## 2 Objective

The objective of this project is twofold:

- To acquire a better understanding of data mining techiniques.
- Familiar with how to complete a given project. by discussing with groupmates, surveying literatures and coding experiments.

## 3 Background

This project is based on the KDD Cup 2017, one of the most famous competitions in data mining area. The more details can be accessed through KDD web and Alibaba announcement.

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.

## 4 Task Description

Here we introduce the major two tasks in details. Please note that **you are only required to chooes one of two tasks as your group task.**

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

For travel time prediction, the initial training set contains data gathered from July. 19th to Oct. 17th, where data structures have been shown in table 1. For every 20-minute time window, please estimate the average travel time of vehicles for a specific route (shown in Figure 1).

- Routes from Intersection A to Tollgates 2 & 3;

- Routes from Intersection B to Tollgates 1 & 3;
- 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).

Table 1: Travel Time from Intersections to Tollgates

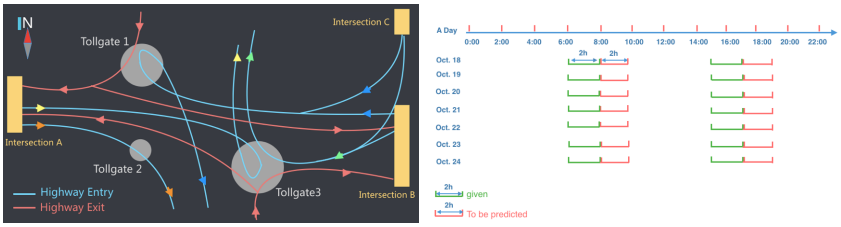
| **Attribute** | **Type** | **Description** |
|---|---|---|
| intersection_id | string | intersection ID |
| tollgate_id | int | tollgate ID |
| time_window | string | e.g., [2016-09-18 08:40:00, 2016-09-18 09:00:00] |
| avg_travel_time | float | average travel time (seconds) |

## 4.2 Task 2: To predict average tollgate traffic volume

For travel volume prediction, the initial training set contains data gathered from Sep. 19th to Oct. 17th, where data structures have been shown in table 2. For every 20-minute time window, please predict the entry and exit traffic volumes at tollgates 1, 2 and 3 (Figure 1). 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.

Table 2: Traffic Volume at Tollgates

| **Attribute** | **Type** | **Description** |
|---|---|---|
| tollgate_id | int | 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 |




(a) Road Network Topology of the Target Area (b) Time Windows for Traffic Prediction

Figure 1: Illustration on tasks.

### 4.3 Training and Test Data Set

The initial training set of travel time prediction contains data gathered from July. 19th to Oct. 17th, while the initial training set of traffic volume prediction contains data gathered from Sep. 19th to Oct. 17th.

For two tasks, you are require to predict specific rush hours from Oct. 18th to Oct. 24th. In the test 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.

### 4.4 Evaluation Metric

In this project, we mainly 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:

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

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

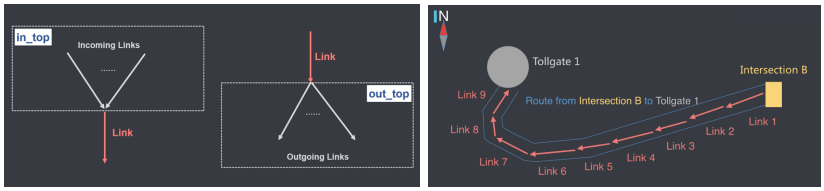
**Task 2:** Let $C$ be the number of tollgate-direction pairs (as aforementioned: 1-entry, 1-exit, 2-entry, 3-entry and 3-exit), T be the number of time windows in the testing period, and $f_{ct}$ and $p_{ct}$ be the actual and predicted traffic volume for a specific tollgate-direction pair $c$ during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

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

## 5 Additonal Data Description

The whole dataset can be downloaded in the link. Except for training and test data, you are provided with a list of extra data sources:

- The road network topology: (table 3) and (table 4)
- Vehicle trajectories (table 5)
- Historical traffic volume at tollgates (table 6)
- Weather data (table 7)




(a) In_top and Out_top for a Road Link (b) Link Sequence for the Route from Intersection B to Tollgate 1

Figure 2: Illustration on road networks.

## 5.1 Road Network

The road network (Figure 1) here used is a directed graph formed by interconnected road links (Figure 2). A route (Figure 2) 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 2 describe road links.

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 2 illustrates the route with 9 consecutive road links from Intersection B to tollgate 1.

Table 3: Road links attributes

| **Attribute** | **Type** | **Description** |
|---|---|---|
| link_id | int | link ID |
| length | float | length (meter) |
| width | float | length (meter) |
| lanes | int | number of lanes |
| in_top | string | incoming road links |
| out_top | string | outgoing road links |
| lane_width | float | lane width (meter) |

Table 4: Vehicle Routes from Intersections to Tollgates

| **Attribute** | **Type** | **Description** |
|---|---|---|
| intersection_id | string | intersection ID |
| tollgate_id | int | tollgate ID |
| link_seq | string | a sequence of link IDs from the intersection to the tollgate |

## 5.2 Additional Information

Table 5 introduces the time-stamped records of actual vehicles along the routes from road intersections to highway tollgates. Table 6 introduces the information

of vehicles through the tollgates. Table 7 introduces the weather conditions.

Table 5: Vehicle Trajectories Along Routes

| **Attribute** | **Type** | **Description** |
|---|---|---|
| intersection_id | string | intersection ID |
| tollgate_id | int | tollgate ID |
| vehicle_id | int | 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 seperated by ";", each trace consists of link ID, enter time, and travel time in seconds, separared by "#". |
| travel_time | float | the total time (in seconds) that the vehicle takes to travel from the intersections to the tollgate |

Table 6: Traffic Volume through the Tollgates

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

# 6 Project Guidelines

The objective of this project is to prompt you to learn how you complete a given project. In other words, what we expect from you is to learn how to solve a problem instead of a well-performed models.

Given any task or project, taking a good survey on literatures is one of the most important steps. You can learn from previous works and implement them to solve your own task. Hence we here provide some related paper:

- Wang, Zheng, Kun Fu, and Jieping Ye. "Learning to estimate the travel time." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

Furthermore, paying attention to some acedemic conferences will make it easy for you to solve problems in career. Some conferences related to data mining are listed as:

- **Data Mining:** SIGKDD, WWW
- **Database:** SIGMOD, VLDB, ICDE
- **Artificial Intelligence:** ICML, ICLR, NIPS