
# KDD CUP 2017

## Highway Tollgates Traffic Flow Prediction

### Travel Time & Traffic Volume Prediction

## Background

Highway tollgates are well known bottlenecks in traffic networks. During rush hours, long queues at tollgates can overwhelm traffic management authorities. Effective preemptive countermeasures are desired to solve this challenge. Such countermeasures include expediting the toll collection process and streamlining future traffic flow. The expedition of toll collection could be simply allocating temporary toll collectors to open more lanes. Future traffic flow could be streamlined by adaptively tweaking traffic signals at upstream intersections. Preemptive countermeasures will only work when the traffic management authorities receive reliable predictions for future traffic flow. For example, if heavy traffic in the next hour is predicted, then traffic regulators could immediately deploy additional toll collectors and/or divert traffic at upstream intersections.

Traffic flow patterns vary due to different stochastic factors, such as weather conditions, holidays, time of the day, etc. The prediction of future traffic flow and ETA (Estimated Time of Arrival) is a known challenge. An unprecedented large amount of traffic data from mobile apps such as Waze (in the US) or Amap (in China) can help us take up that challenge. If the contestants in this proposed KDD CUP could design reliable approaches for future traffic flow and ETA prediction, then the traffic management authorities might be able to capitalize on big data & algorithms for fewer congestions at tollgates.

## Tasks

Available datasets are: the road network topology in the target area (Figures 1, 3, and 4, Tables 3 and 4), vehicle trajectories (Table 5), historical traffic volume at tollgates (Table 6), and weather data (Table 7). The contest consists of two tasks with the details below.

**Task 1: To estimate the average travel time from designated intersections to tollgates**

For every 20-minute time window, please estimate the average travel time of vehicles for a specific route (shown in Figure 1).

**a.** Routes from Intersection A to Tollgates 2 & 3;

**b.** Routes from Intersection B to Tollgates 1 & 3;

**c.** Routes from Intersection C to Tollages 1 & 3.

Note: the ETA of a 20-minute time window for a given route is the average travel time of all vehicle trajectories that enter the route in that time window. Each 20-minute time window is defined as a right half-open interval, e.g., [2016-09-18 23:40:00, 2016-09-19 00:00:00).

**Submission Format (see Table 1)**

The data types used in all tables in this document are *int*, *float*, *string*, *date* and *datetime*. The *date* and *datetime* comply with the formats "yyyy-MM-dd" and "yyyy-MM-dd HH:mm:ss". The time_window field consists of two *datetime* types separated by a comma without any blank, e.g., "2016-09-18 08:40:00,2016-09-18 09:00:00".

Table 1. Travel Time from Intersections to Tollgates

| *Field* | *Type* | *Description* |
|---|---|---|
| *intersection_id* | string | intersection ID |
| *tollgate_id* | string | tollgate ID |
| *time_window* | string | e.g., [2016-09-18 08:40:00,2016-09-18 09:00:00) |
| *avg_travel_time* | float | average travel time (seconds) |

**Task 2: To predict average tollgate traffic volume**

For every 20-minute time window, please predict the entry and exit traffic volumes at tollgates 1, 2 and 3 (Figures 1 and 2). Note that tollgate 2 only allows traffic entering the highway while others allow traffic both ways (entry and exit). Therefore, we need to predict the volume for 5 tollgate-direction pairs in total.

**Submission Format (see Table 2)**

Table 2. Traffic Volume at Tollgates

| *Field* | *Type* | *Description* |
|---|---|---|
| *tollgate_id* | string | tollgate ID |
| *time_window* | string | e.g., [2016-09-18 08:40:00,2016-09-18 09:00:00) |
| *direction* | string | 0: entry, 1: exit |
| *volume* | int | total volume |

