DelayRadar: A Multivariate Predictive Model for Transit Systems

Aparna Oruganti*, Fangzhou Sun†, Hiba Baroud*, and Abhishek Dubey†
*Department of Civil and Environmental Engineering,Vanderbilt University
†Department of Electrical Engineering and Computer Science, Vanderbilt University

Abstract—**Effective public transit operations are one of the fundamental requirements for a modern community. Recently, a number of transit agencies have started integrating automated vehicle locators in their fleet, which provides a real-time estimate of the time of arrival. In this paper, we use the data collected over several months from one such transit system and show how this data can be potentially used to learn long term patterns of travel time. More specifically, we study the effect of weather and other factors such as traffic on the transit system delay. These models can later be used to understand the seasonal variations and to design adaptive and transient transit schedules. Towards this goal, we also propose an online architecture called DelayRadar. The novelty of DelayRadar lies in three aspects: (1) a data store that collects and integrates real-time and static data from multiple data sources, (2) a predictive statistical model that analyzes the data to make predictions on transit travel time, and (3) a decision making framework to develop an optimal transit schedule based on variable forecasts related to traffic, weather, and other impactful factors. This paper focuses on identifying the model with the best predictive accuracy to be used in DelayRadar. According to the preliminary study results, we are able to explain more than 70% of the variance in the bus travel time and we can make future travel predictions with an out-of-sample error of 4.8 minutes with information on the bus schedule, traffic, and weather.**

Keywords-**Transit; Bus Delay; Prediction; Traffic Flow; Weather**

I. Introduction

Emerging Trends and Challenges. With interest in transforming urban planning to address smart cities advancement becoming more prominent, multiple challenges arise as city planners strive to provide smarter and interconnected communities to improve the performance of infrastructure systems and ultimately better serve the society. Transportation systems constitute one of the main focus areas of such challenges as they connect people and commodities between and within cities. In particular, for the case of mid-sized cities, the bus system is considered as the main and often times the only public transportation facility. However, such system is not used very often. For example, according to Census data from 2009 [1], less than 3% of people in Nashville use the public transit for their daily commute, and the main reason why people do not use the transit system in mid-sized cities like Nashville is due to the longer commute time and inefficiency of scheduling [2].

Public transit agencies, such as the Metropolitan Transit Authority (MTA) in Nashville, are currently concerned with identifying the factors causing delays in the bus schedule and determining preparedness strategies to minimize delays in response to such factors. Therefore, as a starting point a number of cities are integrating automated vehicle locators (AVL) in their fleets. With the real-time location information, they can provide a better estimate of the time of arrival. However, addressing the long term and seasonal trends for delay is still a challenge. Identifying such patterns is crucial for designing optimal transit schedules that can be adjusted seasonally to account for delay variations and ensure that overall, the transit system runs punctually. Better planning apps can provide the seasonal schedule change information back to the customers.

There are multiple challenges that city planners and researchers face when addressing these issues. First, identifying and collecting data from multiple sources to represent the different factors is a critical aspect of this work, as different data sources are often not consistent across space and time. Second, modeling traffic data and bus schedule delays on route segments that lack traffic flow sensors can present a challenge to obtain the desired level of accuracy in travel time predictions. And finally, identifying a modeling approach that provides a high level of accuracy while insuring the interpretability of the model to draw conclusions and provide a timely update of the travel time prediction can impose certain trade-offs.

State of the art. Travel time prediction is a well studied problem [3], [4], [5]. Once available, the predicted value can be used for transit operation monitoring, smart trip planning, rough delay time estimation, at-stop displays, among others. Researchers have also studied long term data of bus service. For example, Abkowitz et al. [6] found that trip distance, passenger activity and signalized intersections could greatly affect the mean and variance of bus running time. Kimpel et al. [7] analyzed the bus service performance and passenger demand using Tri-Met Bus Dispatch System data at time point level. They found that the delay variation at previous time points, passenger demand variation, speed and distance contribute to delay variations. However, given that real-time AVL data has only recently started becoming available in communities, we are not equipped to study and understand the correlation between environmental variables such as precipitation, wind speed, visibility, and other factors that vary seasonally and the transit vehicle performance. Understanding such correlations will allow us to develop models

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

that can help the transit agencies adapt their schedules based on seasonal weather variations and traffic information.

