Speed Prediction from Taxi Trajectory Data

Chunheng Jiang

December 7, 2018

Problem

We are given a trajectory data of DiDi Express and DiDi Premier drivers within the Second Ring Road of Xi'An City. All track points are bound to physical roads with resolution about 2-4 seconds. The problem is to predict the average speed of all vehicles running on a segment of road (either north or south bound) at specific time.

Statistics of Trajectory Data

The trajectory data contains 23,654,135 records, each entry has three columns: driver id, order id, timestamp, longitude, and latitude. Both driver id and order id are encrypted and anonymized with long strings. To reduce the size of the data and the memory usage, we map them into integers. As a result, the file size shrinks from 2.2G to 1.0G. The unix timestamp can be converted into readable format, such that accompanied with the location we can calculated the average speed of the vehicle at given time. To compute the distance between a pair of longitude and latitude, we are required to call eviltransform[1] using the same coordinate system.

[1] https://github.com/googollee/eviltransform

There are 18,635 drivers, 111,541 trip orders, 86,495 different timestamps, and 16,224 different locations. To have an overview of the data, we examine the records distribution of each trip order. As seen from Fig. **??**, around 100 trip orders have only one record entry in the trajectory data, 120 trip orders have 10 records in each order, and majority of the records have several records tracked.

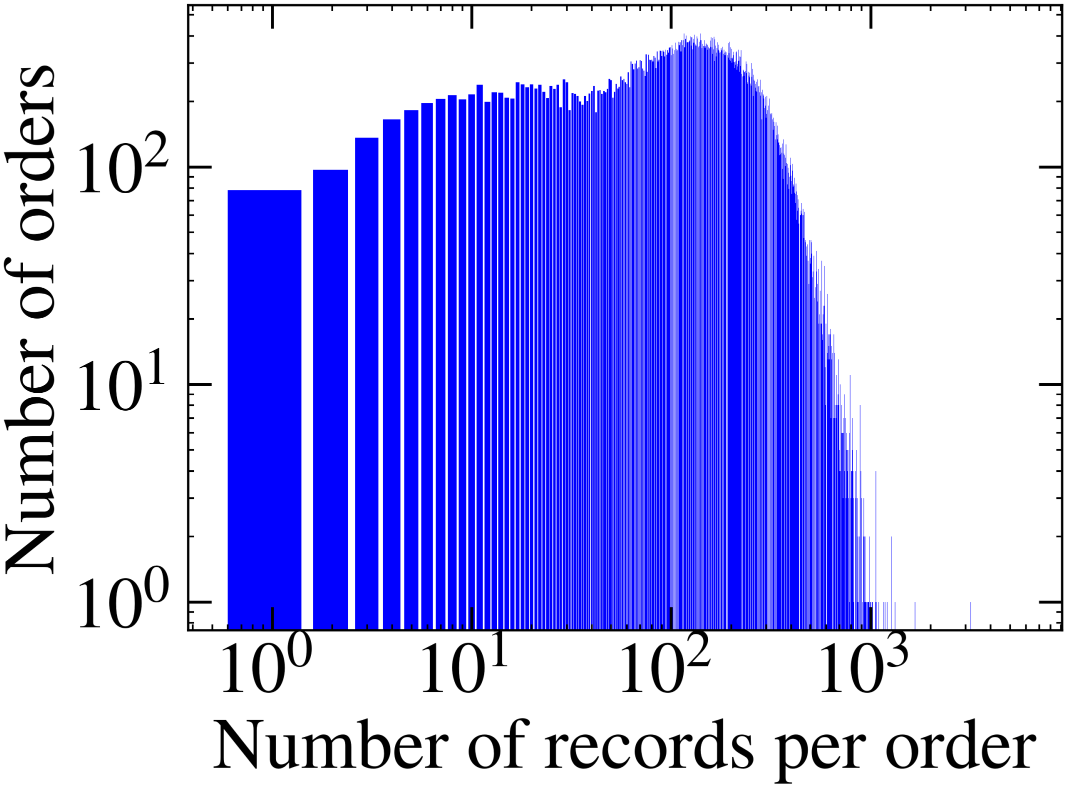

Fig. 1: **Records distribution per order.** The x−axis is the possible number of records in a trip order, and y−axis is the number of orders with specific number of recording entries.

Coordinate Systems and Timezone

There are several different coordinate systems in spatial data processing, e.g. cartography, geodesy, and satellite navigation. As shown in Fig. **??**, OpenStreetMap[2] uses the WGS-84 coordinate system (a.k.a World Geodetic System), but Google map uses GCJ-02 coordinate system (a.k.a Mars coordinate) for Chinese users. Therefore, it requires a conversion from one system to another to visualize the trajectories on Google map.

