Exploring Social Traffic Data for Evaluating Urban Arterial Congestion

Rakesh Nune, Corresponding Author
District of Columbia Department of Transportation
55 M Street, SE, Washington DC20003
Email: Rakesh.Nune@dc.gov

Weisheng Zhong
Department of Computer Science
Virginia Polytechnic Institute and State University
7054 Haycock Rd, Falls Church, VA 22043
Email: zwscn123@vt.edu

Kaiqun Fu
Department of Computer Science
Virginia Polytechnic Institute and State University
7054 Haycock Rd, Falls Church, VA 22043
Email: fukaiqun@vt.edu

Jason X. Tao
District of Columbia Department of Transportation
55 M Street, SE, Washington DC20003
Email: jason.tao@dc.gov

Word count: 4133 word text + 13 tables/figures x 250 words (each) = 7383

ABSTRACT

Monitoring traffic flows and establishing performance measures on roadways are critical tasks for transportation planning and traffic operations. Compared to the freeways, monitoring the roadway performance on urban arterials is much more challenging due to the nature of more dynamic traffic assignment patterns and insufficient traffic monitoring mechanisms on arterials. In the recent years, the emerging social traffic data from the Google/Waze application has provided a different layer for monitoring the real-time traffic conditions. In this study, the possibility of using the social traffic data in developing mobility performance measures on urban arterials is explored. A data mining approach has been developed to retrieve the traffic speed from the Google/Waze application. The study compares the traffic speed data from Google/Waze against the traffic data collected from the on-road vehicle detection stations. The preliminary experiments have shown that the application of social traffic data daily traffic operations and transportation planning is very promising.

Key words: Social Traffic Data, Data Mining, Roadway Performance Measures, Traffic Monitoring

INTRODUCTION

Due to the ever-increasing demand of traffic, mitigation of roadway congestion has become more and more challenging to the transportation researchers and practitioners. Traditionally, people tackle the congestion issue by investing in infrastructure such as constructing new roads and bridges to increase the roadway capacity. With advancement of intelligent transportation systems (ITS), application of high technology has become new strategy for relieving the traffic congestion. The fundamentals of intelligent transportation systems are collection, processing and analysis of real-time traffic data for monitoring roadway conditions and improving traffic management and operations.

In the recent two decades, a wide range of technologies have been developed for real-time collecting traffic data [1]. The common traffic data detectors deployed on freeways and urban arterials include induction loop detectors, infrared detectors, radar detectors, video processing detectors and geo-magnet based detectors. The various detectors are used to generate vehicle counts, vehicle classifications, and traffic speed and roadway occupancy [2]. Each type of detectors has its own advantages and disadvantages. For all these detectors, one major disadvantage is their maintenance which is often costly and complicated. With the advancement of the wireless communication and the global positioning system (GPS) technology, transportation researchers ([3], [4]) explored the usage of the GPS technology for collection of travel time and speed data. More recently, in their explorative work, Valerio at al [5] estimate the road traffic condition on top of the cellular network infrastructure. As of today, the GPS based technology has been widely applied for collecting travel time data on freeways.

Based on the traffic data collected on the freeways and arterials, performance measures can be established to monitor the effectiveness of operational strategies and to assess the success of achieving operational efficiencies [6]. One common performance measure is Level of Service (LOS). In the Highway Capacity Manuals (version 2010) [7], highways are categorized into three different classes which are Class I, Class II and Class III. LOS was established on the basis of these classes by using different parameters, namely, average travel speed, percentage time delay and percent of free flow speed. Six Level of Service classes (LOS A to F) are calculated in that LOS *A* means free flow and LOS *F* means forced or breakdown flow. In addition to the Level of Services, other quantitative indexes can be used as roadway performance measures [8]. These indexes include: (1) Volume/Capacity ratio; (2) Average travel speed and its reliability; (3) Duration of Congestion. Roadways of different classes may use different performance measures to monitor its efficiency.

