Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Projects

Geospatial Analytics

2023/2024

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

General remarks

- The project should be delivered 10 days before the date of the “appello” chosen by the student.

- The project should be delivered as a .zip folder through **this form**. For python code, both .py (scripts) and .ipynb (notebooks) files are allowed.
    - for notebooks, remember to clear all outputs before uploading it on the form (this will save memory space in the zip folder).

- Unless it is strictly needed, do not upload big datasets but provide the link where to download them or, even better, make the code download the dataset directly from a public URL. You can eventually upload small files (e.g., small json, geojson, or shapefile files).
    - For experiments on SUMO, send also the road networks, the traffic demands, and the configuration files.

- The code in the main notebook should run correctly without any modification from our side.

- Please comment adequately on your notebook using markdown. All your steps and decisions should be adequately described in the notebook.

The projects will be evaluated based on their correctness, level of detail and depth, the elegance of coding, and creativity.

During the exam, the student is expected to go through the notebook

In case of questions about the projects, please write an email addressed to:

- luca.pappalardo@isti.cnr.it
- mirco.nanni@isti.cnr.it
- giuliano.cornacchia@phd.unipi.it
- giovanni.mauro@phd.unipi.it

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 1: On the error between GPS traces and mobile phone records

In this project, the aim is to delve into the distinctions between GPS traces and mobile phone records (MPR). To achieve this, select a minimum of two publicly available GPS trajectory datasets and a public dataset containing the positions of mobile phone towers to allow the alignment of GPS points with their corresponding phone towers.

To emulate the characteristics of a Call Detail Record (CDR) dataset, develop a method to sparsify each GPS trajectory leveraging the typical distribution of inter-call times. Following this, implement trajectory similarity metrics to quantitatively assess the similarity between the original, dense GPS trajectory and the sparsified counterpart.

Make the most appropriate visualizations to show the difference between the two types of trajectories in all the datasets.

Make a thorough analysis of how the results change based on the dataset and the parameters of your sparsification method. Discuss in detail the top-k and bottom-k users (choose a reasonable k) based on the similarity between the original GPS trace and the sparsified one.

Conduct an in-depth examination of fundamental mobility metrics (e.g., the radius of gyration, entropy and more), highlighting their alterations between the original GPS trace and the sparsified counterpart.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 2: Nocturnal vs Day mobility patterns

Do nocturnal mobility patterns differ from day mobility patterns? Select at least two public trajectory datasets and investigate mobility patterns during night and day.

Define reasonable times to delimit night and day (night-day intervals), taking into account the peculiarities of the datasets and the associated geographic areas.

Find adequate visualizations to show evident differences between day and night mobility. Then, quantify and characterize how the two mobilities differ in terms of individual mobility patterns, collective mobility patterns, and other aspects that you consider interesting.

Vary the night-day intervals and assess how your results change. Moreover, can you think of a method to infer night-day intervals directly from the data?

Develop a coherent discussion about what you find. Try to answer fundamental questions, such as: Are nocturnal trips more predictable than day ones? Are they typically longer? Are the patterns of day and night consistent across the selected datasets? Are the night-day intervals inferred from the data the same for all selected datasets? And other questions that you believe may be interesting.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 3: Data-driven estimation of urban gentrification

Gentrification is defined as "*The process by which a place, especially part of a city, changes from being a poor area to a richer one, where people from a higher social class live*" (Cambridge Dictionary).

Yelp is a platform for sharing, reading, and collecting reviews about activities like restaurants, pubs, hotels, etc. They provide a dataset here.

OverpassTurbo provides an API for retrieving historical data from openstreetmap, so as to have insights on the evolution in time of the POIs.

Select 2 or more cities for which data quality is particularly good, both for the Yelp data and for the Overpass.

Divide the city in zones according to a tessellation you consider meaningful.
Find a reasonable way to estimate the date of creation/foundation of a facility/POI.

Develop an analysis to study the evolution of the Time Series of the number of facilities, their type, and the number/kind of POIs per zone. Discuss any trend, pattern, seasonality, or breakout date the analysis reveals for each dataset.

Develop a coherent discussion about what you find. Answer to questions like: Are there some neighborhoods with abnormal, consistent, values? Are there neighborhoods with an abnormal increment of Yelp activities and POIs in a time window (i.e. do you observe some sort of synchronization)? Would you define such neighborhoods gentrified? Are there neighborhoods in which, in a time window, only one of the two time-series show an abnormal increase? How would you define them? And other questions you believe may be interesting.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 4: Football matches as mobility networks

During a football match, players move on the field to attack and defend. This generates a series of movements that can be analyzed to understand the players behavior.

