

Transit Bus Travel Time Prediction using AVL Data

Dr. Stephen Arhin, P.E., PTOE, PMP
Assistant Professor
Civil & Environmental Engineering Department
Howard University
Washington DC, USA

Regis Zeola Stinson
Graduate Student
Civil Engineering Department
Howard University
Washington DC, USA

Abstract - The prediction of transit bus travel times along corridors is critical in the planning and operation of buses, especially in urban areas. Bus patrons tend to have more confidence in a transit system if travel times can be adequately predicted, within a certain margin of error. Washington DC's the transit agency, the Washington Metropolitan Authority (WMATA), recently equipped some of its fleet with Automated Vehicle Location (AVL) systems and Passenger Count Systems (PCS) to obtain data as buses travel along corridors.

In this study, data from the AVL/PCS system on transit buses were used to develop a travel time model to predict how long buses travel along selected corridors in Washington DC. AVL and PCS data for a period of one-month during the summer of 2016 for eight arterial bus routes used was in this study. The advertised travel times for the selected corridors from the selected origins and destinations were also obtained. Based on the literature review, a number of variables were selected as input for the prediction of bus travel times.

From the data analysis, it was determined that the number of passengers alighting, passengers boarding, number of access approaches and signalized intersections, significantly predicted transit bus travel time at 95% confidence interval. In addition, the bus travel time prediction model was determined to be statistically significant with validation tests indication model adequacy at 5% level of significance.

Keywords— *Transit Travel Time; Travel Time Prediction, AVL data.*

INTRODUCTION

The District of Columbia (DC) is one the largest Metropolitan Areas in the United States. The City attracts daily commuters from Maryland, Virginia, and West Virginia since it is the capital of the United States, causing most roadways to experience severe traffic congestion. In 2015, a study conducted by the Texas A&M Transportation Institute concluded that the District of Columbia has the worst traffic congestion ranking in the country. On average, commuters in the District spend approximately 82 hours in rush hour traffic per year. As a consequence, the Washington Metropolitan Area Transit Authority (WMATA), which is a government agency that operates and manages transit services, considers the impact of congestion when planning and scheduling their operations. WMATA provides transportation services such as metrorail, metrobus, and paratransit to the DC-Maryland-Virginia (DMV) Area to curb congestion. WMATA's transit buses operate over 300 fix-scheduled routes throughout the DMV area. These fixed routes allow WMATA to advertise timetables for riders to estimate the transit travel times along the routes.

AVL and PCS data can now be obtained from transit buses in the DMV area. Although travel time reliability is a performance metric of WMATA, there is no method employed to validate and/or predict transit travel time in the District. Therefore, validating the scheduled arrival times using Automatic Vehicle Location (AVL) data will help reduce inconsistencies in schedule adherence and allow WMATA to advertise realistic bus schedules based on traffic conditions. To maintain on-time arrivals on a bus route, a model can be developed to predict bus travel time based on bus route characteristics, land use and traffic conditions.

I. OBJECTIVE

The objectives of this research are to:

- develop a transit travel time model using AVL data using eight arterial bus routes
- compare the actual travel time to the predicted (using AVL data), and advertised travel time along the selected routes

II. LITERATURE REVIEW

Travel time is a fundamental measure in transportation that is defined as the time it takes for a vehicle to navigate between two points of interest. It is used by planners and engineers to help to schedule transit bus arrivals at each bus stop along a route. Accurate prediction of bus arrival time can help improve the quality of bus-arrival-time information service, and attract more ridership. Travel time reliability is one of the major measurements of effectiveness that affect mode choice for transportation between two locations within a network. Mobility in urban areas impacts urban livelihood to a great extent. To enhance urban mobility, several research studies on predicting travel time have been conducted to provide passengers (or commuters) with estimations (within a margin of error) of how long a particular trip will take. (1)

In 2014, Feng (2) analyzed bus travel times and the factors that affect its reliability. And included a review of several articles that studied impacting factors on travel time. One of the most influential factors that is associated with travel time is travel distance. Other studies considered the number of signalized intersections to be an impacting factor; however, these factors' impacts varied due to the different geometric characteristics and signal timings of the arterials used in the study. Another important factor that impacts transit travel time is traffic congestion. The author reviewed the impact of congestion on travel time using "time of day" and/or "travel direction" as the independent variable.

