可利用的数据集：目标领域的公路网络拓扑（图1，3和4，表3和4），汽车轨迹（表5），收费站历史车流量（表6）；天气数据（表7）。比赛包括两个任务，详情如下：

任务1：由交叉口通过收费站预计需要的时间

对于每个20分钟的窗口，估计汽车在指定路径上的平均行驶时间（如图1所示）。——

a.交叉口A到收费站2与3的路径

b.交叉口B到收费站1与3的路径

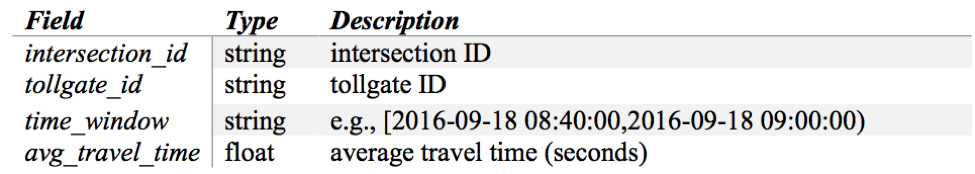
c.交叉口C到收费站1与3的路径

注意：对于一个给定的路径的20分钟窗口的ETA(估计达到时间）是所有汽车轨迹的平均行驶时间，在那个时间到达窗口。每个20分钟窗口定义为左闭右开区间，如 [2016-09-18 23:40:00, 2016-09-19 00:00:00).

提交格式（见表1）

表格中使用的数据类型为Int型，float,string,date和datetime。date和datetime的格式为“yyyy-MM-dd”和“yyyy-MM-dd HH:mm:ss”。时间窗口域包含两种时间类型，由逗号分隔，如 “2016-09-18 08:40:00,2016-09-18 09:00:00”。

表1 由交叉口到收费站的时间

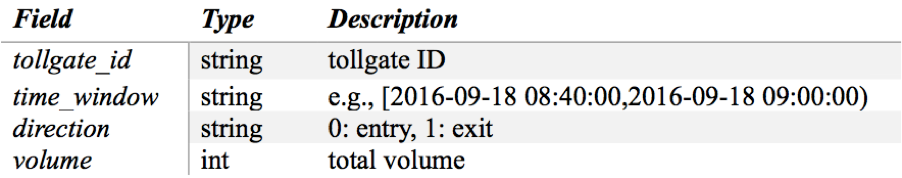


任务2：预测收费站平均车流量

对于每个20分钟窗口，请预测收费站1，2，3进入和驶出高速的车流量（图1和图2）。注意收费站2只允许进入高速，其它的都为双向（进入与驶出）。因此，我们需要预测5组数据。

