**Traffic Volume Prediction at Highway Tollgates**

Mi Tian, Deepan, Dhaval

**Abstract.** This paper tries to answer the following question. How to accurately predict the number of vehicles passing a particular tollgate at a particular time interval? The problem is challenging for two reasons. First, the data provided is not clean, meaning it has outliers in terms of the data being collected has time period which was a national holiday. Second, the time series data provided in the training set is continuous. However, the test data does not contain data in a continuous manner which makes it difficult to apply traditional time series analysis and use it for prediction. We designed a supervised inference model utilizing pre-processed volume data together with link and trajectory data information. We evaluated the proposed approach using the test dataset provided to us on the MAPE metric.

**1 INTRODUCTION**

Modern urban transportation network heavily relies on highway systems. Since highway conditions need to be regularly maintained, tollgates have been widely deployed to collect supporting fund. During peak hours, huge congestions can form at key highway intersections, creating a main bottleneck of traffic. Long queues at highway tollgates not only upset drivers with longer wait time, but also cause large amount of gas consumption and air pollution. Traffic slowdown at highway entrances and exits has been one of the main challenges of managing highway system, and imposes overwhelming pressure on management authorities at a regular basis. Ideally, if the traffic volume at the tollgates can be predicated accurately ahead of time, the management authorities can then arrange preemptive solutions to mitigate the situation, such as opening more lanes or deploying temporary alternative toll collecting options to speed up the process.

Our project is therefore focused on the problem of predicting enter (vehicles entering highway system from local streets) and exit (vehicle exiting highway system back into local streets) traffic volume at any tollgate given any time interval. Predicting tollgate traffic is not a trivial problem, as the result traffic volume is under the influence of many different factors. There are many data resources at hand that can potentially help solve this problem. At a minimum, the traffic volume change at each gate over the temporal dimension can indicate the basic trend of different hours of the day. But the pattern may not be perfectly stationary, with holidays and weekends adding exceptions. Additionally, historical traffic trajectory data near tollgates (aka the time it takes for each car to travel through different road segments) indicates the throughput of the road network passing the tollgates, and can indirectly affect the volume. It can be combined with other factors such as weather and road parameters (width, number of lanes, etc.) to define the profile of the tollgate. Lastly, time of day, day of week, holiday, and so on can be treated as feature as well.

The methodology we took in this paper evolved through several different stages with improvement in the overall prediction accuracy. The eventual goal is to predict the traffic volume at different tollgates during morning and evening rush hours. First of all, we processed the raw temporal volume data collected at different tollgates to aggregate them into 20-minute bins. Basic statistical analysis plotting was then performed to reveal the volume change trend at a high level. The analysis result indicated that the data is stationary from day to day weekly, with an exception of a holiday week which featured abrupt volume increase/decrease. Based on these, we began by fitting the historical volume data into an Auto-Regressive (AR) model and an Auto-Regressive Moving-Average (ARMA) model. We applied the model using all historical data, versus only data of the rush hours and excluding the holiday week. We compared the results using Mean Absolute Percentage Error (MAPE). The result was improved of the latter and showed potential to improve.

Moreover, to try and further improve the accuracy of our predictions, we performed some feature engineering on the data to create dummy variables for the time series predictors. The reason we created these additional dummy variables was to try and see different non-time-series prediction models on our data. Here, we created a dummy variable for the “Day of the Week” from the “Date” attribute. Similarly, we created dummy variables for “Hour” and “Minutes”. We also added new features like the average time it took for a vehicle to reach from intersection A, B and C to the particular tollgate-direction pair. Once we had our data ready with the new set of features, we performed predictive modeling. We tried various models starting with Lasso Regression, Random Forest and SVM Regression. Moreover, we also performed Principle Component Analysis (PCA) to do dimensionality reduction and then applied Random Forest to just compare our results. Finally, we compared the results using Mean Absolute Percentage Error (MAPE). It is seen that different models perform better for different tollgate-direction pair. This observation seemed reasonable and obvious to us because every tollgate would have a different distribution of traffic at different times. Thus, having different models for different tollgate-direction pair makes sense to us.