The Washington D.C. Department of Transportation (DDOT) plans, designs, builds, maintains and operates the transportation infrastructure within Washington, DC. Washington, DC is primarily an arterial system; less than one percent of roadway mileage is freeways. This implies that the efficiency of the transportation system is primarily dictated by the reliability and efficiency of the signal system.

In the recent years, the District Department of Transportation has expanded intelligent transportation infrastructure for collecting traffic data. Expansion of the data detection system on arterials will be very costly and take long time to complete. In addition, maintenance of data detection stations has been an issue in the District. However, the emerging social traffic data such as the Waze data has potential in providing a low-cost alternative solution to meeting the gap.

Waze [9] is a GPS-based geographical navigation application program for smartphones which provides user-submitted travel times and route details. Through the online map interface, the drivers can report accidents, traffic jams, speed and police traps. The speed data is derived by tracking the users' GPS coordinates on the roads. Different from traditional GPS navigation software, the Waze data is community-driven, gathering complementary map data and traffic information from its users. The Waze data includes real-time incident information and travel times from individual drivers. Recently the transportation researchers started using the Waze data for improving incident management. For instance, Fire et. al. [10] trolled the Waze data and plotted traffic accident patterns to determine the hot spots in the city. The results of their research are valuable in optimizing the deployment of police force. Transportation practitioners in Boston have also explored the possibility of using the Waze data to improve their signal operations [11]. Derived from Waze, a specific Participatory sensor network (PSN) is proposed in [12] for sensing traffic conditions and understanding of city dynamics and the urban behavioral patterns of their inhabitants. Iovanovici et. al. [13] focused on utilizing Waze data to monitor the state of roads in urban environments over a large period of time to build a traffic map. Inspired by the Waze driving app, Martelaro et. al. [14] developed a system to measure driver's situation awareness through real time on-road event questions. In the aspect of privacy and authenticity, Jeske et. al [15] evaluated the Google and Waze protocol regarding privacy and authenticity, and proposed a solution that increases the user's privacy and at the same time prevents attacks manipulating the traffic analysis. Brown et. al. [16] proposed a protocol for traffic-update mobile applications that supports the creation of traffic statistics from Waze reports while protecting the privacy of the users. As of today, it is estimated that there are nearly 500 thousands Waze users in the Washington DC metropolitan area. With fast-increasing number of users, the Waze database has accumulated a great amount of traffic data that have potential in using for the traffic management purpose.

The application of social traffic data for traffic monitoring is consistent to the "Asset Lite" Concept that has been adopted by DDOT recently. "Asset lite" solutions refer to strategies geared towards getting to a desired outcome using fewer assets. DDOT has applied the asset lite concept for real-time estimation of parking availability information. Similarly, for travel time information on a transportation system or a segment of roadway, a certain number of assets based on the desired outcome (say accuracy) are required. However, if different data elements are able to be added into the estimation process, there is a likelihood of getting to the same level of accuracy with fewer assets. The accuracy (or outcome gap) is filled with other available data sources. The cost effectiveness stems from having to procure, install and maintain fewer assets.

This paper looks into the possibility of using social media data to fill the outcome gap. A comprehensive approach has been developed to analyze the Waze data for real-time traffic monitoring. A performance measure was developed based on the Waze data for evaluating the major arterials in the District. The rest sections of this paper is organized as below: the next section describes a data mining approach for retrieve the crowdsourced Waze data, followed by a section for numeric experiment on a major arterial in the District of Columbia. The conclusions are drawn in the last section.

A DATA-MINING APPROACH FOR RETRIEVING AND ANALYZING WAZE DATA

The crowdsourced Waze data consists of two parts: the Waze traffic incident alerts data and the Waze traffic jam data. The Waze traffic incident alert data contains all traffic incident information reported by Waze users through the Waze mobile application. The Waze application automatically generates a score of reliability, which is from level 1 to level 10, for each incident alert report. The score increases when more users report the same incident. The Waze traffic jam data source reports real-time traffic slowdown in specific road segments based on data the system gather online. Waze generates its traffic jam information by processing the following data sources: (1) GPS location points sent from user phone; (2) the actual speed vs. average speed; and (3) free flow speed. One the map, traffic jams is represented by a color-coded polyline string which indicates the roadway segment(s) suffering from congestion. A level of congestion from 0 to 5 is created for each report roadway segments. The level of congestion increases if the length of polyline string increases. The data types for the Waze traffic jam report is described in Table 1:

Table 1: Data Types for the Waze Traffic Jam Report

Field	Description	Features	Description
Time	Time when the traffic jam data is report	Length	Length of roadway segment (specified as polyline) in meters
Delay	Delay of jam compared to free flow speed, in seconds	Line	Traffic jam line with starting and end points
Level	Traffic congestion level (0 = free flow 5 = blocked)	Speed	Current average speed on the target segments in meter/second

A program was developed to retrieve the Waze congestion data and save it to the MongoDB [17] database. The retrieved Waze data is stored under a tree structure according to the MongoDB database schema. Each report in the database has an associated "id" with a sub-node called "updates" storing the history of the records evolution. The raw Waze traffic data is updated every 5 minutes.

Let $\boldsymbol{S}$ denote the roadway route for the study. $\boldsymbol{S}$ can be divided into several relatively-short roadway segments s_i, i=1, 2, …, N. The lengths of the segments are defined to be in the range of 0.3– 0.8 miles. Let $P(d,t)$ be the set of all Waze polylines for reporting travel speed on date d at time t. For each Waze polyline $p \in P(d,t)$, there is space-mean speed $v_{\mathrm{p}}(d,t)$ reported by the Waze users. Let $Q_i(d,t)$ be the subset of $P(d,t)$ in which the Waze polylines overlap the roadway segment s_i. The average speed on roadway segment s_i on date *d* at time *t*, denoted as $v_i(d,t)$, can be estimated as:

$$v_i(d,t) = \begin{cases} null & if\ Q_i(d,t) is\ empty \\ \frac{\sum_{p \in Q_i(d,t)} v_p(d,t)}{N_p(d,t)} & Otherwise \end{cases} \quad \text{Eq (1)}$$

Where $N_p(d,t)$ is the number of elements in the set $Q_i(d,t)$. In a typical scenario, roadway segment s_i only overlaps one polyline *p*, then $v_i(d,t) = v_p(d,t)$.

Three types of days are categorized: regular working days, weekends and special days. Let D_w be set of all regular working days within the data collection period when the $v_i(d,t)$ is not null. The mean and variance of travel speed at time t on a regular working day can be estimated in Eqs (2) and (3), respectively:

$$\overline{V_i}(t) = \frac{1}{N_w} \sum_{d \in D_w} v_i(d,t) \quad \text{Eq (2)}$$

$$Var\big(V_i(t)\big) = \frac{1}{N_w - 1} \sum_{d \in D_w} [v_i(d,t) - \overline{V_i}(t)]^2 \quad \text{Eq (3)}$$

Based on the Waze data, the traffic flow profile for each roadway segment can be calculated by calculating the mean value and the variance of the average speed at each time point for a regular working day. It is also useful to calculate the 80^{th} percentile of the segment travel speed $\hat{V}_i(t)$ for each segment s_i and time *t*. The 80^{th} percentile of the travel speed is more appropriate than the average speed in capturing the traffic congestion on arterials where the variances of travel speeds are big.

With the estimated travel speed for each roadway segment in place, the travel speed for the entire route can be easily calculated as:

$$\hat{V}(t) = L / (\sum_i \frac{L_i}{\hat{V}_i(t)}) \quad \text{Eq (4)}$$

Where *L* is the total length of the route and L_i is the length of the individual roadway segment s_i.

