INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Department of Applied Mathematics

Data analysis: Vélib data

Analysis of the loading profiles of the Vélib stations in Paris

---

Juan AYALA

Jeong Hwan KO

Alice LALOUE

Aldo MELLADO AGUILAR

4th Year Applied Mathematics - Groups A&B

Date of submission: May 14th 2021

Contents

1 Descriptive Statistics **1**

1.1 Basic database descriptions 1

1.2 A first look at the behaviour of the loading through time 1

1.3 Elementary conclusions after descriptive statistics 2

2 Multidimensional Analysis **3**

2.1 Correlation on station loading and date 3

2.2 Loading profiles based on the altitude 4

2.3 Main conclusions from multidimensional analysis 5

3 Principal Component Analysis (PCA) **6**

3.1 Reducing data dimension: number of principal components 6

3.2 Interpreting the principal components 7

4 Clustering on the original data **9**

4.1 Ascending Hierarchical Classification 9

4.2 K-Means 10

5 Clustering on the coefficients of a suitable functional basis **12**

5.1 Ascending Hierarchical Classification 12

5.2 K-Means 13

5.3 Gaussian mixture model 13

5.4 Clustering conclusions 14

In this project, we have the data of loading profiles of the 1189 Vélib (public bicycles) stations, scattered all around Paris. The loading corresponds to the filling percentage of a given station: a loading of 1 means that the bike station is full, and 0 means that there are no bicycles left. The loading is measured every hour from September 2nd 2014 at 00:00 am to Sunday 9th at 11:59 pm. The goal of this project is to detect common customer behaviour, what we will call clusters. Our analysis will be done with the help of **R** and **Python**.

1 Descriptive Statistics

1.1 Basic database descriptions

We have two databases. The first one, called `velibLoading`, is the database containing 168 variables and 1189 individuals. Each variable corresponds to the time of measurement of the station's loading, that is, one measurement per hour for 7 days. Every variable in this database is quantitative.

The second one, called `velibAdds`, contains the same 1189 individuals (in the same order), but this time it contains the stations' name, longitude, latitude, and a variable called "bonus" that indicates if the station is located on a hill (1), or not (0). The variable "bonus" and the coordinates of the stations are quantitative variables, and the name of the station is qualitative. Around 10.68% of the stations are on hills.

1.2 A first look at the behaviour of the loading through time

In order to get a general idea of the loading's behaviour, we start by plotting the loading profile evolution through time for the first 16 stations on the list.

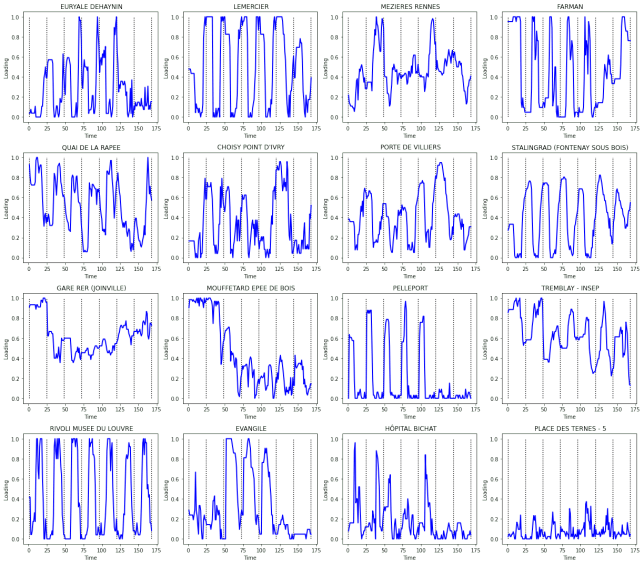

Figure 1: Loading profile evolution through time.

These graphs show that, in general, the loading is periodical, following a similar pattern each day. The seasonality is as following: a high loading at night, but lower during the day. This can be explained by the fact that the bikes are used more during the day to move around in the city (for instance, going to work), whereas during the night the activity is generally lower.

However, the popularity of the stations vary. For example, there is a clear difference in the loading profiles of the 12th and the 16th stations compared to all of the other ones that are plotted above, which we will explore further on.

Then we plot 168 box plots side by side corresponding to each variable (time of the week). These box plots are also separated by blue vertical lines to represent the seven days of the week.