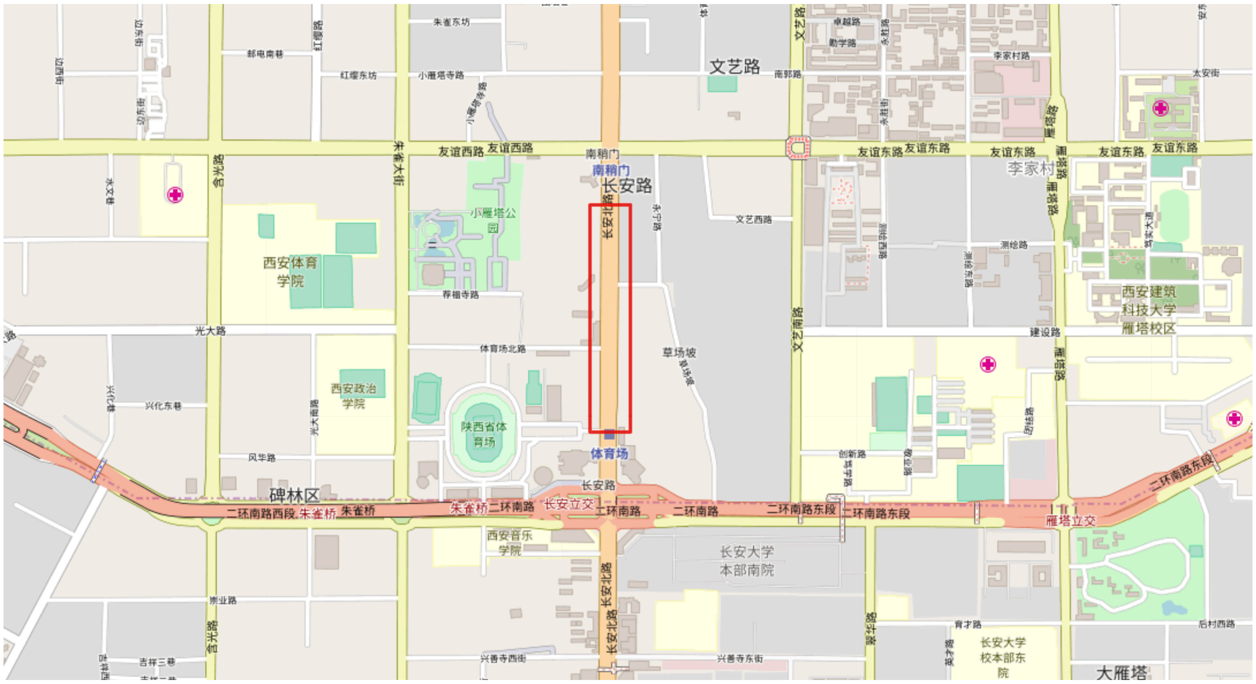

Fig. 2: **Road section to be predict.** The road section selected for prediction. The average driving speeds of all vehicles driving towards both north bound (latitude increasing) and south bound in every 5 minute are required to predict.

The road section in question (see red rectangular in Fig. **??**) is defined by four corners in counter-clockwise:

- Point 1 (upper-right): 34.241, 108.943

[2] https://www.openstreetmap.org/

- Point 2 (upper-left): 34.241, 108.9415
- Point 3 (lower-left): 34.234, 108.9415
- Point 4 (lower-right): 34.234, 108.943

Another important aspect is the underlying timezone, which is supposed to be GMT+8. Therefore, we need to convert the timestamp with specific timezone, i.e. Asia/Shanghai. The prediction will be made based on the assumption that the drivers in the extra day exactly replicates what they behaved in previous day. The trajectory file have a time range from 00:01:37 Dec 1, 2016 to 00:03:55 Dec 2, 2016. Also, the missing average speeds for north bound traffic are morning peak hours from 06:00:05 to 10:55:05, afternoon peak hours from 16:00:05 to 20:55:05, while for south bound traffic the missing average speeds in morning peak hours starts from 06:00:04 to 10:55:04, afternoon peak hours from 16:00:04 to 20:55:04.

Data Preprocessing

To predict the missing components, we need to recovery the average speeds during the same time interval based on the known trajectories and the given average speeds in other time intervals. However, the given trajectories maybe do not contribute any useful information when they do not hit the specific road section. To estimate the average speeds in the given road section, we are required to identify all trips that fall into the road section based on the latitude and longitude information.

The first step will be data filtering, we are figuring out the trajectories that have overlap region with the road section set by the rectangular. We can imagine the data will become sparser, and many trajectories will definitely been truncated. Some of the data points may be useful even though not included by the specific region. For example, the down-stream or up-stream traffic, the entering or exit ramps.