ANALYSIS OF WAZE DATA ON TWO CORRIDORS

As described in above sections, the data obtained from Waze is polyline with average speed and delay attached to each line. The westbound New York Avenue corridor is one the most congested roadway in the nation's capital city. The selected corridor is between Bladensburg and North Capitol St NW. New York Ave is the only corridor in DC where DDOT has installed travel time collection system which can be used as ground truth. New York Ave is divided in to three segments as shown in Figure 1:

a. Bladensburg Rd NE to Kendall St NE
b. Kendall St NE to 4th St NE

c. 4th St NE to North Capitol St NW

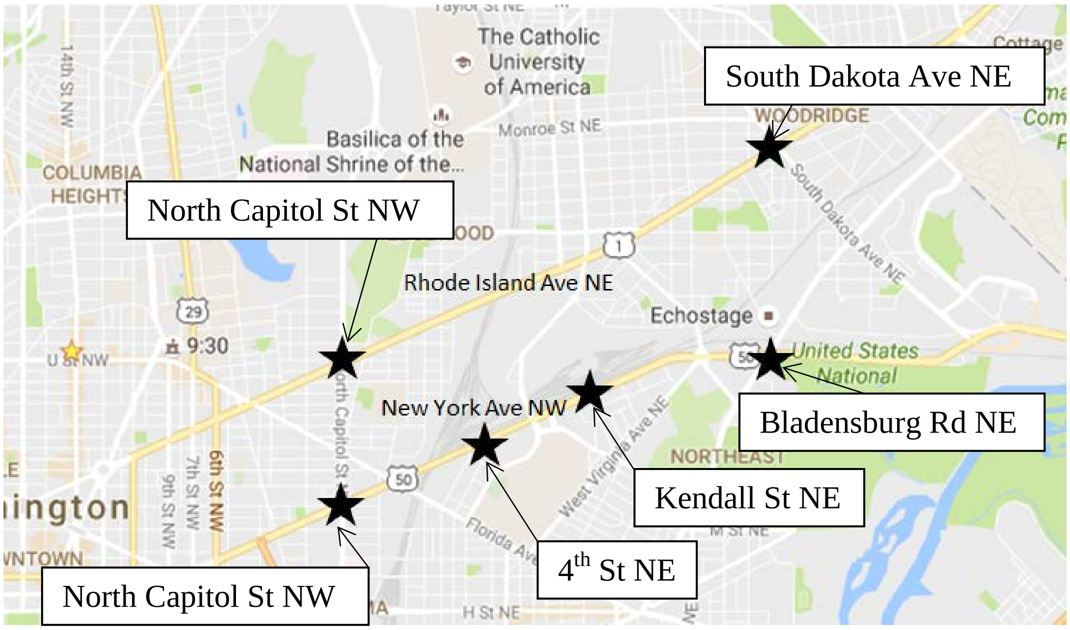

Figure 1: Analysis on New York Avenue (West Bound) and Rhode Island Avenue

Apart from westbound New York Avenue NW, both the eastbound and westbound of Rhode Island Avenue corridor between South Dakota Avenue NE and North Capitol St NW are chosen, as shown in Figure 1.

Data from each Waze record that overlaps with the above segments is transferred to corresponding segment. There are totally 3 scenarios the overlap can take place which are shown in Figure 2 below. In scenario A and B, the average speed of Waze record is mapped to roadway segment and in scenario C, an average of the speeds is taken and mapped to the segment.