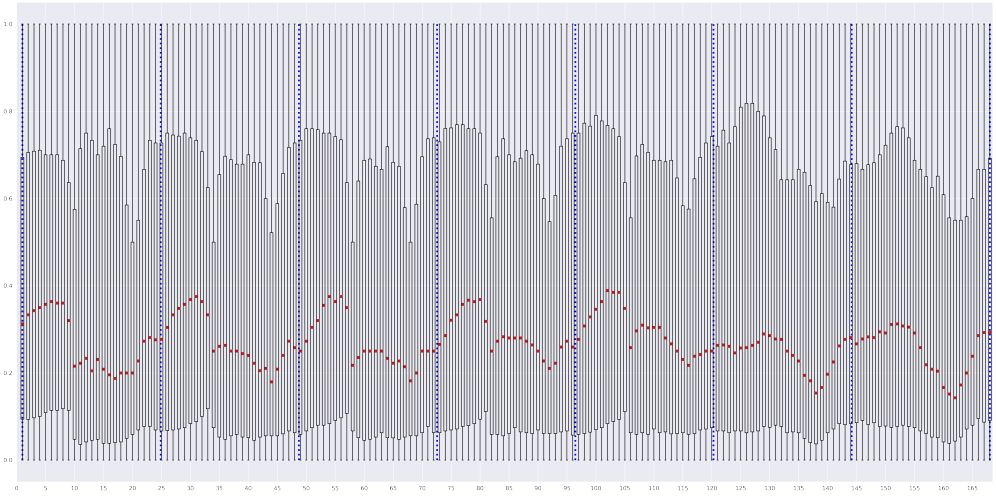

Figure 2: Loading through 7 days.

This graph shows that the median behaves a certain way during the weekdays, and a different way during the weekend. This reflects how bikes are used differently during the week days from the weekends. During the weekdays we see that the median reaches a peak at around 8:00 AM and then it suddenly drops. This might be due to the large amount of people who use the bikes to go to work or to school at that time. During the weekend the median reaches a minimum value at around 5:00 PM, which might occur because of all the people who like to go out on the weekends to have fun.

Furthermore, we try to find any correlation between two chosen variables. Thus, we plot the loading of a given time versus the loading with a slight shift in time to see if there is a linear correlation.

1.3 Elementary conclusions after descriptive statistics

Our first analysis on the variables strongly indicates that the loading varies through time. We can even affirm that there is a clear seasonality, with the loading being generally higher during night times and lower during the day. Finally, the loading on the weekends has different properties.

2 Multidimensional Analysis

2.1 Correlation on station loading and date

To visualize the linear correlation between all of the variables in `velibLoading`, we plotted the correlation table.Two variables will be strongly correlated if the absolute value of their correlation is close to 1, and the opposite for a correlation close to 0.

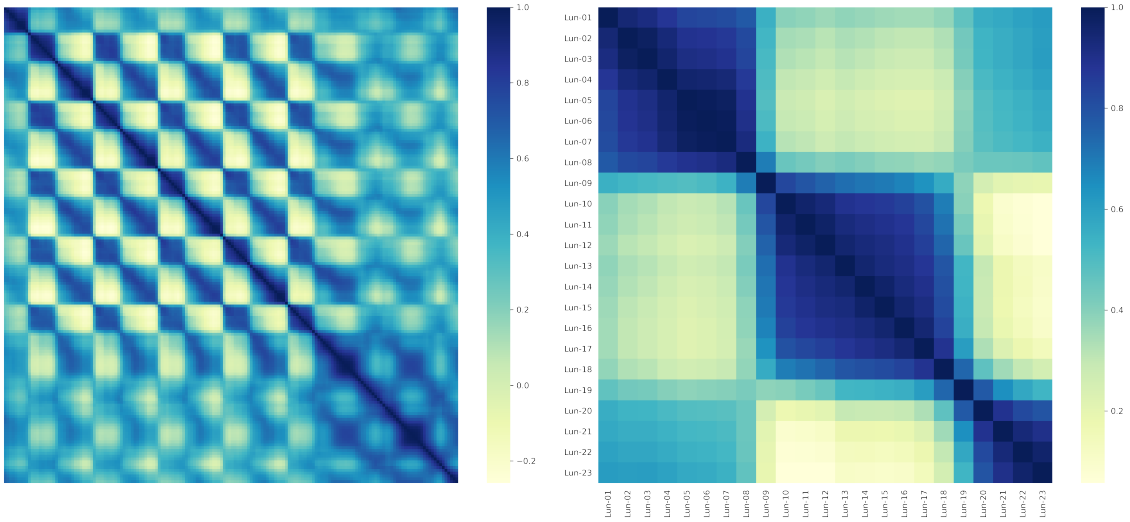

Figure 3: Left: correlation for all 168 variables. Right: correlation between the loading of all of the hours in the first day (zoom).

To research correlation, we use the `corrplot` library from **R** to visually show the correlation strength between variables. The first image corresponding to the whole week and the second one representing only the first day.

