STUDY OF THE SURROUNDINGS OF THE METRO LINE 1
IN THE CITY OF PARIS, FRANCE

1. Intrdoduction

1.1. Business problem

La Défense business district is a very attractive area located West of Paris. Many urban people work there and take the metro line 1 direct to this area every day. They likely ask where they could live in Paris with easy access to their work site that means they exclude losing time with transport connections. What places combine both a neighborhood with their favorite venues and a reduced transport time to go to work?

1.2. Objective

In order to answer that question, we will study the surroundings of all the stations of the metro line 1 for a business man who is looking for a home close to a metro entrance and close to restaurants and cinemas (for example).

So the project aims to cluster the metro 1 stations according to:

(1) transport time between La Défense station and the home's station,

(2) the numbers of restaurants and cinemas around the home's station.

2. Data presentation: acquisition and preprocessing

2. 1. Data source

- Metro line 1 stations

We can collect the coordinates of the metro stations on the RATP website: https://dataratp2.opendatasoft.com/explore/dataset/positions-geographiques-des-stations-du-reseau-ratp/export/?disjunctive.stop_name&location=9,48.86463,2.39738 (Licence ODbL Version Française)
We can determine the locations of all the stations and the distance between La Défense district and the potential home. So we have an estimation of the transport time.

- Venues

We can use the Foursquare API that provides us a list of the venues around each station in a radius of 300 meters. The result is not exhaustive because it depends on the focused place. However we consider that the Foursquare database is enough supplied for a city like Paris.

2. 2. Data cleaning

- Stations

The downloaded csv file from RATP website includes a complete dataset about public transport of Paris. We select the data that interest us: the geographical coordinates and the names of the 25 stations belonging to the metro line 1 (*cf*. fig.1). We drop all the other columns and rows, make the needed conversions of the data's types, and split the column « coordinates » into two columns « Latitude » and « Longitude ». We control the possible redundant or missing data.

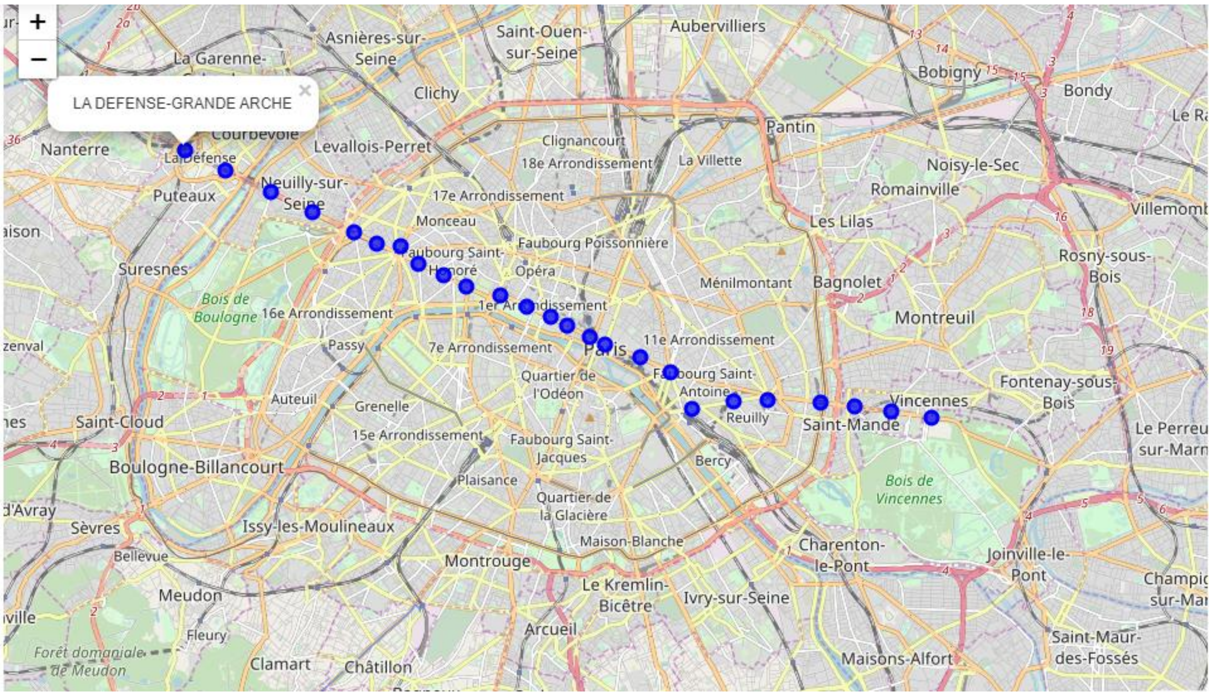


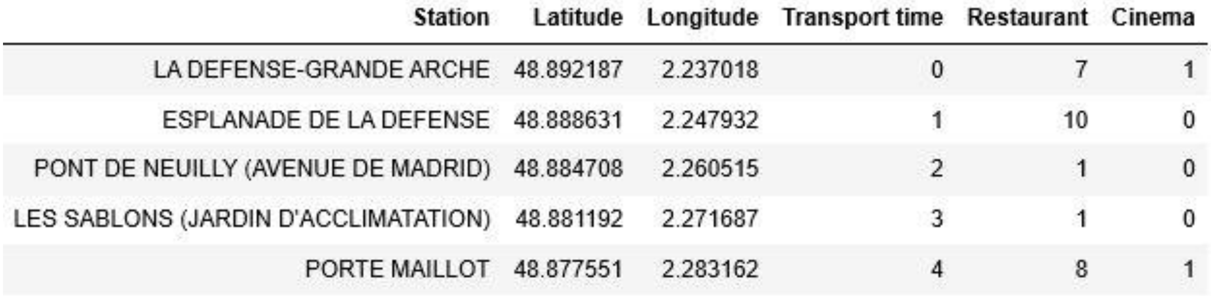
Figure 1. Metro line 1 in Paris, France: location of the 25 stations.
La Défense station is the western end of the line.

We add a column « Transport time » based on the relative distance between each station and La Défense station. The metro line 1 has a W-E orientation, and La Défense station is the western end of the line. So we sort the stations by their latitude and we assign to each station a transport time that is proportional to its position on the line.

- Venues

For each station of the previous stations table, we make a request to the Foorsquare app and get back all the nearby restaurants and cinemas (*i.e.* located in a radius of 300 m). We count the results for the two categories and group the total numbers in a global table.

Our dataset is now ready and contains 25 rows and 6 columns (*cf.* fig. 2). The three variables are : Transport time, Number of restaurants and Number of cinemas.



Station	Latitude	Longitude	Transport time	Restaurant	Cinema
LA DEFENSE-GRANDE ARCHE	48.892187	2.237018	0	7	1
ESPLANADE DE LA DEFENSE	48.888631	2.247932	1	10	0
PONT DE NEUILLY (AVENUE DE MADRID)	48.884708	2.260515	2	1	0
LES SABLONS (JARDIN D'ACCLIMATATION)	48.881192	2.271687	3	1	0
PORTE MAILLOT	48.877551	2.283162	4	8	1

Figure 2. Extract of the stations' dataset after cleaning:
transport time and number of venues for each station of the metro line 1.

3. Data analysis

3.1. Methodology and tools

- Model's choice

The objective is to segment the stations based on similar characteristics they share. We need to find structure in our dataset by grouping the stations according to the transport time and the numbers of venues. The data is unlabelled, it is a case of unsupervised process. So we decide to use a clustering algorithm such as K-Means.

- Tools

We use the Python environment which is very convenient: Pandas, Numpy, Scikit learn to run the modeling process, and the Folium library to create a map of Paris with the clustered stations.

3.2. K-Means clustering

We define our variable X by selecting the features: transport time, number of restaurants, number of cinemas and we normalize the dataset.

- Optimal k

Before applying the K-Means algorithm, we determine the best number of clusters k for our dataset by minimizing the inertia. Indeed a low value of inertia means that the individuals within a cluster are very close to each other. So we look for the smallest acceptable value of k linked to the best minimal inertia by testing various k. We fit the model with X and visualize the resulting inertias *(cf.* fig.3).

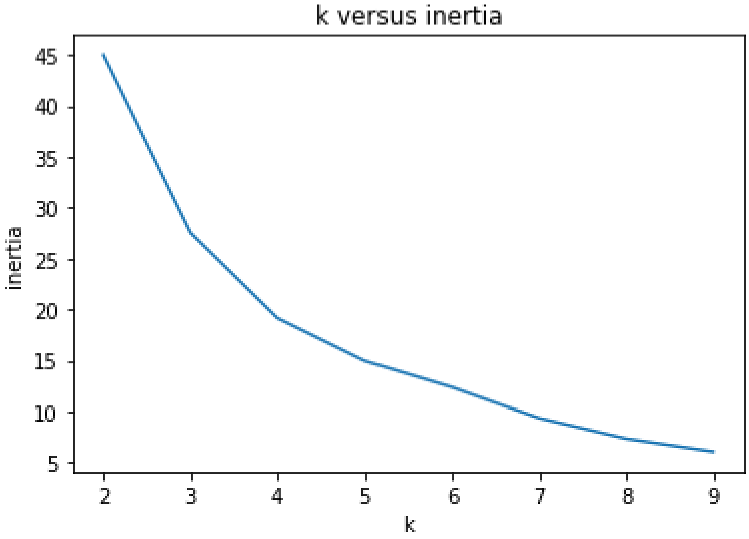
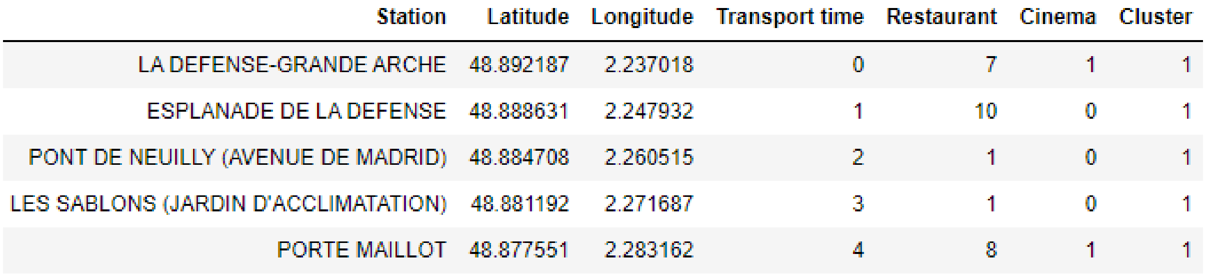


Figure 3. Variation of the inertia for different numbers of clusters. When k increases, the inertia decreases.

We choose k=5 by the elbow method, even if k=7 could give a more detailed clustering. Here we don’t want too many groups.

- K-Means processing

Then we fit the K-Means model the normalized data X.
Finally we assign the predicted cluster to each station and add a column « Cluster » in the dataframe (cf. fig. 4).

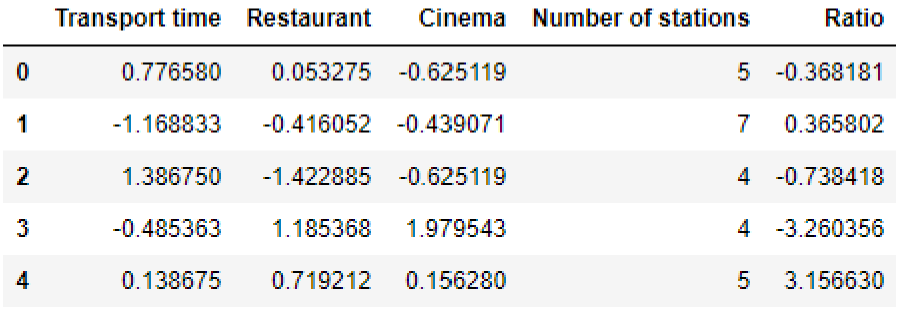


Station	Latitude	Longitude	Transport time	Restaurant	Cinema	Cluster
LA DEFENSE-GRANDE ARCHE	48.892187	2.237018	0	7	1	1
ESPLANADE DE LA DEFENSE	48.888631	2.247932	1	10	0	1
PONT DE NEUILLY (AVENUE DE MADRID)	48.884708	2.260515	2	1	0	1
LES SABLONS (JARDIN D'ACCLIMATATION)	48.881192	2.271687	3	1	0	1
PORTE MAILLOT	48.877551	2.283162	4	8	1	1

Figure 4. Extract of the stations’ dataframe after K-Means modeling: predicted cluster for each station of the metro line 1.

We can present the attributes of the centroides of the 5 clusters in the table below (*cf.* fig. 5): the 3 medians corresponding to the 3 variables (Transport time, Restaurant, Cinema) and the ratio « total venues / transport time ». We also count the stations forming each cluster and ad dit in the column « Number of stations ».

.



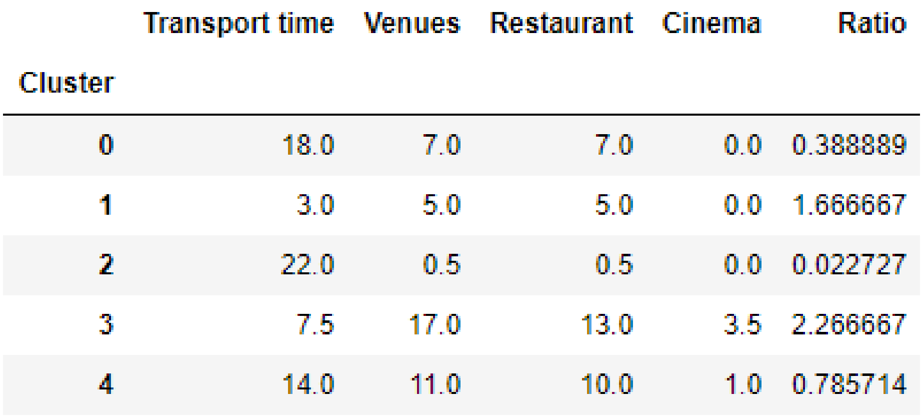
	Transport time	Restaurant	Cinema	Number of stations	Ratio
0	0.776580	0.053275	-0.625119	5	-0.368181
1	-1.168833	-0.416052	-0.439071	7	0.365802
2	1.386750	-1.422885	-0.625119	4	-0.738418
3	-0.485363	1.185368	1.979543	4	-3.260356
4	0.138675	0.719212	0.156280	5	3.156630

Figure 5. Normalized centroïdes for the 5 predicted clusters.

4. Results

The 25 stations are clustered into 5 well-balanced groups: 3 clusters are composed by 4 stations and 2 clusters by 6 or 7 stations. The distribution is rather homogeneous. We can show the median value of the 3 variables calculated after modelisation (cf. fig.6) maybe more concrete and easy to represent than the normalized centers' table.

:



	Transport time	Venues	Restaurant	Cinema	Ratio
Cluster					
0	18.0	7.0	7.0	0.0	0.388889
1	3.0	5.0	5.0	0.0	1.666667
2	22.0	0.5	0.5	0.0	0.022727
3	7.5	17.0	13.0	3.5	2.266667
4	14.0	11.0	10.0	1.0	0.785714

Figure 6. Medians calculated for each cluster after modelling.
Ratio=Venues/Transport time

In order to visualize and understand the similarities and the differences that characterize the clusters we build a comparative bar plot (cf. fig.7).

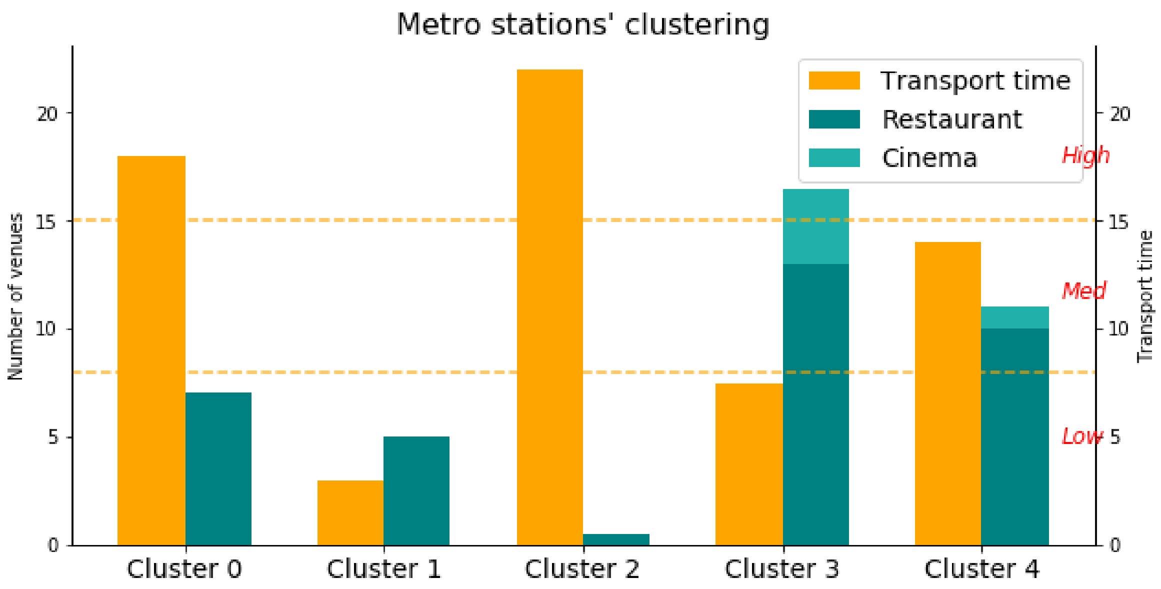
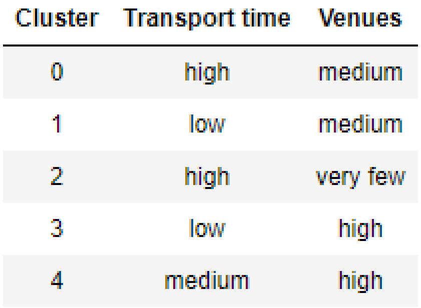


Figure 7. Predicted clusters for the 25 stations of the metro line 1 (K-Means model). Number of venues(Restaurants and Cinemas) and Transport time for the 5 clusters. (the dashed lines delimit the areas of low, mdeium or high transport time)

We see that the farthest stations with very few venues are gathered in Cluster 2. Cluster 0 regroups distant stations with some restaurants. Cluster 4 has an intermediate transport time, many restaurants and some cinemas. Cluster 1 is formed by the closest stations with heterogenous number of venues (some restaurants and very few cinemas). Cluster 3 has a low transport time and many venues, and moreover presents the best ratio Venues/Transport time. We notice that only Clusters 3 and 4 are likely to offer cinemas.

We can resume the attributes of the clusters in the table below (cf. fig.8).



Cluster	Transport time	Venues
0	high	medium
1	low	medium
2	high	very few
3	low	high
4	medium	high

Figure 8. Interpretation of the results: attributes of the 5 predicted clusters. Categorical classification of transport time and number of venues.

We finally visualize the prediction by mapping the 25 clustered stations (cf. fig.9). The stations are represented by circles whose color depends on the cluster and whose radius depends on the number of the nearby venues.

