Clustering stations in the Italian region of Lazio

CapStone Project of IBM Data Science Certification

165 Stations

- The project is focused on the railway stations located in the Italian region of Lazio, a territory populated by 5.9 million people, visited by more than 12 million tourists each year and served by 165 train stations (41 of them inside the city of Rome)

Platinum, Gold, Silver, Bronze

- RFI train stations underwent a significant transformation, from travel hubs to meeting centers and places of aggregation, going beyond their original role of serving travelers by offering diverse services to travelers and non-travelers as well
- The service offering of a station depends on its class, that goes from “bronze” to “silver”, “gold” and “platinum”

Area analysis

- Goal of this study is to use open data about the stations to categorize them depending on a series of factors, such as the number of check-ins and reviews, and characteristics of the nearby area such as traffic and population, and the number of significant venues in a radius of 1.5 km from the station

Supporting decisions

- This categorization will be useful to compare the current classification made by RFI with the one deriving from the above parameters, to understand if some station should deserve a different status, such as a promotion or a demotion
- The main target audience will be, therefore, RFI management

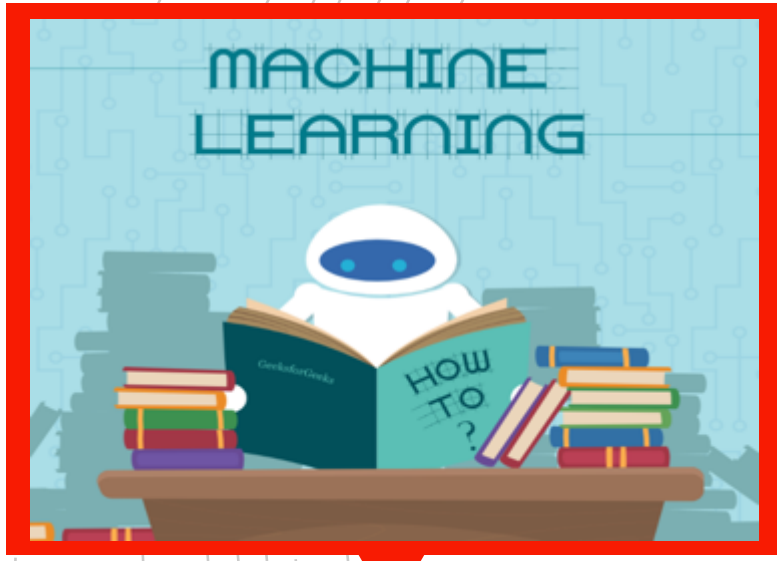
Data Sources

Data	Source	Last updated
List of RFI train stations in the Lazio region, with location and classification	http://www.rfi.it/rfi/LINEE-STAZIONI-TERRITORIO/Nelle-regioni/Lazio	2019
Number of check-ins and reviews of each station	FourSquare API	Daily updates
Top 100 venues in a 1000 meters range of each station, categorized by high-level groups	FourSquare API	Daily updates
Population and density of the neighborhoods hosting the stations	https://it.wikipedia.org/wiki/Municipi_di_Roma https://www.comune.roma.it/web-resources/cms/documents/Territorio2017DEF.pdf https://www.tuttitalia.it/lazio/27-comuni/popolazione/	2017 2017 2019
Pollution levels (as a rough indicator of traffic)	http://www.arpalazio.net/main/aria/sci/qa/misure/PM10.php	2019

