Optimization of Barcelona’s new subway line

Laia Alcaraz, December 2020

1. Introduction

Barcelona is a city located in Spain with a population of 1.6 million people. As other cities in Europe, Barcelona has a subway transportation system which in 2019 was used 411.95 million times. This subway transportation system is called Metro and it is the most popular one in the city.

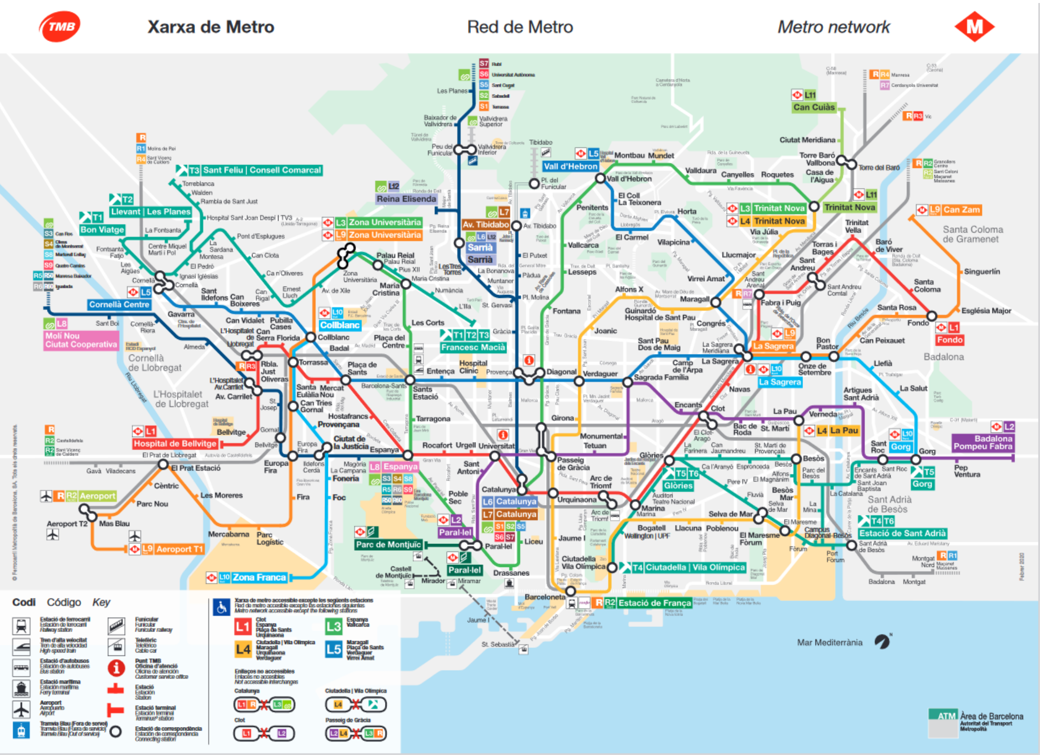


Figure 1: Barcelona’s Metro network (source: www.tmb.cat)

Currently, the line 9 of the Metro, which is expected to be 27.7 km long (this will make it the longest in Europe) is being built, see figures 2 and 3 for more details. The central part of the L9 line includes some stops that already belong to other lines, whose position can of course not be changed. However, it would be of interest to study if the planned coordinates of the stops which are still not built have been selected according to the necessities of the people who live in those areas, that is: proximity to hospitals, universities and schools, places of interest, etc.

The stops that will be studied are the following ones:

- Campus Nord
- Manuel Girona
- Prat de la Riba
- Mandri
- Muntanya
- Sanllehy
- Maragall

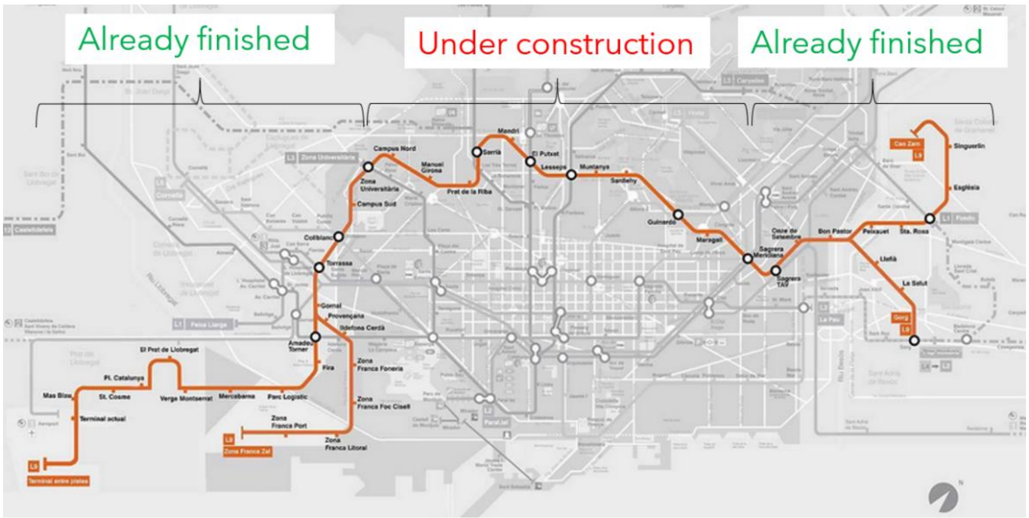


Figure 2: L9 line of Barcelona's Metro (source: Pinterest)

