Università di Pisa	A.A. 2014-2015
Data Mining II	

Project assignment

Taxi cabs in S.F.

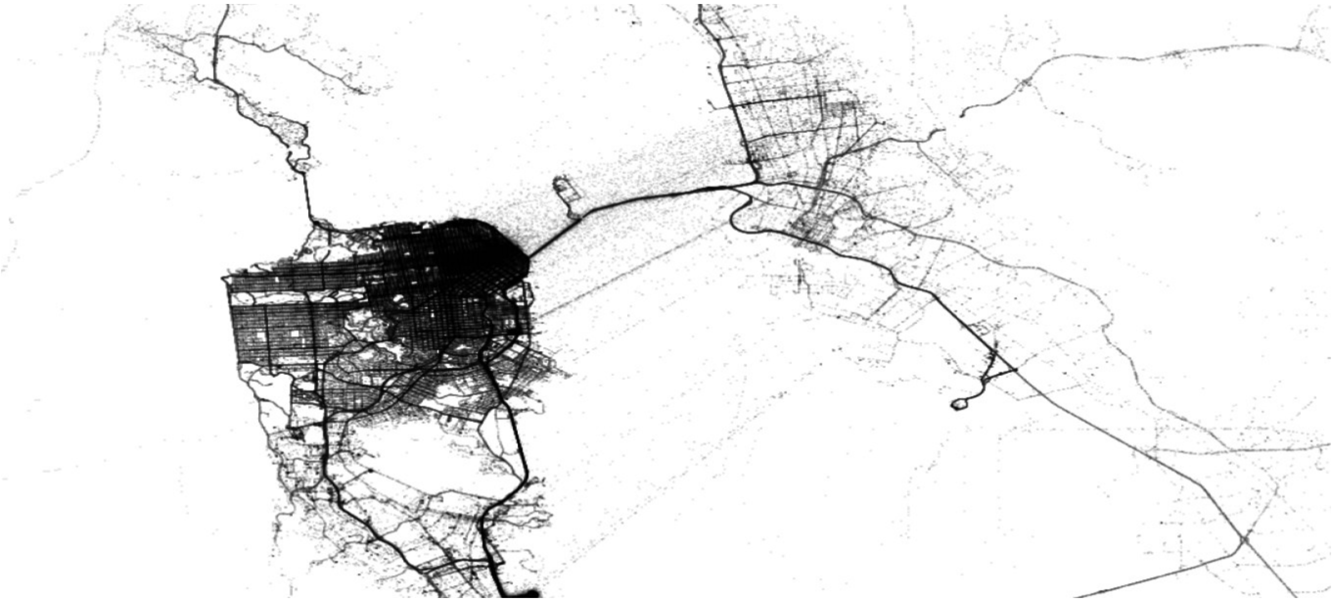
General information

Objective of this project is to perform a few analyses on a dataset of mobility data involving taxis in San Francisco. The general guidelines for this assignment are the following:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the processing and analyses indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers;
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it) and Anna Monreale (anna.monreale@unipi.it).

The dataset

GPS traces of ~500 taxis over 30 days. Each San Francisco based Yellow Cab vehicle is currently outfitted with a GPS tracking device. The data is transmitted from each cab to a central receiving station, and then delivered in real-time to dispatch computers via a central server. This system broadcasts the cab call number, location and whether the cab currently has a fare. The following picture shows (a zoom of) the density distribution of points in S.F. downtown.



The raw dataset includes ~500 files, one per cab, containing <Latitude, Longitude, Passenger?, Unix Timestamp>. E.g.: 37.80246 -122.40186 0 1213034473.
The processed dataset consists of two sets of trajectories: one for taxi trips with passengers on-board, and one for trips without passengers.

Objectives

The following activities should be performed and reported:

1. Exploration of the data, aimed to describe its general characteristics (distribution of key variables, spatial coverage, etc.) and detect possible issues (noise, strange behaviours, etc.). This phase should include the construction of a origin-destination matrix to explore the distribution of flows across areas.
2. Through the O/D matrix built above, select a significant area "A" in S.F. and compare the behaviour of "loaded trips" (i.e. taxi trips with passengers onboard) vs. the "unloaded" ones, adopting the three following approaches:
    a. for each origin, compute the percentage of loaded trips that go towards the selected area "A", and analyze the distribution of the values obtained;
    b. divide the set of trips directed to the area "A" into two separate groups: the dataset "AL" of loaded trips, and the dataset "AU" of unloaded ones. Compute the main access patterns within the two sets, and compare the results;
    c. try to characterize each of the datasets "AL" and "AU" through some set of indicators, such as travel duration, travel length, average speed, etc.