FourSquare

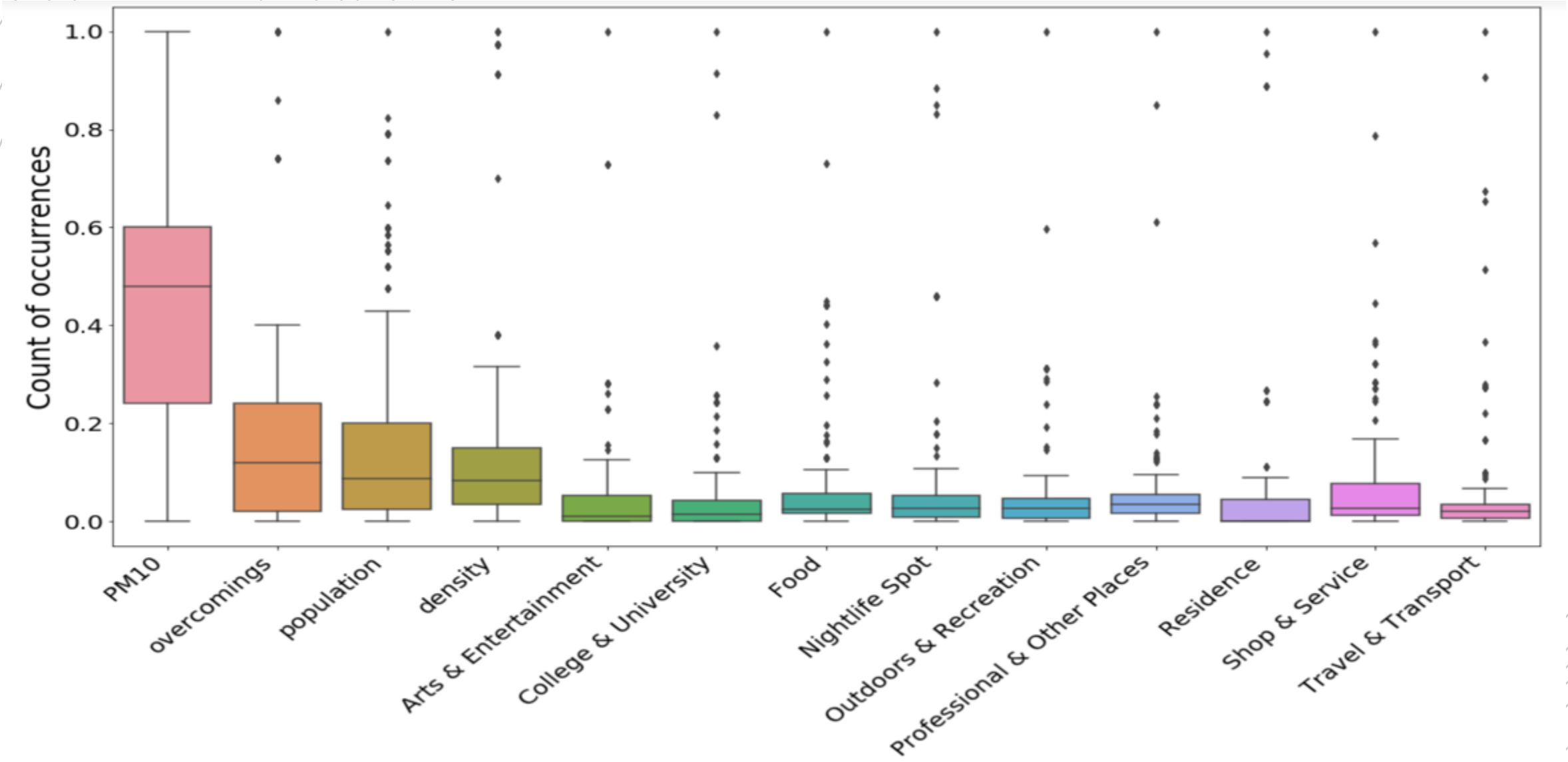
- For each station, from the address the latitude and longitude will be derived, and those in turn will be used to retrieve the FourSquare data and venues
- The match with population, density and pollution data will be done on the basis of the neighborhood hosting the station