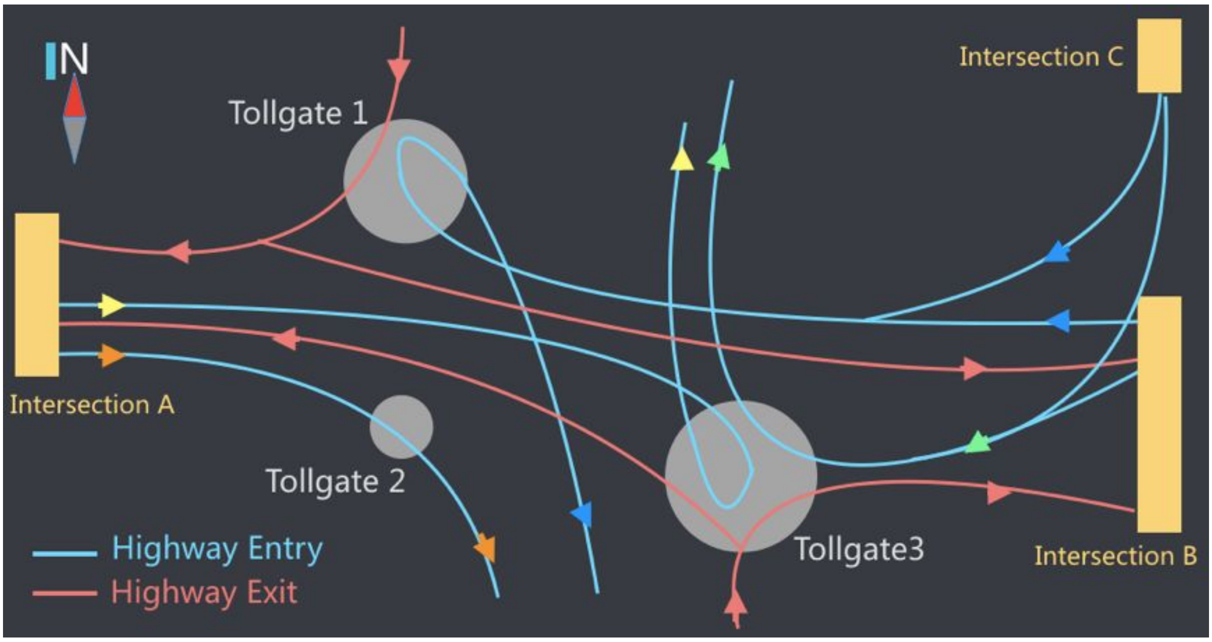


Figure 1. Road Network Topology of the Target Area

**Training & Testing Datasets:**

At the beginning of the contest, traffic predictions for specific rush hours from Oct. 18th to Oct. 24th are to be made by the contestants. On May 25 there will be a data swap, after which the participants need to predict traffic during rush hours from Oct. 25th to Oct. 31st.

Contestants are to predict the ensuing traffic during the red time slots shown in Figure 2, i.e., 08:00 - 10:00 and 17:00 - 19:00, at 20-minute intervals.


Figure 2. Time Windows for Traffic Prediction

For travel time prediction, the initial training set contains data gathered from July. 19th to Oct. 17th. For volume prediction, the initial training set contains data gathered from Sep. 19th to Oct. 17th. After the data swap on May 25, additional training data from Oct. 18th to Oct. 24th will be added for both prediction tasks.

In the testing datasets, contestants are provided with traffic data during the green time slots shown in Figure 2, i.e., 06:00 - 08:00 and 15:00 - 17:00. Contestants can use that information as a leading indicator of traffic in the next two hours, which is to be predicted.

Note: Contestants are not restricted to use only the previous 2-hour data in prediction. However, each prediction is restricted to use only the traffic data before the predicted time window. For example, contestants are NOT allowed use the traffic data from Oct. 20th to predict the traffic on Oct. 19th.

## Evaluation Metrics

We choose Mean Absolute Percentage Error (MAPE) to evaluate the result.

**Task 1:** Let $d_{rt}$ and $p_{rt}$ be the actual and predicted average travel time for route $r$ during time window $t$. The MAPE for travel time prediction is defined as:

$$MAPE = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\frac{d_{rt}-p_{rt}}{d_{rt}}\right|\right)$$

$R$ and $T$ are the number of routes and number of to-predict time windows in the testing period respectively.