Contributions. In this paper, we use the data collected over several months from one such transit system and study the effect of weather and other covariates such as traffic on the transit network delay. These models can later be used to understand the seasonal variations and design an adaptive and transient transit schedule as part of future work. Towards this goal, we also propose an online architecture called DelayRadar. The novelty of DelayRadar lies in three aspects: (1) a data store that collects and integrates real-time and static data from multiple data sources, (2) a predictive statistical model that analyzes the data to make predictions on transit travel time, and (3) a decision making framework to develop an optimal transit schedule based on variable forecasts related to traffic, weather, and other impactful factors (not covered in this paper). This paper focuses on identifying the model with the best predictive accuracy to be used in DelayRadar. According to the preliminary study results, we are able to explain more than 70% of the variance in the bus travel time and we can make future travel predictions with an out-of-sample error of 4.8 minutes with information on bus schedule, traffic, and weather.

The outline of this paper is as follows. Section II is an overview of the current literature on transit travel time analytics. Section III is a summary of the data collection and curation process. The modeling approaches along with the analysis results and discussion are presented in Section IV, and concluding remarks with future research directions are outlined in Section V.

II. Background and Related Work

Travel time and arrival time variation were found to have a great impact on commuters' satisfaction [8]. In the past decade, numerous studies have been conducted to develop models and algorithms to predict bus travel delay and arrival delay. Abdelfattah et al. [9] developed linear and nonlinear regression models for predicting bus delay under normal conditions using simulation data. Regression models measure various independent variables to predict a dependent variable. Williams and Hoel [10] found that daily traffic condition patterns are consistent across the weeks. Jeong et al. [11] presented a historical average model and found that the historical model was outperformed by other models because its prediction accuracy was limited by the reliability of traffic patterns.

Patnaik et al. [4] used distance, number of passengers at stops, stop numbers, and weather conditions for multilinear regression models to predict bus arrival time. However, since the attributes in transit services are often not independent but correlated with each other, the performance of regression models will deteriorate as the dimension of the data increases. Machine learning models can deal with complicated relationships and noisy data.

Bus Schedules		Real-time Transit	
Format	Static GTFS	Format	Real-time GTFS
Source	Nashville MTA	Source	Nashville MTA
Update	Every public release	Update	Every minute
Size	193 MB (used version)	Size	278 GB
Time Points		Real-time Traffic	
Format	Excel	Format	JSON
Source	Nashville MTA	Source	Here API
Update	Every month	Update	Every minute
Size	300,000 entries/month	Size	4.95 GB (compressed)
Weather			
Format	JSON		
Source	Dark Sky API		
Update	Every 5 minute		
Size	17 MB		

Table I
Realtime and Static Datasets Collected in the System.

Elhenawy et al. [3] presented a data clustering and genetic programming approach to predict the travel time along freeways. Artificial neural network (ANN), [12], [13], [14] and support vector machines (SVM), [5], [15], [14], [16] are two of the most widely used machine learning models in bus time prediction. Kalman Filtering models rely on historical data and real-time data and have been employed extensively for bus time prediction, [12], [17], [18], [19], [16].

The prior work in this area has been primarily focused on developing models for predicting delay as a short or long term self-contained process. While some of the approaches have studied the effect of traffic on the travel delay, to the best of our knowledge the effect of other environmental variables and effect of local events has not been extensively studied. In this paper, we present models to understand the effect of weather and traffic speed on travel delays, however, our future extension will look on other factors such as local events and will consider other models as our data collection continues.

III. Data

In this section, we first describe the multiple sources of data that are integrated into the system and then present how the heterogeneous data that are different in types, formats and sample rates are preprocessed and managed.

A. Data collection

We have been collaborating with Nashville Metropolitan Transit Authority (MTA) to access the bus schedules and real-time bus data feeds in Nashville. Also, we are integrating data from multiple other data sources to collect the real-time traffic and weather data in the city. The data sets that we have integrated into the multivariate predictive system are as follows:

- Bus schedule datasets are the static public transportation schedules and associated geographic information of routes, trips, stop times, physical route layout in

General Transit Feed Specification (GTFS) format [20] for all the 57 bus routes in Nashville.

- Real-time transit feeds are the real-time updates of transit fleet information in real-time GTFS format [21], including three types of information: (1) trip updates: bus delays and changes, (2) service alerts: routes and buses that are affected by unforeseen events, (3) vehicle position: bus locations with timestamps.
- Time-point feed provides the historical bus operating details, including each bus's route, trip and vehicle ID, accurate arrival and departure time at time points, etc. Nashville MTA releases the time-point data sets at the end of each month.
- Traffic flow feed contains the speed and congestion ("jam factor") information for 174,339 road segments in Nashville. We collect and store the real-time traffic flow data from HERE Traffic API [22] every minute. The uncompressed daily size traffic flow data for all bus routes in Nashville is about 5.7 GB.
- Weather condition feed provides the current weather conditions in Nashville, such as temperature, wind speed, nearest storm distance, precipitation intensity, visibility, among others.

