Transportation Research Record

MODELING AND PREDICTING THE CASCADING EFFECTS OF DELAY IN TRANSIT SYSTEMS

--Manuscript Draft--

Full Title:	MODELING AND PREDICTING THE CASCADING EFFECTS OF DELAY IN TRANSIT SYSTEMS
Manuscript Number:	19-04994R1
Article Type:	Presentation Only
Order of Authors:	Aparna Oruganti
	Sanchita Basak
	Fangzhou Sun
	Hiba Baroud
	Abhishek Dubey

MODELING AND PREDICTING THE CASCADING EFFECTS OF DELAY IN TRANSIT SYSTEMS

Aparna Oruganti
aparna.oruganti@vanderbilt.edu
Sanchita Basak
sanchita.basak@vanderbilt.edu
Fangzhou Sun, Ph.D.
fzsun316@gmail.com
Hiba Baroud, Ph.D.
hiba.baroud@vanderbilt.edu
Abhishek Dubey, Ph.D.
abhishek.dubey@vanderbilt.edu

Word Count: 1531 words + 2 figures× 250 + 2 tables× 250 = 2531 words

Submission Date: November 15, 2018

PAPER NUMBER
19-04994

INTRODUCTION

An effective real-time estimation of the travel time for vehicles, using AVL(Automatic Vehicle Locators) has added a new dimension to the smart city planning. In this paper, we used data collected over several months from a transit agency and show how this data can be potentially used to learn patterns of travel time during specially planned events like NFL (National Football League) games and music award ceremonies. The impact of NFL games along with consideration of other factors like weather, traffic condition, distance is discussed with their relative importance to the prediction of travel time. Statistical learning models are used to predict travel time and subsequently assess the cascading effects of delay. The model performance is determined based on its predictive accuracy according to the out-of-sample error. In addition, the models help identify the most significant variables that influence the delay in the transit system. In order to compare the actual and predicted travel time for days having special events, heat maps are generated showing the delay impacts in different time windows between two timepoint-segments in comparison to a non-game day.

The existing literature, (1), (2), (3), (4) talks about real-time traffic delay predictions using only traffic information and Nookala (5) studies its dependence only on weather conditions. However, in this paper, we use the data collected over several months from Nashville transit system, the real-time traffic and weather feed, occurrence of special events like NFL(National Football League), NHL (National Hockey League) and Vanderbilt basket ball games and study the cascading effects of delay in transit network by predicting the bus delay at each time point capturing multidimensional aspects in the feature space including traffic measurement quantities, time, detailed weather conditions as well as response of people towards an event. Towards this goal we develop two predictive machine learning models used for analyzing the data to make predictions on transit travel time for all the relatively busier bus routes in the network. Hence, this paper focuses on identifying the model with the best predictive accuracy to be used in DelayRadar (this architecture was proposed in (6)). According to the study results, we are able to explain more than 80% of the variance in the bus travel time and we can make future travel predictions for each timepoint segment with an out-of-sample error of 2.0 minutes with information on bus schedule, traffic, weather and the special events. To the best of our knowledge, there are very few studies (7) that included impacts of special events in predicting traffic delays focusing only on the adjacent arterial of the event location, while in this work we considered all the route segments covering multiple bus trips. We also present the cascading effect of delay in bus network before and after a special event using the Nashville transit system as a case study, showing how far the delay propagates from the actual event location through spatial heat maps.

METHODOLOGY

The real-time traffic data is collected and stored continuously in our database using HERE API and weather data for the city is collected from DarkSky API. We have collaborated with Nashville Metropolitan Transit Authority(MTA) for accessing the static and real-time bus transit data for Nashville. We only excluded the bus routes with just two or three trips on weekdays and having no trips on weekends. To explore the cascading effects of delays in transit system during events, we collected the game data for Nashville manually.

In this paper, we consider two ensemble tree based models, Random Forest and gradient boosted trees to train the data. The dataset is divided into train and validation set. The validation set consists of data for the NFL game on '2016-10-16'. The model is trained using 10-fold cross validation to reduce the model bias towards the in-sample data.

Ensemble Methods : Ensemble methods are techniques that combine many models to get better prediction accuracy (8).

Decision Tree : Decision tree is a regression or classification model in the form of a tree like structure.

Bagging : Bagging (9) is used in statistics to generate confidence values and confidence intervals of estimates and understand the variation due to a particular realization of the dataset.