The number and spacing of bus stops is also a variable that had a positive impact on bus travel time and reliability. Several studies use the number of actual stops made as an independent variable. Other variables such as bus departure delays and dwell time impact bus travel time. Lastly, passenger load, number of passengers boarding and alighting had a significant influence on bus travel time and reliability. However, nearside and farside bus stop types did not have any significant impact on travel time.

A study was conducted in 2013 by Xinghao et. al (3) to develop a short-term prediction model using real-time bus location and radio-frequency identification (RFID) data. The proposed model were based on an augmented self-adapting smoothing algorithm that is used to predict the running speed of transit buses using short-term sample speeds of taxis and buses. In the development of the model, the researchers took into consideration the variation of bus speeds due to traffic controls and other impacting factors. The proposed model, which integrated AVL and RFID data, was tested against the historical data-based model which used only historical AVL data. The results indicated that the relationship between speeds of transit buses and taxis on the same link during the same time period is linear which was determined to be statistically significant with R^2 values ranging from 0.72 to 0.83. Also, the results showed that the combined data model out-performed the AVL-only data model.

Improving a transportation system is one of the prominent public policy issues for any government. Decisions for transportation infrastructure often involve a cost-benefit analysis. It has been established that monetary estimates of travel time savings and travel time reliability are two important components in improving transportation systems. This is usually one of the metrics used in the decision-making processes on public transportation projects pertaining to travel time reliability (e.g., constructing a bypass to reduce congestion) and/or travel time savings (e.g., constructing a faster public transportation mode). A study conducted by Beaud et. al (4), derived practicable measures to determine the extent to which commuters value a reliable travel time (VRT) and savings on travel time (VST). This was illustrated by using the Bernoulli approach to develop a microeconomic model of transportation mode choices which identified each trip by its monetary value and the statistical distribution of its random travel time. The function of a traveler's preference was assumed to be discrete, and was defined as the sum of the linear function of price and the non-linear function of travel time. For this model, VST was defined as the willingness to pay for a reduced travel time, and VRT was defined as the willingness to pay for a consistent travel time. This study explored how these two variables are functions of travel time, and how they are affected by the statistical distribution of travel time and the preferences of travelers in terms of travel time variability.

A study in Ankara, Turkey by Yetiskul and Senbil (5) was conducted to determine which factors influence the variability of bus transit travel time. The causes of inconsistent travel times were identified as both external and internal factors. Re-occurring traffic congestion during peak hours and non-recurring factors such as traffic accidents or roadway maintenance were classified as external factors; whereas, fare collection process, passenger capacity, and number of stops along a route were classified as internal factors. To account for variation caused by service region, highways, and individual bus lines, three models were developed and tested. The outcome indicated that travel time variability in transit systems were caused by temporal dimension (time of day and day of week), spatial dimension (operation system's physical characteristics), and service characteristics (number of stops on a route, dwell time, maximum passenger load, etc.)

Zhang and Xiong (6) employed an agent-based model (ABM) approach that performs multi-step travel time predictions by using historic and real-time traffic data. Each agent in the model represented a domain in a decision-making system that predicts travel time for each time interval based on a historical database and real-time data. A set of agent interactions were developed to preserve agents that correspond to similar traffic patterns to the real-time measurements, then the invalid agents or agents associated with insignificant weights are replaced with new agents. A combination of each agent's prediction results in an output that presents the predicted travel time distribution of the proposed model.

The instantaneous travel time method, historical average method, and the k-Nearest Neighbor (k -NN) prediction method were all compared with the proposed model to evaluate its performance. The instantaneous travel time method was used to predict future travel times with the assumption that the current speed of traffic along a segment will remain constant throughout the trip. The historical average method predicts travel time when the traffic conditions are consistent. The k -NN method was used to predict real-time travel time. Based on the results of daily predictions, the instantaneous and historical average method had large variations in performance compared to the ABM and k -NN methods. Table 1 presents the comparison of each method over a 60-minute duration with their associated mean absolute error and mean absolute percentage error.

