Data Description

The package is comprised of six 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) *Historical fine-grained traffic patterns on each road segment in different time slots of a day.*
3) *Real-time traffic conditions in each region in different time slots of a day.*
4) *Averaged coarse-grained traffic patterns in each region in different time slots of a day.*
5) *Road network features of each road segment and POIs around each road segment.*
6) *Traffic volume ground truth on different levels of road segments at different time slots.*

Download the Data!!

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 15th, 2013, consisting two workdays and two holidays. The traffic information on each road segment, containing the average travel speed $\overline{v}$, the variance of the travel speed dv, 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	$\overline{v}$	dv	n

- Road Segment ID ranges from 0 to 115,559.
- Time Slot ID ranges from 0 to 143. (We partition a day into 144 time slots with 10 minutes per time slot)
- $\overline{v}$ means the average travel speed of all the vehicles traversing the road segment in the time slot.
- dv means the variance 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 M'_r in paper [1] can be constructed based on the data stored in these files.

2) *Historical fine-grained traffic patterns on each road segment:* The second part of data, stored in "Speed History" folder, denotes historical fine-grained traffic patterns calculated based on the data over a long period of time (form Sep. 1st, 2013 to Oct 31st, 2013). We projected the data of all the workdays (34 days) into a day, and calculated the $\overline{v}$, dv and n w.r.t. the same road segment and time slot. The information is then saved in a file, entitled "*history.workday.speed*". We did the same thing to all the 20 holidays in the period and stored the information in "*history.holiday.speed*". The columns, separated by blank, in the two files are defined as follows:

Road Segment ID	Time Slot ID	H$\overline{v}$	Hdv	Hn

- Road segment ID and Time Slot ID have the same meaning as the first part of the data.
- H$\overline{v}$ means the average $\overline{v}$ of all the records, w.r.t. the same time slot of a workday (or a holiday), on the same road segment.
- Hdv means the average dv of all the records, w.r.t. the same time slot of a workday (or a holiday) on the same road segment.
- H*n* means the sum of *n* of all the records in workdays (or holidays) w.r.t. the road segment and time slot.

H$\overline{v}$ is calculated for a time slot and a road segment, only when we can find more than 3 days, in each of which there are at least one taxi traversing the road segment in the time slot, in the historical data. So does the Hdv. The matrix M_r in paper [1] can be constructed based on the data stored in these files.

3) *Real-time traffic conditions in each region:* The third part of data, stored in "Time" folder, represents the real-time traffic conditions of four days, from Sep. 12nd, 2013 to Sep 15th, 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 M'_G in paper [1] can be constructed based on the data stored in these files.

4) *Historical coarse-grained traffic patterns in each region:* The fourth part of data, stored in "Time History" folder, denotes historical coarse-grained traffic patterns calculated based on the data over a long period of time (form Sep. 1st, 2013 to Oct 31st, 2013). We projected the data of all the workdays (34 days) into a day, and averaged number of vehicles of the same time slots and region from different workdays. The information is then saved in a file, entitled "*history.workday.time*". We did the same thing to the 20 holidays in the period and stored the information in "*history.holiday.time*". The columns, separated by blank, in the two files are defined as follows:

Time Slot ID	Region ID	HRn

- Time Slot ID and Region ID have the same meaning as mentioned above.
- HRn means the average number of vehicles of all records in workdays (or holidays) in one region with one time slot.

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

5) *Road network features and POI features:* The fifth 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, 2012. 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	Feature ID	Value

- Road Segment ID is same with the description mentioned above.
- Feature ID ranges from 0 to 17, corresponding to the Length of a road segment, number of Lanes, Speed constraints, Direction, Level, Tortuosity, Number of connections, Schools, Companies & Offices, Banks & ATMs, Malls & Shopping, Restaurants, Gas stations & Vehicle services, Parkings, Hotels & Residences, Transportations, Entertainments & Living Services, total sum of POI. The first 7 features depict road network features f_r and the rest depict POI features f_p. Refer to [1] for details.
- Value is the value of a feature of this road segment. Note, it is not normalized.

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

6) *Traffic volume ground truth:* The sixth part of data, stored in "Volume Ground Truth", 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. 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.

Contact:

Yu Zheng, yuzheng@microsoft.com

Lead Researcher at Microsoft Research

http://research.microsoft.com/en-us/people/yuzheng/