**2 PROBLEM DEFINITION**

The highway network and dataset used in this paper is provided as part of the KDD CUP 2017 challenge [1]. As shown in Figure 1 below, the exemplary problem is focused on a highway network of three local network intersections and three tollgates. The blue lines indicate traffic entering the highway from local road network, while the red lines represent traffic exiting the highway back into local. Tollgate 1 and 3 have both entry and exist traffic, and tollgate 2 only serves as highway entrance. Tollgate 1 and 3 both have an entry route that joins traffic from two different local intersections. And tollgate 3 has one more route from intersection A. These indicate the tollgates’ profile will be very different and can affect the volume differently.

The problem of interest for this paper can be defined in the form of the output as shown in Table 1. For any given tollgate ID and any time window of *20 minutes*, the prediction algorithm needs to output the number of vehicles passing through that gate in that time period, for any applicable enter and exit direction.

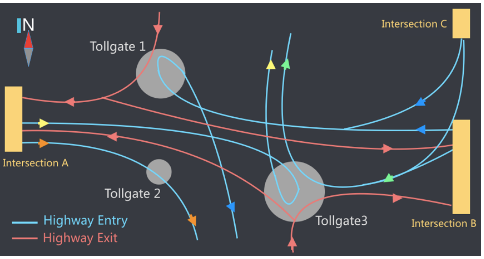
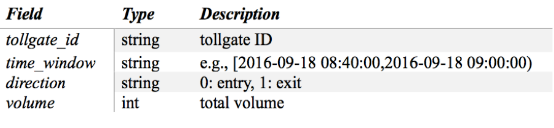


Figure 1. Highway network topology of the target area.

Table 1. Expected output of predicted traffic volume at the three tollgates



As shown in Table 2, the raw dataset contains numerous entry and exit record down to each passing vehicle. For each vehicle passing a tollgate, the time, tollgate ID, direction, vehicle model, type, and whether it uses Electronic Toll Collection (ETC) are provided. The training dataset contains data collected from September 19th to October 17th of 2016.

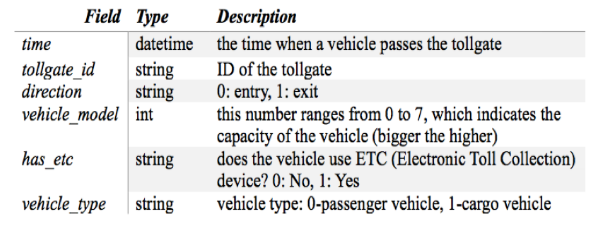
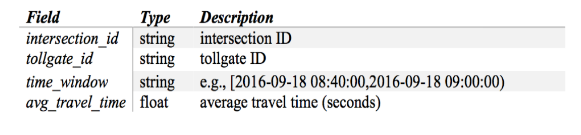
Table 2. The raw traffic data passing each tollgate

Table 3. The raw traffic data passing each tollgate



We have also incorporated additional features such as average travel time in our dataset, for each intersection-tollgate-direction pair. By doing this we are trying to specifically predict the number of vehicles arriving a particular tollgate, by considering the average time it takes for a vehicle to arrive at a tollgate from a particular intersection. As shown in Table 3, we have merged the *avg\_travel\_time* information on the basis of the *intersection\_id*, *travel\_id* and the respective *time\_window*.

As shown in Figure 2. The KDD CUP specifically asks to *predict the traffic during morning (8-10am) and evening (5-7pm) rush hours, for a week following the training dataset (October 18th to October 24th)*. The traffic data of the two hours prior to the target rush hour were also provided, which can affect the volume because they are temporally adjacent. In our analysis, we were able to use the models we developed to predict any given time in the future for any tollgate-direction pair, which also fulfills the goal set by the contest.