Random Forest : Random Forest (RF) is an improved ensemble machine learning algorithm (10). The tuning parameters for the random forests are number of trees and m_{try}, the no of predictors to be considered at each split. Breiman (10) suggests three possible values of m_{try} ($1/3p$, $\sqrt{p}$,$2\sqrt{p}$). He recommends using $1/3p$ for regression and $\sqrt{p}$ for classification.
We use 10-fold cross validation to train the random forest model using 200 trees and m_{try} = 8.

Gradient Boost Method (GBM) : GBM is also an ensemble method used for regression and classification. The commonly used residuals for regression is Mean Square Error (MSE) expressed by

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2 \tag{1}$$

where $\hat{y}_i$ are the predictions of the travel time and can be estimated initially as function:
f(traffic,weather,events,busdata).

These predictions of the individual trees are then eventually added up i.e. $F(x) = f_{tree1}(x) + f_{tree2}(x) + f_{tree3}(x) + \ldots$

To avoid the overfitting problem a regularization term θ is added. Finally the objective function that minimizes the error in predictions is represented as:

$$obj = L(\theta) + \Omega(fx) \tag{2}$$

We chose XGBoost (extreme gradient boosting) for training our models as it uses sparse matrices with sparsity aware algorithms, improved data structures for better processor cache utilization which makes it faster and better support for multi-core processing reducing overall training time. These enhancements make a big difference in speed and memory utilization.

FINDINGS

The model learns at different rate from each input variable to predict the response variable ($Actual\_Travel\_Time$). Through a thorough analysis on the importance of each variable, the most important predictors in this case are found to be as length of the route segment, free flow, traffic speed, jam factor, free flow, traffic speed, hour of the day and the distances from each timepoint to the game location. Other input variables that are important in predicting the response are pressure, visibility and wind speed. Although the main assumption for this study is that events like NFL games affect the travel time, but the categorical variable 'Game' considered for whether it is a game day or not is not found as an important feature. This is because the data is skewed towards non game days as compared to game days. As such, having a binary variable to represent the game day did not add value to the prediction performance.

To understand the performance of the model it is important to evaluate the prediction accuracy and the goodness of fit of the model. We considered two metrics Root Mean Square Error (RMSE) and R^2. RMSE is calculated based on 10-fold cross validation which is the average of the out of sample RMSE for each fold using the formula in Eq. 3

$$RMSE = \frac{1}{k}\sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{3}$$

y_i = Actual travel time between two consecutive time point
$\hat{y}_i$ = Predicted travel time between those two time points
n = Total number of the observations in the dataset
k = Number of cross-validation folds

TABLE 1 Summary of the Models with their Goodness of Fit (R^2) and Prediction Accuracy ($RMSE$)

S.No	Model	R^2	RMSE	Time (*min*)
1	Random Forest	0.78	2.7	265
2	Extreme gradient boost	0.80	2.01	13.13
3	General Additive Models	0.50	2.94	44
4	Linear Regression	0.46	3.05	0.10

Predictive Analysis

By examining the summary of results in Table 1, we notice that XGBoost performs the best in terms of goodness of fit and predictive accuracy. The model explains 80% of the variance in the travel time which is the highest value of R^2 in our experiments. The modeling approach performs well with new data points and provides an average error of approximately 2 minutes in travel time prediction when tested using validation data. We finally applied the XGBoost model on the dataset and validated it on specific NFL games on days('2016-11-23', '2016-12-11', '2017-01-26' and 2 non game days('2016-12-18', '2017-01-13') to assess the model's ability to predict new data. For validation purposes, the game date is chosen as '2016-10-16'.

The predictions for before-game trips performed slightly better in terms of the predictive accuracy, the values of which are shown in Table 2, where the overall difference in RMSE is about one minute as compared to after-game trips. Also from Table 2 we see that the error in prediction is higher for the 0-1 hour time window for both before and after games and the error decreases gradually thereafter. This can be attributed to the fact that there is more congestion during 0-1 hour time window resulting in higher differences between the predicted and the actual travel time.

Analysis of the Cascading Effects of Delay

To study the bus delay patterns we quantify the delay on game day as compared to a non-game day we evaluate the impact of football games on bus delay using eq. 4, where D_{PI} denotes the predicted delay impact, TT_{PG} is the predicted travel time on a game day, TT_S represents the scheduled travel

TABLE 2 RMSE for the predicted values on Dec.11,2016 for each one hour time window before and after the game

S.No	Before/After	one hour time windows before the game	RMSE
1	Before	0-1	1.92
		1-2	1.31
		2-3	1.59
		3-4	1.42
2	After	0-1	2.94
		1-2	2.81
		2-3	2.44

time and TT_{NG} is the actual travel time on a non game day. The results in Figure 1 clearly show the cascading effects of bus delay that are associated with football games.

$$D_{PI} = max(avg(\frac{TT_{PG} - TT_S}{TT_S}) - avg(\frac{TT_{NG} - TT_S}{TT_S}), 0) \tag{4}$$

We collected the data of four National Football League (NFL) games at downtown Nashville between Oct. 10 2016 and Feb. 28 2017. We divided the period before/after football games into four one hour time windows and compared the average bus delay in the time windows for days having a football game vs. the days having no game.

