Report Data Analysis Project

Group: Tran Khoi DANG, Minh Tue TRUONG, Kim Thang VO, Trung Thai DO

1. Introduction

This project is carried out as part of the 4GMM Data Analysis course at INSA Toulouse. Here we study a dataset of the Vélib system (shared bicycle service) in Paris, France. The data are loading profiles of the bicycle stations over one week, collected every hour during the one-week period, from Monday 2nd to Sunday 7th, September 2014.

The loading profile of a station, or simply loading, is defined as the ratio of number of available bikes divided by the number of bike docks. A loading of 1 means that the station is fully loaded, i.e., all bikes are available. A loading of 0 means that the station is empty, all bikes have been rented. From the viewpoint of data analysis, the individuals are the stations. The variables are the 168-time steps (hours in the week). The aim is to detect clusters in the data, corresponding to common customer usages. This clustering should then be used to predict the loading profile.

Guideline: In this report, we will:

- Perform initial exploration data analysis (EDA), e.g., variable identification, univariate analysis, bi-variable analysis, detecting-treating missing values, detecting-treating outliers (or anomalies), visualizing some data samples, …
- Use Principal Component Analysis (PCA) and try to reduce the dimensions of the problem.
- Apply 3 different clustering techniques (K-Means, Agglomerative Hierarchical Clustering, Gaussian Mixture models) to group stations into clusters that we can interpret.
- Compare and choose the best way of defining clusters of stations corresponding to common customer usages.
- Conclude the report.

In the following sections of this report, we focus on the interpretations of the station clusters and link it to common customer usages. Some technical explanations of the methods are also detailed as proof of our conclusions. Detailed codes and results (in R and Python) are available under the format of Jupyter notebooks in the attachment.

2. Exploratory data analysis

2.1. Variable identification

The Velib dataset consists of data from the bike sharing system in Paris. The data are loading profiles of 1186 bike stations over one week, collected every hour, from the period Monday 2nd Sept to Sunday 7th Sept. 2014. Extra information about the 1189 station's name, longitude and attitude, and information indicating if the station is on a hill or not, are also available.

In this report, each station among 1189 stations is considered as an individual, and its loading scores recorded over 168 hours (corresponding to 1 week) are considered statistical variables. These scores are real numbers ranging from 0 and 1, so they are quantitative variables. The extra variable “hill”, indicating if an individual (station) is on a hill or not, is also included for further analysis. Since this extra variable is binary (either yes or no), this is a categorical variable.

Then, during our data preparation, we linked the stations with their corresponding names and found that there are some stations with duplicated names in the dataset. Those duplicated stations however have different longitude and attitude, so we keep them and change the name slightly by adding a dummy index at the end of their name. We have not found any missing values in the tables of loading profiles, so no extra effort on treating missing values is needed.

2.2. Data visualization

We can see below the loading profile over 168 hours of the first 16 stations. Remark that there are different behaviors that we can observe directly from this first figure:

- Some stations with lower loading score over the week (meaning that most bike are rented)
- Some stations with higher loading score over the week (meaning that less bike is rented)
- Some stations with higher loading score in the evening and lower loading score in the morning
- Some stations with lower loading score in the evening and higher loading score in the morning

These observations are not generalized since we observe only a small number of stations, we will try further in the report, to divide all stations in groups of the same behavior and analyze it.

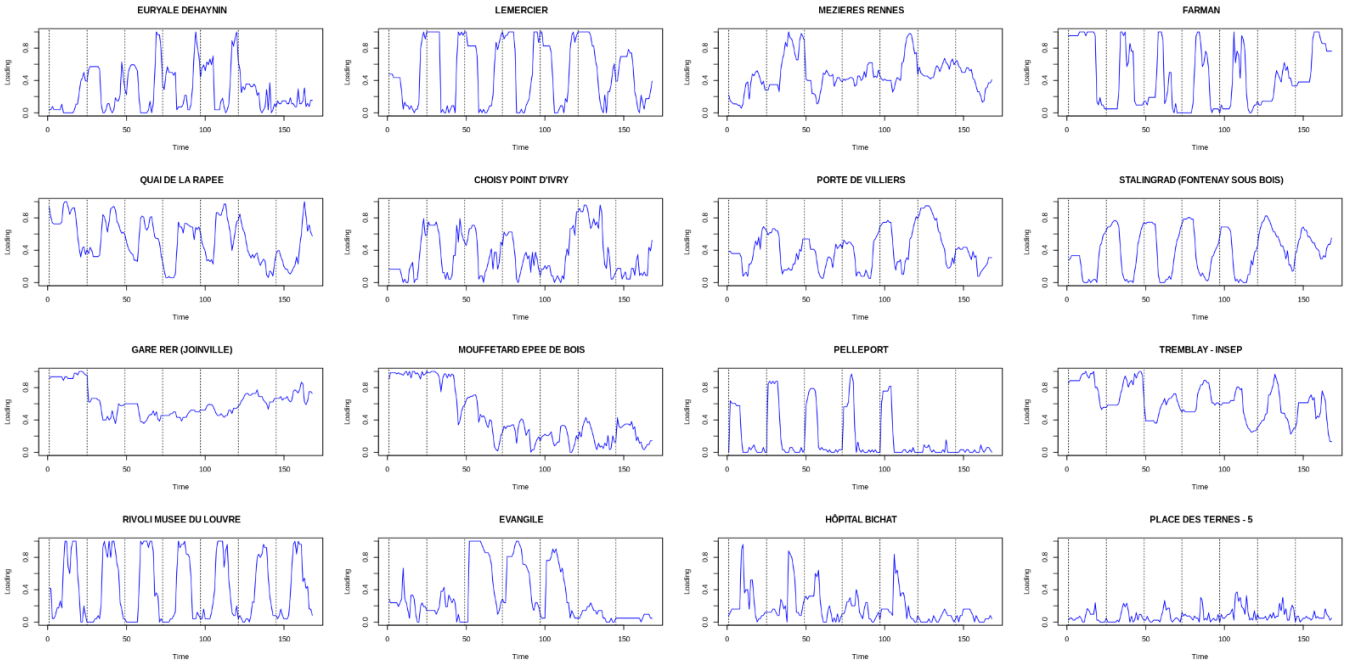


Fig. 1: Loading score over the period of 7 days of the first 16 stations

We draw below the boxplots of all stations over the time period. These boxplots represent locality, spread, and skewness groups of numerical data (here from 0 to 1) with their quartiles. The black lines in figure 2 are median values of each variable (the median loading score of each hour). We can see that the median values are around 0.3, the interquartile (the gray part) is mostly constant, ranges from 0.05 to 0.7, with lower dispersion in hours where people rent more (afternoon). We observe that there is no remarkable difference in days, if not the stations on Saturday and Sunday tend to have smaller loading scores. Difference in hours is interesting to remark: in early morning and late evening, loading scores tend to be higher than in afternoon.

