Designing a mobility model using large-scale data

Leah Kazenmayer[2†], Gabriela Ford[1*], Jiechao Zhang[1*], Rezaur Rahman[1*], Furkan Cimen[1*], Damla Turgut[1‡] and Samiul Hasan[1‡]

[†]kazenml1@tcnj.edu; [*]{gabrielaford1512, zlyjsl123, rezaur.rahman, furkan}@knights.ucf.edu; [‡]{damla.turgut, samiul.hasan}@ucf.edu

[1] University of Central Florida, Orlando, FL, USA [2] The College of New Jersey, Ewing, NJ, USA

Abstract

Being able to predict short-term traffic volume is important for improving transportation systems planning and controlling the overall efficiency of traffic-networks. Traffic prediction can be done by using real time data collected from Internet of Things (IoT) devices. Specifically, the Automated Traffic Signal Performance Measures (ATSPM) collects data containing real-time and historical performance at signalized intersections and reveals spatio-temporal patterns of traffic volume for each signal. For this study, the ATSPM datasets are processed through machine learning algorithms in order to predict future hourly traffic. The hourly predictions are done based on the previous 6 hours volume seen at the selected intersections located in Seminole County, Orlando. Also, factors that play an important role on hourly traffic volume fluctuation were included: peak hour, day of the week, holidays, among others. Multiple machine learning models were applied to the data set to see which one performed the best. Random Forest, XGBoost and LSTM models show the best performance in predicting hourly volumes.

Introduction

Future traffic prediction has shown to be beneficial for:

- Adapting city planning strategies and manage traffic.
- Knowing how the volume of cars changes through time can help to avoid congestions and delays in a particular corridor and/or intersection.

Multiple techniques have been used for traffic flow forecasting, such as statistical and machine learning methods [1]. Machine learning techniques have become more popular due to their demonstrated ability to capture sharp discontinuities in traffic flows using nonlinear functions (e.g., tanh, sigmoid etc...) [2] [3].

Methodology

Accessed raw data from Automated Traffic Signal Performance Measures (ATSPM) database, devices added to traffic signal infrastructures collecting real time data, in Seminole County, Orlando, FL for years 2016 and 2019.

- Corridor of 10 signals was chosen for analysis: SR-426 (Figure 1).

GEH statistic as an evaluation metric was used to thoroughly clean the data (TfL, 2010). Hourly volume of cars was calculated for the through/straight direction at each intersection, for either "upstream" or "downstream" movements.

Factors affecting the fluctuation of volume of cars at every hour were chosen (Figure 2):

- Day of the week
- Peak hour
- Holidays
- Hurricane
- Precipitation

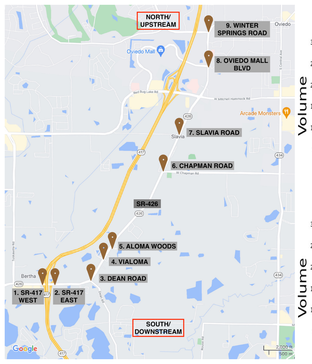

Figure 1. SR-426 Corridor (Google maps 2021)

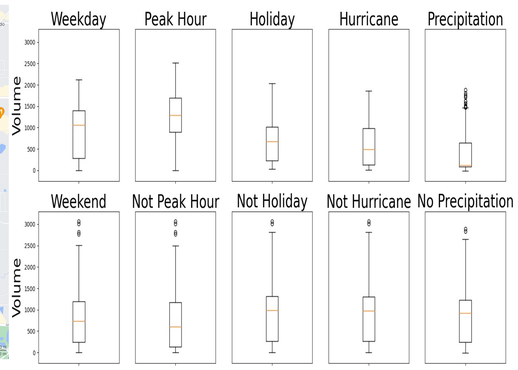

Figure 2. Factors affecting traffic for Signal 1 Downstream

Methodology (Cont.)

The models designed to predict the next hour of volume of traffic based on the past 6 hours of data and its factors were:

- Linear Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbor
- Extreme Gradient Boost (XGBoost)
- Long Short Term Memory (LSTM).

