Data Science – Principles and Applications
Lecturers: Angelo Furno, Tristan Lemoalle (ENTPE)

Final project report

Romain ARNAUD – Maxime DELPLANQUE – Etienne GASTARD

Git link : https://github.com/arnaudromain2003-lang/DSPAP-Project-2025-ARNAUD-DELPLANQUE-GASTARD.git

Analysing TCL validation data to extract typical day profiles and to highlight days that show abnormal usage due to disruptions, events, or changes in mobility habits

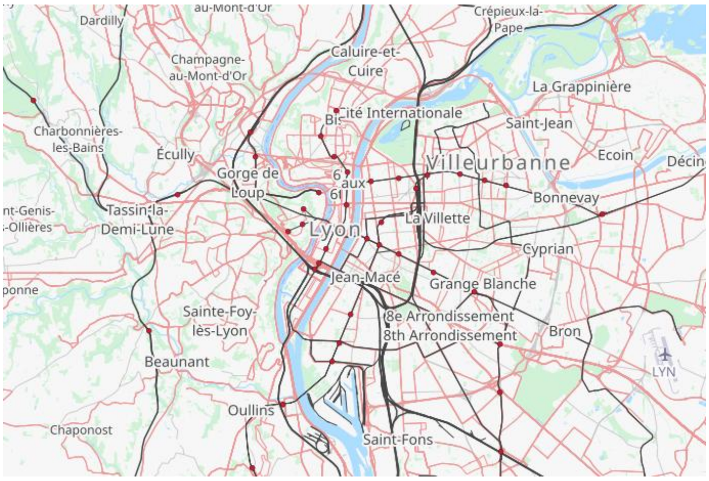


Figure 1 – Extract from OpenStreetMap (© OpenStreetMap contributors, ODbL).

Table des matières

1. Introduction

1.1. Context

This project is part of the final assignment of the lecture: “Data Science : Principles and Applications”. It is based on a part of the **Public Transport Validation Dataset** provided by the Metropole de Lyon. This dataset contains time-series data which reports the number of passenger validations for the three main public transport modes: subway, tramway and bus.

1.2. Dataset structure

The following description is taken from the provided dataset documentation:

> *“The dataset covers the period 1 November 2019 – 30 March 2020 and provides passenger validations aggregated at a 15-minute interval for the three main modes of public transport: bus, tramway, and metro. In addition, spatial reference files for bus/tram stops and metro stations are included.”*
>
> *Source: Presentation of the Public Transport Validation Dataset, A. Furno (ENTPE)*

Among the five original files, we focused on the following three datasets: *bus_indiv_15min.csv, tramway_indiv_15min.csv, subway_indiv_15min.csv*.

For these three modes of transport, we focused on the validation timestamp (“VAL_DATE”) and corresponding flow of passenger validations. For the bus and tramway datasets, only the columns VAL_DATE and Flow have been used. For the subway dataset, we considered VAL_DATE and all station-level flow columns (1 station = 1 column).

1.3. Goals

The goal of this course is to develop a data science project with machine learning (ML) algorithms. The main thread of this project is analysing these datasets to extract typical day profiles (if any) and to highlight anomalies (if any) so we can correlate them with disruption of special events. Then, we will also propose a spatial analysis by municipality of the dataset to check if considering the Metropole as a homogeneous data is relevant.

> ***Research question: Analysing TCL validation data to extract typical day profiles and to highlight days that show abnormal usage due to disruptions, events, or changes in mobility habits***

2. Preliminary work

2.1. Data visualization

First, we explored all three datasets and made some visualizations for each so we can better understand how they are made up and begin to distinguish some noticeable things.

We plotted the whole time series for flow validation so we can highlight a macro-scale pattern. We noticed that there is regularity.

We can clearly see there is a peak during the day and a trough during the night. But, more locally, we can see there are three trends: one for holidays (validations are significantly lower), one for working days, and one for weekends. For all three modes of transportation, the trends looks similar.