By focusing on the first day (on the right), we can see that the values near the diagonal have a very strong correlation. These values correspond to the graphs shown just above the correlation plots (correlation between very close hours). In fact, we observe that daytime hours are strongly correlated between themselves, just as nighttime hours, but these two hour groups are not correlated. Moreover, the correlation plot on the left shows that the weekend loading profiles are less correlated because the bikes are used less and more randomly during the weekend, in contrast to their high activity during the week.

Most of the variables are strongly correlated, except for those comparing time couples (9, 10) and (10, 11). As explained using the box plots, this drastic change from one hour to the next happens when most people go from sleeping to going to work.

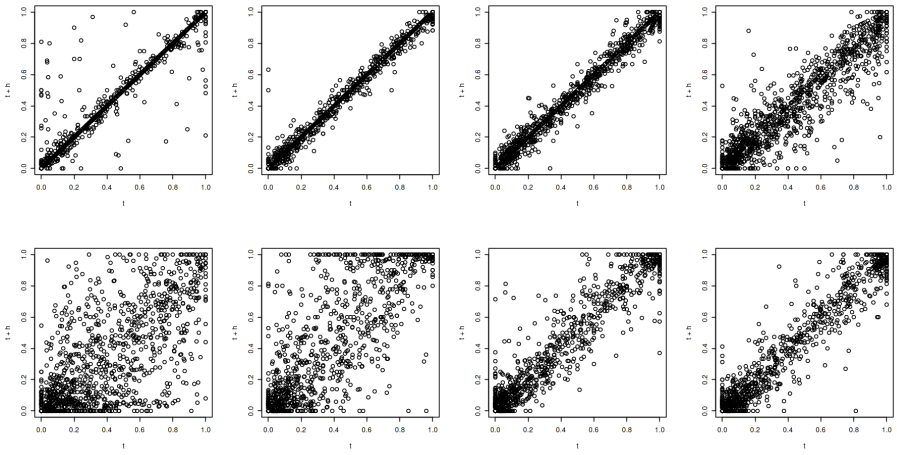

Figure 4: Difference of loading between two consecutive hours.

2.2 Loading profiles based on the altitude

Thanks to the variable `bonus`, we can determine whether a station is on a hill or not. We first suspect that the usage will be lower than the stations that are not on a hill, assuming that customers do not want to ride a bicycle up a hill.

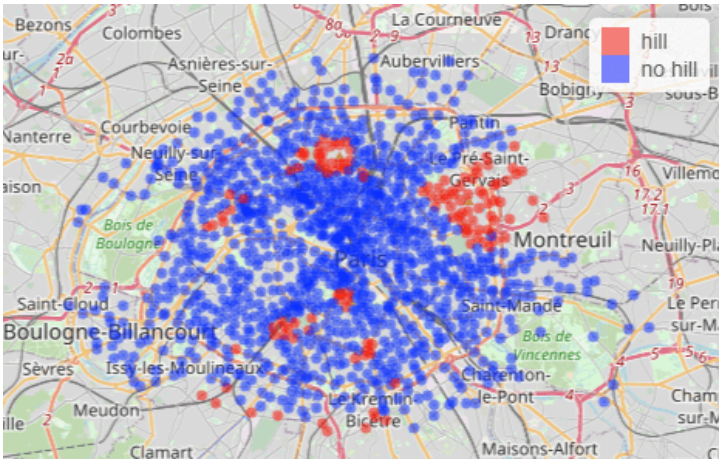

Figure 5: Plot of all stations in Paris.

In order to compare the loading profiles between these groups of stations, let's plot the box plots of the loading profiles throughout the week.

The first graph is not as clear as the second one since it has a large amount of outliers. This is due to the fact that the use of bicycles on the weekends is erratic. However, we can see that there is a certain pattern during the weekdays, and another one during the weekend.