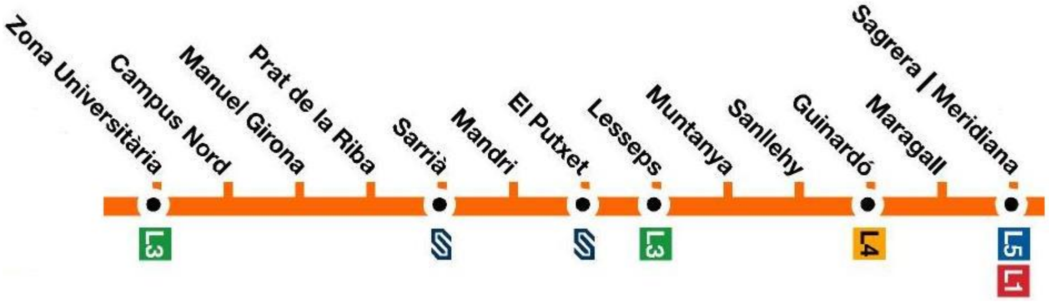


Figure 3: Details of the central section of the L9 (source: www.tmb.cat)

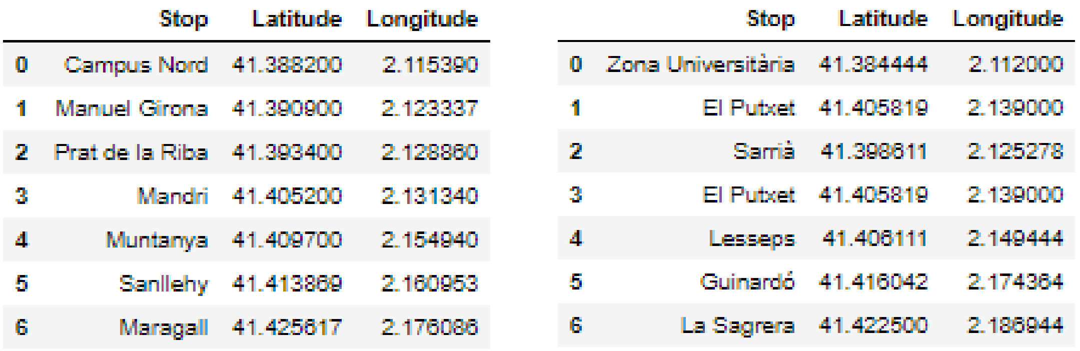
The study shown in this report could be of interest for Transports Metropolitans de Barcelona (TMB) which is the public transport operator from Barcelona and one of the biggest in Spain, because it will bring information on the venues that different parts of Barcelona have and also which is the best way to link them.

2. **Data**

The first step of the study consists in importing the data that will be used. In this case, we need two different types of data:

- Coordinates (latitude and longitude) of the L9 stops. We are mainly interested in the ones that are under construction, but we will also get the coordinates of the ones which are already built.
- Top 100 venues (latitude, longitude, type of venue) located within 500 m of the stops under construction

The coordinates of the Metro stops can be easily found in Wikipedia and can be stored in two different Pandas dataframes:



	Stop	Latitude	Longitude
0	Campus Nord	41.388200	2.115390
1	Manuel Girona	41.390900	2.123337
2	Prat de la Riba	41.393400	2.128860
3	Mandri	41.405200	2.131340
4	Muntanya	41.409700	2.154940
5	Sanllehy	41.413869	2.160953
6	Maragall	41.425617	2.176086

	Stop	Latitude	Longitude
0	Zona Universitària	41.384444	2.112000
1	El Putxet	41.405819	2.139000
2	Sarrià	41.398611	2.125278
3	El Putxet	41.405819	2.139000
4	Lesseps	41.406111	2.149444
5	Guinardó	41.416042	2.174364
6	La Sagrera	41.422500	2.186944

Figure 4: Dataframe containing the stops which will be optimized (left), dataframe containing the stops which have already been built (right).