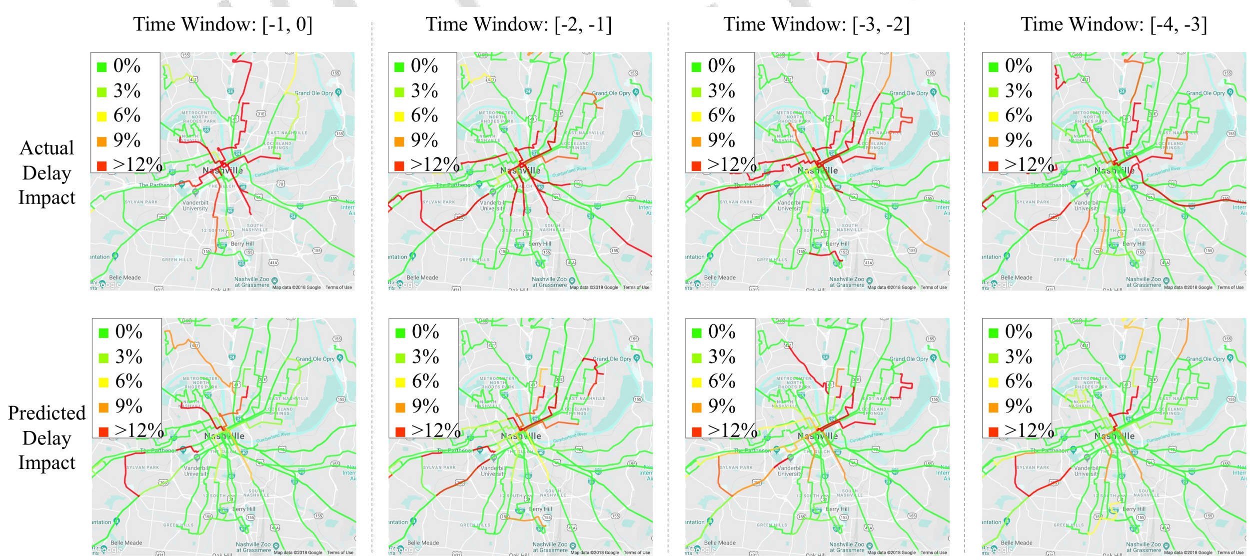

FIGURE 1 Predicted impact of football games on traffic congestion in four one-hour time windows before football games: (a) from 4 hours to 3 hours, (b) from 3 hours to 2 hours, (c) from 2 hours to 1 hour, (d) from 1 hour to 0 hour.

The results are shown in Figures 1 and 2 using heat maps. The actual values and predicted results from the XGBoost model are visualized as shown in Figures 1, 2. It indicates that the predicted results were able to capture the majority of the road segments in the bus network with

