London Graph Dataset

1. Dataset description

The dataset provided here is an output of the Track & Know project, shared with the scientific community. The dataset consists of aggregate origin-destination (OD) flows of private cars in London augmented with feature data describing city locations and dyadic relations between them. The geographical location of each cell in the OD graph is not provided, for privacy protection, since the extension of each area is relatively small.
The dataset was first used in the following publication:

- Gevorg Yeghikyan, Felix L. Opolka, Bruno Lepri, Mirco Nanni, Pietro Lio`. Learning Mobility Flows from Urban Features with Spatial Interaction Models and Neural Networks. 2020 IEEE International Conference on Smart Computing (SMARTCOMP), to appear.

2. How to download the data

In order to access the dataset, please submit a request to:

mirco.nanni@isti.cnr.it

In the request, please specify your institution (if any), and the purpose of using the dataset (Research, Business, etc.).

3. Disclaimer about the Proprietary Data Sources

The vehicle flows of the OD matrix have been inferred from a proprietary dataset, as an aggregation of geographical areas (500 x 500 metres) and traffic flows of 10000 vehicles. The dataset, containing anonymous GPS traces of private vehicles, was made accessible by the data owner to the partners of the Track & Know project, for activities relevant for the project.
The original dataset was stored on a MongoDB hosted and managed by the Sistematica S.p.A. project partner, and the OD matrix flows has been obtained as outputs through a sequence of ad hoc queries to the database server and iterative computations on the client side. Eventually, on the client side no piece of information other than the output matrix was kept.
The proprietary dataset is not accessible to the public.

4. Dataset creation process

The dataset was computed through the steps described in the following:

1. The urban territory has been subdivided into *n* Cartesian grid cells of size 500x500 m, and each such quadratic cell is considered a node in the graph.
2. The GPS trajectories of around 10'000 cars spanning a period of one year, have been superimposed on the grid, and trip origins and destinations have been extracted and associated to cells of the grid.
3. The OD network has been built from the extracted origin-destination pairs by aggregating the flow counts over a year. Since the aggregation spans such a long time period, the OD matrix is approximately symmetric, and thus it has been converted into a symmetric matrix by averaging the matrix with its transpose.
4. The node features have been built by engineering 35 features from various open sources (OpenStreetMap, Airbnb, LondonTransport) and from the GPS data. These features include population density, average Airbnb prices, parking areas, areas covered by residential buildings, number of restaurants, bars, banks, museums, road network density, average radius of gyration, etc. per cell.
5. Similarly, the edge features encode information on 12 dyadic relations such as network distance, average time, average speed, temporal correlation between car incidence in cells, public transport connections, etc. The detailed attribute description is provided below.

5. Content

The London Graph Dataset consists of a file containing the nodes (grid cells) of the graph with associated features, and another one for the edges (flows between two cells).
Dataset size:

- 6791 Nodes, each described by 36 features
- 23062231 edges, each described by 13 features (+ references to start and end cells)

A detailed description of the features follows:

Edge dataset features:

location_1, location_2 : IDs of the London grid cells

1. flows : mobility flow counts between cells location_1 and location_2
2. netw_distance : the physical road distance between the centroids of location_1 and location_2 extracted from OpenStreetMap.
3. total_loc_flow : the total in/out-flow associated to location_1
4. route_factor : the ratio between netw_distance and the Euclidean distance between the centroids of location_1 and location_2. It is greater or equal to 1.
5. subway : the number of subway lines between location_1 and location_2.

6. bus : the number of bus lines between location_1 and location_2
7. airbnb : average of the Airbnb prices of location_1 and location_2
8. speed : average travel speed between location_1 and location_2
9. time: average travel time between location_1 and location_2
10. corr_at_destinations : correlation between the time series of car arrivals at location_1 and location_2
11. corr_incidence : correlation between the time series of car incidences at location_1 and location_2
12. location1_to_neighbourhood : the aggregate car flow count from location_1 to the immediate geographic neighbours of location_2
13. neighbourhood_to_location2 : the aggregate car flow count from the neighbours of location_1 to location_2

Node dataset features:

NodeID : IDs of the London grid cells

1. airbnb_price : average airbnb price in a cell
2. universities : total area of universities in a cell
3. tourism : number of touristic attractions
4. theatres : number of theatres
5. shops : number of shops
6. shopping_malls : number of shopping malls
7. restaurants : number of restaurants
8. residential : total residential area
9. pubs_cafes : number of pubs and cafes
10. post : number of postal offices
11. parking : total parking area
12. offices : number of office buildings
13. museums : number of museums
14. medical : total area of medical facilities
15. schools : number of secondary schools
16. industrial : total industrial area
17. government : number of governmental institutions
18. fuels : number of gas stations
19. fast_foods : number of fast food restaurants
20. commercial : number of commercial firms
21. cinemas : number of cinemas
22. bars_cafes : number of bars
23. banks : number of banks
24. atms : number of ATM machines
25. arts : number of arts venues
26. airport : binary dummy variable denoting 1 if an airport falls into the cell, and 0 otherwise
27. in_total : total inflow to cell
28. out_total : total outflow from cell
29. street_density : number of street junctions per cell
30. gyration_radius : average radius of gyration of cars “housed” in the cell

31. gyration_radius_spatial_lag : average radius of gyration of cars “housed” in the geographic neighbours of a given cell
32. highways : binary dummy variable denoting 1 if the cell is located on the edge of a city on a highway or street leaving/entering the city
33. metro_flow : number of passengers that have entered the subway in the cell
34. avg_betw_centrality : average of the network betweenness centrality values of the street junctions in the cell
35. in_total_spatial_lag : total inflow to the geographic neighbours of a cell
36. out_total_spatial_lag : total outflow from the geographic neighbours of a cell