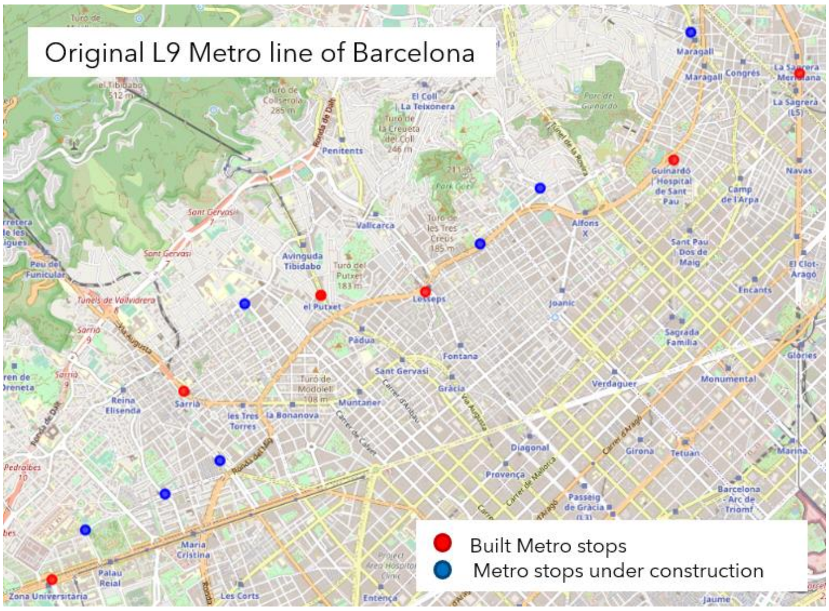
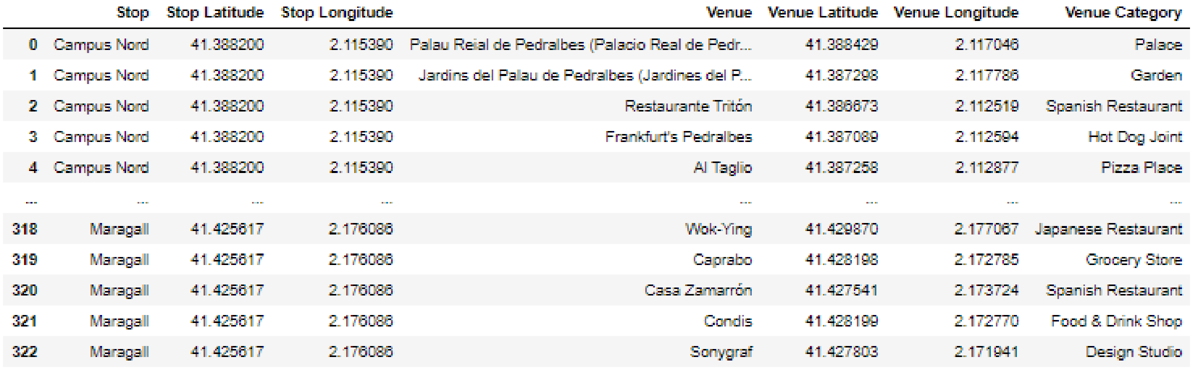


Figure 5: Coordinates of the L9 Metro line.

To get the venues information, we use the Foursquare API and choose to get, for each stop under construction, the top 100 venues that are within a radius of 500 m. After that, we combine the information of the stops and the ones from the venues to build the main dataframe of the study.



	Stop	Stop Latitude	Stop Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Campus Nord	41.388200	2.115390	Palau Reial de Pedralbes (Palacio Real de Pedr...	41.388429	2.117046	Palace
1	Campus Nord	41.388200	2.115390	Jardins del Palau de Pedralbes (Jardines del P...	41.387298	2.117786	Garden
2	Campus Nord	41.388200	2.115390	Restaurante Tritón	41.386673	2.112519	Spanish Restaurant
3	Campus Nord	41.388200	2.115390	Frankfurt's Pedralbes	41.387089	2.112594	Hot Dog Joint
4	Campus Nord	41.388200	2.115390	Al Taglio	41.387258	2.112877	Pizza Place
...	...	...	...	...	...	...	...
318	Maragall	41.425617	2.176086	Wok-Ying	41.429870	2.177067	Japanese Restaurant
319	Maragall	41.425617	2.176086	Caprabo	41.428198	2.172785	Grocery Store
320	Maragall	41.425617	2.176086	Casa Zamarrón	41.427541	2.173724	Spanish Restaurant
321	Maragall	41.425617	2.176086	Condis	41.428199	2.172770	Food & Drink Shop
322	Maragall	41.425617	2.176086	Sonygraf	41.427803	2.171941	Design Studio

Figure 6: Dataframe combining stops coordinates and venues information.