Table 1: Prediction Results By Different Methods (6)

		Prediction Horizon (minutes)						
		0	10	20	30	40	50	60
Instantaneous	MAE (min)	11.52	13.06	14.40	15.78	17.10	18.28	19.38
	MAPE (%)	10.64	12.12	13.47	14.85	16.12	17.29	18.31
Historical Avg.	MAE (min)	13.01	13.01	13.01	13.01	13.01	13.01	13.01
	MAPE (%)	11.46	11.46	11.46	11.46	11.46	11.46	11.46
k-NN	MAE (min)	10.48	11.12	12.10	12.84	13.62	14.24	15.06
	MAPE (%)	9.24	9.95	10.68	11.31	11.98	12.61	13.18
ABM	MAE (min)	7.69	7.92	8.14	8.33	8.62	8.97	9.49
	MAPE (%)	6.75	6.98	7.21	7.53	7.86	8.18	8.57

The research results proved that, compared to other state-of-the art methods, agent-based modeling has a smaller prediction error, and maintained a prediction error less than a 9% for trip departing up to 60 minutes into the future.

Commuters value accurate transit travel time and real-time information. This allows passengers to better plan a trip with minimal waiting time. A study focused on developing an active artificial neural network (ANN) model using global positioning system (GPS) data that could accurately predict travel time of buses. The output is then transmitted into real-time information for a given subsequent bus stop (7). "ANNs learn from patterns and capture subtle functional relationships among data even if the underlying relationships are unknown or hard to explain." The travel time prediction model is based on both real time information and historical data. The proposed model was assessed by comparison to the historical average models, regression models, and Kalman filtering models. The Kalman filtering models encountered several variations and the regression models are not suitable when data is missing; therefore, the ANN model was compared only with the historical average model. The ANN model outperformed the historical average model approach in both prediction accuracy and robustness. The accuracy measure was determined by the predicted travel times average deviation from the actual travel time, whilst the robustness measure was determined by the number of times the algorithm's prediction was far from the actual travel time. The results from this study helped with the implementation of Advanced Public Transportation Systems (APTS).

A relationship between transit travel time and vehicular travel time can be established even though buses and passenger cars have different traveling behaviors. A study conducted by Esawey and Sayed (8), explored the potential of estimating vehicle travel time using transit travel time data. The research hypothesis stated that there is a strong correlation between neighboring roads to have similar traffic conditions. Archived travel time data of links and real-time transit data from adjacent links were analyzed using VISSIM. The overall accuracy of the travel time estimation was 82.4%. This precision was acceptable due to the variation of travel times in the study area. The results showed that the method of using transit travel time to establish a correlation with vehicular travel time of neighboring links was proven to be beneficial for roads that do not have existing travel time data.

Dublin by Gal et. al (9) used the Queueing Theory and Machine Learning methods to predict travel time. These combined methods were purported to be capable of predicting

travel time given a scheduled bus route and an origin and destination. Both real-time and historical transit data were taken into consideration in the process. The model was proposed to compute travel time using a set of predictors and bus stop data. The observed outcome showed that the principles from the Queueing Theory were effective, however, the data contained outliers that impacted the results. The Machine Learning method assisted in identifying the outliers and used historical data for prediction.

In 2009, Pu et. al (10) conducted a study to estimate urban street travel time by using bus probes in Chicago, Illinois. Previous studies fostered the concept of bus probes for Advanced Traveler Information System (ATIS) application; however, the past studies only focused on freeways and arterials using archived bus data which were not apt for real-time forecasting. As a result, real time transit data was used to estimate travel time using multivariate time series state-space modeling. Four state-space model were used for this research: the eastbound morning (EBAM) and evening (EBPM) rush hour, and westbound morning (WBAM) and evening (WBPM) rush hour. The results from each model's travel time estimation is presented in Table 2.

TABLE 2: Travel Time Estimations (Pu, 2009)

	Observed Average Test Vehicle Travel Time (seconds)	Actual Average Test Vehicle Travel Time (seconds)	Difference in Travel Time (seconds)
EBAM	291	296	5
EBPM	293	289	-4
WBAM	291	292	3
WBPM	303	303	0

For the EBAM scenario, based on the t-statistic value of 0.96, it was concluded that vehicular speeds are not significantly related to transit speeds. On the other hand, based on the t-statistic value of 2.18, the authors concluded that transit speeds are significantly related to vehicular speeds. The researcher came to a similar conclusion with the EBPM scenario. However, analysis on the westbound AM and PM scenarios indicated that vehicular and transit speeds are interrelated. For instance, for the WBAM scenario, vehicular speeds were statically significant related to transit speeds (t -statistic = 1.99, $p < 0.05$), and transit speeds were significantly related to vehicular speeds (t -statistic = 3.41, $p < 0.05$). These results supported the notion that a correlation exist between vehicle and bus speeds. Particularly, vehicle operations have a greater influence on transit operations in

the flow of traffic than transit on vehicles. These findings show that for urban roads, buses with AVL systems are acceptable probes for ATIS.