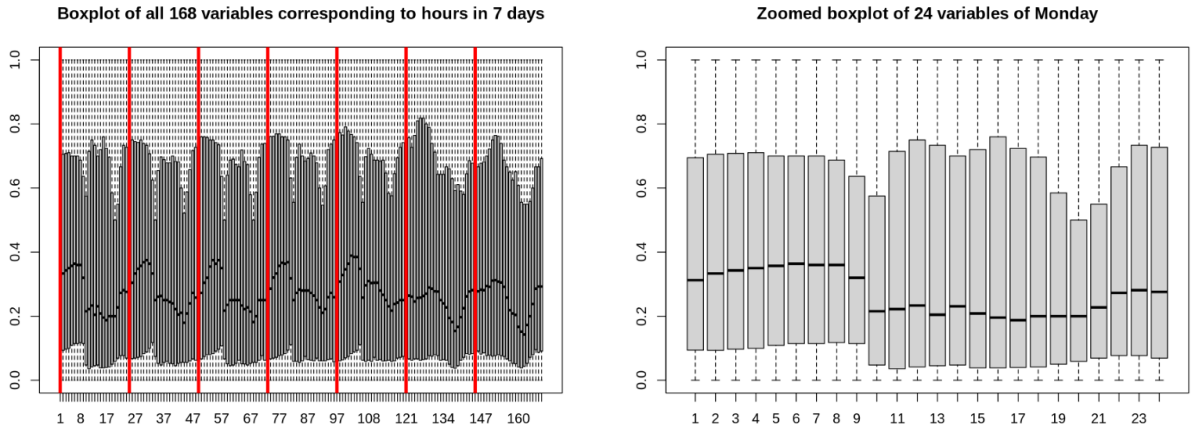


Fig. 2: Boxplot of all stations (left), zoom in first 24 variables (right)

We plot here the average loading score of all stations recorded on each day of the week. We see that the average loading score of weekdays (Monday to Friday) follows the same trend: lower with some significant drops at 10am and 8pm. The average loading score of Saturday and Sunday are lower from 2pm to 8pm.

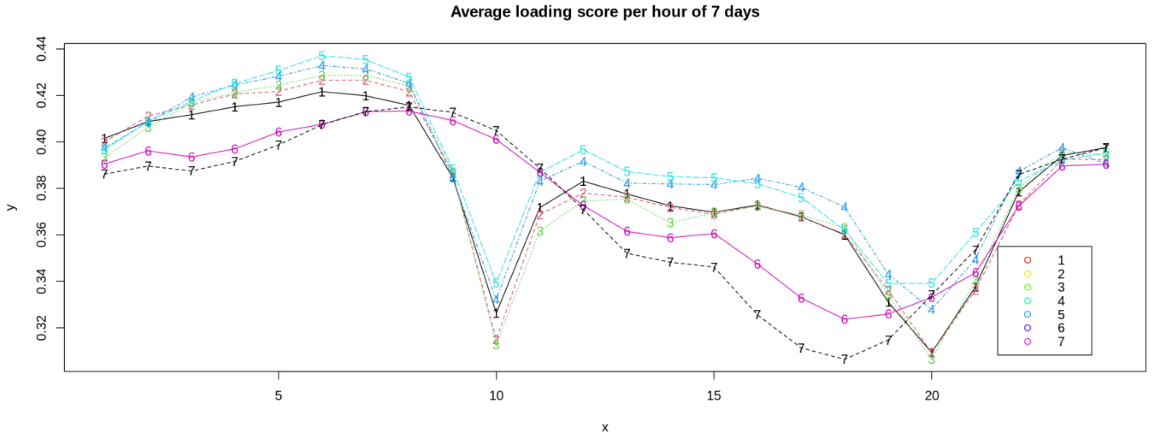


Fig. 3: Average loading score of all stations per hour of 7 days

We plot here the map of Paris and 1189 stations considered with its average loading score over the week represented by the color intensity. The darker the dot is, the lower the loading score of that dot (station), i.e., less bicycles are available at that station.

Fig. 4: Loading score of stations on map of Paris center

Among 1168 stations, there are 127 stations on hills. This group of stations can give us interesting statistics that we can use for further interpretation. Below is the graph of average loading score over the time period of stations on hills and its corresponding box plot. We can see that the mean loading score is low overall, meaning that there are few bicycles in the stations. From the boxplot graph, loading score is very low on Saturday and Sunday while we observe large dispersion on weekdays. A possible explanation can be that people who go out on Friday do not want to put their bikes back on the hills, so the stations are always short of bicycles on Saturday and Sunday.

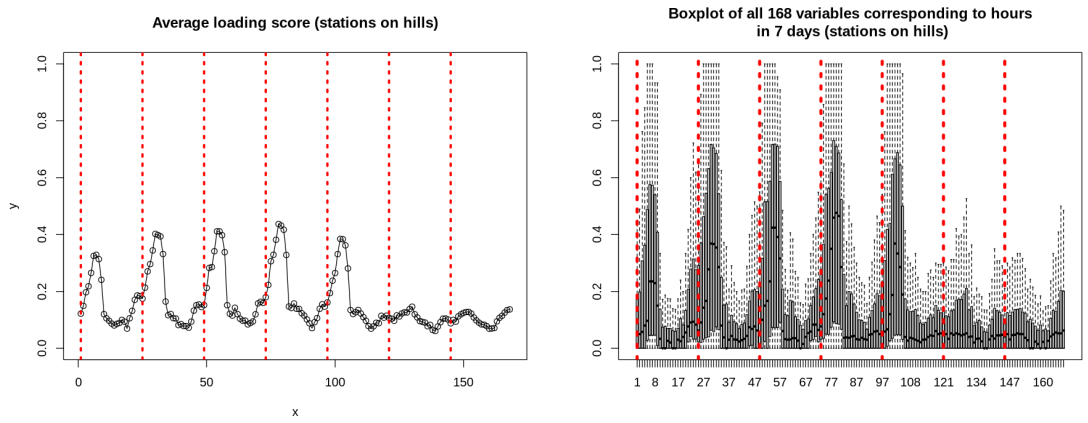


Fig. 5: Average loading score of stations on hills (left), boxplot of stations on hills (right)

3. Feature extraction with PCA

PCA (Principal component analysis) is a dimensionality reduction technique generally used for big datasets containing a high number of dimensions (variables) per observation. By transforming our existing data into a new coordinate system with fewer dimensions, we want to preserve most of the variations (important information) of the old data. In fact, for each station observed among 1189 stations in our dataset, 168 variables (loading score) are measured. To increase the interpretability of our dataset, we can try to use PCA and study trends, jumps, outliers or clusters from the new smaller dataset. Further in the report, we will use different techniques of clustering applied on both the big dataset and the dataset reduced with PCA and analyze the obtained results.