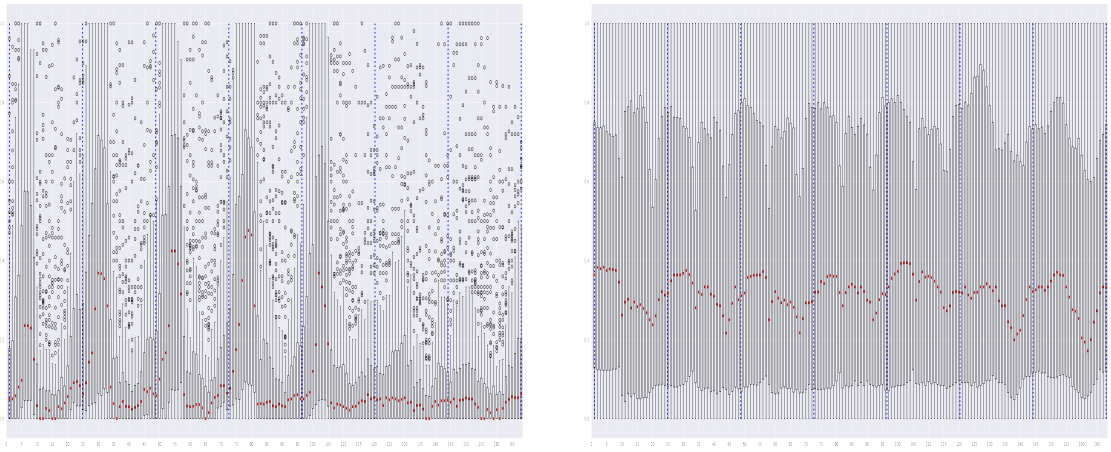

Figure 6: Loading through 7 days of stations on a hill and on plain terrain resp.

The two box plots are fundamentally different, thus explaining a vastly different behaviour. We can notice that the loading of the stations that are on a hill is generally much lower. This can be partially explained by the fact that these stations are close to busy neighbourhoods, for example the ones near the Sacré-Coeur basilica (18th arrondissement), Trocadéro (16th arrondissement), Père-Lachaise graveyard (20th arrondissement), etc. **However, this hypothesis is unproven and needs more substantial data in order to make a conclusion**.

2.3 Main conclusions from multidimensional analysis

3 Principal Component Analysis (PCA)

Given that we have many variables (168 variables), we make a Principal component analysis to better understand the meaning of these variables. Data don't need to be scaled before performing PCA as they are between 0 and 1.

3.1 Reducing data dimension: number of principal components

To determine the number of principal components to keep, we chose to display the explained variance ratios in function of the principal components computed.