3. Methodology

3.1. Data visualization and processing

Since we are interested in optimizing each stop as a function of the venues around it, we must know which is the category of the venues we have. By looking at the last column of figure 6 we can already see that there are many different venue categories, however a bar plot like the following one is more useful:

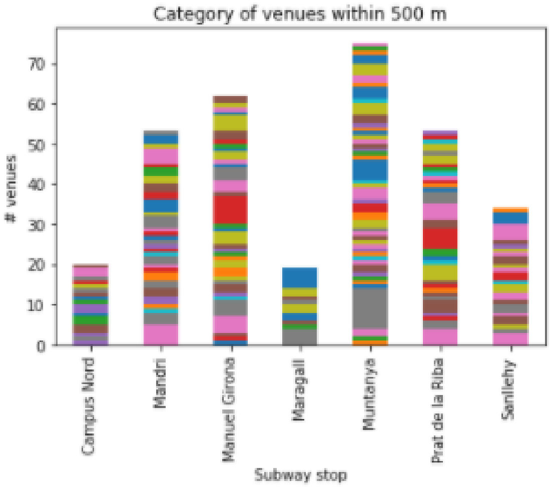


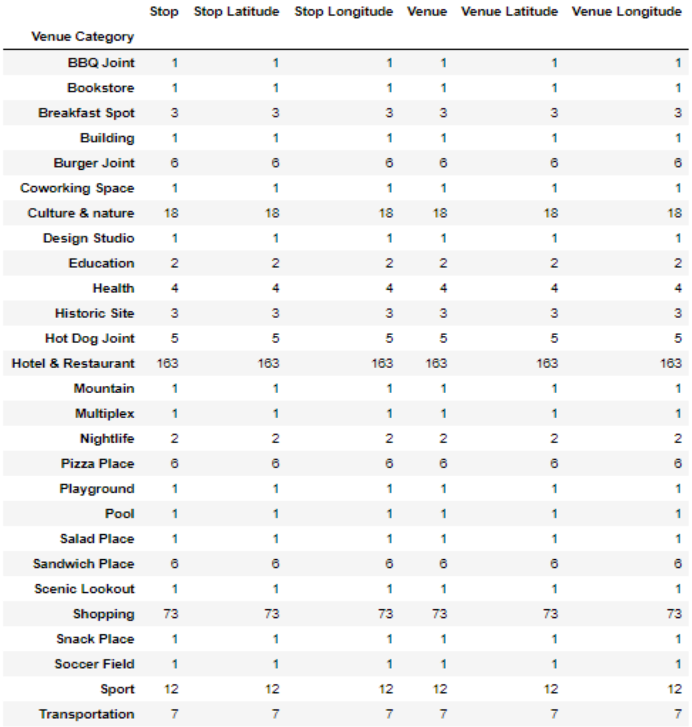
Figure 7: Category of the venues for each Metro stop. The legend has been deleted for the sake of simplicity.

As the plot in figure 7 shows, there are too many different venue categories, therefore a function is defined to help us group them into the following ones:

- Culture & nature
- Education
- Health
- Hotel & restaurant
- Nightlife
- Others
- Shopping
- Sport
- Transportation

The function looks for key words associated to each category and assigns one of the 9 possibilities above to the venue. For instance, in the case of Education, the key words are College, University and School, this means that any venue which contains one of these words will be considered to belong to the Education Category.

A first cleaning step is done, however there are some venues that are not filtered correctly.



	Stop	Stop Latitude	Stop Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
BBQ Joint	1	1	1	1	1	1
Bookstore	1	1	1	1	1	1
Breakfast Spot	3	3	3	3	3	3
Building	1	1	1	1	1	1
Burger Joint	6	6	6	6	6	6
Coworking Space	1	1	1	1	1	1
Culture & nature	18	18	18	18	18	18
Design Studio	1	1	1	1	1	1
Education	2	2	2	2	2	2
Health	4	4	4	4	4	4
Historic Site	3	3	3	3	3	3
Hot Dog Joint	5	5	5	5	5	5
Hotel & Restaurant	163	163	163	163	163	163
Mountain	1	1	1	1	1	1
Multiplex	1	1	1	1	1	1
Nightlife	2	2	2	2	2	2
Pizza Place	6	6	6	6	6	6
Playground	1	1	1	1	1	1
Pool	1	1	1	1	1	1
Salad Place	1	1	1	1	1	1
Sandwich Place	6	6	6	6	6	6
Scenic Lookout	1	1	1	1	1	1
Shopping	73	73	73	73	73	73
Snack Place	1	1	1	1	1	1
Soccer Field	1	1	1	1	1	1
Sport	12	12	12	12	12	12
Transportation	7	7	7	7	7	7

Figure 8: Venues after the first cleaning step.

Therefore, a second cleaning step is done, this one is more specific and focuses on the words from figure 8, that is: BBQ Joint, Burger Joint, Pizza Place, etc are associated to Hotels & restaurant.

After this second cleaning step, the data has the desired structure and can be used to analyse its distribution. For instance, it can be clearly seen that Muntanya and Manuel Girona are the stops which have more venues. On the other hand, Maragall and Campus Nord are the ones with the lowest number of venues. But in the case of Campus Nord, it must be said that it is the one with the highest variety of venues.

