Contents

[1 Project Background 2](#_Toc119770723)

[2 Task 1: Data Preprocessing 2](#_Toc119770724)

[3 Task 2: GPS Point Visualization 4](#_Toc119770725)

# 1 Project Background

Many vehicles such as taxies that are moving in an urban space have their locations sensed in real time nowadays. The sensed data corresponds to sequences of time-stamped points and is called trajectory data. Raw trajectory data usually involves noises (including sensor ones and measurement ones). As a result, the locations of a vehicle as indicated by the raw coordinates are often not on the road networks. A common practice is to map the trajectory data (of GPS points) to road networks before it is visualized and analyzed. This project is to conduct several tasks of preprocessing, visualizing and query processing trajectory data, where one important process is to map the trajectory data to a road network.

# 2 Task 1: Data Preprocessing

This project explores the concept of map matching using taxi trajectory data in the city of Porto, Portugal. Taxi trajectory data was obtained from Kaggle while geospatial data was downloaded from OpenStreetMap database.

1. *taxi trajectory data*

The taxi trajectory dataset consist of 9 variables describing 1,710,670 trajectories of 442 taxis traversing the streets of Porto, Portugal from 1 July 2013 to 30 June 2014.

Table 1: Data Description

|  |  |  |
| --- | --- | --- |
| **Feature** | **Type** | **Description** |
| TRIP\_ID | string | It contains an unique identifier for each trip |
| CALL\_TYPE | char | It identifies the way used to demand this service. It may contain one of three possible values:  ‘A’ if this trip was dispatched from the central;  ‘B’ if this trip was demanded directly to a taxi driver on a specific stand;  ‘C’ otherwise (i.e. a trip demanded on a random street). |
| ORIGIN\_CALL | Integer | It contains an unique identifier for each phone number which was used to demand, at least, one service. It identifies the trip’s customer if CALL\_TYPE=’A’. Otherwise, it assumes a NULL value; |
| ORIGIN\_STAND | Integer | It contains an unique identifier for the taxi stand. It identifies the starting point of the trip if CALL\_TYPE=’B’. Otherwise, it assumes a NULL value; |
| TAXI\_ID | Integer | It contains an unique identifier for the taxi driver that performed each trip; |
| TIMESTAMP | integer | Unix Timestamp (in seconds). It identifies the trip’s start; |
| DAYTYPE | char | It identifies the daytype of the trip’s start. It assumes one of three possible values:  ‘B’ if this trip started on a holiday or any other special day (i.e. extending holidays, floating holidays, etc.);  ‘C’ if the trip started on a day before a type-B day;  ‘A’ otherwise (i.e. a normal day, workday or weekend). |
| MISSING\_DATA | boolean | It is FALSE when the GPS data stream is complete and TRUE whenever one (or more) locations are missing |
| POLYLINE | string | It contains a list of GPS coordinates (i.e. WGS84 format) mapped as a string. The beginning and the end of the string are identified with brackets (i.e. [ and ], respectively). Each pair of coordinates is also identified by the same brackets as [LONGITUDE, LATITUDE]. This list contains one pair of coordinates for each 15 seconds of trip. The last list item corresponds to the trip’s destination while the first one represents its start; |

Initial analysis of the dataset found 80 `TRIP\_ID`s that appeared more than once, with 1,710,589 unique count of “TRIP\_ID”s. However, this project uses only the first 1000 trips and they appear to all have unique TRIP\_IDs. Given that “POLYLINE” gives a list of GPS coordinates for each trip, with the first and last pairs of coordinates representing the start and end of a trip, we expect “POLYLINE” for each trip to be of at least length 2. However, among the first 1000 trips, we observe 10 trips with “POLYLINE” of length lesser than 2, suggesting potential for further data quality issues.

Table 2: Data Anomalies

|  |  |  |
| --- | --- | --- |
| **TRIP\_ID** | **POLYLINE** | **length** |
| 1372642886620000403 | [[-8.610912,41.145786]] | 1 |
| 1372644692620000173 | [[-8.598186,41.145633]] | 1 |
| 1372644011620000463 | [[-8.617662,41.14638]] | 1 |
| 1372650895620000403 | [[-8.615817,41.147298]] | 1 |
| 1372653466620000403 | [[-8.62704,41.15187]] | 1 |
| 1372658129620000397 | [[-8.606484,41.144634]] | 1 |
| 1372660161620000403 | [[-8.580051,41.159394]] | 1 |
| 1372653445620000454 | [[-8.591157,41.153769]] | 1 |
| 1372665673620000353 | [] | 0 |
| 1372669271620000653 | [[-8.628399,41.15772]] | 1 |

1. *geospatial data*

Geospatial data was downloaded from OpenStreetMap in geojson format for Porto City, Portugal. We then visualized the geojson data on geojson.io.