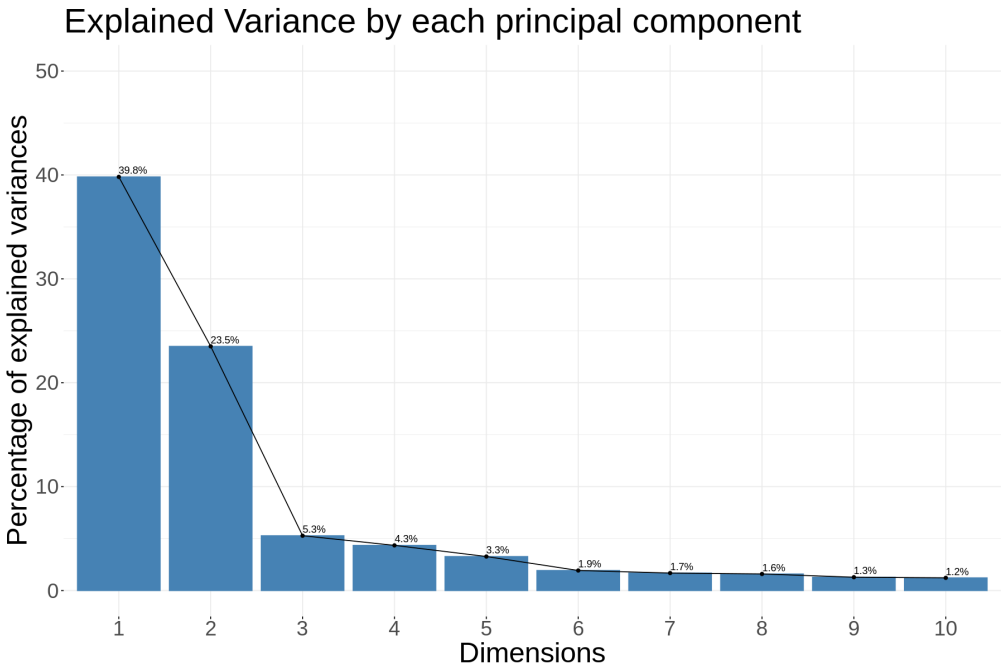


Fig. 6: Decreasing percentage of explained variance per components of PCA

We plot above the percentage of total variance of the initial dataset explained by the first ten principal components from the PCA method applied on our dataset. The two first components account for 63.3% of the total variance. The number of components used are chosen per each method of clustering in the next sections.

Next we plot coordinates of the first four eigenvectors from the PCA method. We can deduce that the first component represents the average score loading, the second component represents the difference in average, the third explains a little the contrasts of score loading between hours (especially afternoon and evening). Overall, we can say that the first 4 components tackle some main characteristics of the initial dataset (to confirm that PCA approximates the best the data in mean square sense).

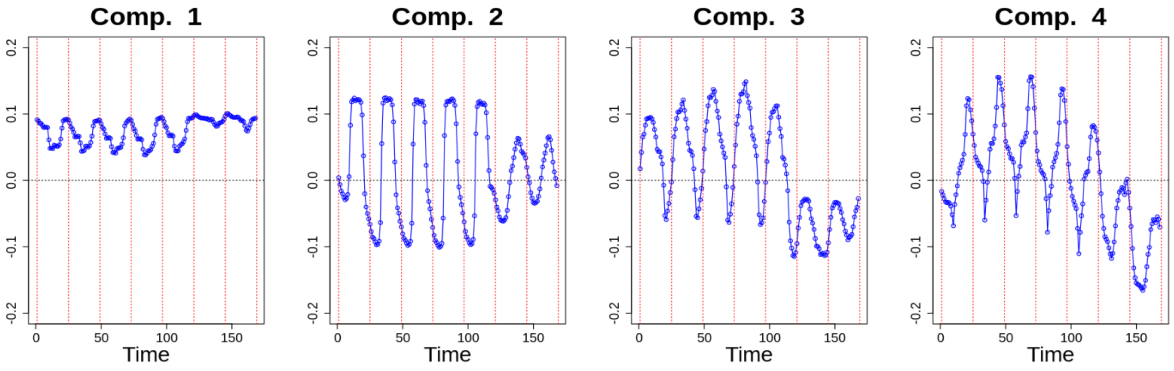


Fig. 7: Coordinates of the first 4 eigenvectors in the analysis of PCA

4. Clustering

4.1. Approach

In this section, we want to use different clustering techniques to group stations into groups of which we can interpret their behaviors and link them to customer usage. To do that we will go through 2 steps:

+ Step 1: Apply each clustering method on both the initial and the PCA-reduced dataset, then select the best model of each one. Involved clustering methods include K-means, agglomerative hierarchical clustering and gaussian mixture models.

+ Step 2: Select the best model among the models selected in step 1 as the final, best model of all methods.

However, as this is an unsupervised problem, it's hard to define what makes a model the "best" model, so we need to define our own criteria. In our analysis, we define the "best" model to be the model that is most interpretable, in other words:

- The resulting clusters should have small intra-class variance, and the loading score of the majority of stations in each group should follow its average behavior.
- The behavior of different clusters should be as distinct as possible.
- There must be a considerable number of stations in each cluster.
- There should not be too many clusters (for interpretability), and also not too few (~2-3, since we discover at least 4 types of behavior in the exploration step).

During and after model selection, we will need to interpret the behaviors of the stations in the clusters. To better grasp and compare these clusters, we consider important characteristics that can for us, perfectly describe the behavior of a cluster.

The 3 characteristics are defined as follows, knowing that we say that a loading score is low if < 0.15; medium if > 0.15 and < 0.8; high if > 0.8.

+ **Weekday-weekend behavior**: The difference between weekday and weekend behavior of a station may suggest interesting information, like its geographical location or the types of building in its vicinity, which will be discussed later.

+ **Off-peak time**: We care about off-peaks because it represents a surge in demand. Note that low value of loading alone is not enough to conclude that a new demand for bicycles has occurred, since it could be that there's always few bicycles in the station, while an off-peak means that a considerable number of bicycles was rented.

Off-peak time is the moment of off-peaks in a day, which are divided into: ***Morning*** (0h - 12h) / ***Afternoon*** (12h - 0h) / (Generally) ***No off-peak***.

+ **Availability of bicycle in the station**: Based on value of the loading score, we can describe the availability of bicycle in the station accordingly:

- ***Available***: loading is generally never low. This means that the station is always available for service.

- ***Medium-active**:* low off-peak with medium peak. This represents moments of lack of bicycles at off-peak time. Since there are off-peaks, there are waves of bicycles being rented, meaning there is considerable activity at the stations, but not as intense as ***Max-active***, as defined below.

- ***Max-active***: low off-peak with high peak. This means that, unlike the previous case, the station is working at its maximum capacity. In fact, while at off-peak, almost all bicycles are rented, at peak time, the station is almost full of bicycles. The demand of the station is high and could serve more customers but it is limited by the maximum number of bicycles slot in the station.

- ***Idle**:* loading is almost always low. This means that the station always has few bicycles, but unlike ***Medium-active*** station, there's no off-peak, meaning there's no considerable daily activity at the station.

We present in the next section results of Agglomerative hierarchical and gaussian mixture models for our clustering problem, since it gives more interesting results and interpretable choice of parameters than the classical k-means algorithm. The code for K-means is available in the attached notebooks. As said before, we will apply those methods on the initial dataset and on the dataset reduced with PCA.

4.2. Clustering with Agglomerative Hierarchical Clustering

In this section, we attempt to use the Agglomerative Hierarchical Clustering (AHC) method to find groups of stations with similar behavior that we can interpret. The AHC method aims to build an indexed hierarchy of clusters, with a "bottom-up" approach: all observations start from the bottom of the dendrogram, then we aggregate two by two the closest parts in the sense of a "linkage function" until we obtain only one cluster. Here, we chose the Ward's linkage function for our implementation. Ward's linkage function has a tendency to build clusters of equal size for a given level of hierarchy, which we can observe in the dendrograms below. Also, Ward's method gives us a minimal increase of the intra-class inertia at each step of aggregating the two closest clusters.

We implement the AHC method on both the full dataset, and the reduced dataset by PCA. The full implementation can be found in the attached notebooks. The resulting clusters from both approaches are pretty the same. However in general, the AHC method is time consuming especially with a high number of observations and applying PCA may be a good strategy to speed up the algorithm. Thus, we show in this section only clusters obtained with AHC on the PCA-reduced dataset.