B. Data Curation and Processing

In this section, we describe how the data from various sources is integrated to get a data set which has traffic, weather, bus schedules, and historical bus operating details for each trip on the bus route.

- Each route segment from the real time feed can be linked to multiple routes.
- The route segments from these routes are selected and matched to their corresponding trips for each bus route.
- Similarly, the weather data is matched to the route segments based on time stamps. The granularity of the weather data is 5 minutes and it is assumed that the weather is constant within this time.
- Finally, the weather data for each trip is taken as the average of each predictor (precipitation intensity, visibility, etc.) of the route segments in the considered trip. The traffic speed is considered as the weighted average of speed and distance traveled (i.e. length of each route segment). The data was processed for the month of May for multiple routes in Nashville.

It is important to point out the key challenges that we encountered and considered using the pre-processing scripts. The most important challenge was combining the real time data with the Nashville MTA bus schedule data. Real-time data has the information for each route segment. Each route segment can correspond to multiple shape IDs. As such, merging the high volume of data needed to be handled carefully in order to match the route segments with their corresponding trips. This was done by relating the trips to their shapes and trip timings.

Model	Formula	RMSE	R^2
1	$\hat{Y}$ = $\hat{\beta}_0$ + $\hat{\beta}_1$TS +$\hat{\beta}_2$visi +$\hat{\beta}_3$pres + $\hat{\beta}_4$humi + $\hat{\beta}_5$WS + $\hat{\beta}_6$ozone + $\hat{\beta}_7$NSD + $\hat{\beta}_8$ppt + $\hat{\beta}_9$temp + $\hat{\beta}_{10}$sch_TT + $\hat{\beta}_{11}$BS	4.916	0.711
2	$\hat{Y}$ = $\hat{\beta}_0$ + $\hat{\beta}_1$TS +$\hat{\beta}_2$visi +$\hat{\beta}_3$pres + $\hat{\beta}_4$humi + $\hat{\beta}_5$WS + $\hat{\beta}_6$ozone + $\hat{\beta}_7$NSD + $\hat{\beta}_8$ppt + $\hat{\beta}_9$temp + $\hat{\beta}_{10}$sch_TT + $\hat{\beta}_{11}$BS + $\hat{\beta}_{12}$(TS + visi + pres + humi + ozone + NSD + ppt + temp + sch_TT + BS)	4.913	0.714
3	Initial Model: $\hat{Y}$ = $\hat{\beta}_0$ + $\hat{\beta}_1$TS + $\hat{\beta}_2$visi + $\hat{\beta}_3$humi + $\hat{\beta}_4$WS + $\hat{\beta}_5$NSD + $\hat{\beta}_6$temp + $\hat{\beta}_7$BS + $\hat{\beta}_8$TS(sch_TT+BS) Final Model: $\hat{Y}$ = $\hat{\beta}_0$ + $\hat{\beta}_1$TS + $\hat{\beta}_2$visi + $\hat{\beta}_3$WS + $\hat{\beta}_4$NSD + $\hat{\beta}_5$BS + $\hat{\beta}_6$TS(sch_TT+BS)	4.882	0.713
4	Initial Model: $\hat{Y}$ = $\hat{\beta}_0$ + $\hat{\beta}_1$TS + $\hat{\beta}_2$visi + $\hat{\beta}_3$humi + $\hat{\beta}_4$WS + $\hat{\beta}_5$NSD + $\hat{\beta}_6$temp + $\hat{\beta}_7$ S + $\hat{\beta}_8$TS(sch_TT) Final Model: $\hat{Y}$ = $\hat{\beta}_0$ + $\hat{\beta}_1$TS + $\hat{\beta}_2$visi + $\hat{\beta}_3$WS + $\hat{\beta}_4$NSD + $\hat{\beta}_5$BS+ $\hat{\beta}_6$TS(sch_TT)	4.882	0.712
5	Random forests	5.79	0.729

Variable Acronyms :
$\hat{Y}$: Predicted Travel Time, TS: Real-time Traffic Speed, ozone: Ozone
pres: Pressure, visi: Visibility, BS: Static Traffic Speed
sch_TT: Scheduled Travel Time, NSD: Nearest Storm Distance
ppt: precipitation intensity, humi: Humidity, WS: Wind Speed

Table II
SUMMARY OF SELECTED MODELS WITH THEIR PERFORMANCE OF FIT (R^2) AND PREDICTION ACCURACY (RMSE)

IV. PREDICTIVE MODELS

A. Modeling Approach

The models considered in this paper fall under the regression and tree-based family of models. In this section, we summarize each model and provide the results of the models in terms of goodness of fit and predictive accuracy. We provide the acronyms used to present the predictors in our models at the bottom of Table II.