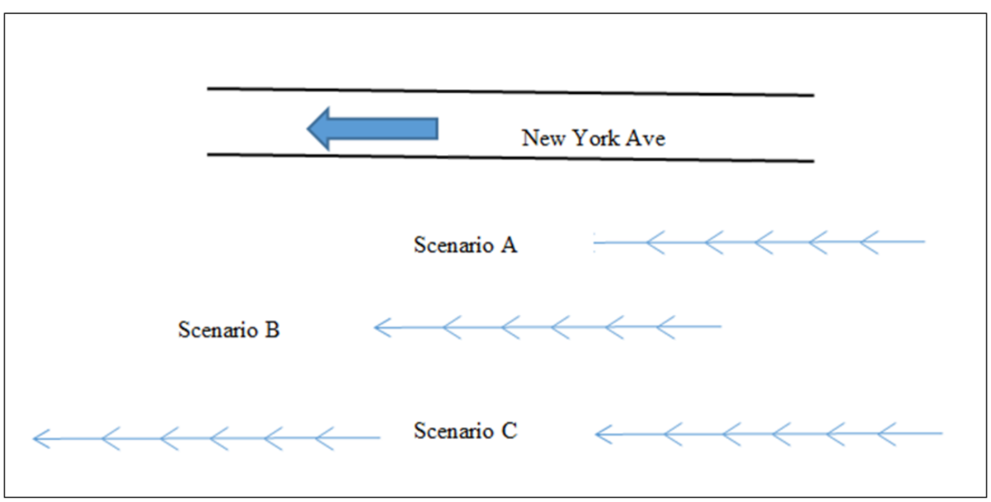

Figure 2: Different Scenarios of Waze Records

Space-mean travel speed is used as performance measures on arterials. Based on the Waze data collected from March, 2015 to June, 2016, the developed approach is applied to calculate the average travel speed for a regular working day. Table 2 shows the travel speed records on each road segment in the Waze data.

Table 2: The number of records on each road segment

Street	Direction	Road Segment	#Records
New York Ave NW	Westbound	Bladensburg Rd NE to Kendall St NE	1258
New York Ave NW	Westbound	Kendall St NE to 4th St NE	2184
New York Ave NW	Westbound	4th St NE to North Capitol St NW	10887
Rhode Island Ave NE	Eastbound	North Capitol St NW to South Dakota Ave NE	19236
Rhode Island Ave NE	Westbound	South Dakota Ave NE to North Capitol St NW	11058

Ground Truth Data

After Waze Data is collected and refined, for each road segment, its travel speed data with the results from the ground truth data. In this paper, two ground truth data sources are considered, one is the Vehicle Detection System (VDS), and the other is the InRix System [18]

The VDS system generates the travel time on the entire corridor. The travel speed $V_{vds}(d,t)$ on date d at time t for road segment S_i can be estimated as:

$$V_{vds}(d,t) = L/T_{d,t} \quad \text{Eq (5)}$$

where L is the length of the corridor and $T_{d,t}$is the travel time on the corridor at time interval t on date *d*. For a regular working day, the 80th percentile of travel speed is calculated, denoted as $V_{vds}(t)$ for the corridor for comparison. Figure 3 shows the comparison results between Waze travel speeds and VDS travel speeds on New York Avenue NW westbound from Kendall St NE to 4th St NE.

