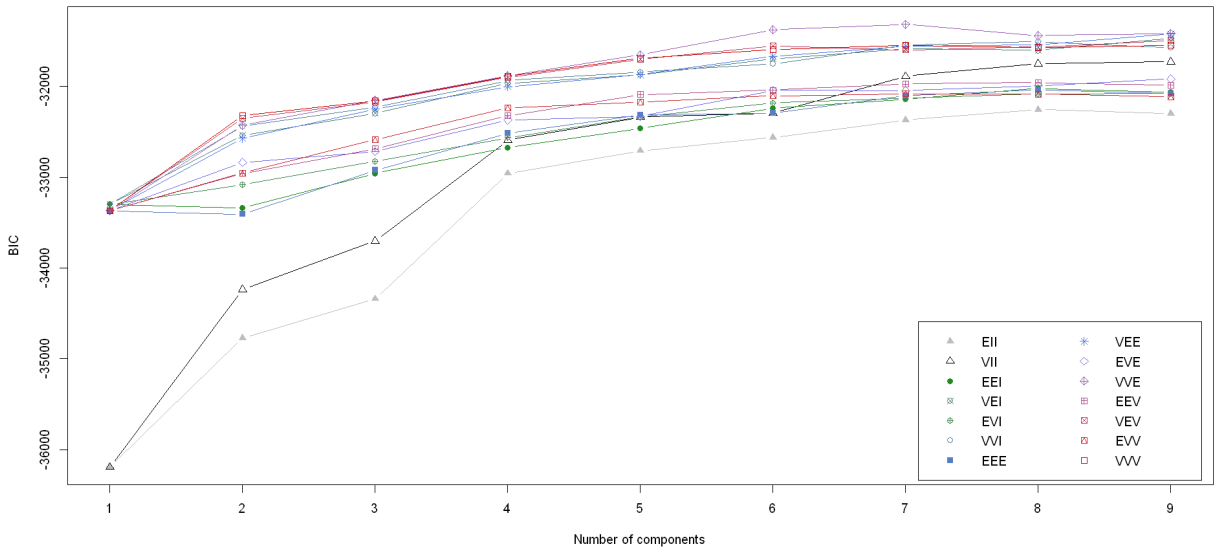


FIGURE 16 – The evolution of BIC on several in-build models on first 5 principal components

We can notice that the VEE model maximizes the BIC value, so it is reasonable that we choose this model for clustering.

Next, we compute the number of individuals in each group when changing the number of groups from 4 to 7. The results are shown in the Table 4.

Nº of classes \ Nº of elements	1	2	3	4	5	6	7
4	241	611	137	200			
5	263	435	150	164	177		
6	238	138	178	308	145	182	
7	213	138	174	271	102	141	150

TABLE 4 – The performance of EEE model on first 5 principal components by changing the number of classes

We observe that we do always have the balance of the number of individuals in each group when changing the number of groups from 4 to 7. On the other hand, from the Figure 16, we see that the BIC value is maximized when the number of groups equal to 6 or 7. But seven is quite many as the number of classes. So it is reasonable to choose 6 as the number of classes.

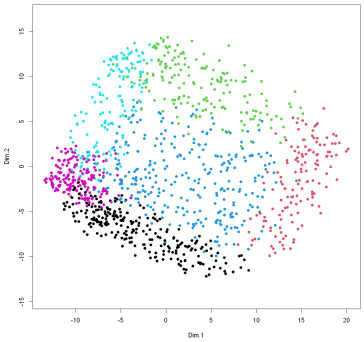


FIGURE 17 – Individual map of VEE Gaussian mixture model on first two principal components with $n\_cluster = 6$

From the Figure 17, we can see that 6 groups are well seperated.

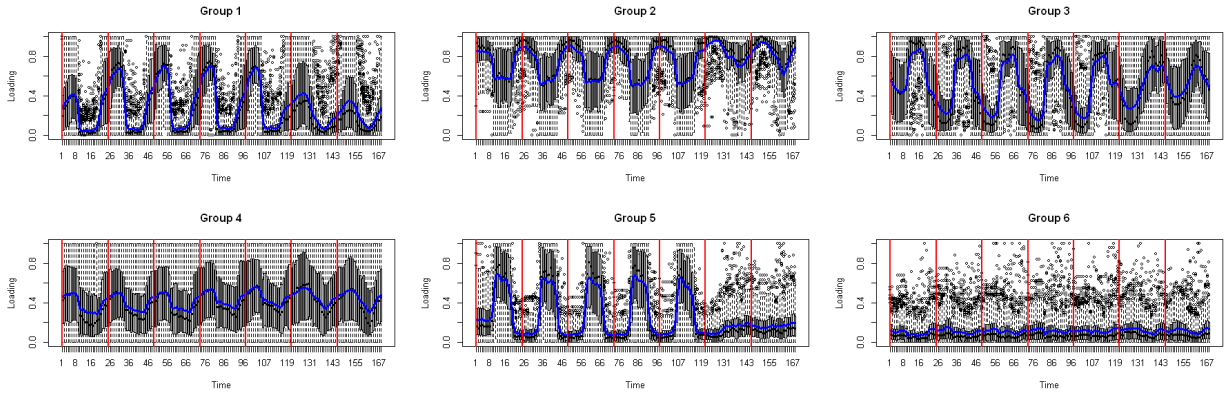


FIGURE 18 – The center of each class separated by VEE Gaussian mixture model on

The Figure 18 shows us the mean of each class. Indeed, the center of each group by using this method is quite similar to Hierarchical clustering method. In particular, we notice that :

— Group 1 of GMM method corresponds to Group 1 of Hierarchical clustering method.