The student should use the Wyscout open dataset, describing the "events" in all matches of seven competitions (e.g., passes, shots, tackles etc.), to analyze pass chains and the mobility of football players. A player's movement is defined by consecutive events made by that player in the match.

Investigate the distances traveled by players during their matches and their distributions. Discuss about the similarity of these distributions with those about mobility trajectories seen during the course.

Relate the pass chains made by teams with the probability of making a shot, a goal, and to win a match. Are long chains more likely to lead to a shot/goal? Are short pass chains more successful?

Quantify the predictability of pass chains based on some division of the football field (tessellation). To what extent can we predict the next tile (field zone) where the ball will be? Use a next-location predictor to quantify the accuracy to predict the next zone the ball will be.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 5: Road Deviation Patterns

It is well known that actual mobility of drivers often does not follow shortest/fastest paths, yet it is not clear what are the reasons. This project aims to dig deeper in the phenomenon.

The basic approach required involves to test two possible (non-exclusive) hypotheses: (1) there are some roads that are almost systematically avoided by real vehicles; (2) there are special events (road works, accidents, exceptional traffic conditions) that make the road unappealing for a limited time – either in a specific day, or regularly every day (e.g. systematic traffic jams). The student is welcome to integrate them with additional ideas to explore.

First, select some public GPS datasets describing the movements of vehicles in a city. For each trajectory, map-match it with the road network to get the trajectory's corresponding route. Compare this route with the fastest/shortest route on the road network.

Develop an analysis of how the road segments differ between the real route and the fastest/shortest route. For example, identify the road segments that more frequently differ between the real and the fastest/shortest route, showing them on a map properly. Are there extreme peaks, namely hugely avoided roads?

Repeat the analysis with specific time slots. Are there road segments that are systematically avoided in the morning, but not in the evening?

Develop a coherent discussion about what you find.

Consiglio Nazionale
delle Ricerche

UNIVERSITÀ DI PISA

Project 6: Road Network Resilience

When a road is closed or becomes extremely slow, its impact on the overall mobility of the city depends on various factors, including its centrality and the existence of alternative routes. The objective of this project is to identify the road edges that are more critical and can thus create resilience issues to the road network.

Select some cities and download the corresponding road networks. Develop a methodology to remove some road edges from the road network and look at how the travel time of the fastest routes changes based on a selection of origin-destination pairs. You are free to decide how many and what road edges to remove and how to select the set of trips (starting and ending road edges of routes).

Find a way to remove a set of road edges from the road network such that: 1) the total length of these road edges is around x kilometers; 2) the typical travel time of fastest routes does not increase more than y%. Vary x and y and discuss the results obtained.

Avoid trivial solutions, e.g., removing road edges that are never visited based on the fastest paths resulting from the selected trips.

Develop a coherent discussion about what you find.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 7: Urban Segregation with Centralized Intelligence

The idea is to simulate a Schelling-like scenario in which an algorithm acts as a Recommender System (RS) with different policies. When an agent is unhappy and wants to relocate (or even without considering the unhappiness and just simulating a probability of willingness of changing house) it will take into account *k* new places suggested by the RS.
As suggested by [1] the RS will implement several policies, i.e.. it will suggests users places with e.g.

- A similar neighborhood composition
- A short distance/High relevance/ Combination of the two (see [2])
- A rich/poor avg wealth (means that you have to assign a wealth to agents)
- A slightly improving place in terms of wealth
- …

Given the increasingly crucial role that housing recommenders like Idealista[4] are starting to play, simulating such a RS is important, so as to have an insight of the impact of the policies of the algorithms used by these platforms.

The simulation will be conducted using MESA[5] Python Library.
Is the use of a RS beneficial? Does RS usage affect segregation time and levels? How, w.r.t to the classical one or the other versions? Can we propose metrics for quantifying the "trustability" of such a RS? What is the impact of the different policies?

[1] Moro, Esteban. "The minority game: an introductory guide." *arXiv preprint cond-mat/0402651* (2004).
[2] Gambetta, Daniele et al.. "Mobility constraints in segregation models". Sci Rep 13, 12087 (2023). https://doi.org/10.1038/s41598-023-38519-6
[3] https://www.idealista.com/
[5] https://mesa.readthedocs.io/en/stable/

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 8: A Routing-based Urban segregation measure

Measuring segregation is a challenging task. Many measures have been proposed in the last years, such as the Freeman Segregation Index, the dissimilarity index and others. Recently, an important work[1] proposed to quantify segregation in a spatial network as the average number of steps that a random walker starting in a node *i* needs to encounter a fraction *c* (a parameter) of all the classes of nodes. In this work, the idea is to replace the random walker with a routed vehicle/individual.

