Geospatial analysis and representation for data science - University of Trento

Riccardo Zorzoni (mat. 220500) - riccardo.zorzoni@studenti.unitn.it

Introduction

The aim of this work is to analyze and represent geospatial data. I decided to use a Python notebook in *Google Colab* (it takes 16 minutes using the *Run all* command in the *Runtime menu*), which is provided with this report. The notebook is full of documentation and comment and I tried to make it self explanatory. All the data that I used for the analyses are available in my Github repository, at this link: https://github.com/Richip9114/GeospExam

Requirements

It is possible to find a *requirements.txt* file inside the repository. This file contain all the version of the libraries that I used. The version of Python is *3.6.9*, the last available on *Google Colab*.

Libraries

Initially, I needed to install some package through pip, the Python package manager. Useful package for the analysis are:

- *contextily*: to retrieve tile maps, it can add tile as basemap to *matplotlib* figure;
- *seabort*: to plot and visualize the distribution of the points;
- *sklearn*: to use *DBSCAN*, a density based cluster algorithm;
- *geopandas*: to manage geospatial dataframes;
- *geopy*: to use the *geocode* function to retrieve points;
- *osmnx*: to retrieve graph network;
- *rpy2*: to use R in a Python notebook.

Data collection and data preparation

AirBnB data (License: CC0 1.0)

(original data from inside AirBnB - http://insideairbnb.com/get-the-data.html).
Firstly, I loaded the data of the AirBnB in Bologna using Pandas. Then, I decided to filter the data and use them only the rooms with at least 90 days in a year, because otherwise the analysis could be with outliers. For example, some rooms/houses available only on New Year's Eve. After that, I created the geodaframe using *Geopandas*.
Furthermore, I plotted all the rooms over a map of Bologna using *Contextily*, thanks to the *OpenStreetMap* provider. This plot allows to check that all the points are in Bologna and there are not mistake in the dataset. I also plotted the distribution of the points using *Seaborn*, and it is possible to see that the distribution is concentric around the central point. So, I decided to go deeper in this result plotting two other map, one using *hexbin* and one using *kdeplot*, and I was not surprised to notice that the previous results was confirmed. Finally, I decided to investigate the distribution using *DBSCAN* algorithm, with 1% of minimum points and 140 as epsilon, and this shows that there are clusters, but all of them are in the center. Increasing the eps by 10, all the cluster collapse in 2 instead of 6.

Neighbourhoods data (License: CC0 1.0)

(original data from inside AirBnB - http://insideairbnb.com/get-the-data.html).
I loaded the data of neighbourhoods from Github and then I converted them into *Geopandas* dataframe. Then, I plotted the area of the neighbourhoods over a map using *contextily*, to show the location of each neighbourhood.

Data analysis

View statistical information on neighbourhoos

Initially, I collected the data about the area of the neighbourhoods from http://dati.comune.bologna.it/node/1183 (License: CC BY 3.0 IT). Then, I inspected all the neighbourhoods' data one by one. In the code it is possible to choose the number of the neighbourhood and look at the distribution of the rooms. Then, I joined the data using spatial join, and I counted how many rooms there are in each neighbourhood. I compared this result with the area, and it is not surprisingly to see that they are uncorrelated.

Identify which are the neighbourhoods with the highest prices in AirBnB

I used the previously created dataframe and I grouped by the neighbourhoods name, calculating the mean of the price. The result shows that Santo Stefano and Porto - Saragozza have the highest prices, and this is discussed later when I analyzed the tourist activities. However, the min average price is 77.75 Euros for night in Naville and the max is 95.22 in Santo Stefano.

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

Initially, I created the convex hull containing all the neighbourhoods, then I created the *geojson* to use in *HOTOSM* (https://export.hotosm.org/en/v3/exports/a6c29254-cbac-4367-abea-4a00ae0a72c3) to retrieve the *pbf* file with all the information from *OpenStreetMap*. Secondly, I created a custom filter to collect data about tourism attraction in Bologna and, after that, I plotted the results. It is interesting to see that the average prices of the neighbourhoods is correlated to the number of tourism's point.

Find the location of 3 AirBnB hosts closest to one of the city's museums

First of all, I retrieved the graph of walking street of Bologna using *osmnx* (*OpenStreetMap* and *networkx*) library and the location of the city's museum (Palazzo Pepoli Campogrande). Then, using the functions presented during the lectures, I calculated the routes from museum and all the rooms, calculating also the travel time and path length. Finally, I extracted the three closest rooms to the museum.

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

I used *osmnx* to retrieve data of supermarkets, pharmacies and restaurants in Bologna and then I investigated how many of them are in a area of 300 meters using *within* function and *buffer*. All the three rooms have the same number of supermarkets, pharmacies and restaurant, due to the fact that they are near to each other. So, I took a random host, far from the others, and I got different result: this host is not in the city-center, so it has a lower number of services.

Analyze and test spatial autocorrelation of price

This part was done in R, using *rpy2*, a Python library that allows to use R inside a Python notebook. I decided to do that to have all the code in the same place. In order to use this library, it is necessary to run this command `%load_ext rpy2.ipython` and then all the cell must start with `%%R`. Initially, I created a geodataframe with the *multipolygon* of the neighbourhoods, the average price, the tourism's point, the number of AirBnB rooms and the number of services (. Then I created the files for the R library (*spdep* and *rgdal*), thanks to the function `gpd.to_file('name_file.shp')`. Firstly, I tried the *global moran test*, but I did not get any good results (p-value is a huge number, so the results are not statistically significant). Secondly, I tried to detect spatial autocorrelation in the residuals of a linear regression model, where I called this function: `lm(formula = avg_price ~ tourism_po + nr_rooms + services, data = bologna)`. Then, I noticed that the p-value of the last model is low in the Moran's I test on residuals. This result is not useful because the first model is not statistically significant. Finally, I did the *local* analysis of autocorrelation, but the results of the `localmoran` test are not good, due to the p-value.

Represent these analyses on maps (web)

In this last part, I represented some results using *folium* and leaflet web map. Firstly, I plotted the map of Bologna with one or more layers. Secondly, I plotted a map with marker of the centroid of the neighbourhoods, and then a map with the three closest hosts and the museum. Finally, I plotted a map with a marker cluster to see all the services in Bologna.

Raster data

I asked to the geoportal of Bologna (https://geoportale.regione.emilia-romagna.it/ - License: CC BY 3.0 IT) the *tiff* data of Bologna and then I tried to plot them using *rasterio*.

Conclusions

I think that all the tools presented during the course are very powerful and easy to manage thanks to python libraries. In some line of code is possible to analyze spatial data coming from different sources. I started my analysis on Venice, but there was a problem with the *pbf* file coming from *HOTOSM*. I tried to get the *pbf* from *BBBike*, but also here the file is compromised. So, I decided to change city and I notice that it is very simple, because all the data from AirBnB have the same shape and same for most of the sources. My initial thought has been confirmed, the neighbourhoods with the highest price are the same of those with more tourism's points and services.