To do so, we chose 10 as the number of components of the PCA method. The dendrogram below (in the left) shows the clusters of stations, the y-axis (height) is the distance (in the sense of linkage function) between pairs of sub clusters.

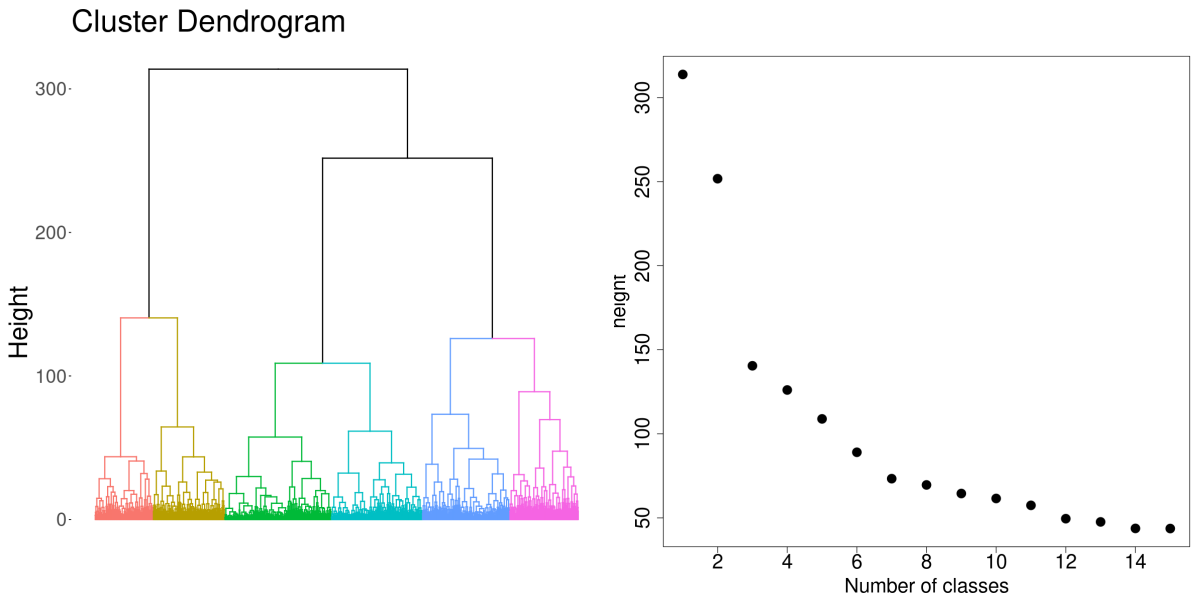


Fig. 8: Cluster dendrogram obtained with AHC method, with color distinguishing 6 clusters taken (left), within-cluster error sum of squares by number of clusters (right)

The cut-off level in dendrogram should be done in a way such that we do not have too many classes (for interpretability) but also a low value of height. In the right graph, we observe a significant decrease in height between 4, 5, 6 and 7 clusters. We still need to choose between these numbers of clusters. To do so, we test AHC with 4 values above and choose the one which gives us the "better" (with definitions above) clusters. Thus, we decided to select 6 clusters in this case.

To further analyze the clusters obtained from AHC, we plot different graphs of the average loading score of stations within each cluster (first column) and the boxplot of loading score at each hour in the time period of each cluster (second column). We also calculate the number of stations belonging to each cluster. By doing so, we want to observe first, the main behavior of the group (mean loading score of each cluster), and second if the loading score of the majority of stations in each group follow its average behavior (in the sense that there is not large dispersion (variability) in the boxplots of each cluster.

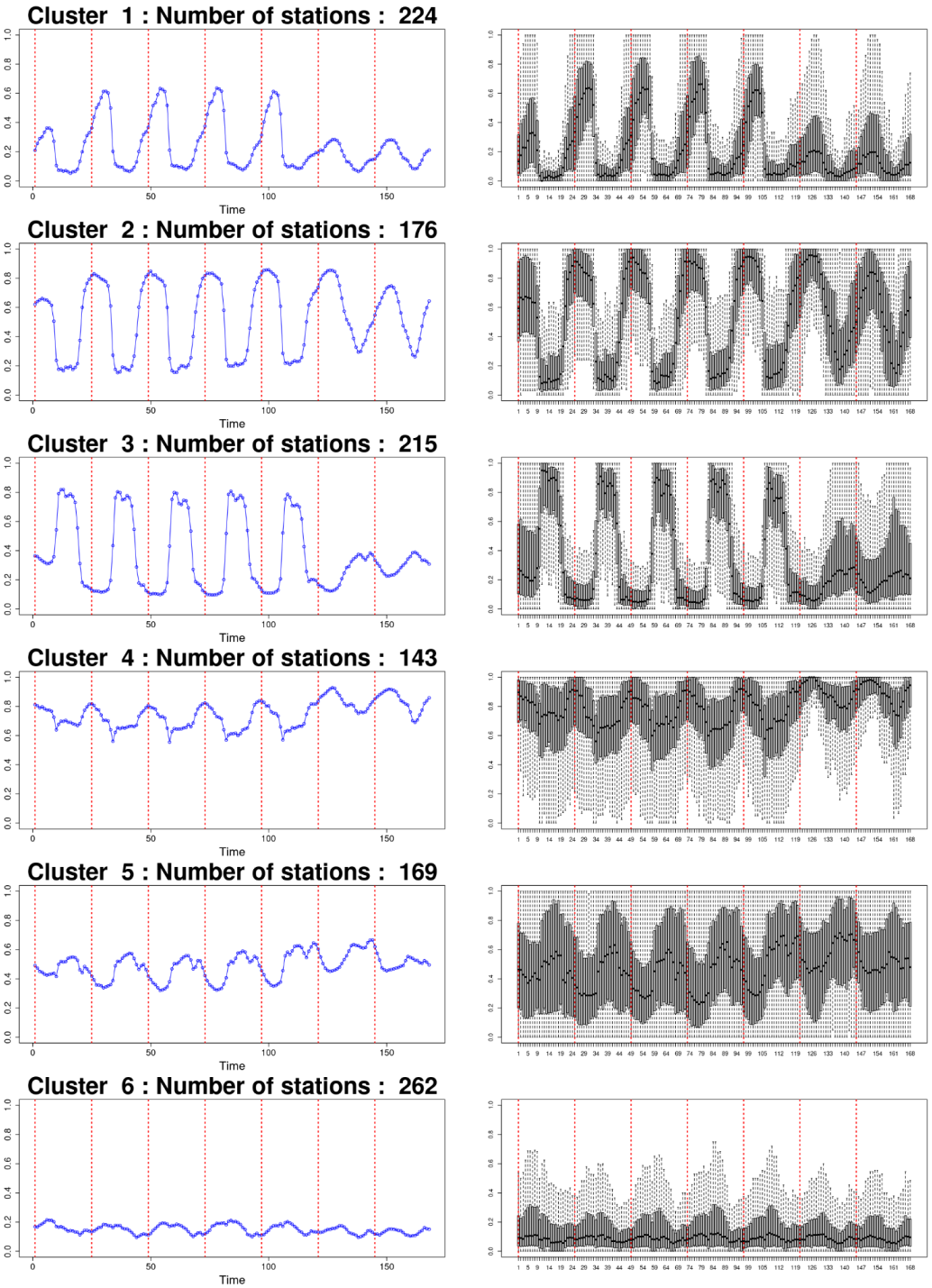
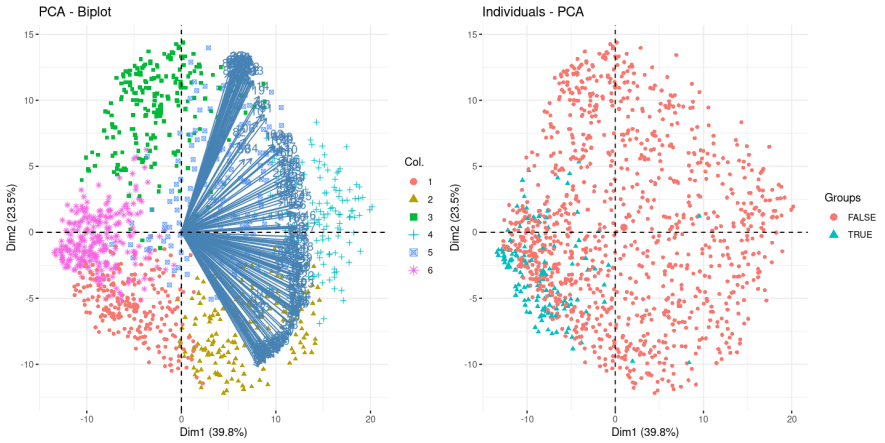