— Group 2 of GMM method corresponds to Group 6 of Hierarchical clustering method.

— Group 3 of GMM method corresponds to Group 3 of Hierarchical clustering method.

— Group 4 of GMM method corresponds to Group 2 of Hierarchical clustering method (We can notice that the amplitude of oscillation for group 4 in GMM method is slightly larger than that of group 1 in Hierarchical clustering method but the interpretation of meaning does not change much).

— Group 5 of GMM method corresponds to Group 4 of Hierarchical clustering method

— Group 6 of GMM method corresponds to Group 5 of Hierarchical clustering method.

So the interpretation of each group is similar to Hierarchical clustering method.

III.4 Plotting the position of each station of each group on the real map

To have more information about each group, we trace the position of each class on the real maps of Paris. We have also created an interactive Google Map that displays the position of each station of each cluster provided by K-means method. For more information about the map click on the following link :
`www.google.com/maps/d/u/0/edit?mid=1AvZY-iASyBBcwQV-nUxYtgiIrn-SOx0R&usp=sharing`
From this map or the Figure 19, we can notice that

— All the stations of class 2 are located near parks and playground

— All the stations of class 3 are in the center of Paris, located along the Seine and distributed around famous tourist destinations (Eiffel Tower, the Louvre Museum, ...)

— All the stations of class 1 and class 4 are widely located throughout Paris. So we have not found an appropriate interpretation yet.

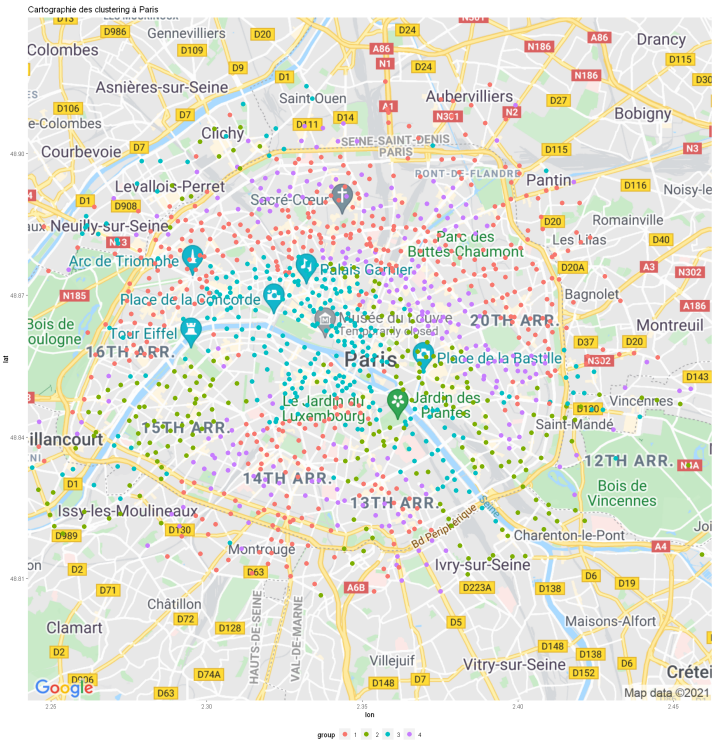


FIGURE 19 – Real map analysis of each class

Conclusion

This report studies the Vélib' system from several aspects, including system characteristics, availability, the location on the hill, and the position of each station of bicycle sharing systems on the map.

In **Descriptive Analysis**,

- We determined the station status which is respectively on-service and off-service according to the available bike level and global service level. It reached a conclusion that no station keeps the status on-service at all times and the improvement on several stations is indispensable.
- We determined the dependence of bicycle using on the different moment of the day that correlated with the shift-work.
- We analyzed the behavior of the system by location (on the hill or not on the hill). It turns out that bicycle using is very stable on the regions that are not on hills.

In **Principal Component Analysis**,

- We reduced the dimensionality of our datasets from 168 to 5. From these 5 principal components, it turns out that we identified new patterns in our datasets based on the correlation between features.
- We determined the dependence of bicycle using on day and night.

In **Clustering Section**,

- We inspected the behavior of three methods (each method with plural in-built models) of classification to find the strengths of each method.
- We realized that although the algorithms are different, some methods could lead to quite similar results.
- We came to figure out that PCA is really useful in the case of a large dataset. When performing on full data or PCA data, the results of classification are quite similar. Besides, if we performed the GMM method on the data reduced by PCA, we could even get better results.
- We find out that each cluster is linked to the presence of nearby infrastructure such as parks, railways stations, rivers, etc. If we had more referenced datasets about inhabitants, jobs, etc..., we would gain more information about each cluster.

Most of the ideas in this report were proposed by Prof O. Roustant. We rely on these big ideas to come up with concrete methods and interpretations. Our coding in Python and R is much inspired by TP-courses given by Prof O. Roustant in program of GMM, INSA Toulouse. So we would like to send a special thank you to Prof O. Roustant for his help during the work for this report.

INSA Toulouse
135, Avenue de Rangueil
31077 Toulouse Cedex 4 - France
www.insa-toulouse.fr

Université Fédérale Toulouse Midi-Pyrénées

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'ÉDUCATION NATIONALE,
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE