Surroundings of a metro line in Paris

CLUSTERING THE METRO STATIONS USING MACHINE LEARNING

December 2019
Coursera – IBM Capstone Project

Claire BRINON

1. Introduction

La Défense business district, West of Paris (France)

- A very attractive area for many business people
- A work site located at the end of the metro line 1

Where are the best places to live if…

- you work in La Défense district
- and you want to live near a station of the metro line 1 …?

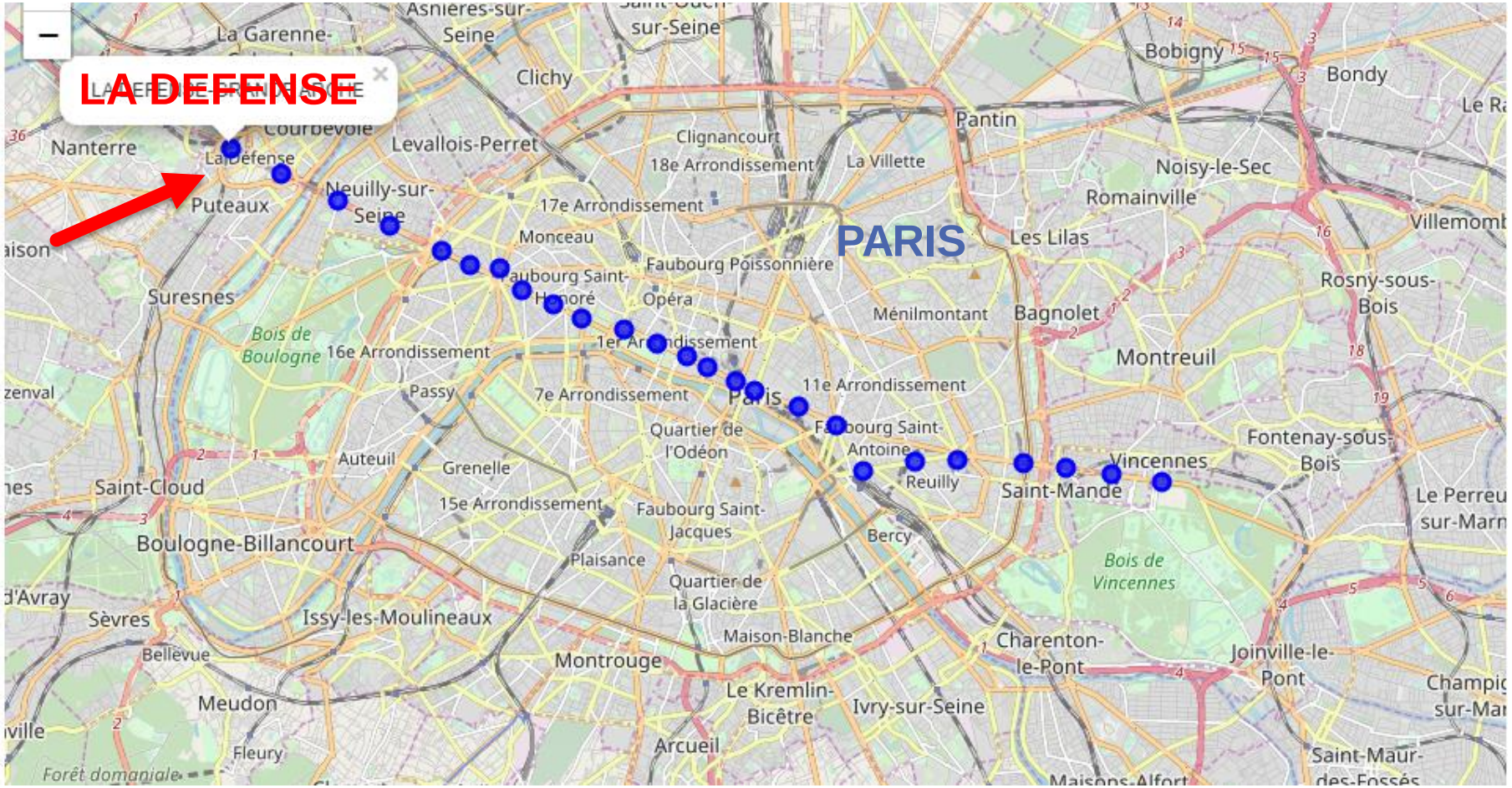
2. Objective
Study of the surroundings of the metro line 1

Criteria for the home's location

- Nearby a metro entrance of the line 1
- Neighbor with restaurants and cinemas

Clustering of the stations according to:

- Transport time between La Défense station and the potential station
- Number of venues (restaurants and cinemas) around the potential station



Location of the 25 stations of the metro line 1 in Paris

Visualization of the studied stations and their neighborhood

3. Data presentation

Sources

- Names and geographical coordinates for each station:

  RATP website (open data)

- Numbers of restaurants and cinemas near each station:

  Foursquare API

Data preprocessing

- Selection and cleaning of the needed data
- Dataframe with 25 rows and 6 columns
- Station/Coordinates/Transport time /Venues

Extract of the dataset after cleaning

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

4. Data analysis

Segmentation of stations based on similar attributes ➡ **Clustering method**

- Model: K-Means algorithm
- 3 variables:
  - Transport time
  - Number of restaurants
  - Number of cinemas
- Optimal number of clusters:
  - k=5

- Model fitting with the data
- Prediction of a cluster assigned to each station
- Python and its libraries (Pandas, Numpy, Scikit kearn, Matplotlib, Folium)

PREDICTED CLUSTER

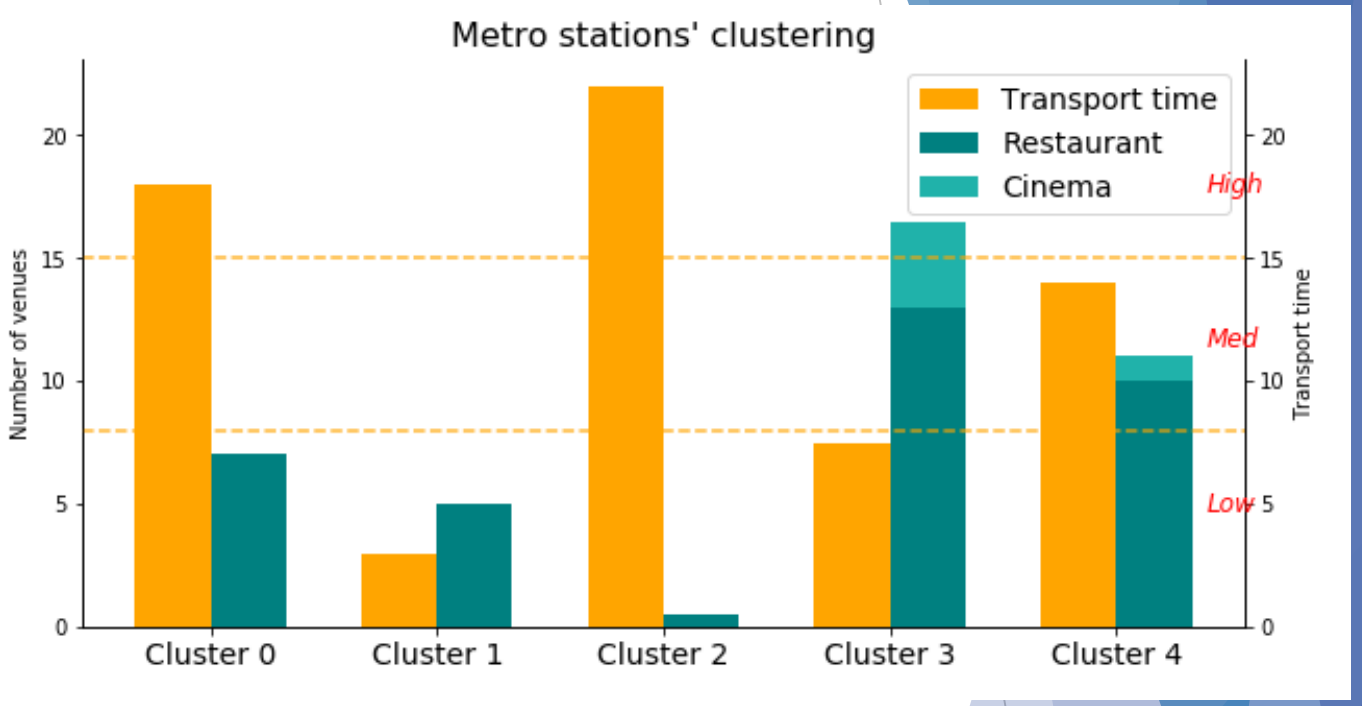
Extract of the dataset after modeling

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

5. Results

- 5 clusters of 4-7 stations each
- Attributes of the groups: what similarities and differences?