Fig. 9: Analysis by mean (left) and boxplot (right) of stations in each cluster

With 4 characteristics defined, we can better describe the station clusters:

- Cluster 1: All week: ***Medium-active*** with off-peak in the afternoon.
- Cluster 2: All week: ***Max-active*** with off-peak in the afternoon.
- Cluster 3: Weekday: ***Max-active***, off-peak in the morning; Weekend: ***Medium-active,*** off-peak in the morning.
- Cluster 4: All week: ***Available*** with off-peak in the afternoon.
- Cluster 5: All week: ***not clearly defined***, since loading score is from 0 to 1.
- Cluster 6: All week: ***Idle*** (*slightly medium-active*) with no off-peak.

In the figure below, we see that there is no big difference between the number of stations in each group. Moreover, we observe that the boxplots of each cluster in the second column, which describe loading score per hour, almost behave like the hourly average loading score (in the first column). Their interquartile range is quite small, i.e., the small dispersion of boxplot, which indicates that the loading score of the majority of stations in each group follow its average behavior (except that there is a bigger dispersion in the weekends of cluster 3 and 5).



Cluster	Number of stations not on hills	Number of stations on hills
1	166	58
2	174	2
3	210	5
4	142	1
5	168	1
6	202	60

Fig. 10: Projection of dataset on first 2 dimensions of PCA with color corresponding to each cluster (left), to stations on hills or not (right)

In the left figure, projections of stations on the first 2-dimension PCA factor map are colored following its cluster. In the right figure, we mark with blue the coordinates of the stations on the hill on this map and the table shows us the number of stations on the hill and not on the hill in each cluster. We realize that almost all stations on the hill are concentrated on cluster 1 and cluster 6, suggesting that there are 2 types of hills (except for variability), **Idle** Hill stations (stations on hills with loading score always low) and ***Medium-active*** Hill station (stations on hills with dispersion of peak and off-peak loading score). The possible interpretation for Idle and Lacking can be found at section 4.1.

4.3. Clustering with GMM

Gaussian Mixture Models (GMM) is another technique to cluster unlabeled data. GMM clustering assumes that the data comes from a mixture of different groups, and each group is represented by a Gaussian distribution. It tries to find the best way to assign data points (in this case: the stations) to these different groups based on their probability of belonging to each Gaussian distribution.

Although our final ultimate criterion is the model's interpretability, there's no method to study this criterion over a large set of models. We first perform an initial model selection according to the BIC score, which is supported by many programming languages, then select the best model among them according to their interpretability with more thorough analysis.

In our study we apply the method on the PCA-reduced dataset. The below figure allows finding the best model based on BIC score. We limit ourselves to a maximum of 20 clusters to have a first look at the evolution of BIC score, then repeat the code with different values of G.

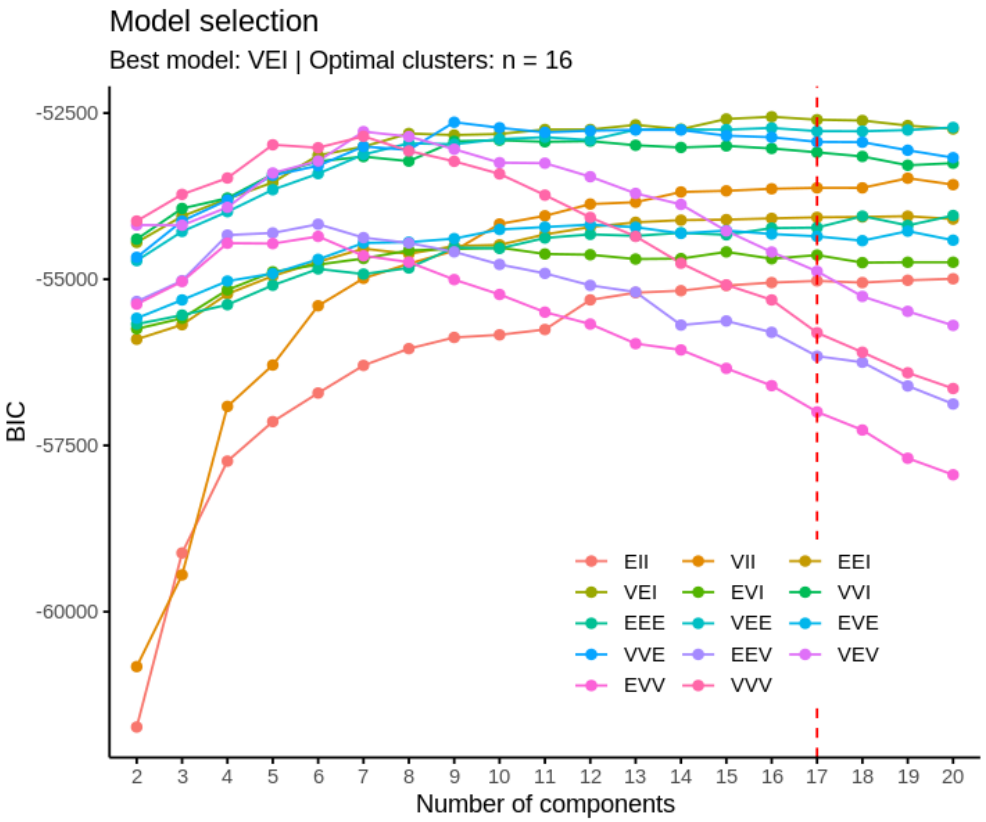


Fig. 11: GMM model selection based on BIC criterion.

We want to have a large BIC score for the model chosen and by testing, we obtain the 6 best models in decreasing optimality order : 16 clusters-VEI, 9 clusters-VVE, 11 clusters-VVE, 10 clusters-VEI, 5 clusters-VVV and 7 clusters-VEV.

Once again, in order to maintain the model's interpretability, we want to assure that there’s not too many clusters and that there's not too few stations (explicit test for each case) in each cluster. We select the best model between 5 clusters-VVV and 7 clusters-VEV.

4.3.1. 5 clusters-VVV vs 7 clusters-VEV

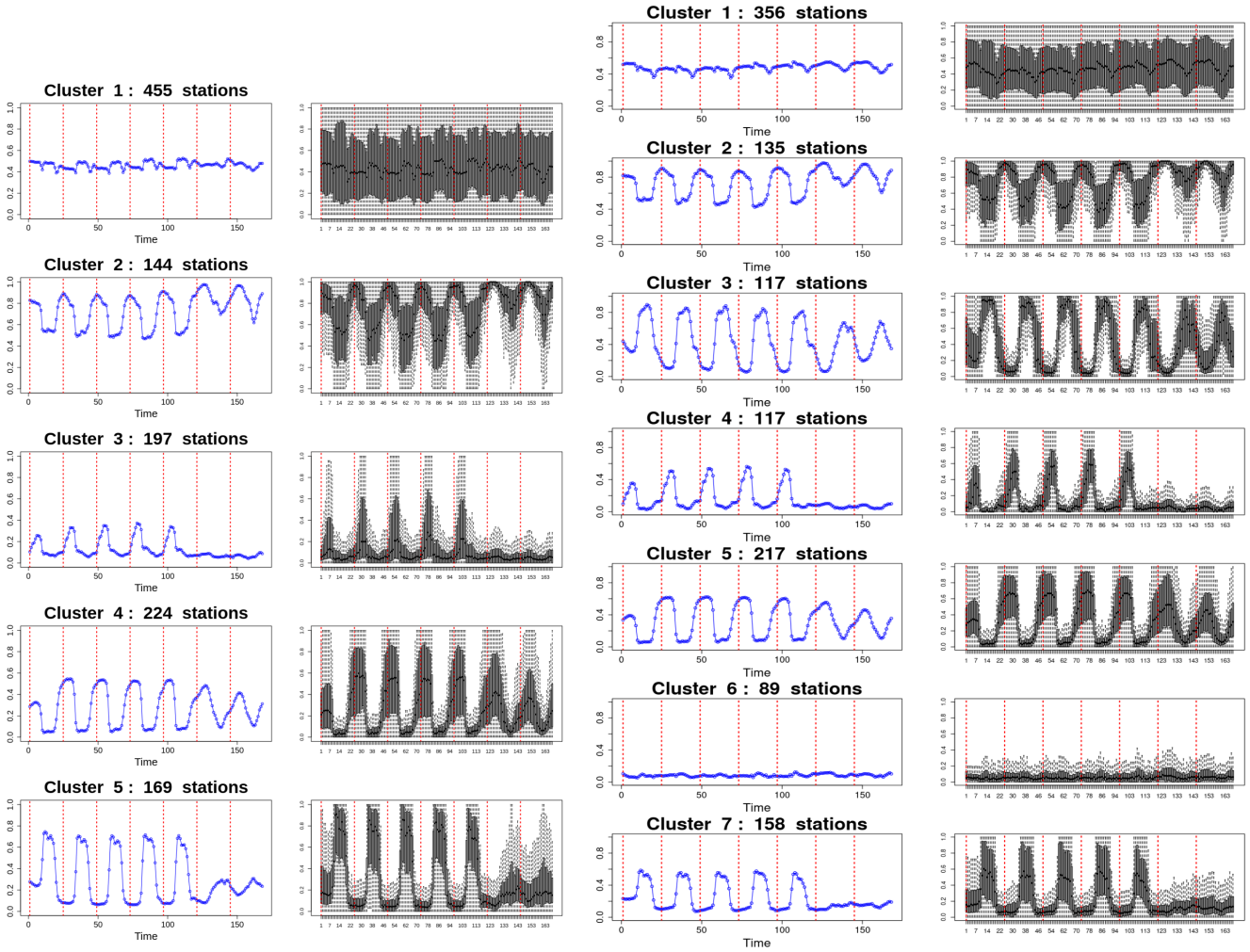


Fig. 12: Model comparison. From left to right : 5 clusters-VVV cluster's mean and boxplot, 7 clusters-VEV cluster's mean and boxplot.

4.3.2. Model selection

Based on the above figures, the main difference of the 2 models is the addition of cluster 3 and cluster 6 of the 7-clusters model. The remaining clusters are roughly the same based on the boxplot figures. In fact, if we denote A~B : cluster A of the 5 clusters-VVV corresponds to cluster B of the 7 clusters-VEV model, we have : 1~1, 2~2, 3~4, 4~5, 5~3.

From the boxplot figures, we can see that cluster 3 and cluster 6 of the 7-clusters model have relatively small intra-group variance, which assures the model interpretability as defined in section 4.1. In addition these 2 clusters represent a significant number of stations (117 and 89), so it makes sense to keep these 2 clusters and select the 7-clusters model.

With these characteristics defined, we can better describe the station clusters of the selected model:

+ Cluster 2 : All week : ***Available*** with off-peak in the ***afternoon***.

+ Cluster 3 : All week : ***Max-active*** with off-peak in the ***morning***.

+ Cluster 4 : Weekday : ***Medium-active*** with off-peak in the ***afternoon***; Weekend : ***Idle*** with no off-peak.

+ Cluster 5 : All week : ***Medium-active*** with off-peak in the ***afternoon***.

+ Cluster 6 : All week : ***Idle*** with ***no off-peak***.

+ Cluster 7 : Weekday : ***Medium-active*** with off-peak in the ***morning***; Weekend: ***Idle*** with ***no off-peak***.

Finally, according to the above figures and the below individual plots, the remaining cluster - cluster 1 contains stations of very different behaviors. We can call it an 'undefined' group with the only characteristic being that it's relatively available all week.

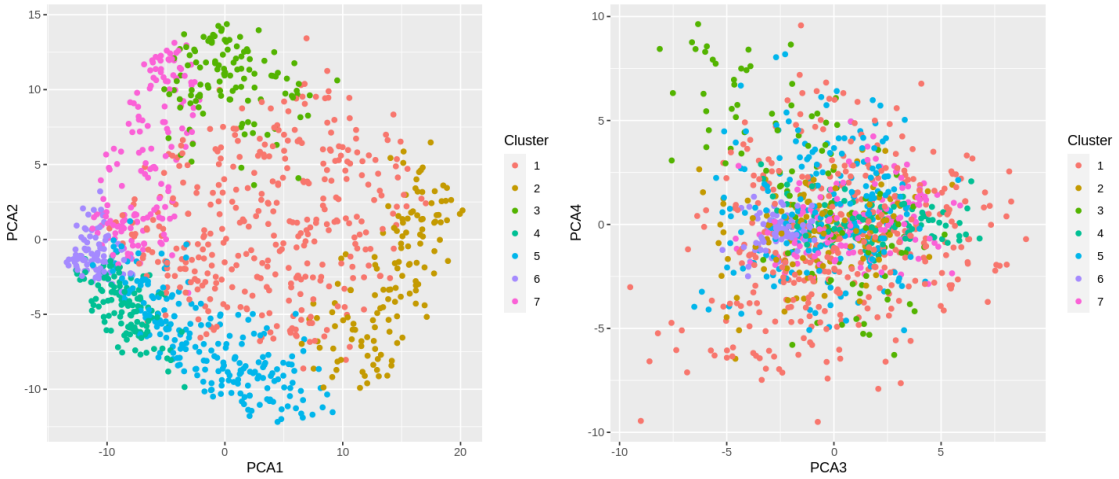


Fig. 13: Coordinates of stations on the factor map of first and second principal components (left), of third and fourth principal components (right)

4.3.3. Extra : Multiple correspondence analysis (MCA)

MCA can be considered the equivalence of PCA but applying to qualitative variables. In order to apply this method to our data, we need to discretize our variables to turn the original variables into new qualitative variables.

We thought it might be interesting to apply the same study over MCA-reduced dataset to see if the method is improved. Unfortunately, the result is worse than our selected 7 clusters-VEV model, so we won't include it in this report, but the code used to discretize the dataset and other studies can be found in the attached notebooks.

5. Comparison of cluster methods

To compare the methods, we will first take a look at the correspondence between the methods with the help of MCA (Multiple Correspondence Analysis).

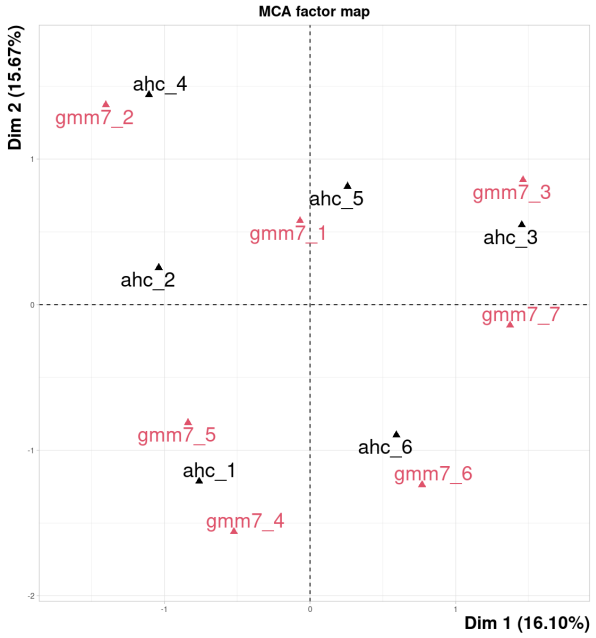


Fig. 14: Correspondence analysis on the disjunctive table for AHC (6 clusters) method and GMM (7 clusters) method

Most clusters of the 2 methods are quite close, which suggest that the interpretation we obtain from the 2 methods would be more or less the same. We notice that there are 3 major differences between the 2 groups.

Difference 1: Cluster 2 of AHC is far away from other GMM clusters, which might mean that the stations in cluster 2 of AHC are badly classified by GMM. We can check this by looking at the mean-boxplot figures and confirm that stations with behavior described by cluster 2 of AHC are represented by cluster 2 and 5 of GMM.

Difference 2: Cluster 7 of GMM is far away from other AHC clusters. Similarly we can see that stations with behavior described by cluster 7 of GMM are represented by cluster 6 and 3 of AHC. However, we notice that this representation is worse than that of GMM, in the sense that cluster 6 and 3 of AHC describe more or less behavior as cluster 7 and 6 of GMM, but with more dispersion. As defined in section 4.1, smaller dispersion allows us to interpret the behavior of stations in the cluster with more confidence, so this is a plus point for GMM.

Difference 3: Cluster 4 and 5 of GMM are both very close to cluster 1 of AHC, which might mean either cluster 4 and 5 of GMM is redundant, or cluster 1 of AHC is better explained by dividing it into 2 clusters like cluster 4 and 5 of GMM. Once again we look at the mean-boxplot figure and notice that cluster 4 and 5 describe 2 different station behaviors (especially in the weekend) while cluster 1 of AHC is roughly the mix of both, which is less interpretable, this is a point in favor of GMM.

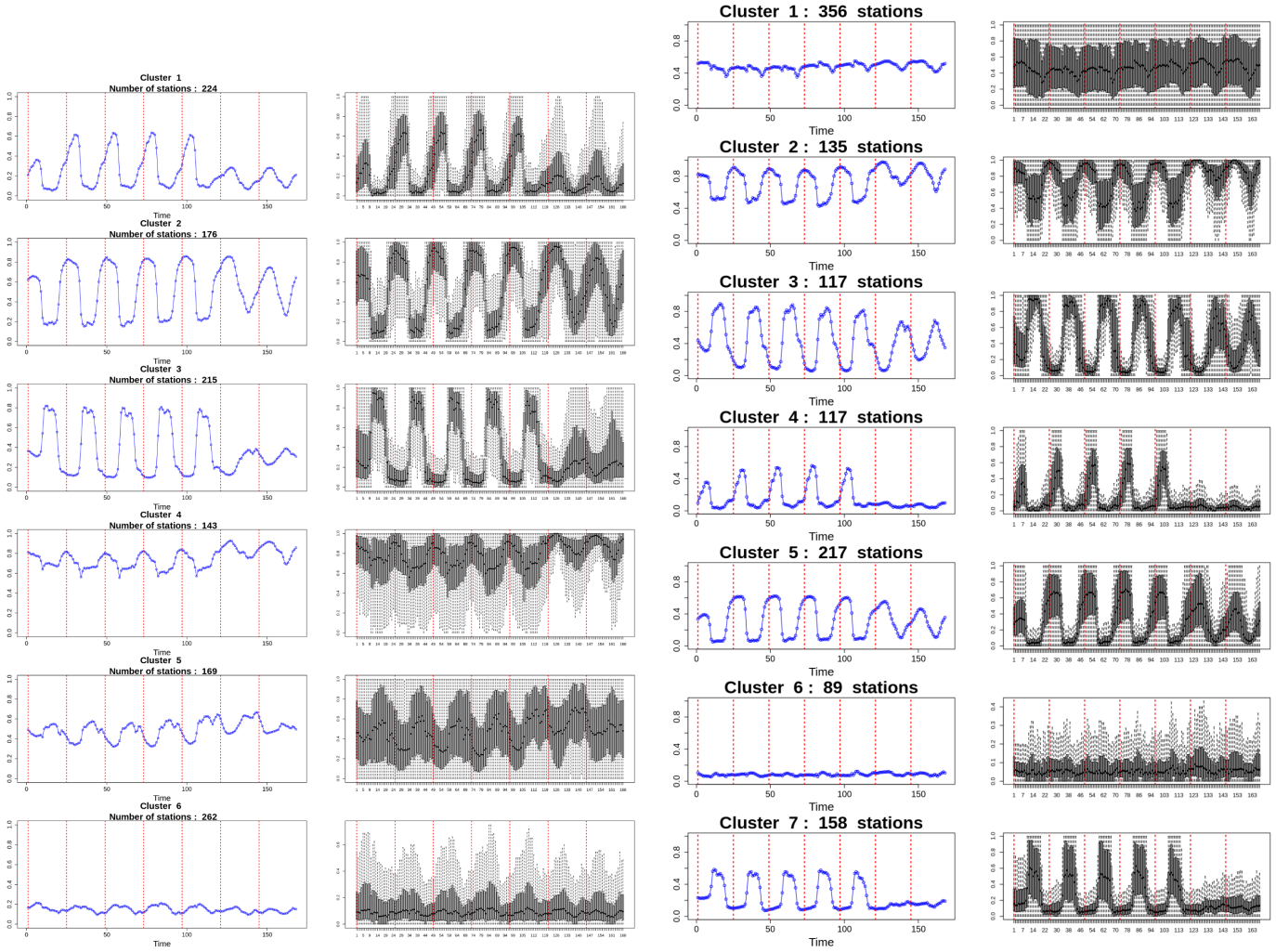


Fig. 15: Side-by-side comparison of AHC (6 clusters) method and GMM (7 clusters) method

Next, we take a look at the quality of the clusters of each model. In general the clusters of GMM have smaller dispersion, which is a plus point for GMM. However the “undefined” cluster of the GMM model (cluster 1 : 356 stations) is considerably bigger than that of AHC (cluster 5 : 169 stations), which is a point in favor of AHC.