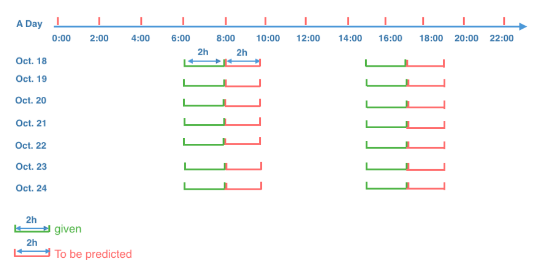


Figure 2. Target time windows for traffic volume prediction

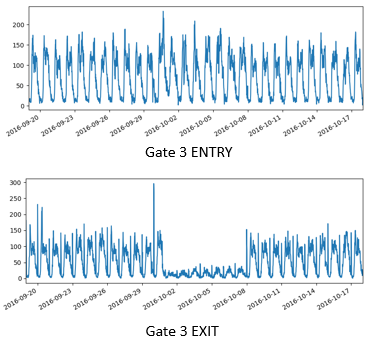
**3 METHODOLOGY**

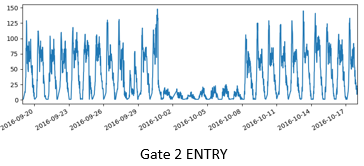
The steps of our methodology is presented in this section. We start off aggregating the tollgate vehicle pass data into 20-minute bins, followed by plotting the temporal volume change of all tollgate and direction combinations. From the trend we were able to discover a stationary pattern of the traffic volume at a daily basis, except for a holiday week which introduced large increase and decrease for all traffic of that week. Based on these discoveries we split the temporal data into a training set and a test set. As an initial attempt, we treated the prediction problem as a regression problem. The training data was used to train an AR and an ARMA model to predict the volume for the time range of the test set.

After coming up with the final dataset, we created a similar test dataset based on the approaches we mentioned. To come up with the prediction of the volume at different time of the day, we came up with various approaches starting with Lasso, because it performs feature selection for us from the large dimension of data we have. To further improve our accuracy we used Regression Tree approach, and to specifically solve the problem of variability in our data and overfitting problem, we extended our approach to Random Forest. Moreover, just to compare our results and identify which method works best we performed SVM Regression using linear and radial basis kernels.

3.1 Data Exploration

A Python script was written to aggregate the raw data into temporal volume datasets for each tollgate and direction combination in time slots of 20 minutes. The volume data was then plotted as shown in Figure 3. Two important patterns were discovered from the graph. First, the traffic volume change throughout a day is mostly similar of each day during a week. For example for “Gate 3 Entry”, each day of the first week started with a high peak of morning rush hour traffic, followed by a drop until another increase later in the evening rush hour.





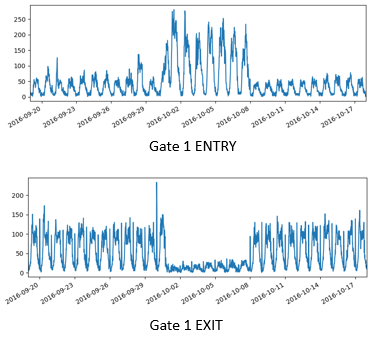


Figure 3. The traffic volume of different tollgate and direction combinations, from 09/20/2016 to 10/17/2016 (four weeks)

The day-to-day stationary pattern hinted that auto-regression can be attempted to fit the historical training data towards prediction of future volumes. In addition, the evening rush hour peak also looked higher on Fridays compared to other days of the week in a way dependent to the week of the month. This indicated that day-of-week and week-of-month could be treated as features in a more complex model.