提交格式（见表2）

表2收费站平均车流量



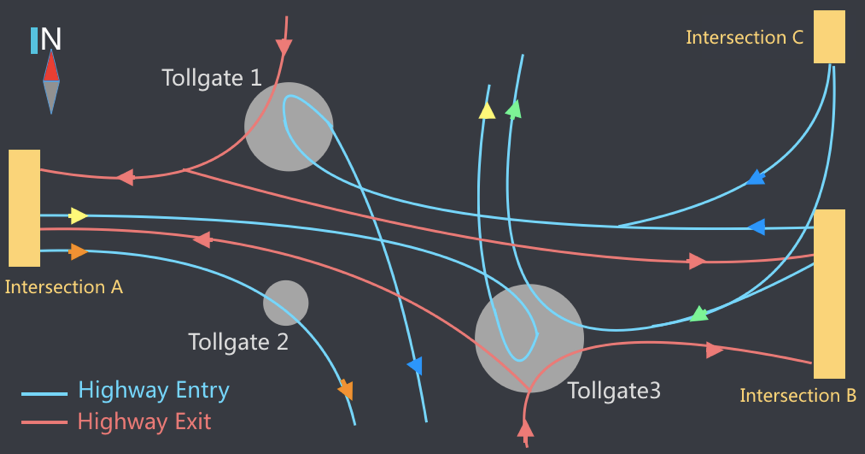


图1 交叉口区域的公路网络拓扑

方向对：标红为需要预测的通过时间

A-2

A-3

1-A

3-A

C-1

C-3

B-1

B-3

1-B

3-B

问题：是否堵塞只存在与收费站与交叉口之间？收费站是否一个时刻只能通过一个方向的车？是否图中红色的通行线会对蓝色的通行线造成影响？

训练和测试数据集：

比赛的最开始，交通预测指定的高峰时段为10月18号到10月24号，在5月25号，将会交换数据，需要参赛者预测时间段为10.25-10.31

参赛者必须预测的时间间隙标红色，由图2所示，如08:00 - 10:00 和 17:00 - 19:00，时间间隔为20分钟。

2小时6个时间段。1天12个时间段。7天共84个时间段需要预测

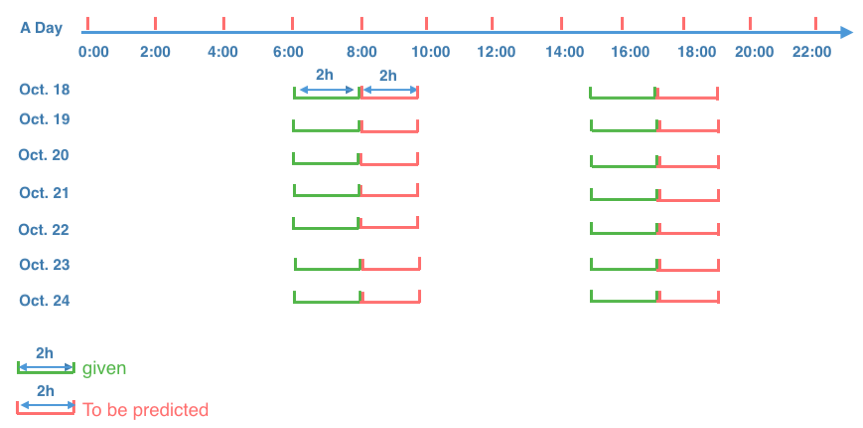


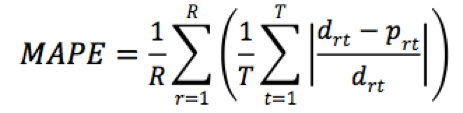
图2 交通预测的时间窗口

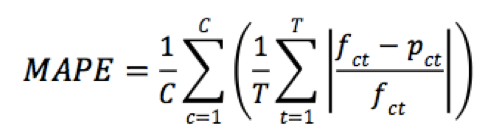
对于行驶时间预测，初始训练集包含数据集7.19-10.17.对于流量预测，初始训练集包含数据集9.19-10.17。在数据集交换后，添加数据集10.18-10.24到这两个预测任务中。在测试集中，提供了标绿色的时间段，如图2所示，如 06:00 - 08:00 和15:00 - 17:00。参赛者可以使用这些信息作指示指导。

注意：没有限制参赛者只使用前面的2个小时数据来进行预测。然而 ，每个预测必须使用时间窗口之前的数据。例如，参赛者不可以使用10月20号的数据来预测10月19号的交通情况。

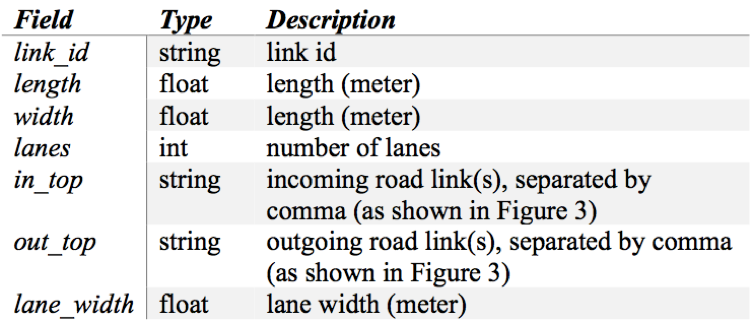
评价

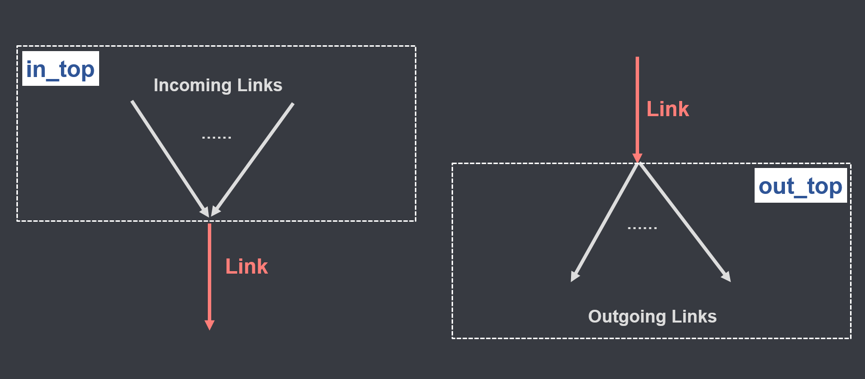
选择MAPE来评价结果

**Evaluation Metrics**  
We choose Mean Absolute Percentage Error (MAPE) to evaluate the result.   
**Task 1:**Let drt and prt be the 真实 and 预测 average travel time for 路径 r 在窗口时间t内. The MAPE for travel time prediction is defined as:  
  
R= 6and T=84 are the 数量 of 路径 and 数量 of 需要预测的时间窗口 in the testing period respectively.