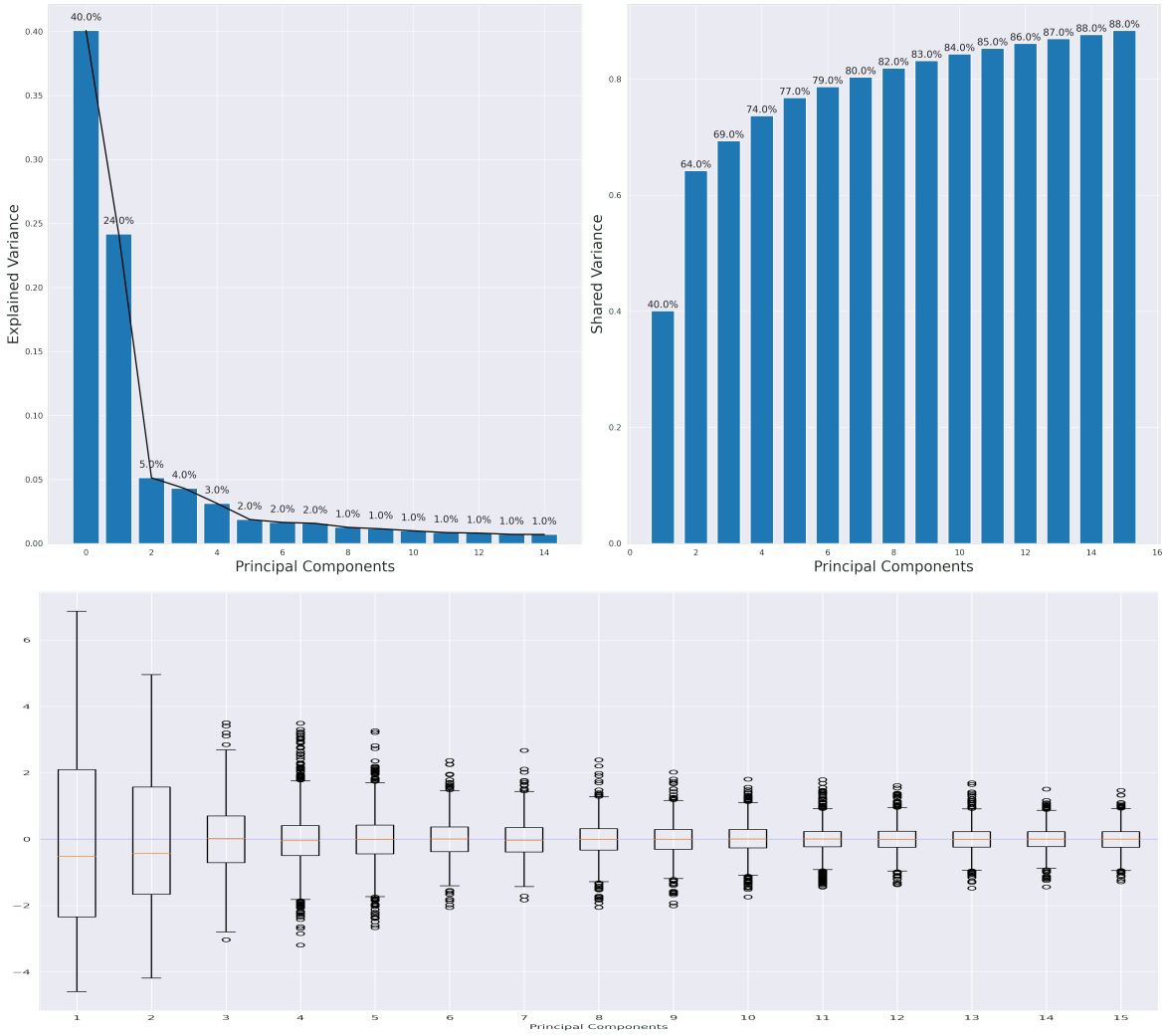

Figure 7: Top: Variance explained by the first fifteen principal components of the PCA applied on the loading data. Bottom: Box plots of the first fifteen principal components.

As shown by figure 7, most of the variance is contained in the first two components (64%), but some important information still remains in some of the following components. Indeed, 80% of the

variance is explained by the first seven components.

On the graph of explained variance ratios (fig. 7, top figure), the curve of explained variance ratios is bent over the sixth principal component.

Figure 7 (bottom) shows that the first two components are the ones with the largest variance. After the sixth principal component, the explained variance is low and decreases very slowly. The medians are centered around 0 after the third principal component and the extent of the box plots is constant after the sixth principal component.

Therefore we chose to work with the first six principal components, given that after the sixth one, the accumulated sum of explained variance increases very slowly.

3.2 Interpreting the principal components

The circles of correlation give a first interpretation of the first three principal components.

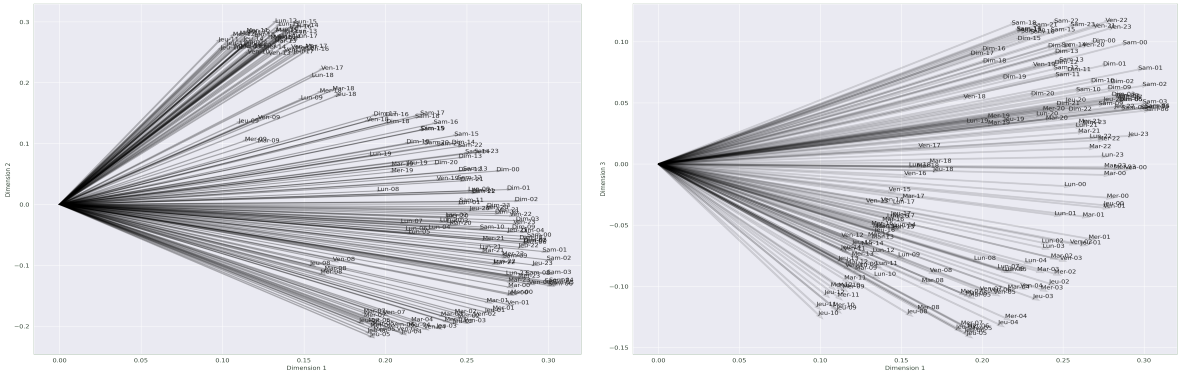

Figure 8: Left: Circle of correlation between variables on the first and second principal components. Right: Correlation on the first and second principal components.

On the circle of correlation between the first and second components, all variables take positive values along the first dimension. As for the second dimension, positive values tend to correspond to daytime hours (from 2pm to 6pm), while negative values correspond to nighttime hours (11pm to 5am). All variables are very far away from the edge of the circle of radius one.

On the circle of correlation between the first and third components, most variables with positive coordinates correspond to weekends, while most variables with negative coordinates correspond to weekdays. However, this analysis is mostly inconclusive because all arrows are very short in norms, which means that the correlation with either dimension is not very strong.

Therefore we can deduce that :

- The first principal component corresponds to the overall loading of the Vélib stations,
- The second principal component is the contrast between night and day,
- The third principal component corresponds to the difference between weekdays and weekends.

Additionally, on figure 9, the graph of individuals projected on the first two components highlight a potential cluster in the bottom left corner mainly composed of bike stations located on a hill.

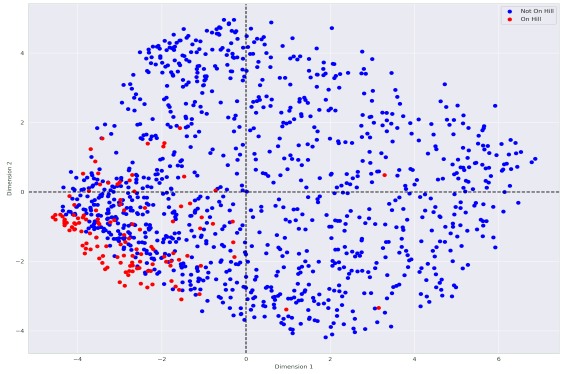

Figure 9: Individuals projected on the first two dimensions of the PCA computed on loading data. In red (blue), Vélib stations (not) located on a hill (resp.).

4 Clustering on the original data

This section aims to detect clusters in the original data, corresponding to common customer usages.

4.1 Ascending Hierarchical Classification

The Ascending Hierarchical Classification (AHC) method allows us to divide our dataset (here, the loadings) into similar clusters, based on several metrics. Here, we use the `Euclidean distance` metric. We explain why later in this section.

The AHC method is first applied on the raw data in order to detect clusters and to determine the best number of clusters to retrieve.

We apply the AHC algorithm using the Ward method and the euclidean metric: first, we chose Ward's method for clustering because it consists in evaluating the distance between clusters as how much the sum of squares increases when merging them. We want to keep this cost as low as possible. Then, we chose the euclidean distance since it corresponds here to the squared distance between two curves (station loading).

According to the cluster dendrogram in figure 10, five, six or eight clusters can be retrieved from the raw data. Indeed, a great step in cophenetic distance is observed between eight and seven clusters, but also between five and four clusters. We obtain six clusters by fixing the distance threshold at 30.5 in order to get homogeneous groups and to better explain the behaviour of Vélib stations.

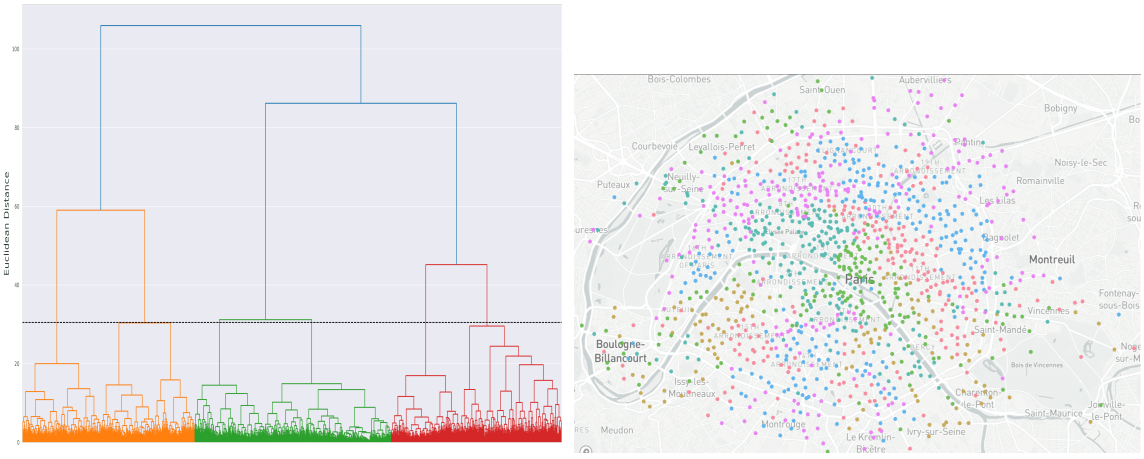

Figure 10: AHC applied on the data using Ward method and euclidean distance. Left: Cluster dendrogram, threshold distance fixed at 30.5. Right: Vélib stations coloured according to their respective cluster (total of 6 clusters).

4.2 K-Means

Secondly, the K-Means algorithm is also applied on the raw data in order to detect clusters of Vélib stations. This method allows us to separate the data with straight lines.

According to the silhouette coefficients and inertia of classes on figure 11, we could choose either $k = 4$ or $k = 6$ classes. Indeed, almost all classes have a a silhouette coefficient above the mean silhouette coefficient and they also have a similar width. These figures allow us to have a basic idea of how many clusters to choose, but further research is needed, because this is only a visual method and could be erroneous.

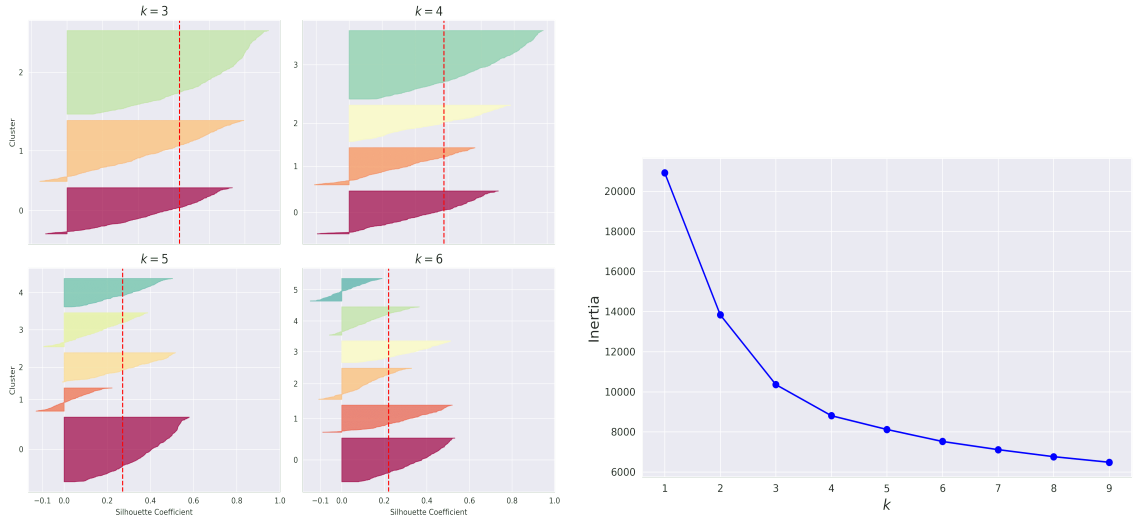

Figure 11: K-Means applied on the raw data. Left: Silhouettes of classes for $k = 3, 4, 5$ and 6. Right: Inertia of classes.

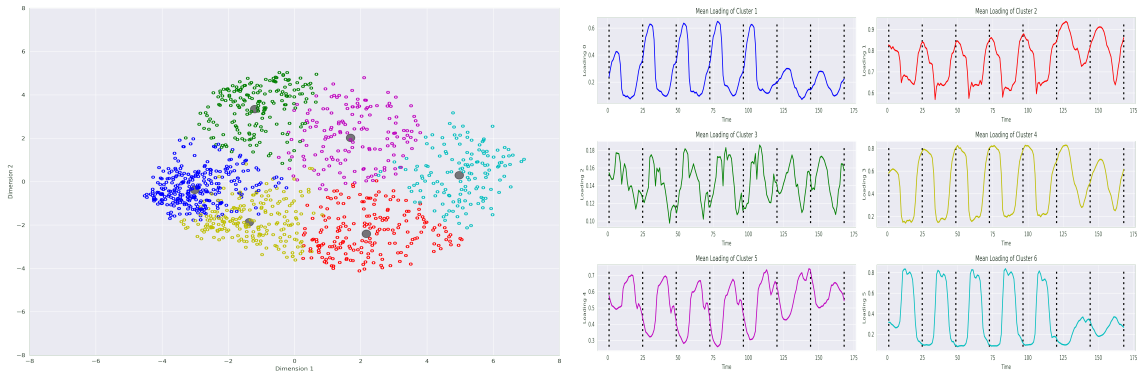

Figure 12: K-Means applied on the raw data with $k = 6$. Right: Loading of the cluster barycenters.

On figure 12, the left plot shows the Vélib stations coloured according to their respective cluster (total of 6 clusters). The plots on the right depict the median loading of each cluster. We can observe that these are different from each other, which highlights the fact that the clusters are separated well.

Thanks to our analysis, we obtain 6 clusters. As we said at the beginning of this section, these clusters correspond to similar loadings and uses of Vélib stations.

5 Clustering on the coefficients of a suitable functional basis

In this section, we use the same methods as before, but now on the projected data found through the PCA.

5.1 Ascending Hierarchical Classification

This time, we apply the Ascending Hierarchical Classification method on the projected data using the PCA. We only use the first 6 principal components, since they are the ones that contain most of the information. Figure 13 depicts the results, and to be consistent with the previous clustering, we decide to keep 6 clusters.

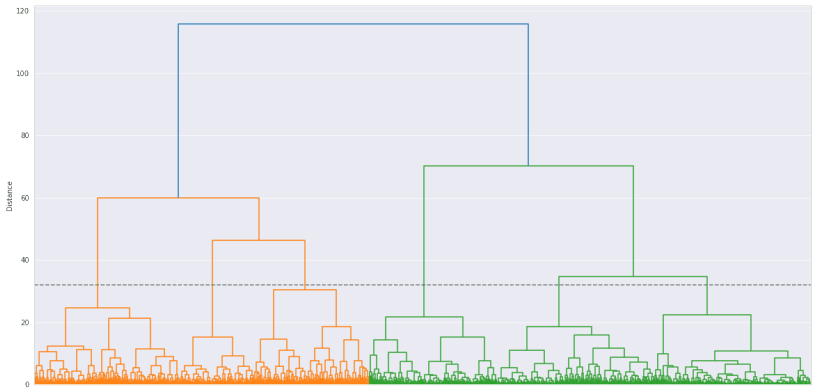

Figure 13: AHC applied on the projected data using Ward's method and the euclidean distance.