Results

Machine and deep learning models were built for upstream and downstream traffic flow through the corridor. The results of each model were compared using statistics:

- Root Mean Squared Error (RMSE): measures the difference between the predicted values and the actual values for the seventh hour prediction (The lower, the better the model) (Figure 3).
- Coefficient of determination (R2): the proportion of the variance in the dependent variable that is predictable from the independent variable. (Table 1).

Figure 3. Comparison of RMSE values for each model built: upstream (top) and downstream (bottom).

Table 1. R2 Statistic for models applied at every signal: mean ± one standard deviation & (max)

R2 values from applied models		
Model	**R^2**	
	Upstream	**Downstream**
Random Forest	0.974 ± 0.009 (0.981)	0.977 ± 0.006 (0.984)
XGBoost	0.972 ± 0.009 (0.981)	0.975 ± 0.007 (0.983)
LSTM	0.969 ± 0.009 (0.977)	0.970 ± 0.008 (0.978)
Decision Tree	0.965 ± 0.009 (0.973)	0.965 ± 0.01 (0.976)
KNN	0.959 ± 0.008 (0.966)	0.964 ± 0.007 (0.971)
Multiple Linear Regression	0.922 ± 0.01 (0.936)	0.893 ± 0.02 (0.919)

Results (Cont.)

Actual vs Prediction graphs were developed for the best performing models-Random Forest, XGBoost and LSTM-to understand the deviation from the prediction to actual values from the dataset (Figure 4).

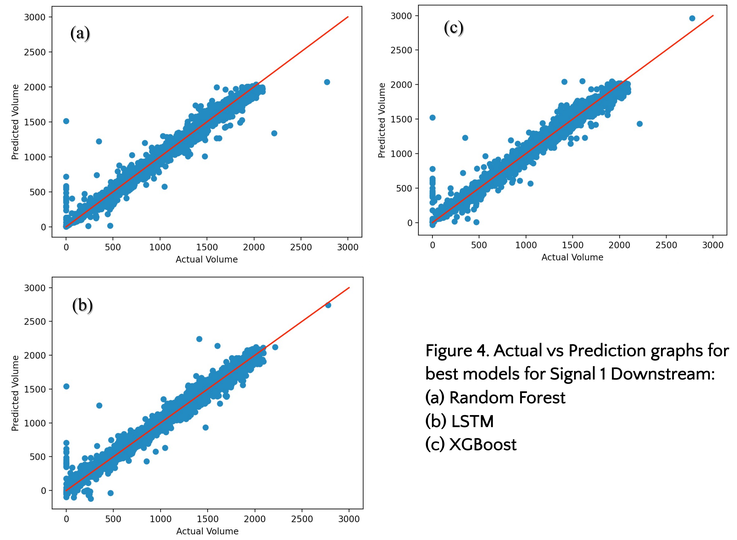

Figure 4. Actual vs Prediction graphs for best models for Signal 1 Downstream:
(a) Random Forest
(b) LSTM
(c) XGBoost

Conclusion

The machine and deep learning algorithms that demonstrated to most successfully predict traffic: Random Forest, XGBoost and LSTM.

- Machine learning algorithms with the best performance are composed of multiple distinct decision trees that allow for better predictions to be made.
- Deep learning algorithm, LSTM, uses short term memory and understands time-dependent patterns which allows it to adapt to traffic patterns.

Prediction of traffic volume at intersections is beneficial for:

- Improving traffic flow by either suggesting possible alternate routes or improving the signal's efficiency in an overflowing intersection.
- Improve quality of life for the population by reducing commuting time and reducing the number of cars on the road to potentially reduce emissions.

References

[1] L. Breimanet al., "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," Statistical science, vol. 16, no. 3, pp. 199–231, 2001.
[2] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," Transportation Research Part C: Emerging Technologies, vol. 79, pp. 1–17, 2017.
[3] B. D. Ripley, Pattern recognition and neural networks. Cambridge university press, 2007.

Acknowledgements

The support for this work was provided by the National Science Foundation REU program under Award No. 1852002. Any opinions, findings, and conclusions and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.