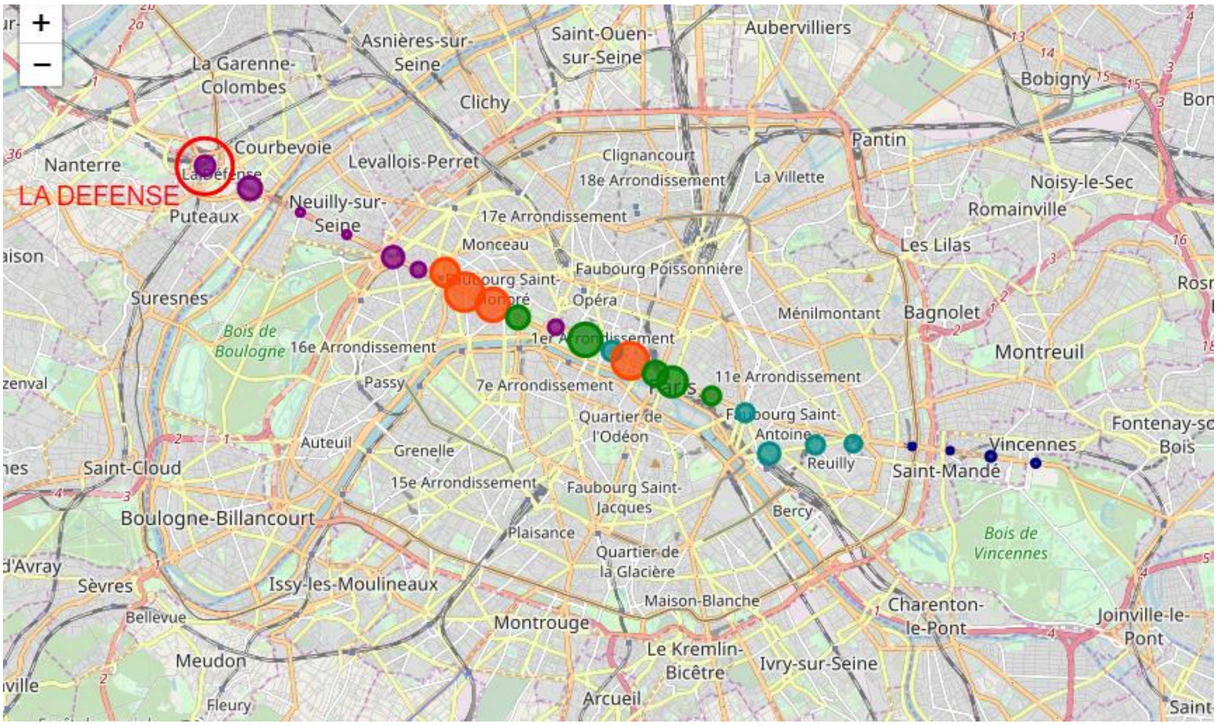


Figure 9. Map of Paris with the clustered stations of the metro line 1.
Each station is figured by a circle. Its radius is proportional to the number of venues and its color is linked to the predicted cluster (cyan: 0, purple:1, blue: 2, orange: 3, green: 4).

5. Discussion

Based on the previous results, we can recommend Cluster 3 to the business man who works at La Défense and looks for a home near a metro station of the line 1 and a neighborhood with facilities such as restaurants and cinemas.
Indeed the best location after modelling is a station in Cluster 3, with proximity to La Défense and a large choice of venues. An alternative solution can be found in cluster 4 that presents a good compromise between venues and distance (many venues and intermediate transport time). If a shorter distance is preferred, cluster 1 is convenient in spite of less venues' choice. Stations in clusters 0 and 2 can obvioulsly be ruled out.

However we must remember that the result here is highly dependant to the Foursquare database. It would be judicious to check with additional data about venues, in particular in the eastern part of the line in the Vincennes surburb that seems under-represented via Foursquare API.

6. Conclusion

Here we use the K-Means modelling in order to help people to choose a place with some initial criteria. We made predictions in a particular case by clustering stations (a specific metro line, a preference for some venues). We could generalize this approach and handle more complex issues, with more initial variables and more data. The Foursquare database does not provide sufficient information and additional geographical data would be welcome. We could widen the problematic by including data on rent in Paris, or different kinds of transport, etc. These are examples for future interesting investigations lying on the potential of Machine Learning.

C. BRINON
December 2019