**Task 2:** Let $C$ be the number of tollgate-direction pairs (as aforementioned: 1-entry, 1-exit, 2-entry, 3-entry and 3-exit), $T$ be the number of time windows in the testing period, and $f_{ct}$ and $p_{ct}$ be the actual and predicted traffic volume for a specific tollgate-direction pair $c$ during time window $t$. The MAPE for traffic volume prediction is defined as:

$$MAPE = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\frac{f_{ct}-p_{ct}}{f_{ct}}\right|\right)$$

## Data Description

The road network (Figure 1) here used is a directed graph formed by interconnected road links (Figure 3). A route (Figure 4) in the network is represented by a sequence of links. For every road link, its vehicle traffic comes from one or more “incoming road links” and goes into one or more “outgoing road links”. Table 3 and Figure 3 describe road links.

Table 3. Road Link Properties

| Field | Type | Description |
| --- | --- | --- |
| *link_id* | string | link id |
| *length* | float | length (meter) |
| *width* | float | length (meter) |
| *lanes* | int | number of lanes |
| *in_top* | string | incoming road link(s), separated by comma (as shown in Figure 3) |
| *out_top* | string | outgoing road link(s), separated by comma (as shown in Figure 3) |
| *lane_width* | float | lane width (meter) |

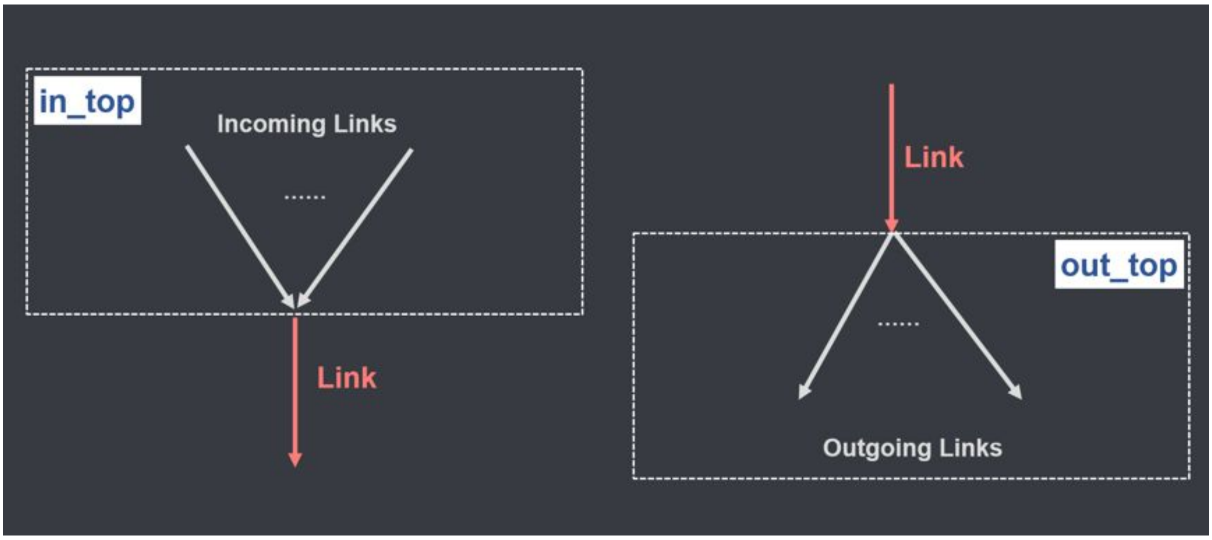


Figure 3. In_top and Out_top for a Road Link

Vehicles traveling from road intersections to highway tollgates have limited route options. For each intersection-tollgate pair, we selected only the most important one into Table 4. For example, Figure 4 illustrates the route with 9 consecutive road links from Intersection B to tollgate 1.

Table 4. Vehicle Routes from Intersections to Tollgates

| Field | Type | Description |
| --- | --- | --- |
| *intersection_id* | string | intersection ID |
| *tollgate_id* | string | tollgate ID |
| *link_seq* | string | a sequence of link IDs from the intersection to the tollgate separated by commas (shown in Figure 4) |