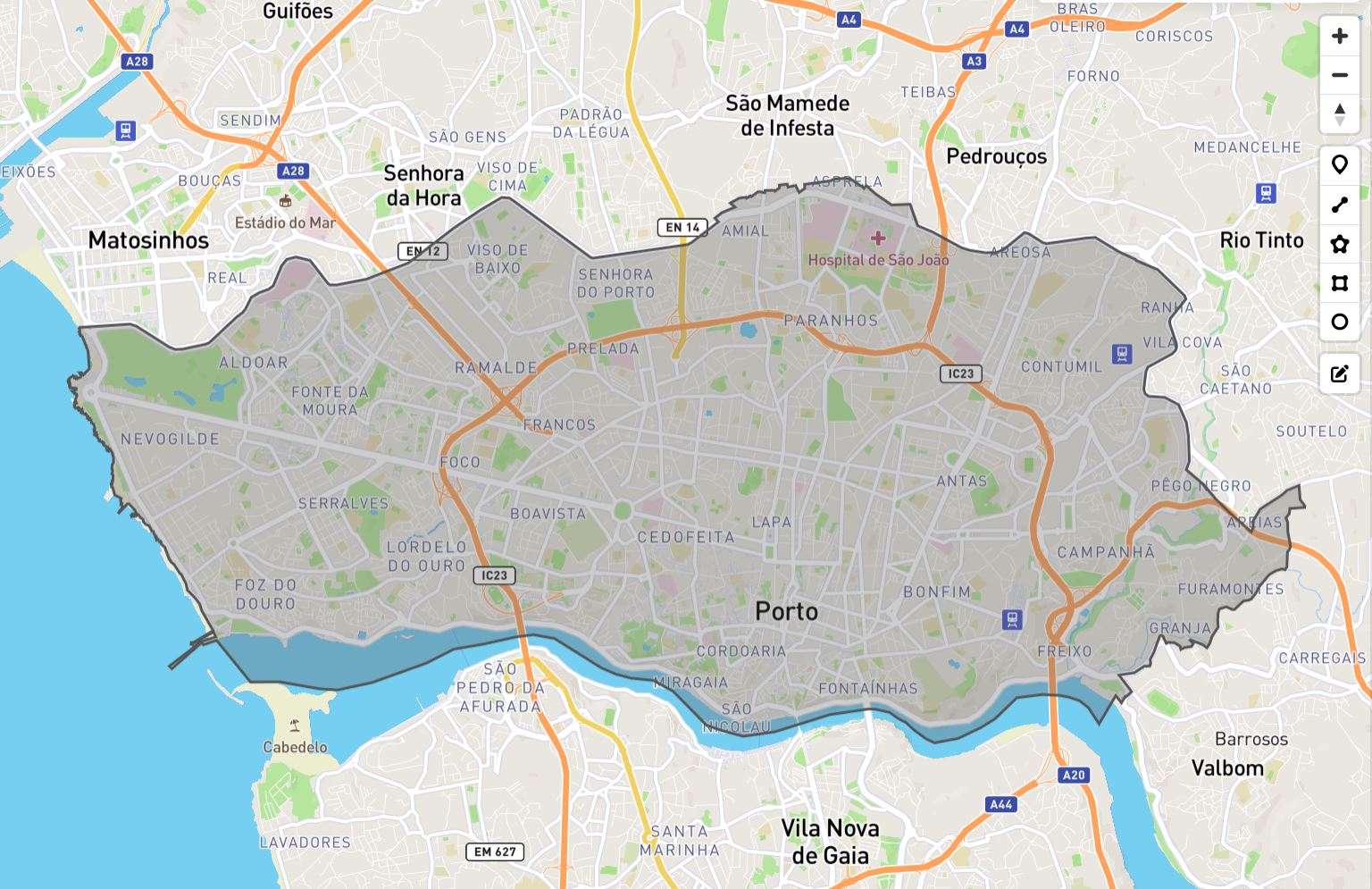


Figure 1: Extracted Road Network Data

Using the OSMnx Python package, the boundary polygon from the geojson data is used to create the road network graph.

# 3 Task 2: GPS Point Visualization

To create the GPS Point Visualisation, Folium Python package was used to complement OSMnx as it allows plotting of points and lines onto an interactive Leaflet map which improves readers’ visualisation experience. Figure 2 illustrates the differences between the two visualisation methods.

|  |  |
| --- | --- |
| 1. OSMnx plot | C:\Users\yuanq\AppData\Local\Microsoft\Windows\INetCache\Content.Word\trip_8.jpg  (b) Folium Plot |

Figure 2: Plots of a sample trip using different visualisation methods

Figure 3 shows the individual GPS points of the first 10 trajectories in the dataset, plotted separately on a map with the Porto road network.

|  |  |
| --- | --- |
| (a) Trip 1 | (b) Trip 2 |
| (c) Trip 3 | C:\Users\yuanq\AppData\Local\Microsoft\Windows\INetCache\Content.Word\trip_4.jpg  (d) Trip 4 |
| (e) Trip 5 | (f) Trip 6 |
| (g) Trip 7 | (h) Trip 8 |
| (i) Trip 9 | (j) Trip 10 |

Figure 3: Individual plots for each of the 10 GPS trajectories with rough lines connecting each point

In preparing and studying these plots, we noted two key observations:

1. Trips 3 (Figure 3c) and 4 (Figure 3d) are examples of unaligned trajectory lines as map matching has not been performed on the raw LineString data;
2. Zooming into the view of the GPS coordinates in Trip 4 (Figure 4), we can see that these GPS points are not aligned to the actual road segments.

As a result, the lines within these plots are not always aligned to the road segments and should serve only as a visual aid. This iterates the importance of map matching to align these GPS points to the most appropriate road segments.



Figure 4: Unaligned GPS points in Trip 4 circled in green