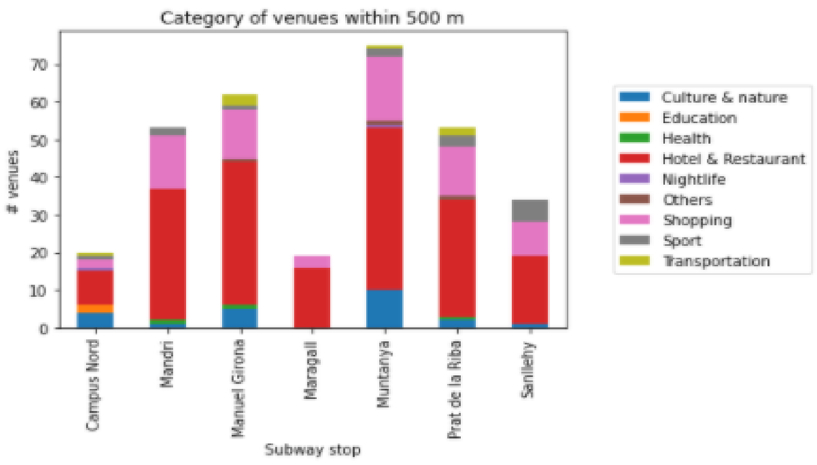


Figure 9: Category of the venues for each Metro stop after two cleaning steps.

3.2. Machine learning algorithms

The K-means algorithm is an unsupervised clustering algorithm that aims to partition unlabelled data into a k number of clusters which share important characteristics. Figures 10 and 11 show a scheme and a pseudocode of how this algorithm works.

K-means is very popular and has a big number of applications, for instance:

- Customer segmentation: it might be of interest to group the costumers that have a similar profile, so that specific products or campaigns can be offered to them.
- Fraud detection: it is possible to isolate new claims based on its proximity to clusters that indicate fraudulent patterns.

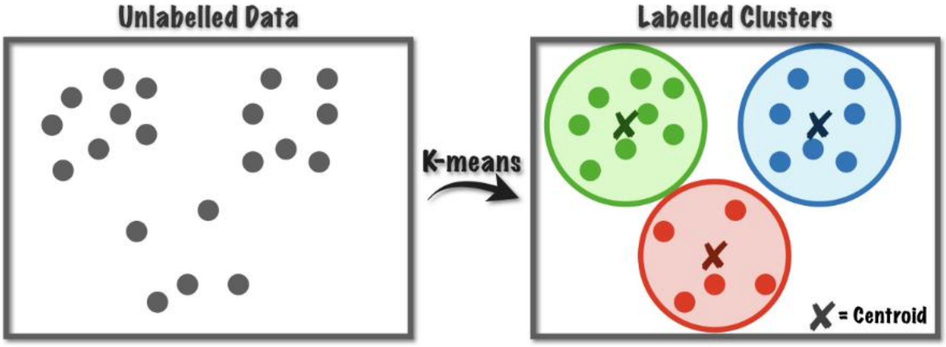
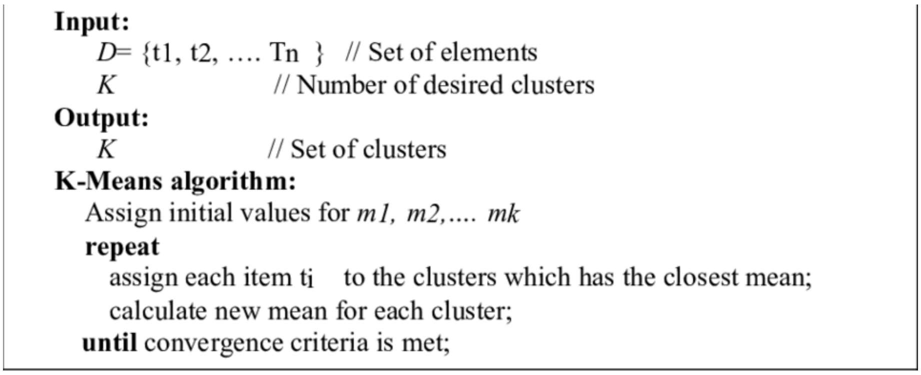


Figure 10: K-means scheme (source: https://www.towardsdatascience.com)



```
Input:
    D= {t1, t2, .... Tn }   // Set of elements
    K                       // Number of desired clusters
Output:
    K                       // Set of clusters
K-Means algorithm:
  Assign initial values for m1, m2,.... mk
  repeat
    assign each item ti   to the clusters which has the closest mean;
    calculate new mean for each cluster;
  until convergence criteria is met;
```

Figure 11: K-means pseudocode (source: Research Gate).

Since our goal is to find the optimal position of the Metro stops, we can use the K-means algorithm and cluster the venues based on their latitude and longitude. By working this way, the centroids of each cluster will be the optimized stop.