Multivariate Linear Regression. We first consider multivariate linear regression models in which the travel time in minutes is the outcome variable, Y, and is modeled as a linear function of the predictors. Mathematically, a linear regression is outlined as follows,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \qquad (1)$$

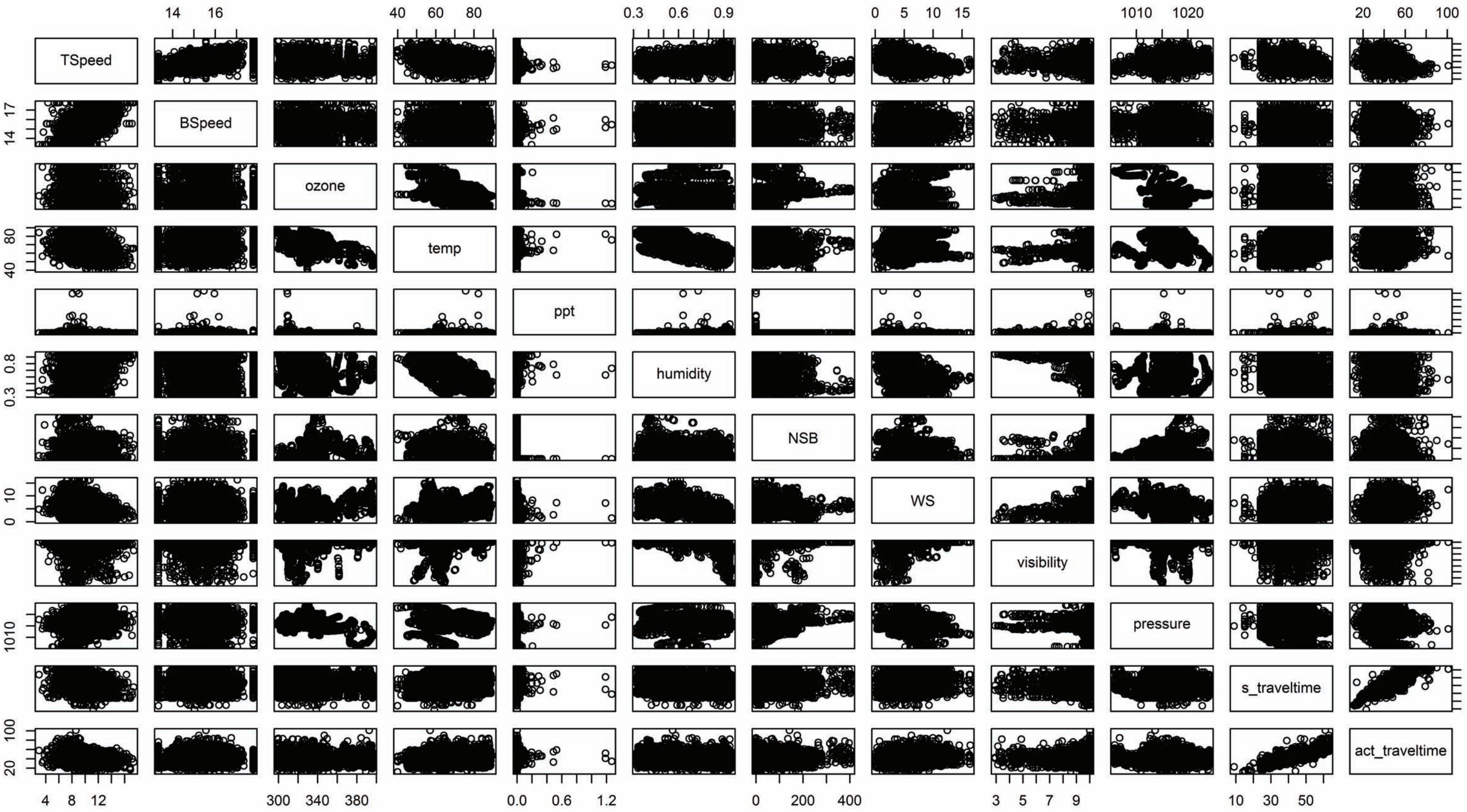

Figure 1. Scatter Plot of the Variables in the Data: Real-time and Static Traffic Speed, Ozone, Temperature, Precipitation Intensity, Humidity, Nearest Storm Distance, Wind Speed, Visibility, Pressure, Schedule Travel Time, Actual Travel Time. Each plot represents the pairwise relationship between the two variables corresponding to the x-axis and the y-axis. For example, if we are looking for a potential correlation between temperature and humidity, we can examine the intersecting plot of these two variables and observe that as the temperature increases, humidity decreases, we can then infer a negative correlation between humidity and temperature.