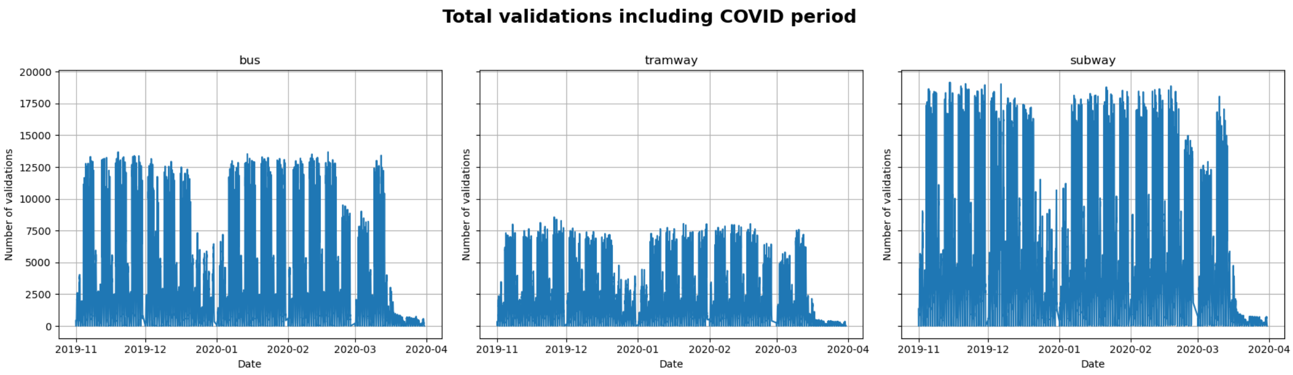


Figure 2 – Total validations including period

We also noticed that there is no longer regularity during the Covid period. Moreover, as Covid was an abnormal period on Earth, we decided to discard all the data that coincide with the lockdown period. The French President announced the lockdown on March 16, 2020, and the country was locked down from March 17. All the data corresponding to this date or after have consequently been discarded from the study so we can capture more regularity and normality in the data.

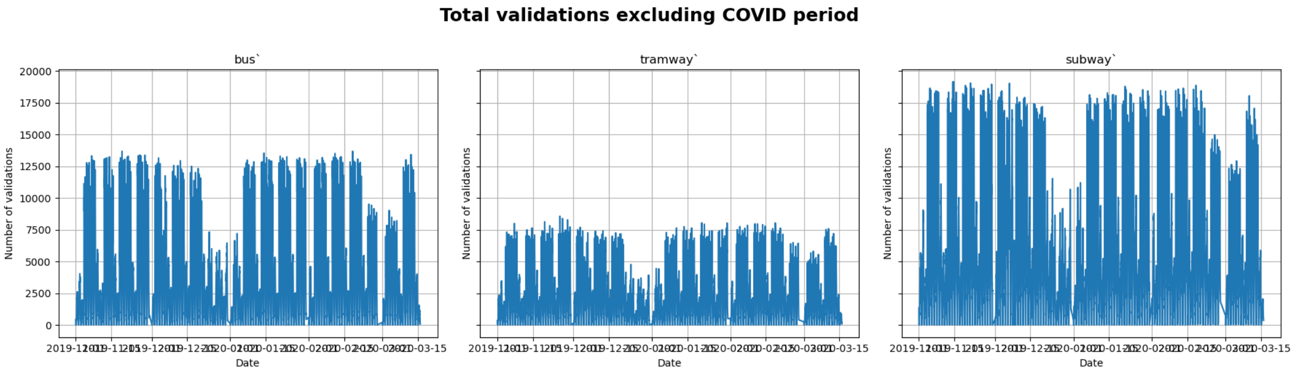


Figure 3 – Total validations excluding COVID period

We then focused on a more micro-scale (a one-week scale). To do so, we needed some “normal weeks”. We arbitrarily assumed that an abnormal week is a week which is far from public holidays, school holidays or strike days. For instance, the week number 6 (February 3rd to February 9th) is considered as normal.

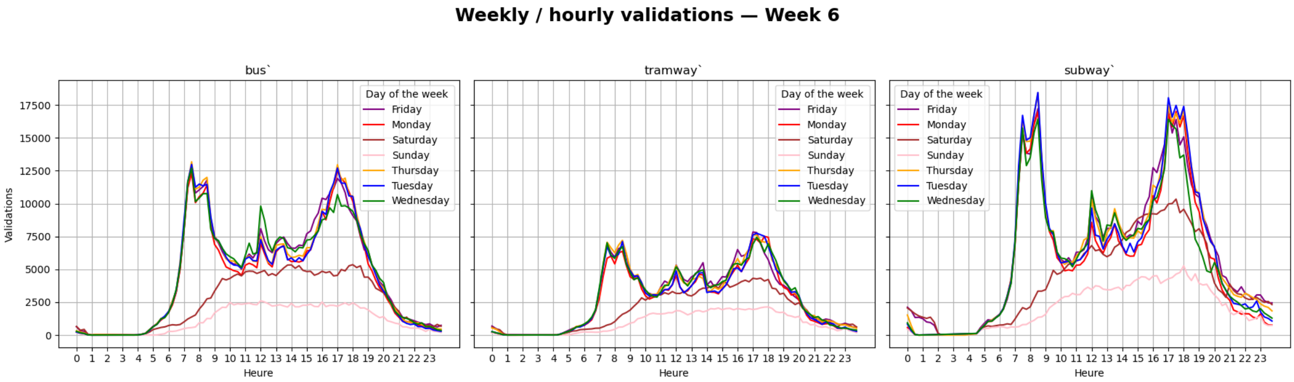


Figure 4 – Weekly / hourly – Week 6

2.2. Pre-processing steps and feature engineering

Our analysis considers public transport at a global level, with no distinction between bus, tramway and subway (except for the spatial analysis, for which we will only use bus data as we don't have enough for the others). We note that the overalls structure of the bus and subway dataset is identical. The subway dataset must be processed first so all three datasets are made up of a compatible structure. As the subway dataset contains the flow for each station, we merged all-station level flow into a unique time series flow columns.

3. Data-analysis and ML methods

3.1. Data-analysis

Based on the previous graphs (visualization step), we noticed that there is regularity when we focus on some specific time periods which follow the local calendar. These observations motivated us to create a calendar-based feature so we can better explain the variations in mobility habits across these time periods. We chose five different time periods*:* *Regular week* (default, when the date doesn't correspond to any other period), *Fete des Lumières* (from 05/12/2025 to 08/12/2025), *Christmas holidays* (school holidays from 21/12/2019 to 05/01/2020), *Winter holidays* (school holidays from 15/02/2025 to 01/03/2025) and *Covid Period* (From 16/03/2020 to 31/03/20). The features are assigned using the *add_week_regularity_feature* of the *df_operations* module.

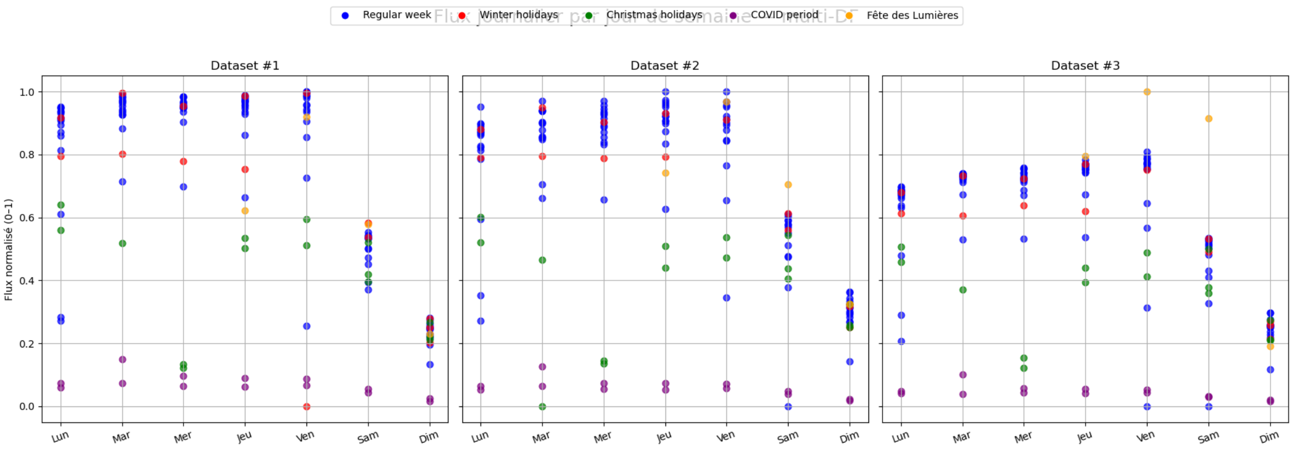


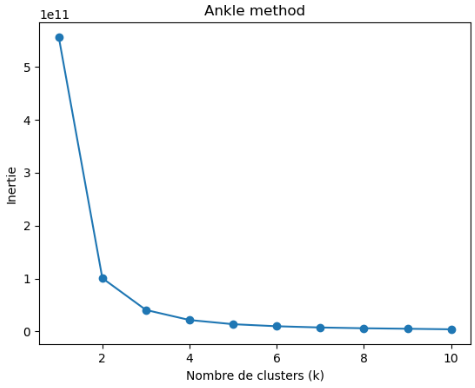
Figure 5 – Normalized daily scatter plot for each transport mod.

First, these graphs demonstrate again the dip during the COVID period. Overall flows are almost zero. But it also highlights that the habits are different depending on which school holiday it is. During Christmas holidays there are less validations than during Winter holidays, which means people use public transport less during Christmas holidays and could travel less during Christmas holidays. During winter holidays days there is only a slight difference with a regular day. Also, overall figures are the same for the three transport modes, but if we take a closer look at the graphs, the effect of Fête des Lumières on public transport usage is clearer for the subway when compared to the other time periods. Moreover, for regular weekdays, public transport usage is least than any other weekday and reaches a maximum on Thursdays.

3.2. ML methods

To get groups of days with similar shape of daily public transport usage, we decided to do a machine learning study. For instance, we decided to do clustering thanks to the *KMeans* method usable thanks to *sklearn.cluster* library.

To select the k number of clusters to do, we did first a study with the ankle method to get the inertia of the clusters and wanted us to get is as low as possible. But we need to select the last one with a significant difference of inertia with the k-1 cluster.



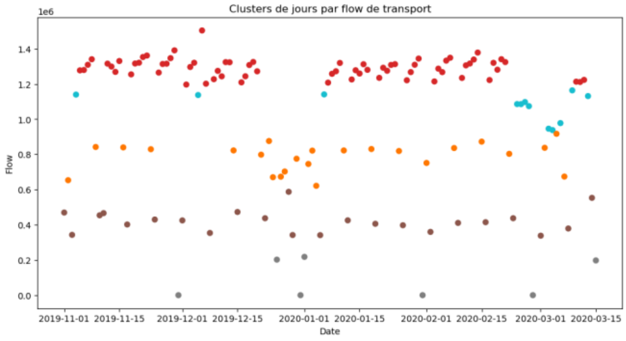


Figure 6: Ankle method

Figure 7: Day clusters by transport flow

With these results, we decided to use 5 clusters as we think that it's the last one corresponding to these criteria shown above. After doing the clusters, we finally have these results.

3.3. Results and interpretation

With the results obtained at the end of the last part, we can se that one cluster seams to be wider to the others (red one), with significantly more days inside of it. To be able to get a better understanding of the clusters, we decided to create a plot with day repartition inside of the clusters, the corresponding plot can be seen just under.

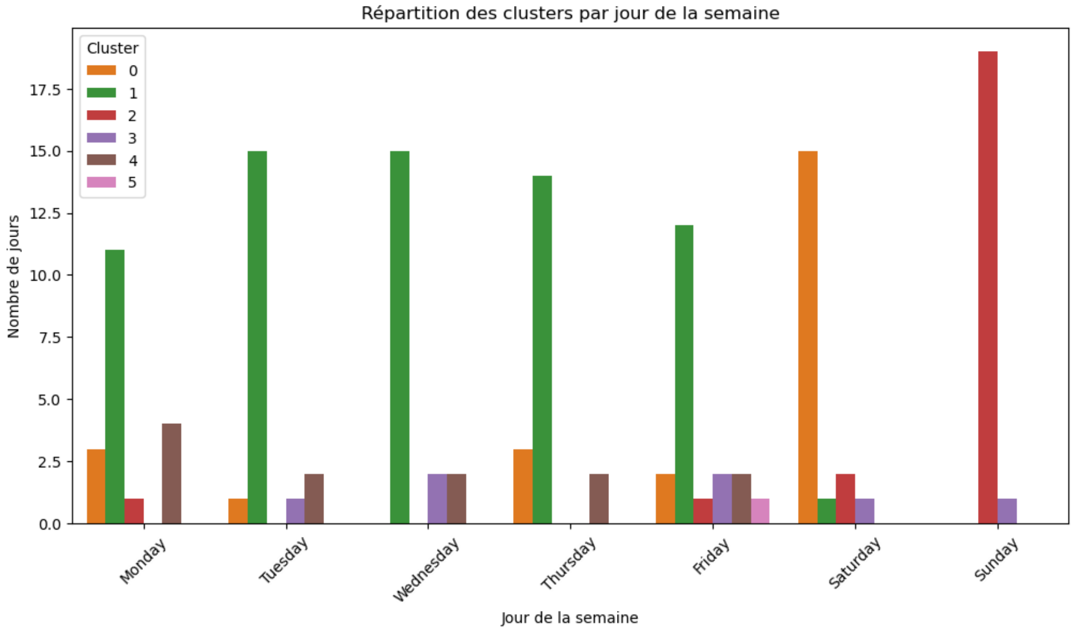


Figure 8: In-cluster day counting

According to this plot, we can start to do assumptions of the cluster significance. For example, cluster n°0 is mainly present on Saturday and cluster n°2 is almost entirely present on Sunday. Both clusters seem to represent weekend situation with cluster 0 for Saturday and cluster 2 for Sunday. Oppositely, cluster n°1 seems to represent weekdays as he is present only in them (except once a Saturday). But for the other clusters, we cannot give a significance right now. For that, we have to create other analysis.

In order to obtain more information, we decided next to get the information on the dates of the clusters. For that we decided to create calendars of each concerned month to be able to understand more things. By using the calendar library, we were able to build these calendars that can be seen just under here.

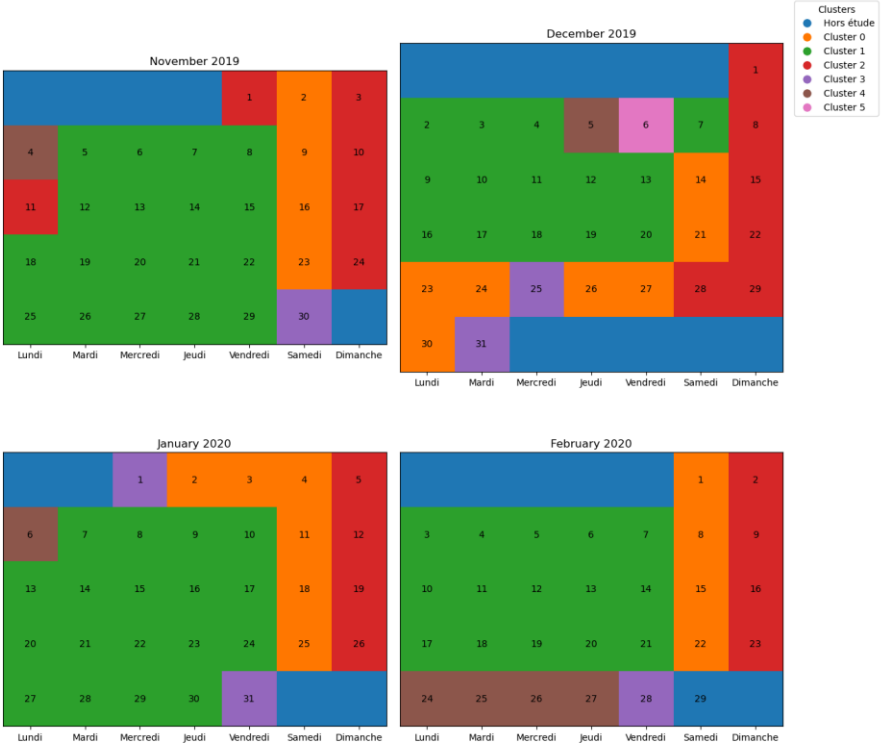


Figure 9: Cluster value shown in calendar

To be able to see thing clearly, the plot of March on the same plot wasn't possible, so March's plot will be put in the appendix.

With these results, we can see again the results shown before as we see again the week cluster, the Saturday cluster and the Sunday cluster. Even if there are some differences with what could have been understand

- Saturdays of December are not in cluster 2: understandable due to the Christmas shopping that make more people use PT to go to the shopping areas, often, these shops are also present in Lyon, where using PT is almost mandatory
- Christmas Holidays: They are mainly in cluster 0 which is like Saturdays
  Winter Holidays: They are mainly in cluster 4, which is a new one,
- 6th of December: Main day of "Fête des Lumières", where everyone is taking PT to go to centre of Lyon to see the numerous lights shows. One of the busiest days (the busiest of the census) of the year for PT usage. This day is the only one in the cluster n°5, showing even more that this day is unique.

- Cluster 4: made mainly of Winter holidays with additional days, seems to be a lower demand compared to classic days but still with peaks as many people are still going to work during winter holidays (at least more than for any other holiday).
- Cluster 3: made of 30th November, Christmas day, New Year Eve/Day, January 31st, February 28th and March 15th
    - For Christmas and New Year period, this can be understood as very low demand period as it's mainly family and friends' period with almost no reasons to use PT throughout the day.
    - March 15th can be understood as the pre-covid lockdown weekday with first demands to reduce trips to reduce the spread of the COVID-19 pandemic
    - February 28th: as we don't have values for February 29th, this can be linked to this with fewer values obtained.
    - November 30th: unless this is linked to the hospital strike that happened this day, no explanations are finable, it may be due to lack of date on this day.

3.4. Typical day per cluster

After getting the cluster in the previous part, to be able to predict how will the daily public transport flow will be, we decided to create typical days per cluster using the mean of the days in each cluster. The results are shown just under this:

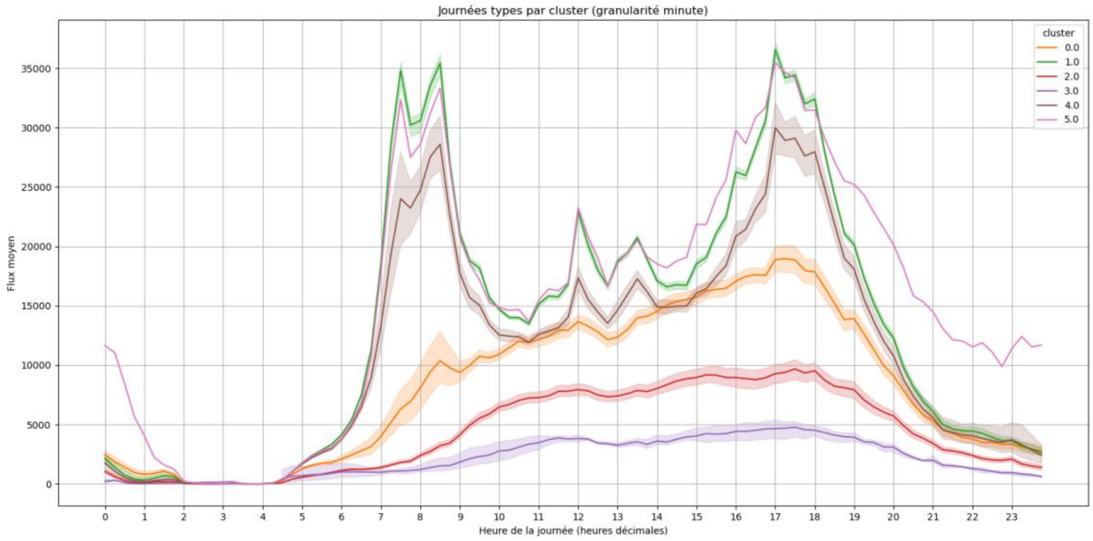


Figure 10: Average day per cluster

As we can see on the graph, we can separate into 2 shapes:

→ Typical weekday shape for cluster 1, 3 and 4 so the classic weekday cluster, the winter holiday cluster and the Friday of "Fête des Lumières" cluster, these were explained before in the previous part.