Machine Learning

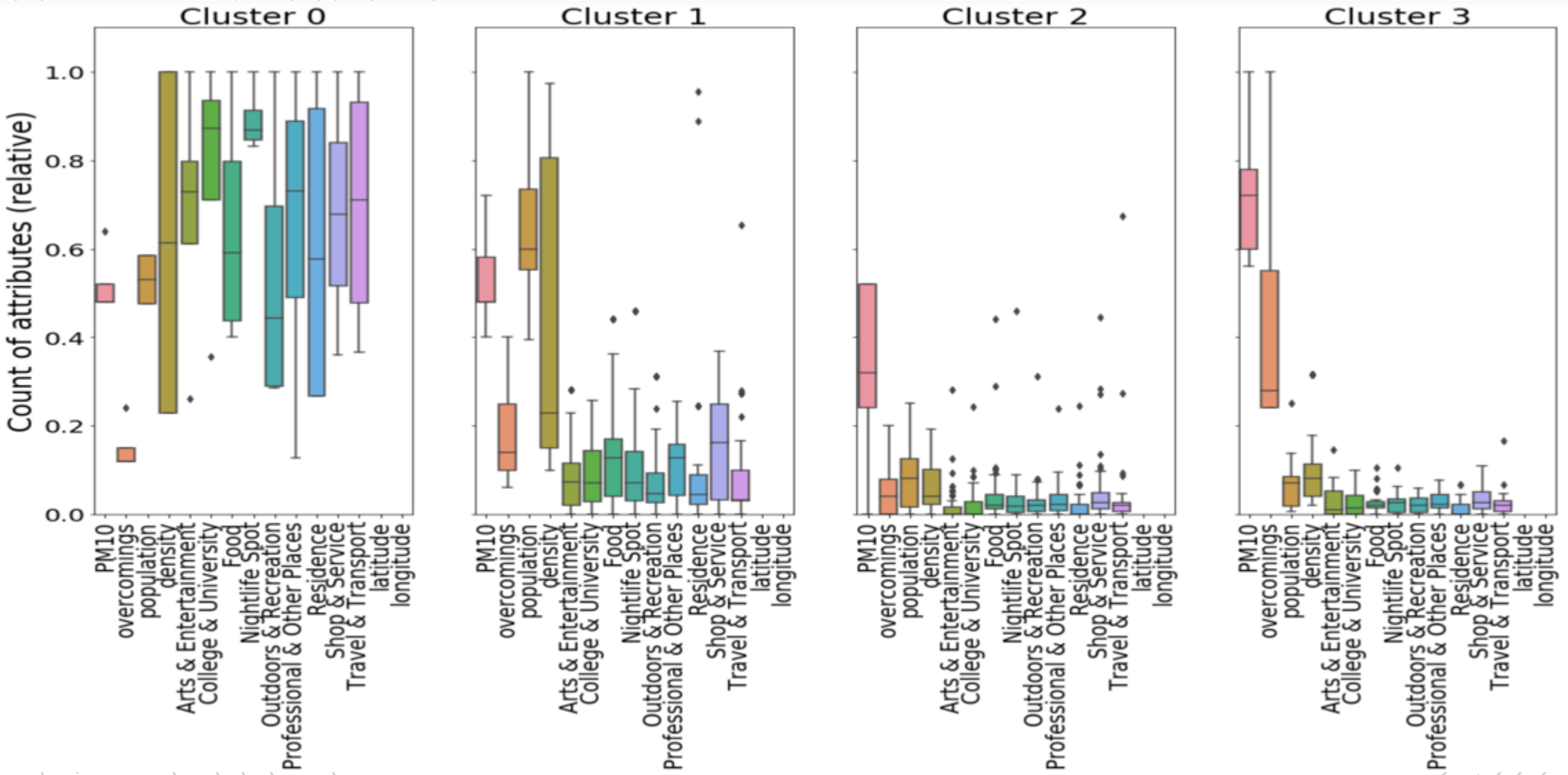


- The features will be standardized for better manipulation of the clustering algorithm, and then a clustering will be performed with 4 clusters, the same number of the classes used by RFI
- The original classes and the new clusters will then be compared, using both tables and maps, to check for correlations between the two