In this model, X_1, ..., X_p represent the independent variables, β_0 is the intercept, β_1, ..., β_p are coefficients of the respective predictors, and ϵ is the random error. Linear regression models can be extended to account for non-linear relationships between the predictors and the outcome variable through polynomial regressions, and they can account for correlations among the predictors through the consideration of interactions between the predictors, [23].

Several linear models were considered that included different combinations of the predictors, polynomial factors, and interactions. The model with the best fit for predicting travel time is identified and shown below using a stepwise regression analysis which is a semi-automated process of model building in which the choice of predictors is successively carried.

$$\begin{aligned} \hat{Y} = \hat{\beta_0} + \hat{\beta_1} TS + \hat{\beta_2} visi + \hat{\beta_3} WS + \hat{\beta_4} NSD \\ + \hat{\beta_5} BS + \hat{\beta_6} TS(sch\_TT + BS) \end{aligned} \tag{2}$$

Random Forests. Random forests [24] are an ensemble learning method for classification and regression. It is an approach where a number of decision trees are constructed and voting is done to define the best classifier. The underlying principle is that a group of weak learners can be combined to form a strong learner. In a decision tree for regression, the outcome variable is fitted for a regression model using each predictor. For each predictor the data is split at split points and the Sum of Squared Errors (SSE) is evaluated at each split point. The predictor resulting in the minimum SSE is selected for the node. This process is repeated until the entire data is covered. In random forests, each tree is trained for N samples by selecting m predictors randomly. Breiman [24] suggests three possible values for m: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$, and $2\sqrt{m}$. The value of m used in this work is $\sqrt{m}$. Our data set consists of 4700 samples and we use a 10-fold cross validation to calculate the test error.

Performance: To assess the performance of the models we identified, we consider two metrics, the Root Mean