The student should:

- Download the road Network of Barcelona, Spain and its neighborhoods
- Associate to each neighborhood an income level among *k=10* class (choose a proper binning strategy)
    - Find the income distribution here or a measure of the richness here. Pick the one you think is more suitable.
- Repeat the following procedure several times for statistical robustness:
    - Connect each neighborhood with all the others, selecting an edge (street) randomly within the origin and destination areas.
    - For each origin-destination pair, compute a route using the following routing strategies:
        - fastest path, a perturbed fastest path
        - shortest path, a perturbed shortest path
        - at least two Alternative Routing algorithms (select one route randomly among the alternatives)
    - For every route and strategy, compute EC, i.e., how many distinct neighborhood classes (income levels) the route passes through. Normalize this count by dividing it by the length of the route, expressing it as EC per kilometer (EC/km)
    - Associate to each neighborhood a segregation score *S* computed as the average *EC/km* of the paths departing from that neighborhood

Develop a coherent discussion about what you find. One question is: How does the routing strategy impact segregation? Does the spatial distribution of *S* change when changing the routing strategy?

Bonus question: What if people walk instead of driving?

[1] Sousa, Sandro, and Vincenzo Nicosia. "Quantifying ethnic segregation in cities through random walks." *Nature Communications* 13.1 (2022): 5809.

Project 9: Mobility Flow Acceleration

The dataset described in [1] offers, with several hierarchies of high spatial resolution, an interesting insight into the human mobility flows in the US during pandemic years (2019-2021).

Following the instruction on their GitHub, select some major cities/areas in the US, focusing on the spatial hierarchy you think is more suitable.

In the literature, it is known that the attractivity of a place is correlated with the in-flow of the place itself. Nevertheless, in some kinds of mobility (e.g. mobility during pandemics, residential mobility) it is interesting to analyze how the trend of in-flows and out-flows of a place changes.

The challenge of this project is to build a measure that can capture the flow "acceleration" in the evolving mobility network. Conceptually, the student should define, for a temporal window *h* (the parameter of the analysis), something like: $\frac{flow_t - flow_{t-h}}{h}$. Of course, a negative flow would indicate a "deceleration" so a force that pushes people out of the place, while a positive one would indicate an attracting force. Investigate both inflow and outflow acceleration.

Perform an analysis in which the measure is clearly defined and studied. Is a zone with a positive acceleration associated with one or more zones with a negative acceleration? (i.e. do their acceleration/deceleration grow together or other strategies). If we analyze the acceleration in terms of both origin and destination (i.e. we observe an acceleration of 5 between place A and place B) is this acceleration correlated with more acceleration towards near or relevant places? Develop a coherent discussion about what you find.

[1] Kang, Yuhao, et al. "Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic." *Scientific data* 7.1 (2020): 390.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 10: Safest Near Shortest Path

Navigation systems usually suggest the fastest path to reach a user's desired destination, starting from a given location. While this optimization at an individual level is undoubtedly advantageous, the aggregate effect of all recommended paths may lead to a growing negative impact, particularly in emissions, if many vehicles converge on the same route.

A way to overcome this problem is using Alternative Routing algorithms that help to dilute traffic across more roads. However, they do not consider dynamic events in traffic for which we may want to avoid some areas not to congest them, e.g., a stadium when there is an event in progress, a hospital to clear the path for ambulances, schools when kids are going out, etc.

Develop an alternative routing algorithm that, given an origin o and a destination d and a set of POIs P:

1. Generates k alternative paths from o to d;

2. Each alternative must avoid edges that are m-meter (select a proper value and discuss it) closer to any POI in P.

3. The generated alternatives should be diverse enough. Think of a way to assess their diversity and select a proper diversity threshold to categorize them as diverse.

How does the path cost change with respect to the optimal path without considering the POIs? How does the distribution of POIs across the city influence the quality of the alternatives?

Develop a coherent discussion about what you find.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 11: Mobility Hubs in 15 minutes

Nowadays, the "15-minute city" [1] is a concept guiding more and more urban strategists and planners. According to this concept, a citizen should be able to perform six essential functions within a 15-minute walk or bike ride from their homes: living, working, commerce, health care, education and entertainment.