The second important fact revealed by the graph is that the week from October 1st to October 7th shows day-to-day stationary pattern but at a vastly different magnitude. This is because the first week of October is the longest national holiday of China, and a massive population choose to travel out of the city for vacation. For example, “Gate 1 Entry” traffic volume for the week almost quadrupled that of the previous and following weeks. On the other hand, the opposite “Exit” direction had a large drop to only one fifth of the normal traffic. Interestingly, “Gate 3 Entry” traffic volume did not demonstrate obvious difference compared to other weeks. The reason may be that the highway entered from Gate 3 is not connecting to point of interests destinations for tourists. Overall, data from the holiday week should be treated specially. Since our target is to predict the volume of a normal week from October 18th to 24th, it is probably a good idea to completely filter out the holiday week data from the regression model.

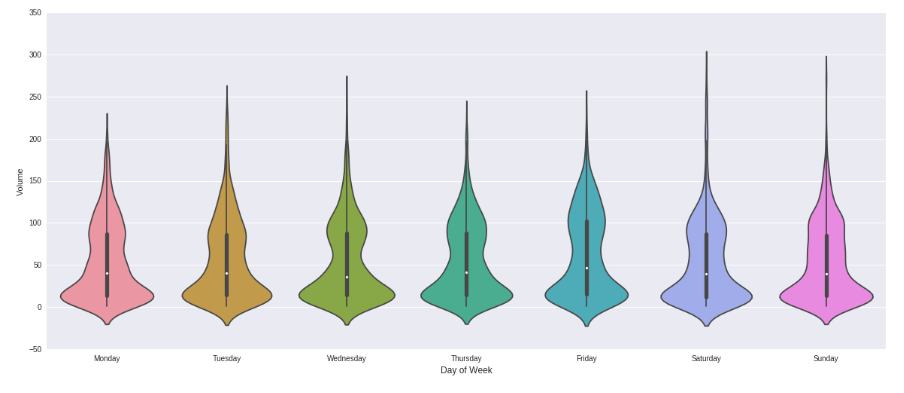


Figure 4. Distribution of traffic at different days of the week

As we can see from Figure 4, the distribution of traffic is roughly similar throughout the week. Also, the median traffic is of 45, which is stays more or less constant through different days of the week.

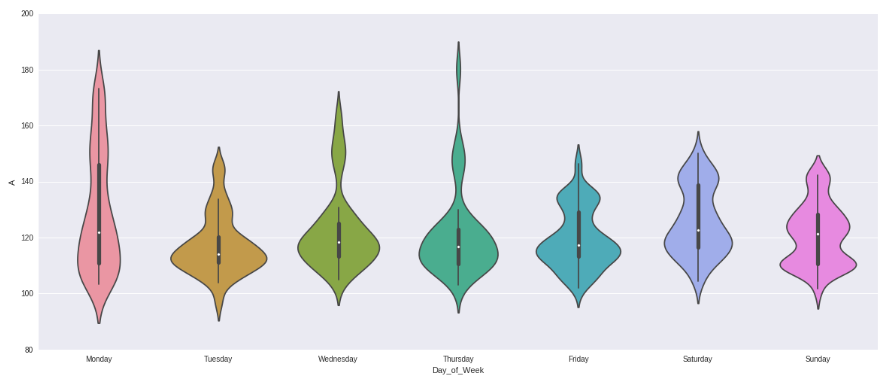


Figure 5. Distribution of average arrival time across different days of the week for vehicles coming from intersection A to Tollgate 3

From Figure 5, we can see that the distribution of average time from intersection A to tollgate 3 varies throughout the week, which led us to think that there must be some activity leading to different traffic times in the days. This was an interesting pattern that we observed in the data exploration.

3.2 AR and ARMA Regression

The temporal traffic volume plot showed repeated patterns from day to day indicating stationarity for regression model fitting. To further validate this assumption, we plotted the correlation several correlation plots using the *Python pandas statistics* *package* to highlight the linear relationship between lagged data points. Figure 6 below shows an example of “Gate 3 Entry”. The first “Lag Plot” graph showed the correlation between the volume data at *t* and *t-1* (i.e. lag = 1 indicates 20 minute earlier). It shows a strong correlation which hints that the traffic volume of the previous time slot should be consulted to predict the current.