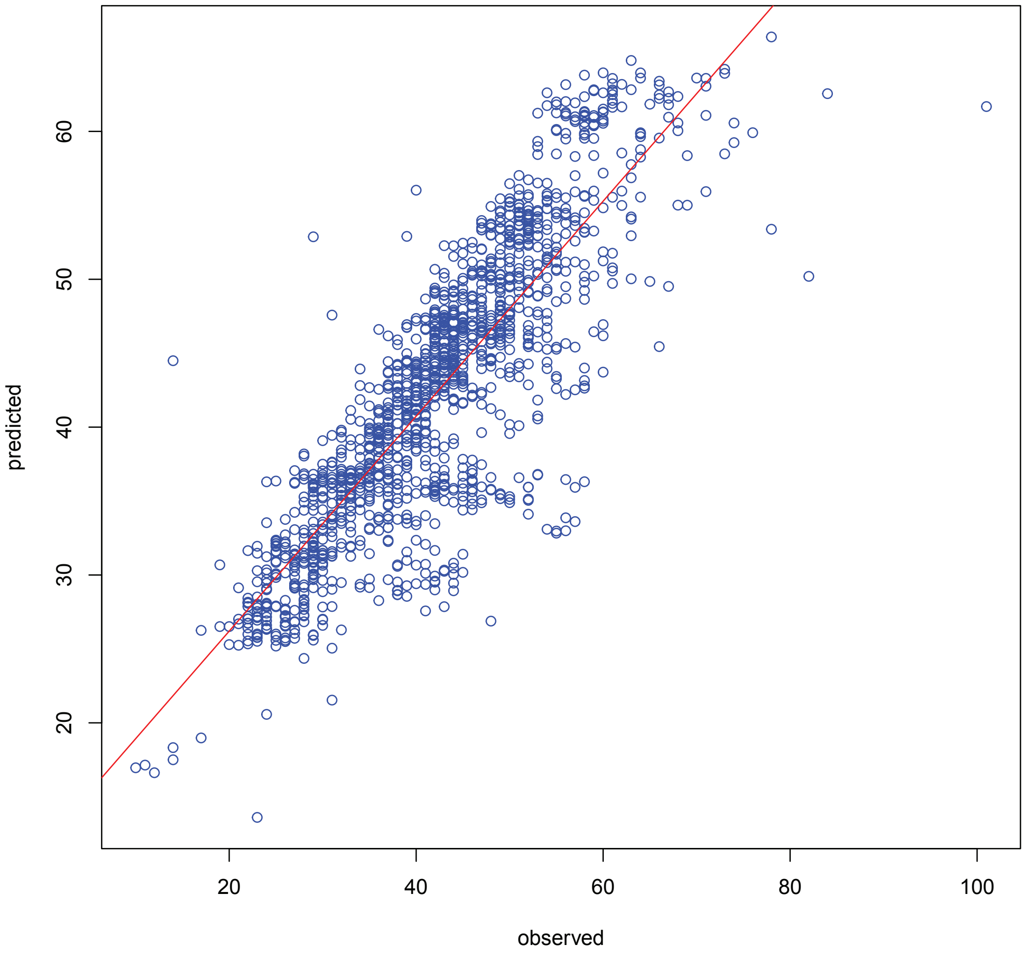

Figure 2. Observed Versus Predicted Travel Time.

Squared Error (RMSE) to assess the predictive accuracy and R^2 to assess the goodness of fit. RMSE calculates the square root of the average of the squared differences between the predicted and observed travel time for new data points, this metric is in minutes, which is the unit of the outcome variable, Y.

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

R^2 is a measure of how well the model fits the set of observations. More specifically, it accounts for the amount of variability in Y that has been explained by the predictors, it takes values between 0 and 1 and is independent of the scale of Y.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{4}$$

TSS $=\sum(y_i - \bar{y})^2$ is the Total Sum of Squares which is a measure of the error in the absence of a statistical model, and *RSS* is the Residual Sum of Squares which measures the discrepancy in the model considered.

These metrics are evaluated for various statistical models using linear regression and random forests and the results of selected models are shown in Table II. RMSE values range between 4.91 and 5.79 minutes with an average error of 4 minutes and 50 seconds in travel time prediction. The value of R^2 ranges between 0.711 and 0.729, with random forests providing the best fit for the data. Models 3 and 4 are the results of a stepwise regression analysis, whereby the initial model represents the full model and the final model represents the best selected model based on multiple iterations considering different subsets of the predictors.