- To go more into details, the night part of the "Fête des Lumières" cluster can be understood as light shows are during the evening and the first part of the night, and public transport offer is improved throughout the night to allow this passenger flow.

→ Nonpeak shape for Saturday, Sunday and Special days cluster where we don't see any peak.

We still have some error values at some point in the plot, this can be explained as where these errors are high, the number of values are pretty low, not allowing to limit the error values.

4. Critical analysis of the spatial homogeneity hypothesis (add-on)

To process the homogeneity analysis among municipalities, we need to use a second database, **IGN BD TOPO**, as it contains spatial information by municipality, that can be joined with the bus table. Once the databases are merged, we can have, for the bus database, information about which municipality contains each station. We also have to remove some municipalities that does not contain any station. We consider the previous timelines before filtering them, as we need to maximum of data to visualize homogeneity.

Our aim is to check if we can consider the whole Metropole as a unique entity, or if we can detect differences in bus validations through time related to the municipality localization. To do so, and to take into account the fact that we don't have the same population in each city but that we can the same relative flow, we decided to apply for each flow the following normalization:

$$\frac{x - x_{min}}{x_{max} - x_{min}}$$

Then, we are able to plot the relative flow for each municipality, and some of them can be seen below :

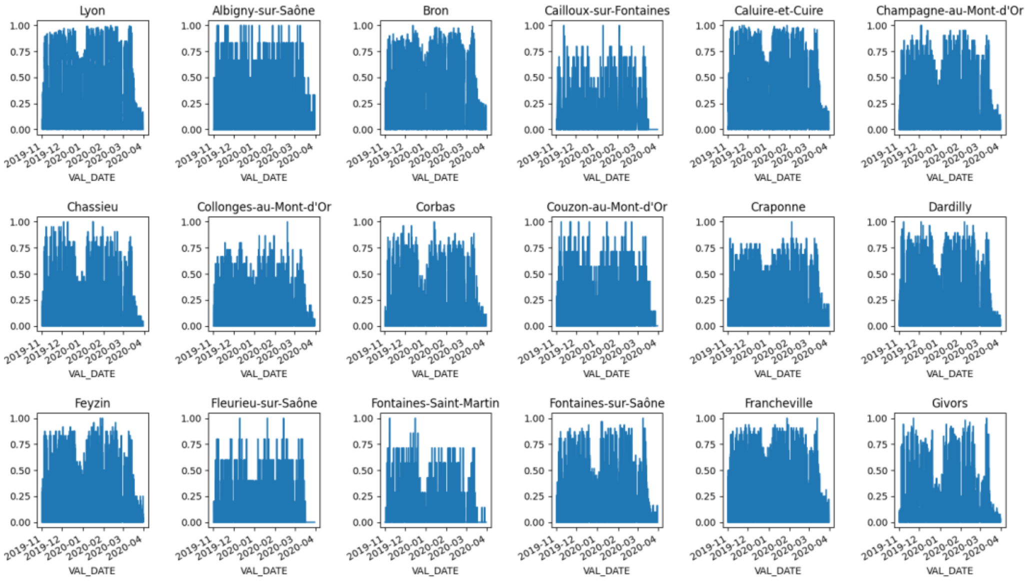


Figure 11: Relative flows of some municipalities during the whole period (from 2019-11 to 2020-04)

We can observe that even if we have normalized our results to avoid strong differences between large and small municipalities, there are several types of timelines in each city, so we need to process a method that allows to identify some clusters of timelines to perform our analysis.

For spatial analysis, in order to check if we can regroup the municipalities into several clusters with a homogeneous timeline, we need to create a matrix that calculates the Euclidian distance between each couple of cities for each date we have. Then, to have normalized results that can be used for the clustering, we divide each distance by the sum of each maximum value between our two municipalities. We then have the following formula:

$$\frac{\sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}}{\sqrt{\sum_{i=0}^{n}\max(x_i, y_i)^2}}$$

Once we have our distance matrix, we are able to use a hierarchical clustering method (only to have a better overview of each municipality with their names) using *scipy.cluster.hierarchy* package and the *linkage* function.