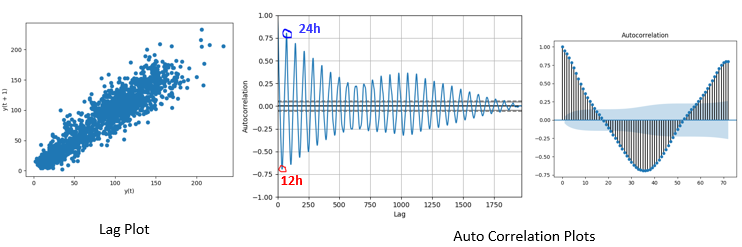
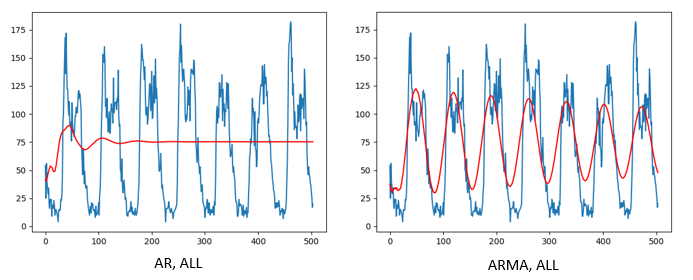
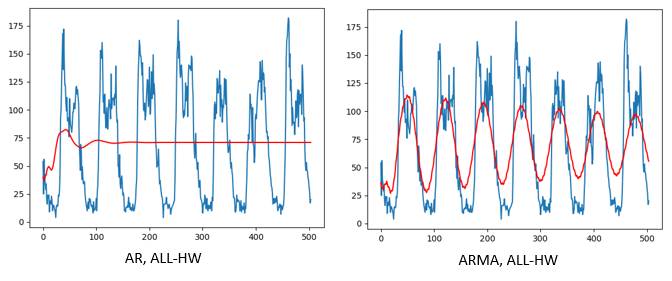


Figure 6. Different correlation plots indicating the relationship between traffic volumes at different time points

The next two auto-correlation plots revealed more temporal correlation with different lag values. In both plots, the x-axis corresponds to the lag value. Specifically, lag = 72 means 72 of 20-minute lags, which is exactly the data point of a day ago. From the second plot, it can be observed that the correlation is strongly positive at daily intervals, and negative at half a day. The last plot shows a zoomed in view of the correlation trend in a 24-hour period.

Based on these correlation pattern, it is promising to predict by fitting the training volume data into linear regression models. We tried AR and ARMA using the *Python sklearn package.* Given training data, they are capable of trying different lag values, and pick the one with the best correlation outcome. We separated out the last week’s dataset (October 11th to 17th) as test data to verify the accuracy of the prediction, and used the rest three weeks (September 19th to October 10th) as the training dataset.





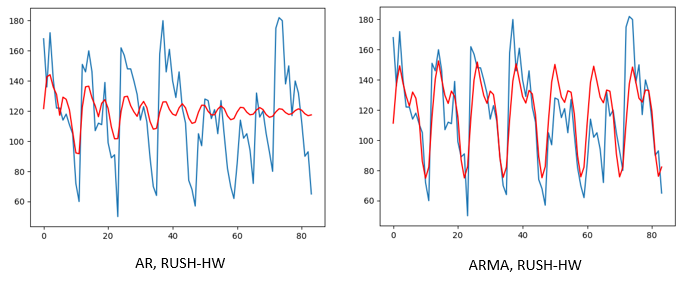


Figure 7. Comparison of real (blue) versus predicted (red) traffic volume for Gate 3 Entry direction, using three differently filtered training datasets

Considering the special traffic volume during holiday week and rush hour, we performed three different AR and ARMA model fitting, using all training data (ALL), all training data minus holiday week (ALL-HW), and rush-hour training data minus holiday week (RUSH-HW). Figure 7 shows the comparison of the real volume data (blue curves) versus predicted volumes (red curves) for each of the three methods, using “Gate 3 Entry” as an example. To quantify the prediction error, we calculated Mean Absolute Percentage Error (MAPE). The results will be presented in Section 4 together with other methods attempted in the next few sections.