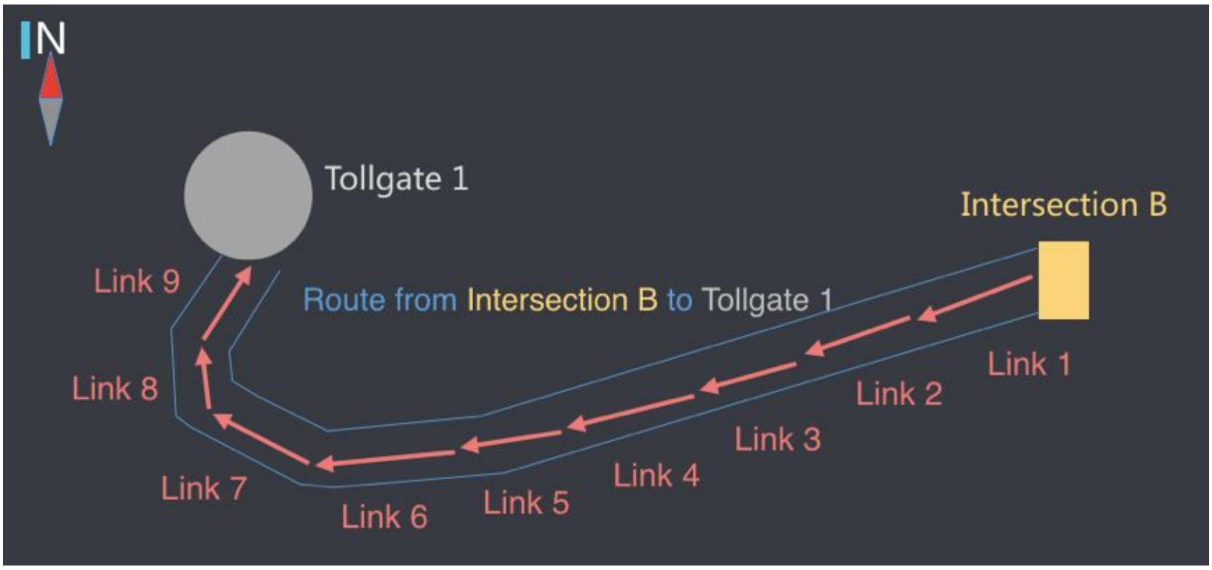


Figure 4. Link Sequence for the Route from Intersection B to Tollgate 1

Table 5 introduces the time-stamped records of actual vehicles along the routes from road intersections to highway tollgates.

Table 5. Vehicle Trajectories Along Routes

| Field | Type | Description |
|---|---|---|
| *intersection_id* | string | intersection ID |
| *tollgate_id* | string | tollgate ID |
| *vehicle_id* | string | vehicle ID |
| *starting_time* | datetime | time point when the vehicle enters the route |
| *travel_seq* | string | trajectory in the form of a sequence of link traces separated by ";", each trace consists of link id, enter time, and travel time in seconds, separated by "#" |
| *travel_time* | float | the total time (in seconds) that the vehicle takes to travel from the intersection to the tollgate |

Table 6. Traffic Volume through the Tollgates

| Field | Type | Description |
|---|---|---|
| *time* | datetime | the time when a vehicle passes the tollgate |
| *tollgate_id* | string | ID of the tollgate |
| *direction* | string | 0: entry, 1: exit |
| *vehicle_model* | int | this number ranges from 0 to 7, which indicates the capacity of the vehicle (bigger the higher) |
| *has_etc* | string | does the vehicle use ETC (Electronic Toll Collection) device? 0: No, 1: Yes |
| *vehicle_type* | string | vehicle type: 0-passenger vehicle, 1-cargo vehicle |

Table 7. Weather Data (every 3 hours) in the Target Area

| *Field* | *Type* | *Description* |
|---|---|---|
| *date* | date | date |
| *hour* | int | hour |
| *pressure* | float | air pressure (hPa: Hundred Pa) |
| *sea_pressure* | float | sea level pressure (hPa: Hundred Pa) |
| *wind_direction* | float | wind direction (°) |
| *wind_speed* | float | wind speed (m/s) |
| *temperature* | float | temperature (℃) |
| *rel_humidity* | float | relative humidity |
| *precipitation* | float | precipitation (mm) |

Table 3 and 4 are time-invariant. Therefore, they are only provided in the training set. Table 5, 6 and 7 are provided both in the training set and testing set according to the aforementioned description.

We also provide two sample python scripts, which can process tables 5 and 6 and generate results conforming to the structure of tables 1 and 2.

## KDD 2017