To compare these results with the ones corresponding to the raw data, we created a cross-table to see which individuals are in the same cluster for both methods, and which ones are on different clusters (figure 14). We observe that, overall, the clusters are almost the same, but there are still several differences between the two methods.

	0	1	2	3	4	5
class...						
0	129	0	45	0	0	1
1	67	111	2	0	0	25
2	1	0	147	5	2	0
3	0	0	37	211	29	1
4	0	0	1	0	134	17
5	27	2	10	23	30	132

Figure 14: Cross table of AHC method applied to raw data and PCA data.

5.2 K-Means

We continue using the first 6 principal components, and we apply the K-Means method to these data to find 6 clusters. We choose to compare both K-Means clustering methods by using a cross table (figure 15), and we conclude that there is no significant difference between the K-Means method applied to the raw or the PCA data.

	0	1	2	3	4	5
class...						
0	189	0	0	0	0	2
1	0	294	5	0	0	0
2	1	2	209	0	0	0
3	0	0	0	151	1	1
4	0	1	0	1	186	0
5	0	0	0	1	0	145

Figure 15: Cross table of K-Means applied to raw and PCA data.

5.3 Gaussian mixture model

Finally, we use the Gaussian mixture model to have a better separation of the data. This method classifies data by the probability of it belonging to a certain cluster. It also allows us to separate the data with curves, instead of straight lines, and it can adapt to different shapes of clusters, be it a sphere, an ellipse or another shape.

As with the other clustering methods, we decide to separate the data in 6 clusters. To do so we have to work on two dimensions, so we use the data projected into the plane formed by the first two principal axes. As for the covariance matrices, we use the default method, which does not apply any constraints to the matrices, therefore each cluster can take any particular shape and size. These clusters are shown in figure 16.

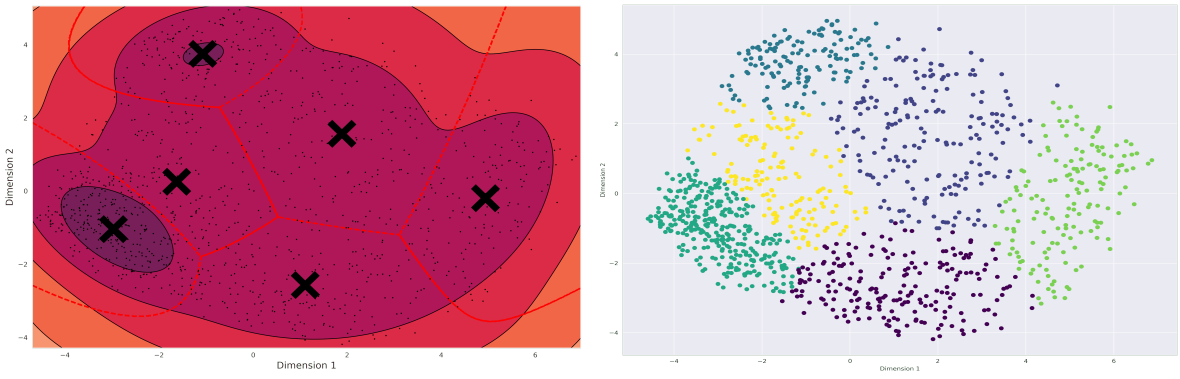

Figure 16: Clustering by Gaussian mixture model.

Even though we had already established that we would use 6 clusters, we could also find the best number of clusters and the most appropriate type of covariance by minimizing the BIC criterion. By

doing so, we find that to minimize the BIC criterion we must use 8 clusters and the diagonal model (parallel ellipsoid clusters, but of different dimension).

5.4 Clustering conclusions

After doing all of the classical clustering techniques, we generally find that 6 clusters is the most efficient way of classifying our data. Different methods give different cluster numbers. For instance, the Gaussian mixture model provides 8 clusters in order the minimize the BIC criterion, whereas the K-Means method suggests 6. We also find that applying these clustering methods on the raw data and on the PCA data might give different results. For the K-Means method we find practically the same result for both sets of data. However, there is a more noticeable difference between the clusters created with the two sets of data when we apply the AHC method.