Based on the data of the MobiDataLab Codagon or similar datasets, you must propose an intelligent deployment of the mobility hubs in Louvain so as to ensure a maximum coverage of the 15-minute measures.

Technically, you may choose to approach the 15-minutes walkability-rideability measure with an isochrone curve approach, or to simulate it using tools like SUMO or other routing approaches.

Of course, different policies can be implemented: deploy mobility hubs so to minimize the travel time as much as possible for a set of zones and discard the others, or sacrifice a bit some zones so to have an average travel time from an higher number of zones that is lower than 15 minutes, or the deployment can respect the population density (i.e. is better/worse to leave more isolate a zone with an higher/lower population density).

Develop a coherent discussion about what you find.

[1] Moreno, Carlos, et al. "Introducing the "15-Minute City": Sustainability, resilience and place identity in future post-pandemic cities." *Smart Cities* 4.1 (2021): 93-111.

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 12: Spatial fairness in urban environments

Is Louvain's Mobility Hubs deployment fair? In other words, do they contribute to Louvain's segregation?

Imagine that a way for measuring the segregation of a place is to estimate the average (among different simulations) amount of time an user needs, starting from a Mobility Hub for:
i) getting out of a neighborhood (picking a random starting street per simulation)
ii) reach the city center (picking a random arriving street per simulation).

Each Mobility Hub configuration leads to different Segregation measure distribution (according to both of the metrics).

Based on the data of the MobiDataLab Codagon or similar datasets, perform an analysis that shows how different deployments of Mobility Hubs can mitigate/exacerbate the segregation scenarios and average values in the city. Is a particular type of deployment preferrable with respect to another? Can you find some generalities in the deployments that lead to higher average segregation levels? Is a measure of segregation more sensitive to some kind of the others w.r.t the other one?

Develop a coherent discussion about what you find.

Project 13: Optimistic segregation

The Schelling segregation model evaluates the happiness of agents based on the number of similar neighbors in comparison to a predefined threshold for segregation acceptance. Unhappy agents relocate to either a random or a specific vacant cell.

Over time, the model may converge to a scenario where all agents are happy. This project seeks to address a fundamental question: What happens during the convergence process and the ultimate level of segregation when introducing an "optimistic behavior" in agents? Specifically, what if unhappy agents, instead of instantly moving to a new cell, are willing to remain in their current position for a while, anticipating future satisfaction?

To explore this, various parameters can be considered, such as the level of the threshold for accepted optimism (under which agents move according to the classic model), the number of steps agents are willing to wait (which could be fixed or a function of the threshold). For instance, agents might possess an internal variable indicating their patience, increasing during happy steps and decreasing during unhappy ones. Alternatively, a timer could be implemented to count overall waiting steps. Various other implementations are conceivable.

The project aims to answer the question: How do time of convergence and final segregation level change, if agents are available to wait some steps in an unhappy place, hoping that they will become happy there in future?

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DI PISA

Project 14: The 30 km/h speed limit in cities. Beneficial or not?

In the last month some cities started to impose a speed limit of 30 km/h, especially in the city center. The rationale behind this decision is to limit the number of accidents.

The goal of this project is to check whether the application of this measure impacts the urban scenario in terms of collective impact (e.g., traffic congestion and CO2 emissions).

Using the mobility simulator SUMO, develop a framework to assess the impact of travel speed limits changes in terms of collective impact.

Apply the simulative framework in the city of Bologna (Italy):

- As to do so you should, first of all, estimate the mobility demand (the OD-matrix) of Bologna
    - Set up a Gravity Model. The population of Bologna is reported here at different spatial hierarchies. Focus on the "Area Statistica" column. For the relevance of each Area Statistica you can either use a measure of distance from the center, the number of roads/intersections, or the number/type of facilities in OpenStreetMap that you think are suitable for the analysis
- Based on the typical OD-matrix, compute a traffic demand of N (select a reasonable value) vehicles moving in Bologna, reflecting the flows described by the OD-Matrix.
- Simulate the traffic flowing in Bologna with the speed limits of the current road network and with all speed limits set to 30 km/h.
- Measure the impact of imposing a speed limit to 30km/h, such as CO2 emissions, accidents, travel times, etc.

Develop a coherent discussion about what you find. You may answer questions like: Is the application of this measure beneficial w.r.t to a scenario with no speed limitation or with the current speed limitation? How does the traffic change? What is the best speed limit to apply to reduce congestion? Should all the roads share the same speed limit?