Tuscany City Features

1. Dataset description

The dataset provided here is an output of the Track & Know project, shared with the scientific community. The dataset consists of mobility- and road network-related aggregates computed for each municipality in Tuscany (Italy), based on private cars GPS traces and the OpenStreetNetwork graph.
The application that originated this dataset is the study of geographical transferability of mobility models, in order to understand under which conditions a model (e.g. to predict traffic) built in one area can be used in another one, or what kind of adjustment is needed.
The dataset and the application mentioned above appeared in the following publications:

- Mirco Nanni, Agnese Bonavita, Riccardo Guidotti. City Indicators for Mobility Data Mining. In Proceedings of BMDA 2021: 4th International Workshop on Big Mobility Data Analytics. http://ceur-ws.org/Vol-2841/BMDA_10.pdf
- Leonardo Longhi, Mirco Nanni. Car telematics big data analytics for insurance and innovative mobility services. In Journal of Ambient Intelligence and Humanized Computing (JAHC), p. 1-11, Springer, 2017. 10.1007/s12652-019-01632-4

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 mobility indicators of the shared dataset have been inferred, as aggregation of geographical areas (10km x 10km, each corresponding to a municipality in Tuscany) of ca 107000 vehicles, from a proprietary data source. The proprietary 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 mobility indicators have 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 Content

The Tuscany City Features Dataset consists of 5 files, each providing a set of features of the same family for each municipality of Tuscany.
Dataset size of each file:

- 1_concentration.csv: 14 features x 276 municipalities
- 2_od_matrix_network.csv: 25 features x 276 municipalities
- 3_IMNs.csv: 14 features x 276 municipalities
- 4_static_roads.csv: 7 features x 276 municipalities
- 5_road_traffic.csv: 104 features x 275 municipalities

A detailed description of the features follows:

1_concentration.csv
The variables in this section all refer to spatial concentration/dispersion:

1. entropy_10*10: Entropy of the geographical distribution of activities in a 10*10 grid. The end of each trajectory (e.g. a car that is parked) in the database is considered as an activity. All activities that are in the database are aggregated in the grid. Each activity has the same weight.
2. entropy_20*20: The same for a 20*20 grid.
3. entropy_50*50: The same for a 50*50 grid.
4. entropy_norm_10*10: Normalized Entropy of the geographical distribution of activities in a 10*10 grid. The end of each trajectory (e.g. a car that is parked) in the database is considered as an activity. All activities that are in the database are aggregated in the grid. Each activity has the same weight. Normalization is obtained by dividing the entropy by the information length (amount of fields in the grid).
5. entropy_norm_20*20: The same for a 20*20 grid.
6. entropy_norm_50*50: The same for a 50*50 grid.
7. moran_i_10_10: Moran's I measure for spatial autocorrelation of activities in a 10*10 grid. The end of each trajectory (e.g. a car that is parked) in the database is considered as an activity. All activities that are in the database are aggregated in the grid.
8. moran_i_20_20: The same for a 20*20 grid.
9. moran_i_50_50: The same for a 50*50 grid.
10. avg_hour_annd: Average of the hourly average nearest neighbor distance values. Explanation: For every hour of the time window of the database, the locations of ongoing

activities are considered ("Snapshot"). An activity is considered to happen, whenever a trajectory ends in the database (e.g. a car is parked). The activity is considered to be terminated, when the car moves again or after the duration of 24 hours. For every hour, the average nearest neighbor distance between those activities is calculated, if there are more than 100 activities. Otherwise, the hour is ignored. The value in this field is the average of those values.

11. avg_hour_entropy_10_10: Average of hourly entropy values measured in a 10*10 grid. The hourly entropy are calculated for "Snapshots" for each hour available in the data which have more than 100 ongoing activities. (see avg_hour_annd, where the same "Snapshot" method is used)
12. avg_hour_moran_i_10_10: Average of hourly Moran I values measured in a 10*10 grid. The hourly entropy are calculated for "Snapshots" for each hour available in the data which have more than 100 ongoing activities. (see avg_hour_annd, where the same "Snapshot" method is used)
13. n_snapshots: The amount of "snapshots" that were used for the calculation of the previous 7 columns. (They were only used if they contained more than 100 activities).
14. avr_snapshot_size: Relevant for the previous 8 columns: How many activities are on average in a snapshot

2_od_matrix_network.csv

The variables in this section describe the directed weighted network obtained by considering the origin-destination matrix of trips within a grid across the municipality. The grid fields are nodes in the network. Trips are represented with weight 1 from the field of origin to the destination field.