All factors considered, we decided to choose the GMM model (7 clusters), as we are willing to accept a larger number of “undefined” stations in exchange for a good clustering and interpretation for the rest of the stations.

6. Further interpretations and conclusion

We decided in the last section to use the GMM model and obtain 7 different groups of stations. Basic graph interpretations about the overall behavior of each group are described with the mean behavior and the intra-group variance in section 4.3.2. The position of those clusters on a map of Paris center are illustrated in Fig 16.

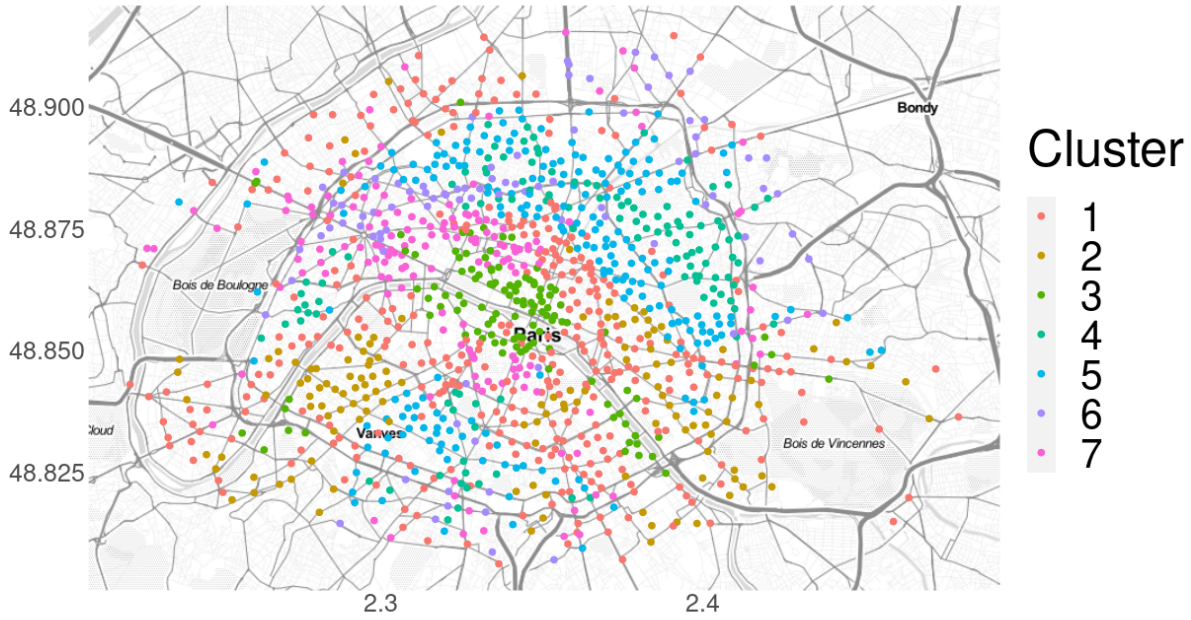


Fig. 16: Stations of each cluster on map of Paris center

With further interpretations and inferences, we can deduce quite confidently that:

Cluster	Loading score interpretation	Further interpretation (include stations that are…)
1	Undefined. Relatively available all week.	Undefined
2	All week : Available with off-peak in the morning.	closed to residential areas, always available for service.
3	All week : Max-active with off-peak in the afternoon.	closed to workplaces, working at maximum capacity.
4	Weekday : Medium-active with off-peak in the afternoon; Weekend : Idle with no off-peak.	closed to workplaces, many uphill stations, no weekend activity.
5	All week : Medium-active with off-peak in the morning.	closed to residential areas.
6	All week : Idle with no off-peak.	unpopular and uphill stations.

7	Weekday : Medium-active with off-peak in the afternoon; Weekend: Idle with no off-peak.	closed to workplaces, no weekend activity.

In fact,

- Identifying workplace / resident area

Based on high-demand time, we can guess which stations are likely to be closed to workplaces and which are closed to the resident area. Stations close to the workplace would receive bicycles in the morning (as people arrive at work) and lose bicycles in the afternoon (as people go home), which fits the description of cluster 3,4 and 7. Stations close to the resident area, in contrast, fit the description of cluster 2 and 5.

We can reaffirm our interpretation by inspecting the difference between weekday and weekend of these clusters. Cluster 4 and 7 is only medium-active during weekdays and have almost no activity during weekends. This strongly agrees with our guess that they must be closed to workplaces, as they have some activities when offices open (weekday) and no activity when offices are closed (weekend). Cluster 3 is the only max-active cluster during the entire week, so they are possibly located in important, more “popular” nodes of the traffic network, which make sense if there are workplaces in its vicinity where it’s easy for employees to commute. Cluster 2, on the other hand, is especially full of bicycles during weekends, which suggests that many people stay there (or rather stay in its vicinity) during weekends, which makes sense if they are really close to residential areas like what we guessed.

- Stations on hills

Based on the table below, 53 among 129 stations on hills are in cluster 4. The remaining stations on the hill are distributed in the other clusters, except for cluster 2 (always available) and 3 (always max-active). This suggests stations on hills generally do not have too many bicycles, which makes perfect sense.

If we take one step further and refer to our analysis of the workplace/resident area, cluster 4 could be stations that are near workplaces. These hill stations that are near workplaces have afternoon bottoms during weekdays as people go to work in the morning and leave in the afternoon, and have very few bicycles as well as no drop of loading score whatsoever in the weekend as people won't ride uphill to these stations in the weekend.

Cluster	Number of stations not on hills	Number of stations on hills
1	341	15
2	134	1
3	117	0
4	64	53
5	198	19
6	65	24
7	143	15

To further exploit theses groups of stations behavior and to potentially adapt some operations that can be used, we give possible direction below:

- Expanding station

Suppose we decided to launch a station expansion operation. According to the clustering, the best stations to target are the stations of cluster 3, as they are working at their almost maximum capacity and the factor limiting them the most is probably their parking capacity.

- Transporting unused bicycles

Suppose we wanted to transport unused bicycles to stations that are lacking to ensure the service is available at a certain rush hour. Depending on the targeted time period, we can identify the receiving and giving stations according to the clustering. We show below some example scenarios of what station to give and to receive bicycles.

Suppose we target the weekday afternoon period. We look to transport bicycles to stations with afternoon demand, which suggests taking clusters 3,4 and 7.

We could also look for receivers in clusters of type *Idle* (in this example of weekday afternoon : cluster 6) since there are always few bicycles in the station. However, if the goal of the operation is only providing for stations idle in high demand, it might be important to study more precise data to determine the situation of these stations because, in fact, *Idle* could mean that the station is in high demand but not reloaded enough (which makes sense to transport extra bicycle to), but it could also mean few bicycles are taken or returned to the station, like some stations on hills for example (which is not worth transporting bicycles to).

Technically, any of the remaining clusters can be the giver, but if we avoid stations having high demand at the same time (cluster 2), stations working at maximum capacity (cluster 3) and undefined stations (cluster 1), we can select stations from cluster 5 to be givers. Finally, we select appropriate givers and receivers based on other criteria (traveling distance, actual loading capacity, etc.)