Previous studies have analyzed factors that may influence transit travel time and reliability. From the literature, time of day, distance, dwell time, and passengers boarding and alighting are influencing factors of travel time. Other variables such as number of bus stops, presence of traffic signals, passenger load, and direction also affect bus travel time. This research used known variables that impact travel time to develop a travel time prediction model for arterial roads in the District of Columbia. Also, variables known to interrupt traffic flow, such as access approaches and mid segment crosswalks, were analyzed to determine their significance or influence on bus travel time.

III. RESEARCH METHODOLOGY

The dataset used for this research was obtained from WMATA's AVL system. The variables from the data that were used are: number of passengers boarding, number of passengers alighting, the total passenger load, dwell time, segment length, number of bus stops, access approaches within the segment, number of signalized intersections and the number of mid-segment crosswalks along the segments.

The AVL data used in this study was obtained from the WMATA's Bus Planning, Scheduling, and Customer Facilities Department AVL system. WMATA's Metrobus operates over 300 bus routes and has a service area of nearly 1,500 square miles. In 2012, WMATA equipped their fleet with on-board systems that provided Computer Aided Dispatch and Automatic Vehicle Location (CAD/AVL). These devices or equipment included next stop annunciation, boarding and alighting passenger counting, location of bus along route, etc., are provided by Clever Devices. (11).

The data was first filtered to capture only weekday events during AM peak hours (6:00 AM - 9:00 AM) and PM peak hours (3:00 PM - 6:00PM). Specifically, the routes that were directed toward the Central Business District (CBD) of the city were observed for the morning events, while the evening events were extracted from routes leaving the City. An event is when a bus route leaves the identified origin and arrived at the selected destination. Table 3 presents the selected corridors, the direction of routes, and the origin and destination observed for the study for both AM and PM.

TABLE 3: SELECTED CORRIDORS AND SEGMENTS

Corridors	Segment Length (miles)
Connecticut Avenue, NW (WB)	1.4
14 th Street, NW (SB)	0.6
16 th Street, NW (SB)	0.6
Georgia Avenue, NW (SB)	2.0
7 th Street, NW (NB)	1.3
14 th Street, NW (NB)	0.1
16 th Street, NW (NB)	1.3
Georgia Avenue, NW (NB)	1.4

The prediction model to estimate bus travel time was developed using independent variables that are not correlated to each other (multicollinearity). Multivariate regression analysis was conducted to develop the model using SPSS. The resulting R^2 value was used to assess how well the predictor variables are estimating the dependent variable, while the F -statistic and significance value (p -value) were used to examine the strength of relationship between the dependent variable and the respective independent variables. The generalized regression model for Transit Travel Time was determined to resemble the following form:

$$TT = \beta_0 + \beta_1 P_a + \beta_2 P_b + \beta_3 P_l + \beta_4 D_t + \beta_5 S_l + \beta_6 B_s + \beta_7 S_i + \beta_8 A_a + \beta_9 X_w + e \quad (1)$$

where:

TT = Travel Time

P_b = Passengers Boarding

P_a = Passengers Alighting

P_l = Passenger Load

D_t = Dwell Time

S_l = Segment Length

B_s = Bus Stops

S_i = Signalized Intersections

A_a = Access Approaches

X_w = Mid-segment Crosswalks

TT is the dependent variable while the independent variables are P_b , P_a , P_l , D_t , S_l , B_s , S_i , A_a , and X_w . The values, and β_n are the regression coefficients with an associated error of e .

IV. RESULTS

As part of the analysis, curve estimation analyses between the dependent variable and each independent variable were performed to identify the best functions that relates them. Due to multicollinearity, only six of the nine variables were used to predict the transit bus travel time. Table 4 presents the optimal functions that relate the dependent variable and each of the independent variables.