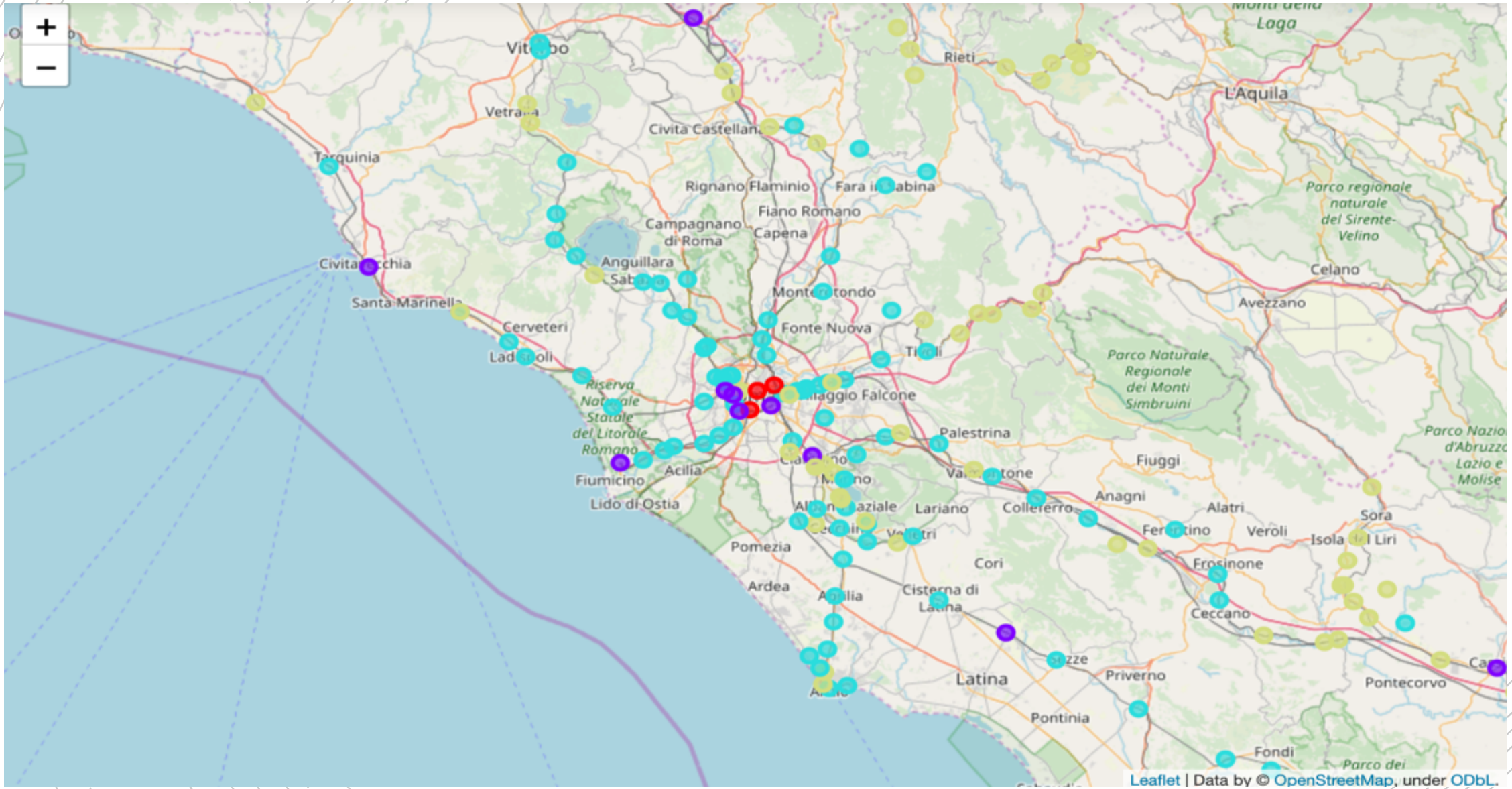
Results – Categories distribution among stations