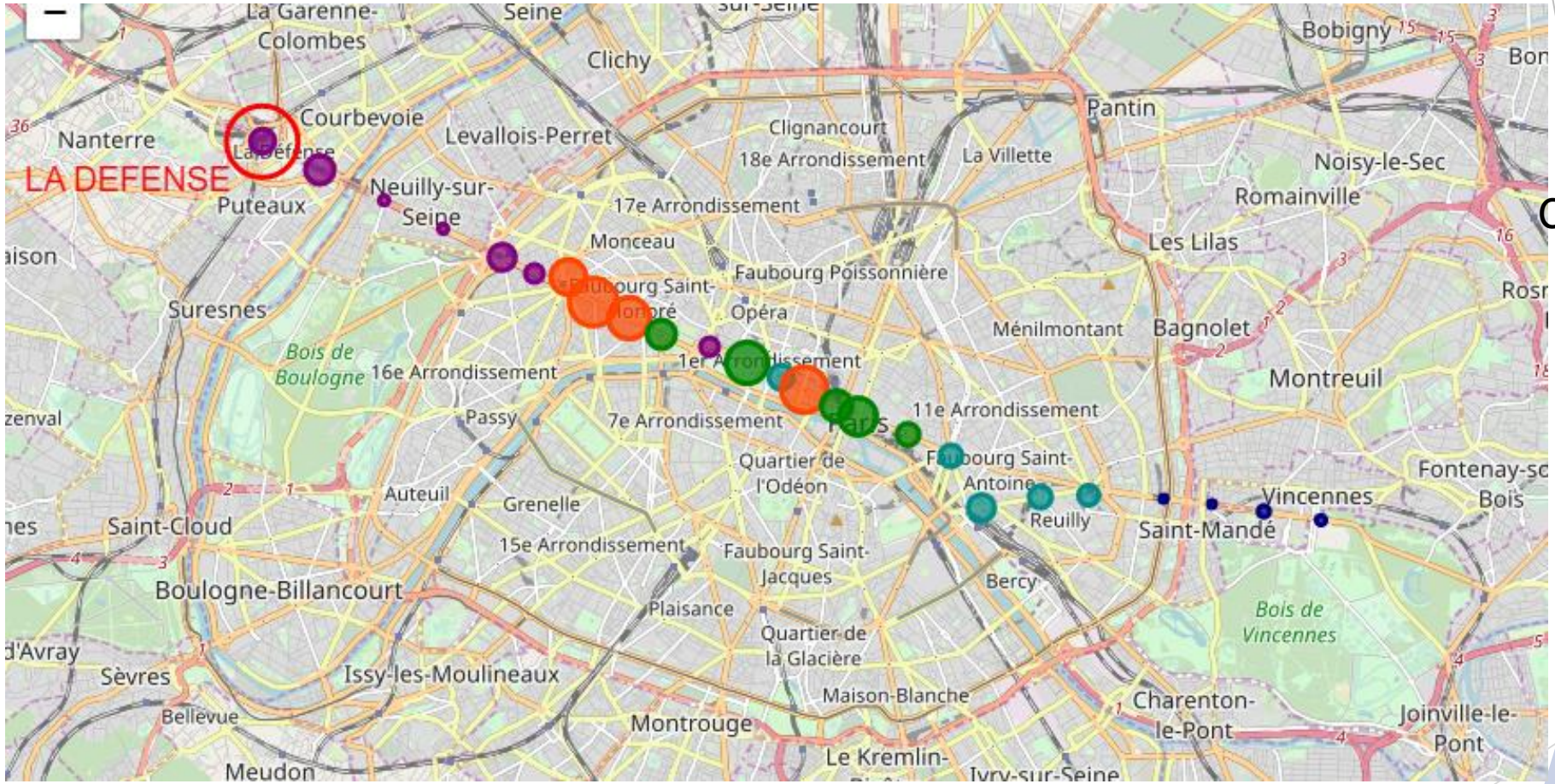
Cluster	Transport time	Venues
0	high	medium
1	low	medium
2	high	very few
3	low	high
4	medium	high



The **best locations** regarding the transport time and the neighborhood !!

- Clusters 0-2: distant stations with few venues
- Cluster 1: nearby stations with restaurants
- Cluster 4: medium distance with larger choice of venues
- Cluster 3: nearby stations and a lot of restaurants and cinemas

6. Map of the clustered stations



CLUSTER

- 0
- 1
- 2
- 3 ★
- 4

Location of the clustered stations of metro line 1

The radius of each circle is proportional to the number of venues around the station.

Recommended cluster

7. Conclusion and perspectives

- The unsupervised clustering helps to choose the best home's location on metro line 1 with initial criteria.
- Recommendations can be provided to the client.

★★★

- Limit of Foursquare database
- Further modelling and refinement with additional data about venues and variables (transport, rent in Paris…)
- Application of machine learning process to solve more complex issues

THANK YOU FOR YOUR ATTENTION !!!