Data Description

The package is comprised of seven parts of data that were extracted from the GPS trajectories of taxicabs, road networks, POIs of Beijing, and video clips recording real traffic on roads:

1) *Real-time traffic conditions on each road segment in different time slots of a day.*
2) *Real-time traffic conditions in each region in different time slots of a day.*
3) *Road network features of each road segment and POIs around each road segment.*
4) *Road network connections of Beijing.*
5) *Traffic volume ground truth on different levels of road segments at different time slots.*

The data package has been used in paper [1] and can be widely used in many urban computing scenarios introduced in [2]. Detailed information about the generation of each part of the data can be found in [1].

1) *Real-time traffic conditions on each road segment:* The first part of data, stored in "Speed" folder, represents the real-time traffic conditions of four days, from Sep. 12nd, 2013 to Sep 14th, 2013, consisting two workdays and two holidays. The traffic information on each road segment, containing the average travel speed $\bar{}$, the standard deviation of the travel speed s , and the number of vehicles, is derived from the GPS trajectories generated by taxicabs traversing the road segment in the time slot. If there is no taxicabs traversing a road segment, the travel conditions information is empty. These information are stored in a stand-alone file by day with each row stands for a road segment in one time slot. The columns, separated by blank, are defined as follows:

Road Segment ID	Time Slot ID	$\bar{}$	s	*n*

- Road Segment ID ranges from 0 to 81,592.
- Time Slot ID ranges from 0 to 143. (We partition a day into 144 time slots with 10 minutes per time slot; time slot 0 stands for 12am-12:10am)
- $\bar{}$ means the average travel speed of all the vehicles traversing the road segment in the time slot.
- s means the standard deviation of the travel speed of all the vehicles traversing the road segment in the time slot.
- *n* means the number of vehicles traversing the road segment in the time slot.

The matrix ′ in paper [1] can be constructed based on the data stored in these files.

2) *Real-time traffic conditions in each region:* The second part of data, stored in "Time" folder, represents the real-time traffic conditions of four days, from Sep. 12nd, 2013 to Sep 14th, 2013. Real-time traffic conditions in each region, containing the number of vehicles in different time slots, is derived from real-time traffic conditions on each road segment in the same day. The information on one day in the period is stored alone, with each row stands for a region in one time slot. The columns, separated by blank, are defined as follows:

Time Slot ID	Region ID	Rn

- Time Slot ID is as the same of mentioned above.
- Region ID ranges from 0 to 15. (We partition a city into 4 X 4 disjoint grids)
- Rn means the number of vehicles traversing the region in the time slot.

The matrix ′ in paper [1] can be constructed based on the data stored in these files.

3) *Road network features and POI features:* The third part of data, stored in "Road" folder, represents the road network features of each road segment and POIs around each road segment. Road network features is derived from Beijing's Road Network in 2012 and POIs is extracted from Beijing's POI file in Quarter 3, 2013. Each row in this file stands for one feature of one road segment. Every 18 rows formulate a group belonging to the same road segment. The columns, separated by blank, are defined as follows:

   Road Segment ID, Length of a road segment, Number of Lanes, Speed constraint, Direction, Level, Tortuosity, Number of connections, Schools, Companies & Offices, Banks & ATMs, Malls & Shopping, Restaurants, Gas stations & Vehicle services, Scenic spot, Hotels & Residences, Transportations, Entertainments & Living Services, sum of POIs.

   Please refer to [1] for details. Note, it is not normalized.
   The matrix Z in paper [1] can be constructed based on the data stored in these files.

6) *Road network:* The fourth part of data, stored in "Road" folder, denotes the structure of Beijing's road network. The columns, separated by blank, are defined as follows: Road segment ID, Start Node ID, End Node ID. The road network can be reconstructed based on the three entries. That is, if two road segments share a common node ID, then they are connected. The direction information of a road segment can be found in geospatial feature.txt. 0 means one-way, start node id to end node id. 1 means bi-directional.

   Traffic volume ground truth: The fifth part of data, stored in "Volume Ground Truth" folder, denotes the real traffic volume on different levels of road segments at different time slots (both in workdays and holidays). We manually recorded 358 videos with 5 minutes as a period and counted the number of vehicles traversing these road segments by replaying the videos. The statistics of traffic volume ground truth is displayed bellow. Detail information please refer to the excel document.

Time	7:00 ~ 10:00			10:00~16:00			16:00~20:00			after 20:00			total
Lev.	0,1	2	3	0,1	2	3	0,1	2	3	0,1	2	3	
Holi	0	0	0	6	1	4	6	8	1	4	6	0	49
Work	7	2	8	29	7	9	28	9	7	6	1	4	309
Total	43			136			142			37			358

Reference:

Please cite the following two papers when using the dataset.

[1] Jingbo Shang, **Yu Zheng,** Wenzhu Tong, Eric Chang, Yong Yu. Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City. In the Proceeding of the 20th SIGKDD conference on Knowledge Discovery and Data Mining (**KDD 2014**).

[2] **Yu Zheng,** Licia Capra, Ouri Wolfson, Hai Yang. Urban Computing: concepts, methodologies, and applications. ACM Transaction on Intelligent Systems and Technology, 5(3), 2014.