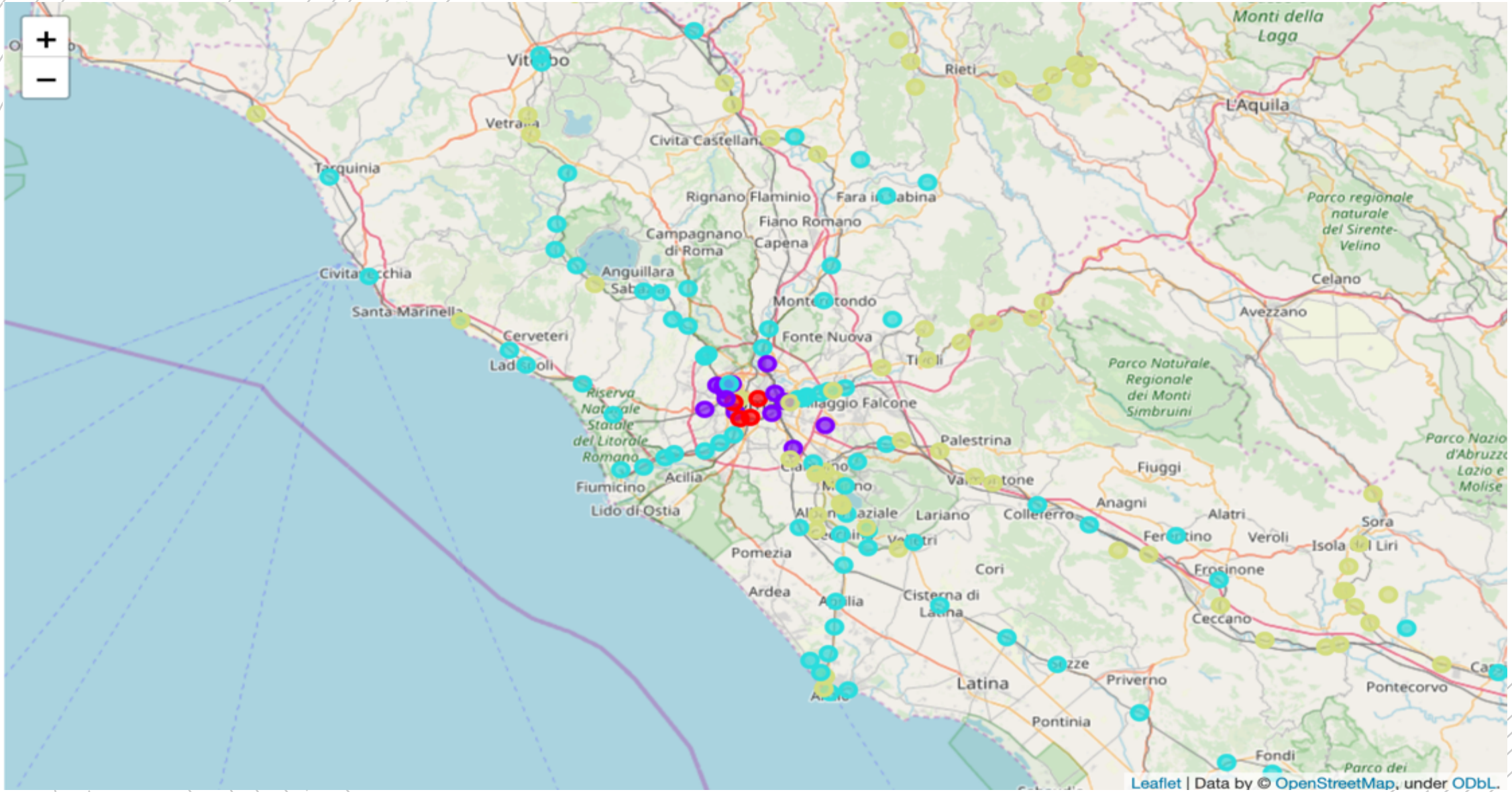
Results – Characteristics of the four clusters