Before training the algorithm, two important questions must be answered:

- From figure 9 it can already be seen that the stops with less venues will be penalised by the k-means algorithm. However, the Campus Nord station is the one with the most variety of venues, something that should be considered when building a Metro stop. Therefore, it is natural to wonder if there is a possible way to take the variety of the venues into account.
- Which is the number of clusters into we should group our data?

The answer to the first question is to use the weighted K-means algorithm, which consists in assigning weights to the different points of the dataset. The higher the weight is, the more the centroid of the cluster is pulled to its direction, as shown in the following figure:

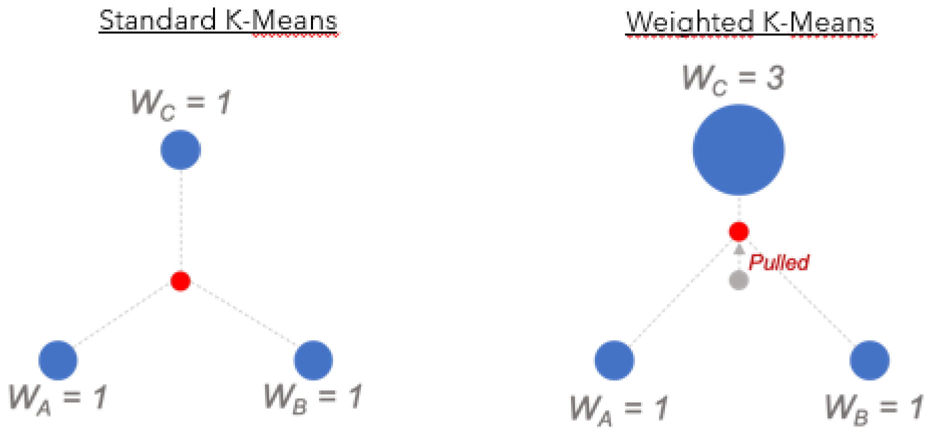


Figure 12: Standard k-means (left) vs weighted k-means (right). Source: https://www.towardsdatascience.com.

To use the weighted k-means algorithm, the following weights are defined:

Venue Category	Weight (W_i)
Education, Health and transportation	10
Culture & nature	2.5
Hotel & Restaurant, nightlife, shopping, sport, others	1

Table 1: Weights for the weighted k-means algorithm.

The answer to the second question is to use the Elbow method, which consists in running the algorithm on the dataset for different values of k, and computing the distortion for each k. Then, a plot of the distortion as a function of k is done, and the optimal k is the one where the elbow of the plot can be seen. Plots of the Elbow method for both standard K-means and weighted K-means can be seen in figures 13 and 14, respectively, which show that the optimal clusters are 4 and 5.

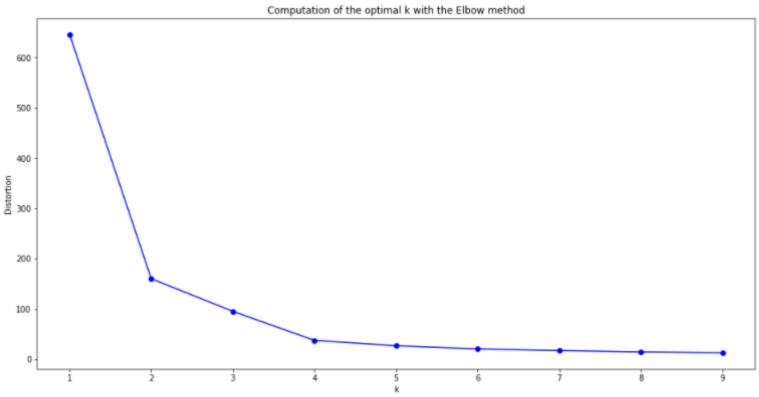


Figure 13: Elbow method for standard K-means.

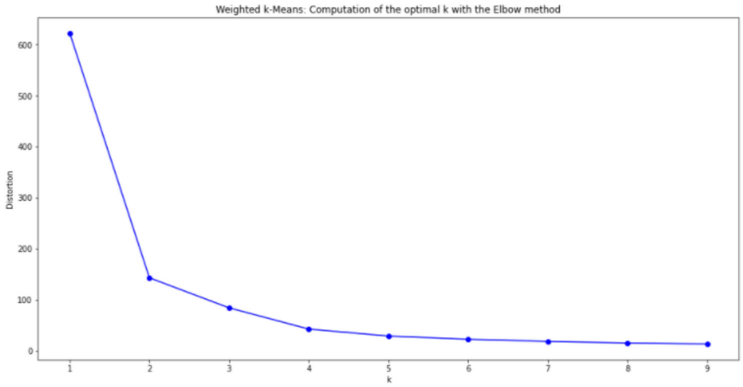


Figure 14: Elbow method for weighted K-means.