TABLE 4: DATA TRANSFORMATION

Variable	Selected Relationship with Dependent Variable	Transformation Formula
D _t	Logistic	$f(x) = \frac{1}{1 + e^{-x}}$
P _b		$f(x) = x^3$
S _i	Cubic	$f(x) = \frac{1}{1 + e^{-s}}$
A _a	Power	$f(x) = x^2$
P _a	Linear	$f(x) = x$

Table 5 presents the summary results of the resulting model. The R^2 is the multiple correlation coefficient that measures the quality of the prediction of the dependent variable and is the coefficient of determination that is the proportion of variance in the dependent variable explained by the independent variables. This means that 69% of the variations in the dependent variable is explained by the independent variables. Table 5 also presents the F -statistic from the ANOVA test that determines whether the overall regression model is statically significant. The result shows that the independent variables significantly predicts the dependent variable, $F (6, 59) = 21.932, p < .05$ at 5% level of significance.

TABLE 5: MODEL SUMMARY

Model	R ²	Adjusted R ²	F	Sig. of F
	0.69	0.59	21.93	.000

The general form of the equation to predict transit travel time based on passengers boarding, and alighting, dwell time, passenger load, access approaches, and signalized intersections, is:

$$TT = 1160.973 - 227.832 \frac{1}{1 + e^{P_b}} + 13.358 P_a - 2152.89 \frac{1}{1 + e^{D_t}} + 67.619 \log_2 P_i + .066 A_a^2 + 1258.106 \frac{1}{1 + e^{S_i}}$$

Variables whose coefficients have associated p -values less than 0.05 deemed to be statistically significant in predicting transit travel time. Of the six independent variables, only passenger boarding and dwell time do not significantly contribute significantly to the predicting transit travel time.

V. MODEL ERROR

Figure 1 presents the output of the results of the Kolmogorov-Smirnov (K - S) Test for travel time model. The test statistic for K - S Test is the D -statistic which is defined as the distance between the two distributions. At a 5% level of significance, the maximum D value is 1.36. For this model, the computed D values is 0.2121, which confirms the validity of the model.

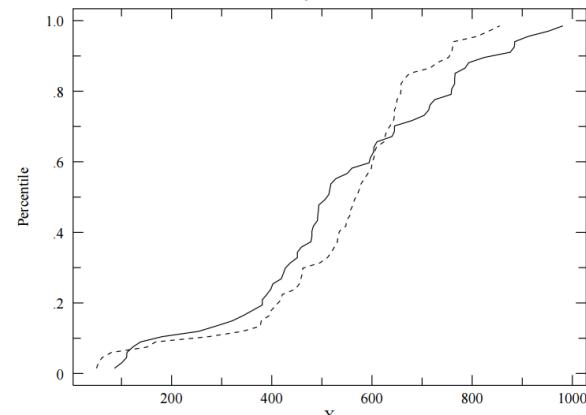


Figure 1: K-S Test Comparison Plot

The Mean Absolute Percent Error (MAPE) was used to determine the size of the deviation of the predicted TT values from the actual in percentage terms. It is computed as the average of the unsigned percentage error using the following formula:

$$\left[\frac{1}{n} \sum \frac{|Actual - Predicted|}{|Actual|} \right] \times 100$$

From the data, the MAPE for this model was determined to be 22.4%, which confirms a relatively low percent error in predicting the actual travel time of buses on arterial streets in Washington, DC.

VI. COMPARING AVERAGE ACTUAL, PREDICTED, AND ADVERTISED TRAVEL TIMES

Table 7 presents a comparison of the average actual, predicted, and advertised travel times for each of the eight routes analyzed in the study.

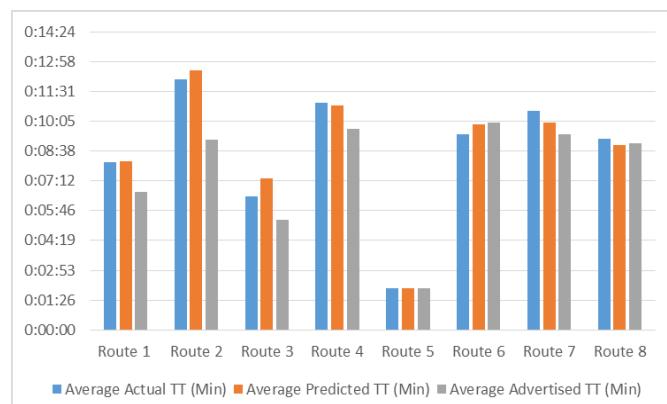


Figure 2: Comparison of Travel Times

From the graph, it can be concluded that the average advertised travel times are generally lower than the actual average travel times for the corridors used in this study. Also, the average predicted travel times are closer to the actual than the advertised travel times.