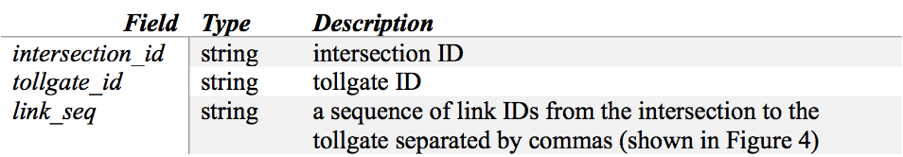
**Task 2:**Let C = 5 be the number of交叉口方向对(as aforementioned:形式 1-entry, 1-exit, 2-entry, 3-entry and 3-exit), T =84  be the number of时间窗口的数量 in the testing period, and fct and pct be the actual and predicted traffic volume for a specific tollgate-direction pair c during time window t. The MAPE for traffic volume prediction is defined as:  


**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  


  
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  


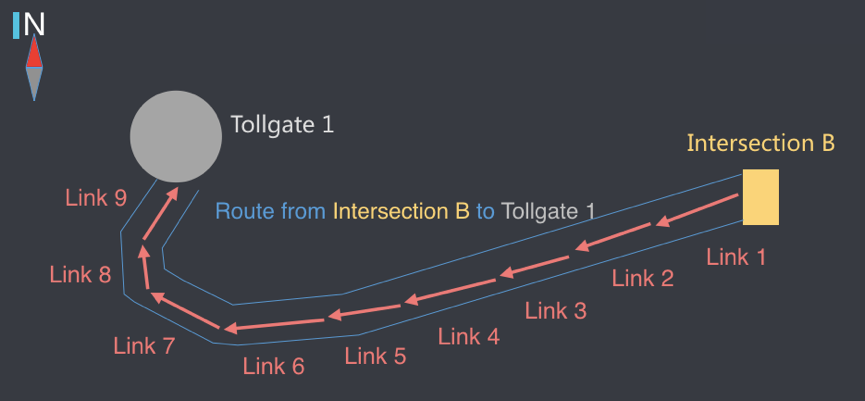
  
Figure 4. Link Sequence for the Route from Intersection B to Tollgate 1

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

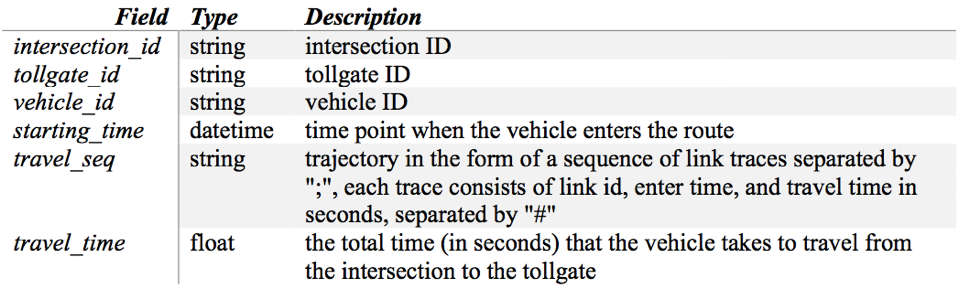
Table 5. Vehicle Trajectories Along Routes  


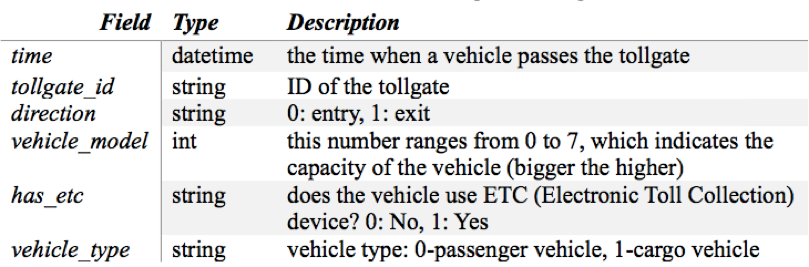
Table 6. Traffic Volume through the Tollgates  


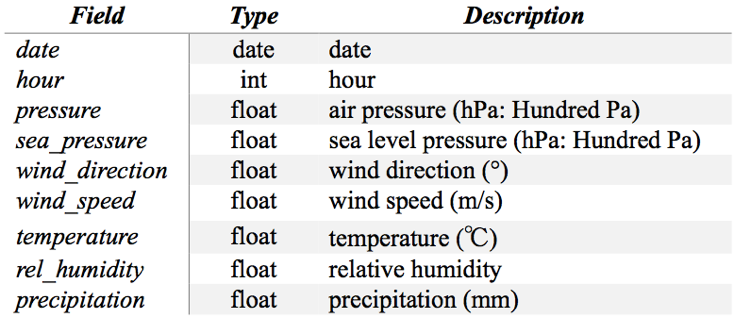
Table 7. Weather Data (every 3 hours) in the Target Area  


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**](http://www.kdd.org/kdd2017/)