Airbnb Analysis in Milan

Geospatial Analysis and Representation for Data Science - University of Trento

Introduction

The aim of this report is to describe the different steps of the project developed for the course *Geospatial Analysis and Representation for Data Science*, which focuses on the analysis of Airbnb data in the city of Milan. The tasks performed are:

- choose one of the cities available in InsideAirbnb
- retrieve data on neighborhoods (city data portal)
- view statistical information on neighborhoods (city data portal)
- identify which are the neighborhoods with the highest prices in AirBnB
- identify which are the districts with the greatest number of tourist activities (city data portal and/or openstreetmap)
- find the location of 3 AirBnB hosts closest to one of the museums (on walking distance) (city data portal and/or openstreetmap)
- of the three hosts, identify which one has the greatest number of services (supermarkets, pharmacies, restaurants) in an area of 300m (city data portal and/or openstreetmap)
- analyze and test spatial autocorrelation of price
- represent these analyses on maps (web and not)

For each of these tasks, the report describes the idea followed to implement it.

Data Sources

- *listings.csv*: it comes from *Inside Airbnb* and it contains data about each Airbnb in the city (name, host, location…)
- *neighbourhoods.geojson*: it contains the name of the neighbourhoods and the geometry
- *datiquartierimilano.csv*: it comes from the city data portal and it contains information about the neighbourhoods (number of inhabitants, areas…). It can be found here: https://dati.comune.milano.it/dataset/ds205-sociale-caratteristiche-demografiche-territoriali-quartiere/resource/084457a7-ec4b-4a6b-b463-d8ab53c64fbb
- *milano.gpkg*: it is created in the Python notebook, merging information about the average price of the Airbnb in the neighbourhoods to the file *datiquartierimilano.csv*
- *bbox_Milano.osm.pbf*: file created with the website export.hotosm.org to retrieve data from OpenStreetMap with a bounding box

How to read the files

All the tasks are implemented in the *Python notebook*, except the one related to the analysis and test of spatial autocorrelation of prices, which is implemented in R, in the *R markdown* file or in the *.html* file, which increases readability.

Data Pre-Processing

The first step was to prepare data for the analysis. This includes to parse data in order to merge dataframe, to check the correctness of the Coordinate Reference System (CRS) and to create GeoPandas dataframes.

Analysis

View statistical information on neighborhoods (city data portal)

The idea in this step is to show the distribution of people in the territory of the city of Milan with maps. Thanks to the information available in the file *datiquartierimilano.csv*, it was possible to retrieve the *number of inhabitants* and to compute the *population density* for each neighbourhood. After that, it was possible to plot them in choropleth maps in two ways: with a web map, thanks to the package *folium*, and with a static map. The neighbourhoods with more inhabitants are *Buenos Aires - Venezia* and *Loreto*, while the neighbourhoods with the highest population density are *Selinunte*, *Viale Monza* e *Loreto*. Of course, it would be possible to plot in the same way many other information about neighbourhoods, such as the number of schools, the number of tram/metro stations, and compare which neighbourhoods are more connected to the others.

Identify which are the neighborhoods with the highest prices in Airbnb

To perform this task, it was necessary to retrieve data from the file *listings.csv*, to group data by neighbourhood and to compute the mean, finding the average price of the Airbnb for each neighbourhood. The neighbourhoods with the highest prices in Airbnb are *Pagano*, *Magenta - San Vittore* e *Brera*.

Identify which are the districts with the greatest number of tourist activities

To perform this task, it was necessary to retrieve data from OpenStreetMap. After having created a bounding box, containing the shape of Milan, data were downloaded from the website *export.hotosm.org*, in the file *bbox_Milano.osm.pbf*. This permits to have all the data in OpenStreetMap and to select only tourist activities. Then, performing a spatial join between the tourist activities and the neighbourhood dataframes, it was possible to identify which neighbourhoods have the greatest number of tourist activities, which are *Duomo* and *Brera*.

Find the location of 3 AirBnB hosts closest to a museum (on walking distance)

For this task, I have selected the museum *Museo Del Design 1880-1980*. Performing a street network analysis, it was necessary to find the travel time from the Airbnbs to the museum. Since the Airbnbs were more than 18'000 and the computation was slow, I have filtered only Airbnbs in the same neighbourhood of the museum, which is *Navigli*. Then, computing the nearest nodes of the museum and the Airbnbs, it was possible to find the shortest paths and the travel time. The Airbnbs which are closest to *Museo Del Design 1880-1980* are *Siddharta house area Navigli Tortona 3 bedrooms* (Vincenza), *Unique Experience in the heart of Milano* (Giovanni) and *Navigli Nightlife* (Chiara).

Of the three hosts, identify which one has the greatest number of services (supermarkets, pharmacies, restaurants) in an area of 300m

For this task, since the number of services in an area of 300m were not enough, I have selected services in an area of 1km. Using the same file to retrieve data from OpenStreetMap, which was *bbox_Milano.osm.pbf*, I have selected the services related to supermarkets, pharmacies, restaurants, bars, cafes, bakeries and malls. From this analysis, the host with the highest number of services in an area of 1 km is *Chiara*, with Navigli Nightlife, with 9 services.

Analyze and test spatial autocorrelation of price (in R)

To test spatial autocorrelation of prices I have created five different definitions of neighbours between spatial units: with a k-nearest neighbour approach with k equal to 1, with a contiguity-based approach with a Queen criterion and with three critical cut-off neighbourhood, which are 1.8km, 2.5km and 3.2km. For each definition, I have created the spatial weights matrix to perform the Moran's I test. All the tests performed with these definitions seems to confirm the evidence of spatial autocorrelation of prices, with Moran's I index from 0.33 to 0.41 and very low p-values, rejecting the null hypothesis of no autocorrelation. Then, to double-check this evidence and to see which neighbourhoods contribute more to the autocorrelation, I have plotted the Moran Scatterplot for the spatial weights matrices with knn = 1, with the contiguity-based approach and the cut-off of 2.5km. All the three scatterplots show a positive autocorrelation. Moreover, the neighbourhoods that seem to contribute the most are 9, 40, 48, which are *Brera*, *Magenta - San Vittore* and

Pagano, those with the highest prices. To test statistical significance for the local entities, I have computed the local Moran’s I and plotted the p-values in maps. As seen in the final plots of the p-values, neighbourhoods in the city center contribute more and create a cluster of local spatial autocorrelation.