Finally, we were able to obtain several clusters of municipalities that have a similar relative flow timeline for bus stations. As mentioned before, we decided to use a dendrogram to list all the municipality names:

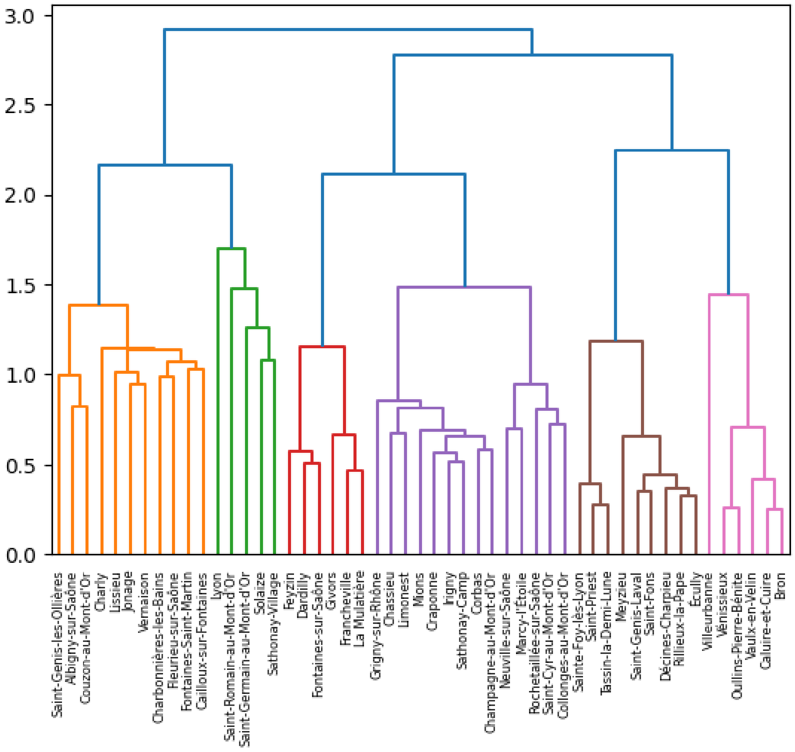


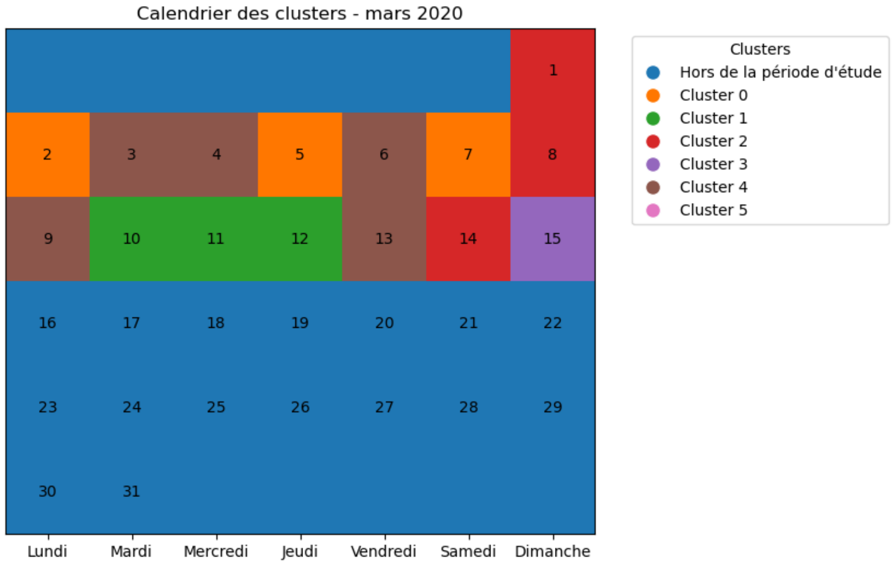
Figure 11: Dendrogram representing the main clusters of municipalities of Lyon based on their bus validation flows

We then have obtained six different clusters into which the relative flows can be considered as homogeneous. In each cluster, the distances between each city are quite low, except for one of them. Indeed, the green cluster contains Lyon which has strong distances with the other cities of the cluster, that are quite small. We can explain that by the fact that Lyon is strongly distant from all the cities of the database, so maybe we can add a new cluster that would contain only the Lyon city. Moreover, we can observe that there is an important correlation between the distances in each cluster and the size of each municipality. Indeed, if we sort these ones by order of size, we can see that the closest cities are the ones that have a similar population size, except for really small or really important cities (this is why Lyon is not in a relevant cluster).