* 1. Random Forest

As our aim is to predict the volume of traffic arriving at a particular tollgate for a given time we wanted to perform regression and had decided to use regression tree approach to solve the problem. Tree Ensembles have different advantages over Linear Regression. One main advantage is that they do not expect linear features or even features that interact linearly. Linear Regression can hardly handle categorical (binary) features. Tree Ensembles, because they are nothing more than a bunch of Decision Trees combined, can handle this very well. The other main advantage is that, because of how they are constructed (using bagging or boosting) these algorithms handle very well high dimensional spaces as well as large number of training examples. We have also taken a different approach in solving this problem when applying Random Forest, here because there is an abrupt change in volume in the holiday period; namely October 1st – October 7th, we have removed these data points treating them as outliers for our model. We have used the entire training dataset that we created by doing feature engineering and to evaluate our model, we used test dataset provided in the competition. We used cross validation to tune the hyper parameters. To quantify the prediction error, we calculated Mean Absolute Percentage Error (MAPE). The results for different tollgate-direction combinations can be seen from Figure 8.

Tollgate-1 Entry

Tollgate-1 Exit

Tollgate-2 Entry

Tollgate-3 Entry

Tollgate-3 Exit

Figure 7. Prediction results for different tollgate-direction pairs using Random Forest.

* 1. SVM Regression

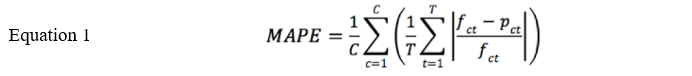
In SVM the basic idea is to map the data x into a high-dimensional feature space F via a nonlinear mapping, and to do linear regression in this space. We used cross validation to tune the hyper parameters like epsilon and C. Also, it makes sense to use different kernels like polynomial, linear, radial and Gaussian based on the distribution of traffic in different scenarios. Again, to quantify the prediction error, we calculated Mean Absolute Percentage Error (MAPE).

* 1. PCA + Random Forest:

As our data was of high dimension, we tried performing Principle Component Analysis (PCA) to do dimensionality reduction and try observing whether there is any discernible pattern in the lower feature space. Although, after performing dimensionality reduction, 34 principle components covered around 98.2% of the variance in the data. Also, we performed Random Forest by using 34 principle components as our features in the lower feature space. For the evaluation, we used cross validation method on training set to validate our model and we got the lowest MAPE that was 0.12. But, there was a problem when we tried testing the model on testing set. The test dataset consisted of data for selected time intervals and PCA requires same number of columns as in training set. Thus, we cannot perform PCA because the variance of some of the features would be zero in the test set. But, if test data set consist of proper distribution or some data for each time interval, our claim is that this model will outperform all other model results. As KDD organization will publish the new test data latter, we can try our model on that and try to get the lowest test error.

**4 EVALUATION**

The metric used to evaluate the prediction accuracy is the Mean Absolute Percentage Error (MAPE) as shown as Equation 1. In the equation, *f* refers to the fact value and *p* stands for prediction, *C* is the number of tollgate-direction pairs, and *T* is the number of time windows).



As mentioned in the last section, we separated out the last week’s dataset and used the rest three weeks as training data. Using the fitted model we were able to predict the volumes of the last week, and compare it to the real data to verify the accuracy of the prediction. The results are summarized in Table 4 below. From the results we can see that excluding the holiday week and focus on the rush hour data both significantly increased the accuracy of prediction.