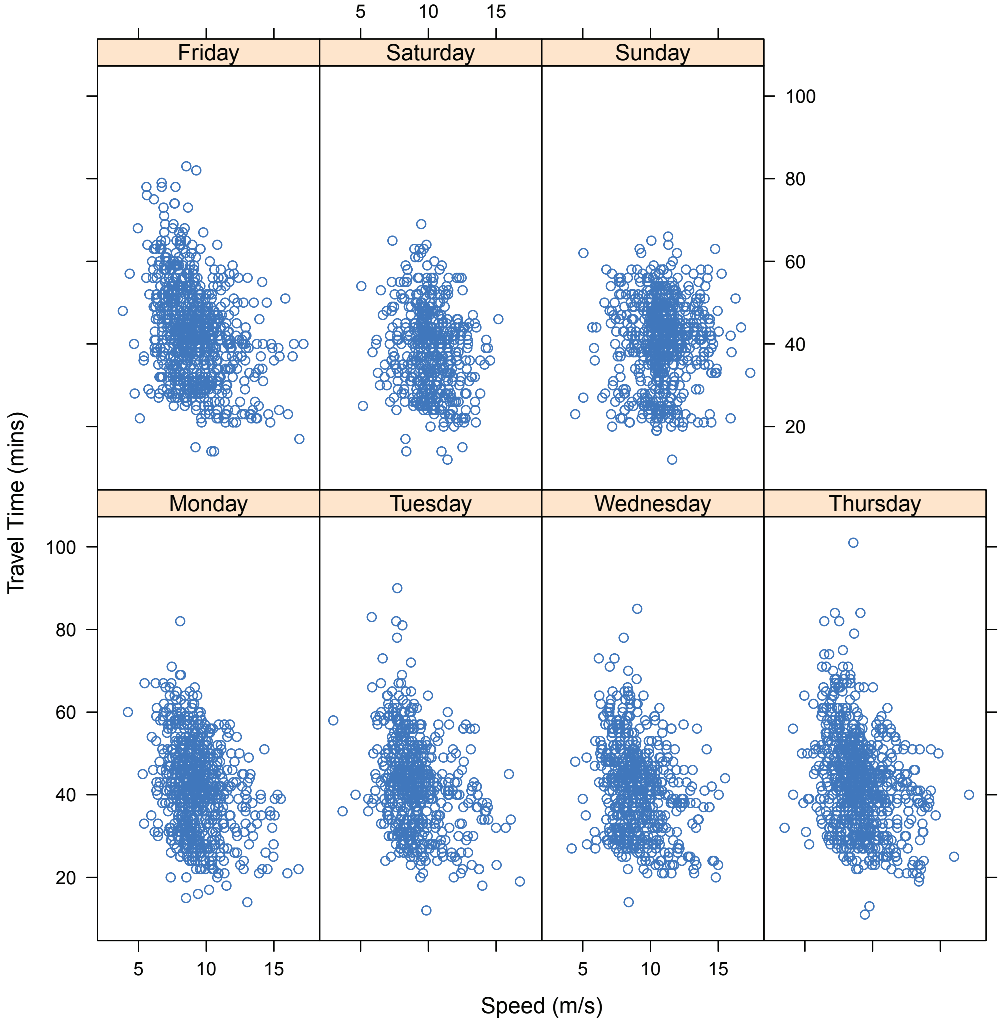

Figure 3. Scatter Plot of the Travel Time as a Function of Traffic Speed for Each Day of the Week Separately.

B. *Predictive Analysis*

We first examine the scatter plot of all the variables in our data set, Figure 1. Based on this initial examination of the data, we notice that there is a strong positive correlation between the scheduled travel time and actual travel time, which is expected. In addition, we notice a negative correlation between humidity and temperature and a positive correlation between wind speed and visibility, meaning that not all predictors are needed in our model due to potential collinearity. Besides these observed relations, the scatter plots of the data collected do not convey any other information in terms of correlation between the travel time and the weather and traffic factors considered, suggesting that there might be confounding or interaction effects among the predictors.

By examining the summary of results in Table II, we notice that the model with the best goodness of fit expressed by the highest value of R^2 is the random forests model which explains approximately 73% of the variance in the travel time. However, this modeling approach does not perform well with new data and provides a discrepancy of approximately 5.8 minutes in travel time prediction. RMSE was computed on new data points by training the model on a subset and testing it with a different set to assess the model's ability to predict new data, the process is done using 10-fold cross validation. Two linear regression models perform the best in terms of the predictive accuracy, and these are models 3 and 4 in Table II, where both provide an improvement in the error by 1 minute, with a predictive error of 4.8 minutes.

In order to visualize the effectiveness of the multivariate linear regression model in predicting travel time, we plot the observed and predicted travel time in Figure 2 for model 3.

We notice that for most of the data points, the observed and predicted values are close to each other with the exception of about 20 points that represent large discrepancies, which, compared to 1000s of data points, represents a reasonable amount of error.

As such, while we are interested in identifying the model with the best fit in order to make inferences on the significance and impact of predictors on the travel time, given the ultimate goal of developing DelayRadar, we want to strive for a high prediction accuracy in order to inform a dynamic and transient transit system scheduling in the future.

C. Attribute Analysis and Discussion

Based on the results in Table II of models 3 and 4, which have been shown to provide the best accuracy and a comparable goodness of fit to random forests, we can identify the set of predictors that are most significant in the estimation of travel time. These predictors include, the traffic speed, the visibility, the wind speed, the nearest storm distance, the speed limit, with an interaction term between the traffic speed and the scheduled travel time for model 4, and a similar result was obtained in model 3 where the interaction was between the traffic, the scheduled travel time, and the speed limit. As such, the correlation between humidity and temperature and their impact on travel time is expressed through other weather predictors. Also, while the precipitation might be considered an important factor that will affect travel time, it was not a significant predictor in the best model, rather the visibility and wind related factors were more important.

We considered the potential effect of a weekday versus a weekend on the travel time and we plot the travel time as a function of traffic speed for each day of the week. According to Figure 3, Wednesdays, Thursdays, and Fridays showed a significant correlation between the traffic speed and travel time, suggesting that the day of the week could have an impact on the prediction of the travel time. However, in our analysis, we considered a dummy variable for the weekend, and that did not improve the predictive accuracy and did not show any significance in the coefficient estimate.