- n_trips: Number of trips in the databased. All trips were used to calculate the network.
- n_real_nodes_10_10: Number of nodes in the network that have at least one in- or outgoing trip/degree. A 10*10 grid is used.
- degree_percXX_10_10: XXth percentile of the node degrees in ascending order. Also nodes with a degree of 0 are counted. A 10*10 grid is used.
- louvain_score_10_10: Louvain modularity score of the network. A 10*10 grid is used.
- louvain_partitions_10_10: Amount of partitions obtained by the louvain algorithm for optimizing modularity.

The same variables are computed on other grid sizes, with the suffix 20_20 and 50_50 that refer to a 20*20 grid and a 50*50 grid, respectively.

3_IMNs.csv

The variables in this section describe average properties of Individual Mobility Networks (IMNs). The nodes in the IMN represent a point on a map. All origins/destinations that lie within a distance of 125 meters are grouped to that node. The IMNs are grouped to the cities by the location of the individual's most frequent location.

- n_networks: Amount of IMNs
- avg_n_nodes: Average amount of nodes in the network.

- avg_n_trips: Average amount of trips.
- avg_rad_gyr: Average radius of gyration.
- avg_entropy: Average uncorrelated entropy.
- avg_variance_arrival_time_minutes: Average of the mean temporal variance of the node's arrival times.
- avg_total_driven_meter: Average of the total meters driven.
- avg_total_trip_time_driven_hours: Average of the total hours driven.
- distinct_days: On how many distinct days all of a person's trips take place in the database on average.
- avg_louvain_score: Average Louvain modularity score of the IMN.
- avg_louvain_partitions: Average amount of partitions obtained by the Louvain modularity algorithm.
- avg_n_nodes_regular: Average of amount of regular nodes. Regular nodes are defined as nodes that are visited at least once per week and that have been visited at least twice in absolute terms.
- time_regular_ratio: The average value of the time spent on regular trips divided by the time spent on all trips. Regular trips are defined as trips that take place at least once per week and take place at least twice in absolute terms.
- avg_driven_regular_ratio: The average value of the meters driven during regular trips divided by the meters driven in total. (See above for definition of regular trips.)
- avg_ratio_regular_trips: The average value of the amount of regular trips divided by the total amount of trips. meters driven in total. (See above for definition of regular trips.)

4_static_roads.csv

The variables in this section refer to the network of roads (without the information about trips or traffic):

- amount_nodes: Amount of nodes in the road network. Nodes are intersections or dead-ends.
- amount_intersections: Amount of intersections in the network.
- amount_edges: Amount of directed edges. Directed edges are road segments leading from one node to another. If such a road segment is bidirectional, two edges are counted.
- amount_edges_undirected: Amount of undirected edges. An undirected edge is a segment between two nodes, representing a road segment, and it is counted once, no matter if it is bidirectional or one-way.
- total_edge_length: The length of all directed edges. (e.g. bidirectional roads are counted twice)
- total_edge_length_undirected: The length of all undirected edges. (e.g length of all roads, regardless if one-way or not)
- avg_centrality: Average closeness centrality of the network nodes. A nodes closeness centrality is here defined as the reciprocal of the average distance to all other nodes. The distance is measured as the road length of the shortest path in km.

5_road_traffic.csv

The variables in this section refer to trajectories in the database that were matched to the road maps of the municipalities. The trajectories are a random sample out of all trajectories in the database which have either their starting point or their destination in the municipality's bounding box. Variables that do not contain any values ("NA") did not have enough samples to calculate the values (min: 2000).

- sample_size: How many trajectories were used to calculate the values. (In these data, a minimum and maximum amount of 2000 was used, therefore this value is always either 2000 or "NA").
- n_in_in: How many trajectories of the samples start inside the bounding box and end within.
- n_in_out: How many trajectories of the samples start inside the bounding box and end outside.
- n_out_in: How many trajectories of the samples start outside the bounding box and end within.
- traffic_in_lowest_X: The percentage of traffic that lies within the X percent least busy roads segments (measured in road length). Only roads that are used within the overall sample are considered for the calculation of the percentage.