VII. DISCUSSION

From the linear transformation and regression analysis, passengers boarding, passengers alighting, passenger load, dwell time, signalized intersections and number of access approaches were used as independent variables for the model. This is due to the presence of multicollinearity.

The regression model yielded an R^2 value of 0.69 and a statistically significant F -value of 21.93 at 5% level of significance. The model was validated using several ad-hoc analyses, such as a homoscedasticity test, the K - S test, and MAPE. The homoscedasticity test showed that the residuals were randomly and evenly distributed about the mean line. In addition, from the K - S test, the maximum D -value of 0.2121 from the model was less than the critical D -value of 1.36. Hence, the transit travel time model was deemed to be statistically significant at 95% confidence interval.

The analysis shows a promising future for the AVL data in predicting actual transit travel times of buses along arterial corridors in an urban area with a relatively low MAPE. The data used in this study combined the travel times for both morning and evening peak periods since it was assumed that the corridor transit travel time for each peak period will remain unchanged. In addition, the combination provides variability in the data for statistical analysis purposes.

VIII. CONCLUSION

This research explored the development of a model to predict transit travel time. Based on the results of the statistical findings, the model effectively predicts the actual travel time at 5% level of significance. The model explains 69% (R^2) of the variation in transit travel time values. Furthermore, the K-S test showed that the predicted values are comparable to the actual travel times. The variables that significantly contributed to the transit travel time were the number of passengers alighting, the passenger load, and the number of access approaches and signalized intersections. The results and comparison showed the AVL data can be used to help improve planning, forecasting of transit travel time in urban areas.

REFERENCES

- [1] Carrion, C., & Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, 46(4), 720-741. doi:10.1016/j.tra.2012.01.003
- [2] Feng, W. (2014). Analyses of Bus Travel Time Reliability and Transit Signal Priority at the Stop-To-Stop Segment Level (Doctoral dissertation, Portland State University, 2014). Portland, Oregon: Portland State University PDXScholar.
- [3] Xinghao, S., Jing, T., Guojun, C., & Qichong, S. (2013). Predicting Bus Real-time Travel Time Basing on both GPS and RFID Data. *Procedia - Social and Behavioral Sciences*, 96, 2287-2299. doi:10.1016/j.sbspro.2013.08.258
- [4] Beaud, M., Blayac, T., & Stéphan, M. (2016). The impact of travel time variability and travelers' risk attitudes on the values of time and reliability. *Transportation Research Part B: Methodological*, 93, 207-224. doi:10.1016/j.trb.2016.07.007
- [5] Yetiskul, E., & Senbil, M. (2012). Public bus transit travel-time variability in Ankara (Turkey). *Transport Policy*, 23, 50-59. doi:10.1016/j.tranpol.2012.05.008
- [6] Zhang, L., & Xiong, C. (2016). A novel agent-based modelling framework for travel time reliability analysis. *Transportmetrica B: Transport Dynamics*, 1-18. doi:10.1080/21680566.2016.1142400
- [7] Gurmu, Z., & Fan, W. (2014). Artificial Neural Network Travel Time Prediction Model for Buses Using Only GPS Data. *Journal of Public Transportation*, 17(2), 45-65. doi:10.5038/2375-0901.17.2.3
- [8] Esawey, M. E., & Sayed, T. (2011). Travel time estimation in urban networks using limited probes data. *Canadian Journal of Civil Engineering*, 38(3), 305-318. doi:10.1139/111-001
- [9] Gal, A., Mandelbaum, A., Schnitzler, F., Senderovich, A., & Weidlich, M. (2015). Traveling time prediction in scheduled transportation with journey segments. *Information Systems*. doi:10.1016/j.is.2015.12.001
- [10] Pu, W., Lin, J., & Long, L. (2009). Real-Time Estimation of Urban Street Segment Travel Time Using Buses as Speed Probes. *Transportation Research Record: Journal of the Transportation Research Board*, 2129, 81-89. doi:10.3141/2129-10
- [11] Clever Devices. (n.d.). Retrieved September 04, 2016, from <http://www.cleverdevices.com/products-overview.htm>