V. Conclusion

In this paper, we showed that we can develop a model that helps us predict the affect of weather and traffic on the transit system delay. We collected and integrated data from multiple sources and used statistical models to predict long-term patterns in the bus travel time. According to the preliminary study results, we are able to explain more than 70% of the variance in the bus travel time and we can make future travel predictions with an out-of-sample error of 4.8 minutes with information on bus schedule, traffic, and weather. As part of our ongoing work, we are using these models for (a) developing the overall decision framework for DelayRadar that will help the transit agency develop an optimal transit schedule, and (b) integrate the results in the transit-hub application [18] that provides the delay estimates to the residents and visitors using the transit system.

Acknowledgments

This work is sponsored in part by the National Science Foundation under the award number CNS-1528799 and in part of the Vanderbilt Initiative on Smart City Operations and Research, a trans-institutional initiative funded by the Vanderbilt University. We acknowledge the support and suggestions provided by our partners from Nashville Metropolitan Transport Authority.

References

[1] B. McKenzie and M. Rapino, “Commuting in the united states: 2009,” *US Department of Commerce, Economics and Statistics Administration, US Census Bureau*, 2011.

[2] A. Semuels, “Why people don’t ride public transit in small cities,” *The Atlantic*, October 28, 2015. [Online]. Available: http://www.theatlantic.com/business/archive/2015/10/nashville-charlotte-public-transit/412741/

[3] M. Elhenawy, H. Chen, and H. A. Rakha, “Dynamic travel time prediction using data clustering and genetic programming,” *Transportation Research Part C: Emerging Technologies*, vol. 42, pp. 82–98, 2014.

[4] J. Patnaik, S. Chien, and A. Bladikas, “Estimation of bus arrival times using apc data,” *Journal of public transportation*, vol. 7, no. 1, p. 1, 2004.

[5] C.-H. Wu, J.-M. Ho, and D.-T. Lee, “Travel-time prediction with support vector regression,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 5, no. 4, pp. 276–281, 2004.

[6] M. D. Abkowitz and I. Engelstein, “Factors affecting running time on transit routes,” *Transportation Research Part A: General*, vol. 17, no. 2, pp. 107–113, 1983.

[7] T. Kimpel, “Time point-level analysis of transit service reliability and passenger demand, urban studies and planning,” *Portland, OR: Portland State University*, p. 154, 2001.

[8] J. Bates, J. Polak, P. Jones, and A. Cook, “The valuation of reliability for personal travel,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 37, no. 2, pp. 191–229, 2001.

[9] A. Abdelfattah and A. Khan, “Models for predicting bus delays,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1623, pp. 8–15, 1998.

[10] B. M. Williams and L. A. Hoel, “Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results,” *Journal of transportation engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[11] R. Jeong and L. R. Rilett, “Bus arrival time prediction using artificial neural network model,” in *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*. IEEE, 2004, pp. 988–993.

[12] M. Chen, X. Liu, J. Xia, and S. I. Chien, “A dynamic bus-arrival time prediction model based on apc data,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 19, no. 5, pp. 364–376, 2004.

[13] R. H. Jeong, “The prediction of bus arrival time using automatic vehicle location systems data,” Ph.D. dissertation, Texas A&M University, 2005.

[14] B. Yu, W. H. Lam, and M. L. Tam, “Bus arrival time prediction at bus stop with multiple routes,” *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1157–1170, 2011.

[15] Y. Bin, Y. Zhongzhen, and Y. Baozhen, “Bus arrival time prediction using support vector machines,” *Journal of Intelligent Transportation Systems*, vol. 10, no. 4, pp. 151–158, 2006.

[16] C. Bai, Z.-R. Peng, Q.-C. Lu, and J. Sun, “Dynamic bus travel time prediction models on road with multiple bus routes,” *Computational intelligence and neuroscience*, vol. 2015, p. 63, 2015.

[17] L. Vanajakshi, S. C. Subramanian, and R. Sivanandan, “Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses,” *IET intelligent transport systems*, vol. 3, no. 1, pp. 1–9, 2009.

[18] F. Sun, 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.

[19] S. Shekhar, S. Pradhan, F. Sun, A. Dubey, and A. Gokhale, “Empowering the next generation city-scale smart systems,” *IEEE 22nd International Conference on High Performance Computing Workshops (HiPCW)*, 2015.

[20] Wikipedia, “General transit feed specification,” 2015, [Online; accessed 31-January-2016]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=General_Transit_Feed_Specification&oldid=693322749

[21] “General transit feed specification (gtfs) real-time overview,” https://developers.google.com/transit/gtfs-realtime/, 2016, accessed: 2016-09-18.

[22] “Here traffic api,” https://developer.here.com/rest-apis/documentation/traffic/topics_v6.1/flow.html.

[23] G. A. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012.

[24] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324