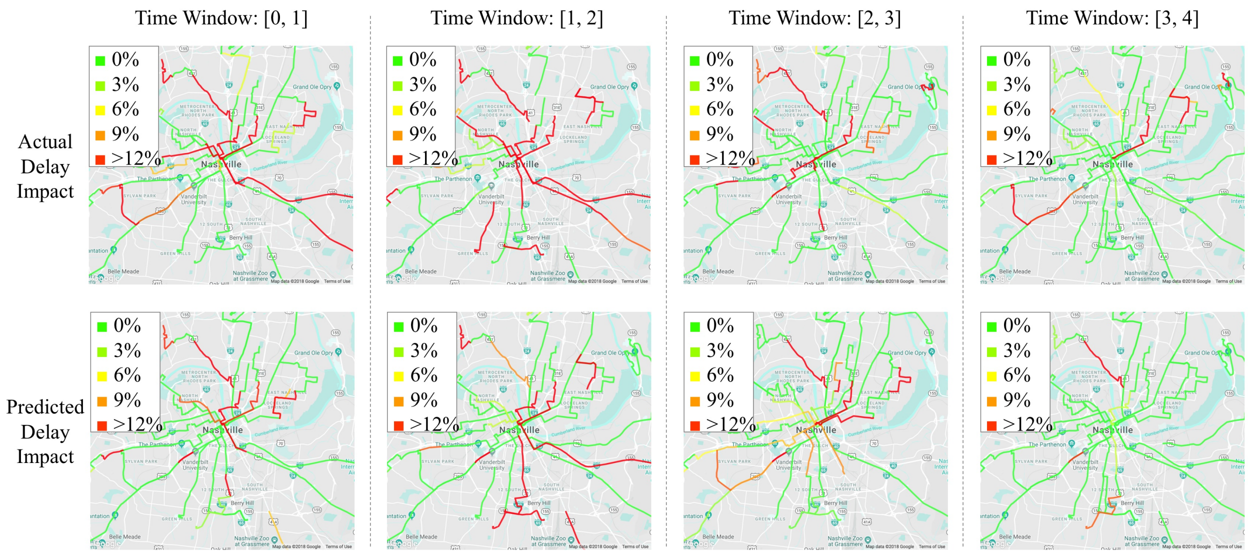

FIGURE 2 Predicted impact of football games on traffic congestion in four one-hour time windows after football games: (a) from 0 hour to 1 hour, (b) from 1 hour to 2 hours, (c) from 2 hours to 3 hours, (d) from 3 hours to 4 hours.

medium($> 6\%$ and $< 12\%$) and high delay impact($> 12\%$) between 0-3 hours before and after the game. However, the model could not capture the high delay impacts(i.e. $> 12\%$) in the time window 3-4 hours before and after the games accurately. The high delay impacts in the time window 3-4 hours might be caused due to unforeseen circumstances like bus break down, huge accident causing traffic congestion which are not considered in our current analysis. From this we can infer that bus traffic delays are affected between 0-3 hours before and after the games which changes our initial hypothesis a bit that bus traffic delays occur between 0-4 hours before and after games.

CONCLUSION

In this paper, the cascading effects of delay in transit systems are studied. It was observed that the impact of delay during events such as NFL games occurs between 0-3 hours before and after the games and cascades up to a radius of six miles. According to the study results, we are able to explain more than 80% of the variance in the bus travel time at each segment and can make future travel time predictions during special events with an out-of-sample error of 2 minutes with information on bus schedule, traffic, weather, scheduled events and participation of people in an event. The model with the highest performance in terms of goodness of fit and predictive accuracy is the XGBoost.

The main contribution of the paper lies in an efficient real time estimation of travel time assimilating multifaceted feature space and analysis of its cascading implications. The outcome of this work can be integrated in different transportation analytics initiatives. Using this information we generate heat maps that can be used in (1) a decision framework for DelayRadar (6), a process that assists the transit agency in developing a dynamic transit schedule during the special events, and (2) in the transit-hub application (11) that provides the delay estimates to the residents and visitors using the transit system.

REFERENCES

[1] Patnaik, J., S. Chien, and A. Bladikas, Estimation of bus arrival times using APC data. *Journal of public transportation*, Vol. 7, No. 1, 2004, p. 1.

[2] Zarei, N., M. A. Ghayour, and S. Hashemi, *Road Traffic Prediction Using Context-Aware Random Forest Based on Volatility Nature of Traffic Flows*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 196–205, 2013.

[3] Elhenawy, M., H. Chen, and H. A. Rakha, Random Forest Travel Time Prediction Algorithm using Spatiotemporal Speed Measurements, 2014.

[4] Sun, F., Y. Pan, J. White, and A. Dubey, Real-time and Predictive Analytics for Smart Public Transportation Decision Support System. In *2nd IEEE International Conference on Smart Computing*, 2016.

[5] Nookala, L. S., *Weather impact on traffic conditions and travel time prediction*. Ph.D. thesis, Citeseer, 2006.

[6] Oruganti, A., F. Sun, H. Baroud, and A. Dubey, Delayradar: A multivariate predictive model for transit systems. In *Big Data (Big Data), 2016 IEEE International Conference on*, IEEE, 2016, pp. 1799–1806.

[7] Yang, J.-S., Travel time prediction using the GPS test vehicle and Kalman filtering techniques. In *Proceedings of the 2005, American Control Conference, 2005.*, 2005, pp. 2128–2133 vol. 3.

[8] Dietterich, T. G., Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 1–15.

[9] Breiman, L., Bagging Predictors. *Machine Learning*, Vol. 24, No. 2, 1996, pp. 123–140.

[10] Breiman, L., Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32.

[11] Shekhar, S., F. Sun, A. Dubey, A. Gokhale, H. Neema, M. Lehofer, and D. Freudberg, *Transit Hub: A Smart Decision Support System for Public Transit Operations*, Hoboken, NJ, chap. 36. John Wiley & Sons, 1st ed., 2017.