Also, comparing our results after applying Random Forest and SVM Regression, with different hyper tuning parameters selected based on cross validation on individual tollgate-direction pair we can see that coming up with different models for different situation is a better idea. For example, for tollgate 2-entry we can see that the MAPE value is comparatively one of the best with 0.154 for random forest. We found out that after performing SVM regression works good when choosing linear kernel for our problem, however it is not an improvement over random forest because of the obvious advantages random forest has of bagging approach and minimizing variability among the data over SVM.

Table 4. The MAPE result of our own evaluation

|  |  |
| --- | --- |
| **Method** | **MAPE** |
| AR (ALL) | 1.703 |
| ARMA (ALL) | 1.045 |
| AR (ALL-HW) | 1.610 |
| ARMA (ALL-HW) | 0.761 |
| AR (RUSH-HW) | 0.228 |
| ARMA (RUSH-HW) | 0.191 |
| RF (T1-Entry) | 0.330 |
| RF (T1-Exit) | 0.210 |
| RF (T2-Entry) | 0.154 |
| RF (T3-Entry) | 0.176 |
| RF (T3-Exit) | 0.263 |
| SVM-Regression | 0.210 |
| PCA-RF | 0.121 |

In addition to our own evaluation, the KDD CUP has also provided training and testing dataset of seven days between 6AM and 8AM and 3PM and 5PM. The prediction target interval is two hours after the provided time range. Although the real data during the rush hours (8-10AM for morning and 5-7PM for evening) will not be released till May 25, we were able to upload our predicted values to the website, and receive a calculated result of MAPE. In this case, we were able to use the full four weeks of given volume dataset as the training data to further increase the accuracy of our prediction. We picked some of the better methods to upload to the contest website, and Table 4 below shows the result reported by the contest server based on calculation using the unannounced test data.

Table 4. The MAPE result of KDD CUP’s evaluation

|  |  |
| --- | --- |
| **Method** | **MAPE** |
| **ARMA (RUSH-HW)** | 0.2647 |

*(Mi) I will upload the results on the website tonight, I should get the values tomorrow.*

1. **CONCLUSION**

The approach we used aims to predict the volume of traffic at a particular place in a twenty minute time interval by using the trajectory and other temporal information. We believe the proposed framework is general enough to be applied to the inference and prediction of volume at different places other than tollgates. For example, it can be applied to the monitoring of traffic flows at arbitrary locations and time. Several reasons lead to the success of the proposed model. First, the use of different random forest and other machine learning techniques at different tollgates, lead to better predictions at different locations. Finally, the hyper-tuning parameters were selected for different models on the basis of cross validation leading us to find the best combination of parameters for our model. In the future, we will focus on improving the accuracy of our prediction through gradient boosting algorithms. Moreover, we will seek for more applications of our model, in particular in the area of traffic monitoring and business customer frequency in urban areas.

**REFERENCES**

[1] KDD CUP 2017, “Highway Tollgate Traffic Flow Prediction – Travel Time & Traffic Volume Prediction”, Alibaba Cloud, http://www.kdd.org/kdd2017/News/view/announcing-kdd-cup-2017-highway-tollgates-traffic-flow-prediction

[2] Toto, Ermal, Elke A. Rundensteiner, Yanhua Li, Richard Jordan, Mariya Ishutkina, Kajal Claypool, Jun Luo, and Fan Zhang. "PULSE: A Real Time System for Crowd Flow Prediction at Metropolitan Subway Stations." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 112-128. Springer International Publishing, 2016.

[3] Bengio, Yoshua. "Gradient-based optimization of hyperparameters." *Neural computation* 12, no. 8 (2000): 1889-1900.

[4] L. Breiman. Bagging predictors. Machine Learning, 24 (2):123–140, 1996.

[5] Aizerman M.A., Braverman É.M., and Rozonoér L.I. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25: 821-837.