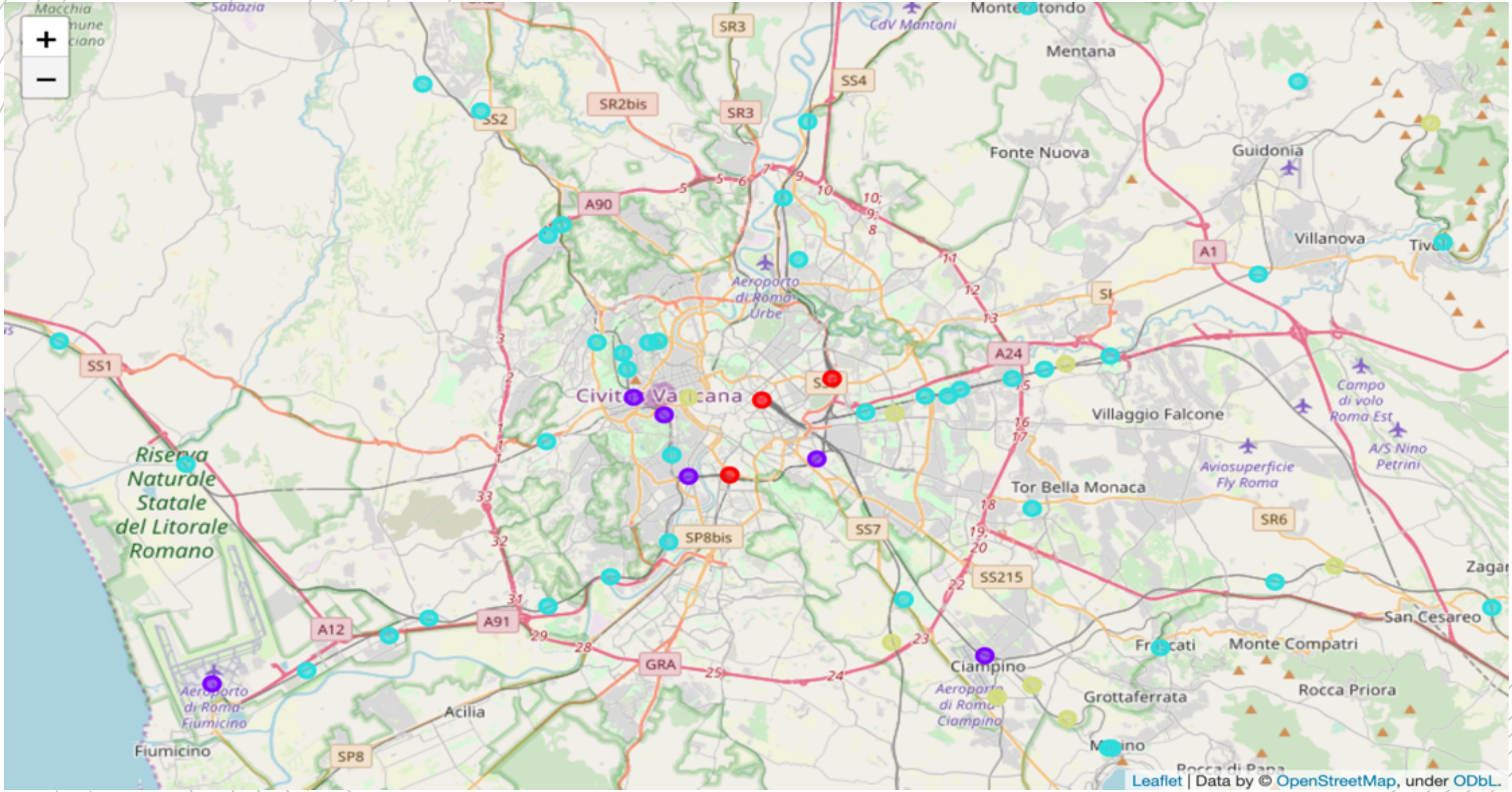
Results – Map of the RFI categories



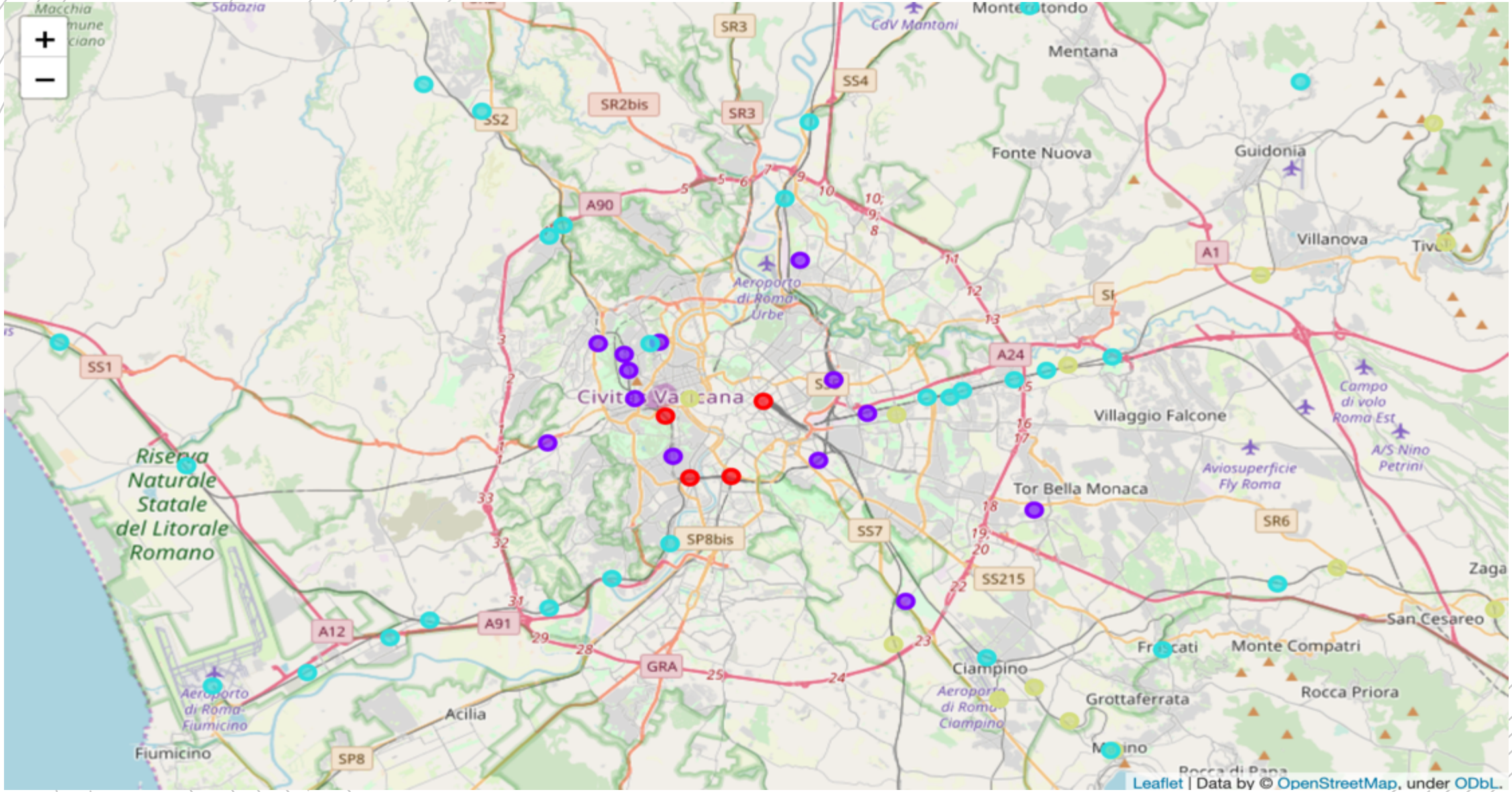
Results – Map of the computed categories



Results – Map of the RFI categories, zoomed



Results – Map of the computed categories, zoomed



Results

- The stations that received a different classification (15%) fell mostly into 2 categories:

- stations in the centre of Rome, with a venue-rich neighbourhood, promoted to the upper category (such as Trastevere, S. Pietro, Aurelia, Quattro Venti, Balduina, Monte Mario)

- stations playing a crucial role as a node in the transportation network, but with poorer neighbourhood, that were demoted to the category immediately below (such as Tiburtina, Fiumicino, Ciampino, Civitavecchia)

Results

- Therefore, notwithstanding the fact that FourSquare Data is un-balanced, with some categories, such as food, over-represented, the clustering made quite sense, with the stations distribution matching broadly the investments made by the network manager on each station, as reflected by the category level

Results

- The non-matching labels made sense as well, showing that some stations correctly offer higher level of service than the one suggested by the nearby area, given their strategic position (a close-by airport or harbor), while others should deserve a promotion considering the amount of venues and activities in the nearby area

Conclusions

- The project made use of websourced data and a non-commercial FourSquare account
- Notwithstanding these limited resources, it proved strong insights into the areas surrounding each station, supporting the choices made by the rail network manager and giving some valuable suggestion
- Using a commercial FourSquare account could give a more granular description of venues and increase the understanding of stations' neighbourhoods

Further Developments

- Exploring the predicted impact of a station promotion/demotion on its reviews
- Exploring the correlation of check-ins and reviews to other parameters, like availability of certain venues, such as parking or bus stops or restaurants, or the lack of others, such as coworking places or professional structures

Further Developments

- Creation of a recommendation system able to advice citizens and tourists about the nearby venues of each station, depending on their needs and past choices of similar users
- Forecasting the profitability of parking places offered by the station, analyzing the demand and availability in the nearby area
- Extension to the whole Italian network
- Comparing foreign stations hosted in similar neighbourhoods (Paris, London, Berlin, Amsterdam)

THANKS!