4. Results

In the case of standard k-means we get the following information:

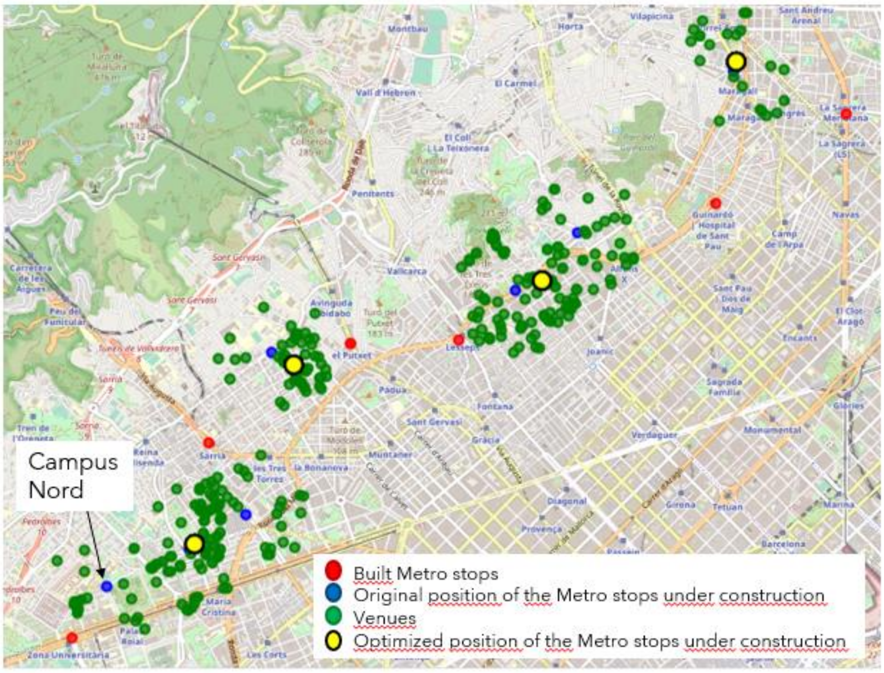

Campus Nord
● Built Metro stops
● Original position of the Metro stops under construction
● Venues
● Optimized position of the Metro stops under construction

Figure 15: Optimized Metro stops after running standard K-means.

And we can compute the distances between the original coordinates of each stop (from figure 4) and the optimized ones:

Metro stop under construction	Distance to the optimized stop (m)
Campus Nord	789.53
Manuel Girona	64.72
Prat de la Riba	475.16
Mandri	204.85
Muntanya	231.64
Sanllehy	475.69
Maragall	77.26

Table 2: Distance between original stop and optimized stop after running standard K-means.

Now let us run weighted K-means:

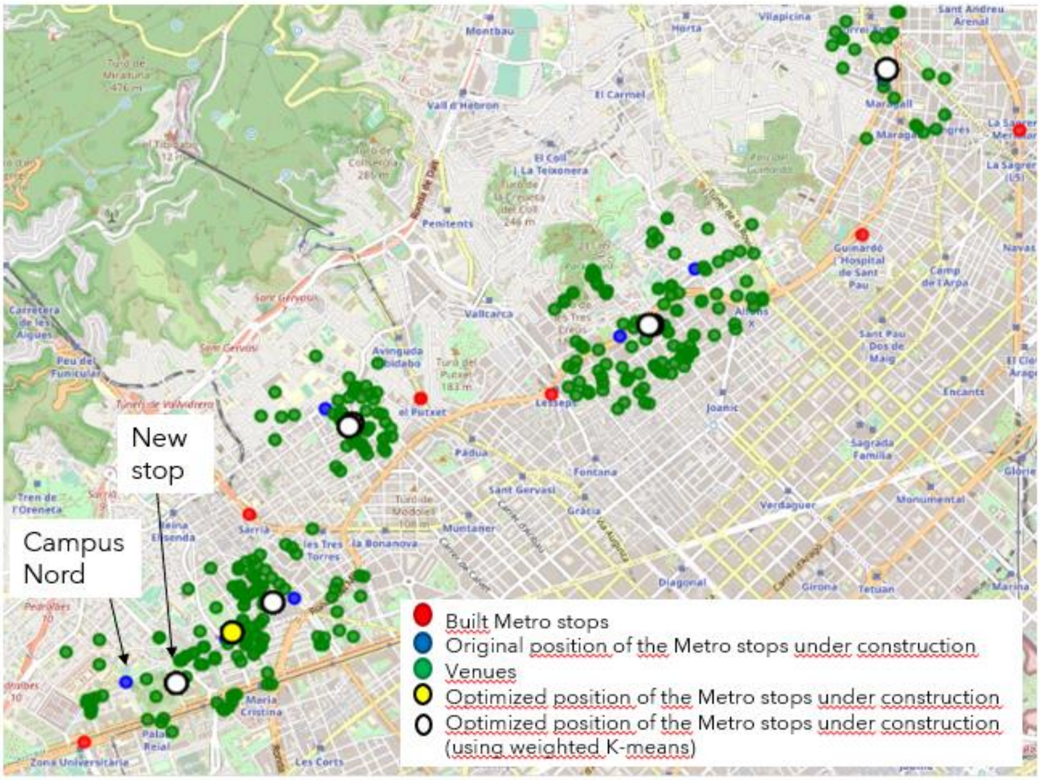

New stop
Campus Nord
Built Metro stops
Original position of the Metro stops under construction
Venues
Optimized position of the Metro stops under construction
Optimized position of the Metro stops under construction (using weighted K-means)

Figure 16: Optimized Metro stops after running weighted K-means.

Metro stop under construction	Distance to the optimized stop (m)
Campus Nord	338.87
Manuel Girona	408. 46
Prat de la Riba	143.58
Mandri	200.08
Muntanya	211.96
Sanllehy	489.84
Maragall	77.26

Table 2: Distance between original stop and optimized stop after running weighted K-means.

5. Discussion

It has been seen that all Metro stops have been modified due to the algorithm (i.e., there is no case where the distance between the optimized stop and the original one equals 0).

For the Maragall stop, both algorithms suggest a new stop which is less than 80 m away from the original one, so in this case, the original coordinates of the stop can be kept. This is something good because the coordinates of the stop that are found in Wikipedia are for sure computed according to things that are not considered in this study, such as the geological properties of that area of Barcelona, etc. Hence, we are now sure that this stop is optimal not only from the “construction” point of view, but also from the “venues” point of view.

In the case of the Campus Nord stop, the weighted K-means provides a solution that is much better than the standard K-means, because it takes into account the variety of its venues and decides to place a stop that the standard k-means algorithm does not.

To sum up, it can be said that the results provided by the weighted K-means are more realistic than the ones provided by standard K-means, so the L9 Metro line of Barcelona should look like:

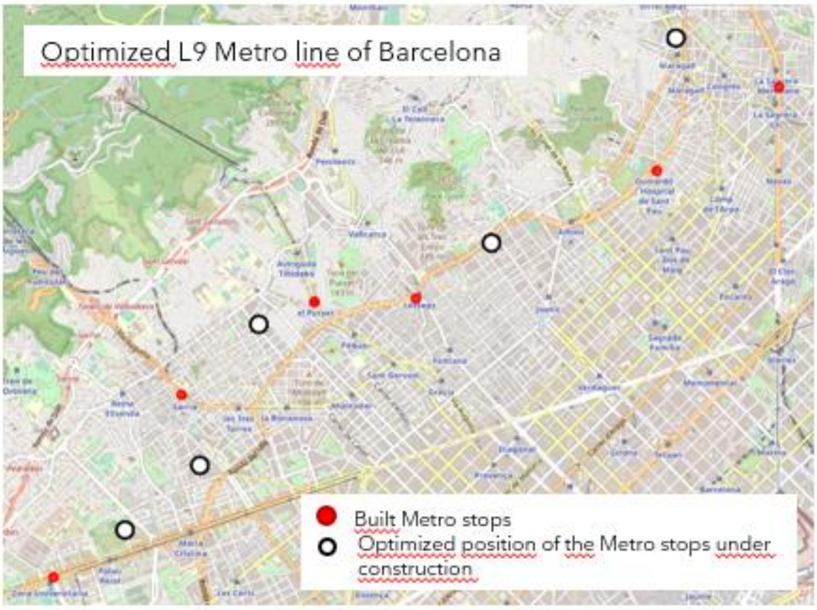

Optimized L9 Metro line of Barcelona
● Built Metro stops
○ Optimized position of the Metro stops under construction

Figure 17: Optimized L9 Metro line of Barcelona.

6. Conclusions

For the L9 of Barcelona's Metro, a study to find the optimal position of the stops under construction based on the venues around them has been done.

The algorithm chosen to optimize the Metro stops is K-means, with the aim of setting the centroids of the clusters as optimized Metro stops.

- As a first option, the standard K-means algorithm has been used, however this algorithm only focuses on the number of venues around a metro stop and does not consider its variety, which is also of interest (e.g., a University or a hospital are more important than 20 restaurants).
- As a second option, the weighted K-means algorithm has been chosen. This algorithm takes the variety of the venues into account and suggests that the 7 original Metro stops could become only 5. It can also be noticed that the optimal position of the Maragall stop is very closed to the original one (less than 80 m distance), therefore it

is a better idea to keep the original coordinates when building the station.

In the future, the optimization study presented in this report could be improved by taking into account the popultion density and also the proximity of bus stops, other Metro lines, train stops, etc.