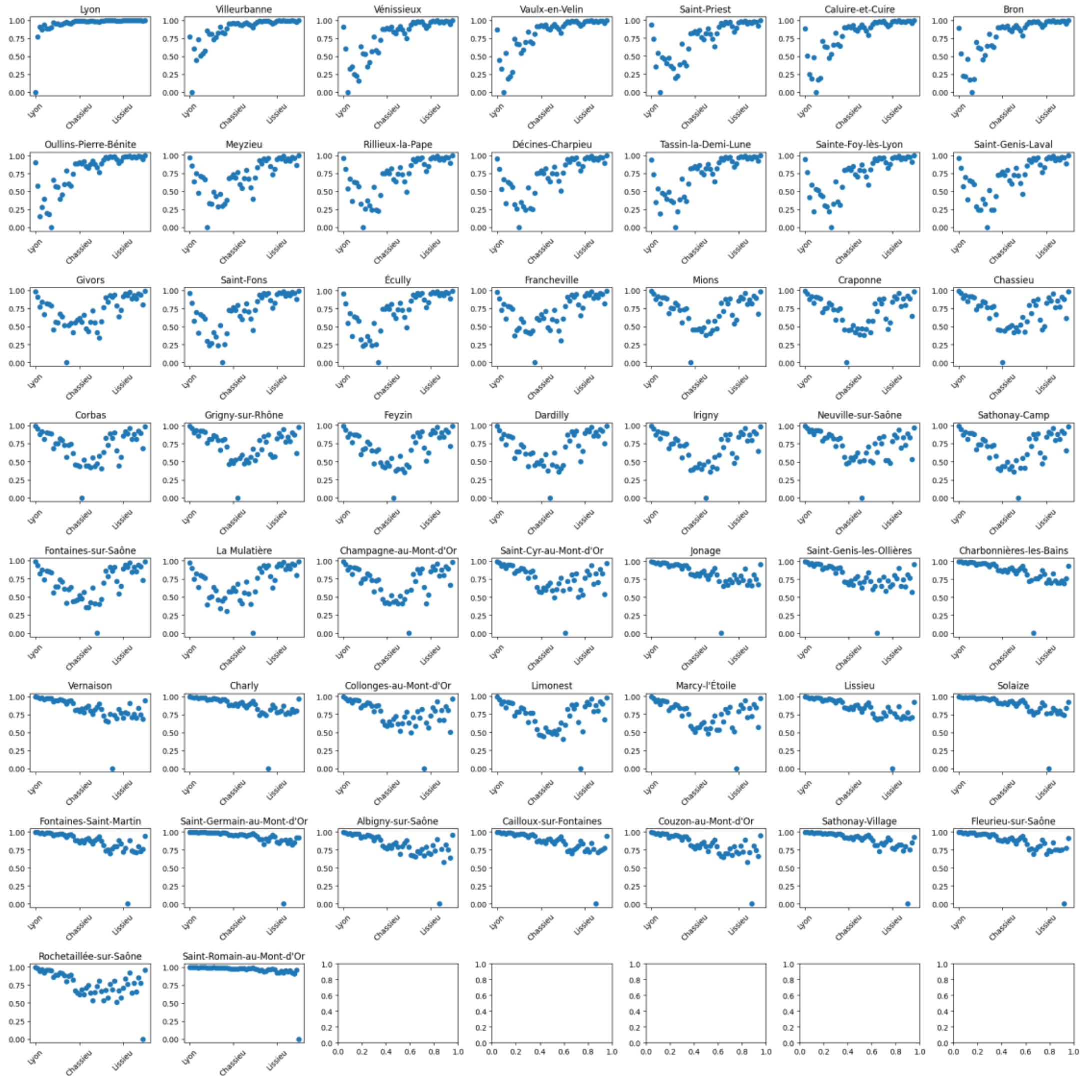
In conclusion, even if we can detect global homogeneity in our dataset, we can repeat our study for 6 or 7 different clusters, in order to have a better understanding of the global population repartition and flow heterogeneity.

5. Appendix

March cluster calendar



Correlation between municipality size and distance among them


Lyon
Villeurbanne
Vénissieux
Vaulx-en-Velin
Saint-Priest
Caluire-et-Cuire
Bron
Oullins-Pierre-Bénite
Meyzieu
Rillieux-la-Pape
Décines-Charpieu
Tassin-la-Demi-Lune
Sainte-Foy-lès-Lyon
Saint-Genis-Laval
Givors
Saint-Fons
Écully
Francheville
Mions
Craponne
Chassieu
Corbas
Grigny-sur-Rhône
Feyzin
Dardilly
Irigny
Neuville-sur-Saône
Sathonay-Camp
Fontaines-sur-Saône
La Mulatière
Champagne-au-Mont-d'Or
Saint-Cyr-au-Mont-d'Or
Jonage
Saint-Genis-les-Ollières
Charbonnières-les-Bains
Vernaison
Charly
Collonges-au-Mont-d'Or
Limonest
Marcy-l'Étoile
Lissieu
Solaize
Fontaines-Saint-Martin
Saint-Germain-au-Mont-d'Or
Albigny-sur-Saône
Cailloux-sur-Fontaines
Couzon-au-Mont-d'Or
Sathonay-Village
Fleurieu-sur-Saône
Rochetaillée-sur-Saône
Saint-Romain-au-Mont-d'Or
Lyon
Chassieu
Lissieu