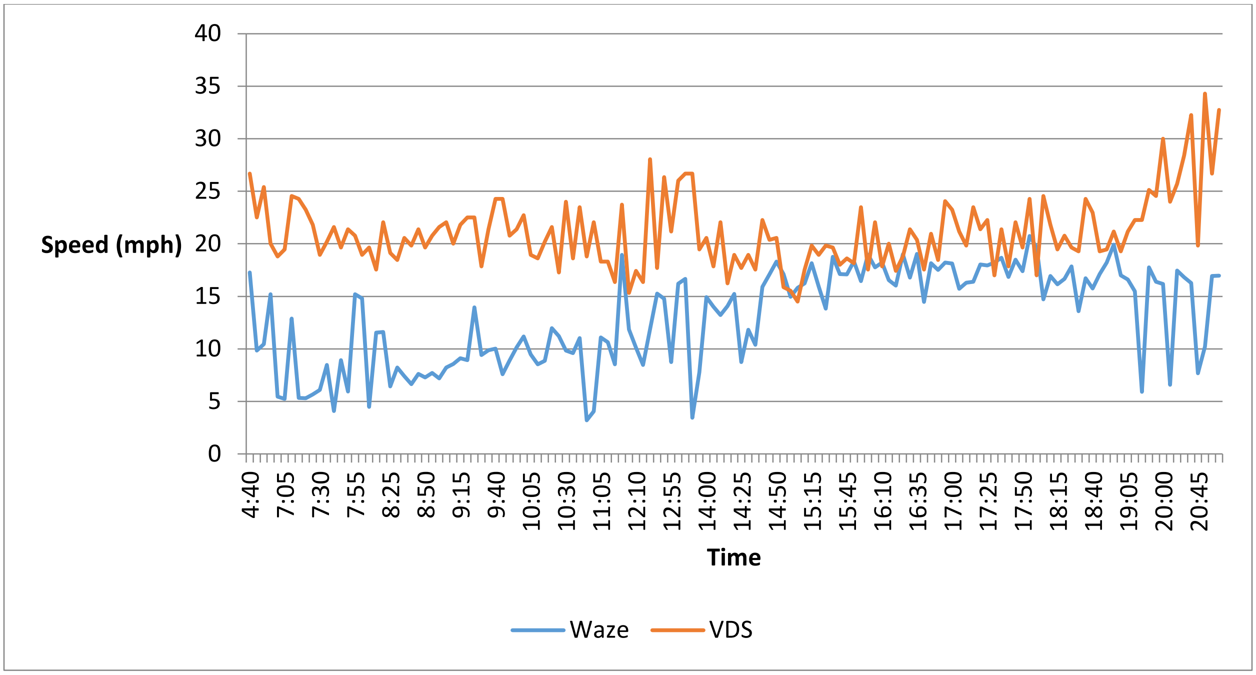

Figure 3: Comparison of Travel Speed on New York Avenue NW westbound between Kendall St NE and 4th St NE

The traffic pattern on this westbound New York Ave corridor is different from other typical roadways where there exist double peak periods, one in the morning and another in the afternoon. Figure 3 shows that vehicles move at free flow speeds on westbound New York Avenue before 6:00am in the morning and after 7:30pm in the evening. From 6:00am to 7:30pm, the pattern of travel speed data from the Waze application is consistent with that from the VDS system. The travel speed data from the VDS system is a direct measure of the vehicle movements that can be used as ground truth. Compared to the VDS data, the travel speed from the Waze database is substantially underestimated and suffers from different sampling bias at various time periods. To mitigate the effect of sampling bias, morning rush hour and afternoon rush hour are considered separately for developing bias calculation, which is discussed in the next section.

For Rhode Island Avenue, InRix data is procured as DDOT does not have sensor along the corridor. Similar to New York Ave, 80th percentile of the travel speed is used in this analysis.

Mitigate Sampling Bias

It is noted that in Waze data, compared with congestion period, fewer travel speeds are reported during free flow period. In statistics, this is issue of sampling bias. In this section, a comparison will be made between the estimated travel speed based on the Waze data and the measured ground truth travel speed data from VDS and InRix. Using the VDS and InRix data as ground truth, a correction factor is calculated to improve the estimation.

Compared to ground truth data, as shown in Figure 3, the travel speed from the Waze database is substantially underestimated. To mitigate this effect, a correction factor b is introduced to improve the Waze-derived travel speeds. The new estimates of travel speeds can be computed as:

$$\tilde{V}(t) = \hat{V}(t) + b \qquad \text{Eq(6)}$$

Where $\tilde{V}(t)$ is the corrected travel speed at time t and $\hat{V}(t)$ is the travel speed from Waze data at time t. In this paper, the following empirical formula is used to evaluate b:

$$b = \frac{\sum_{t\in T}(V_g(t) - \hat{V}(t))}{N_T} \qquad \text{Eq(7)}$$

Where $V_g(t)$ is the ground truth travel speed at time t, T is the set of total time points during a time period and N_T is the size of set T. It is noted that when more people report travel speeds to Waze, the correction factor b will become smaller.

Analysis of Road Segments on New York Ave NW Westbound

In this section, the Waze-corrected speeds are compared with the ground truth data on road segment level. The morning rush hour is defined from 6:00am to 9:30am during a typical working day, while afternoon rush hour is defined from 3:00pm to 7:30 pm. For each rush hour, the corrected speed at every time point during the rush hour is calculated by applying Eq(6).

For each road segment on New York Ave NW, its correction factor b in the morning rush hour and afternoon rush hour is calculated respectively, and the Mean-Square-Error (MSE) between corrected speeds and ground truth is also calculated, as shown in Table 3.

Table 3: Correction factor b and MSE on each road segment

Time Period	Road Segment	b	MSE
6:00am to 9:30am	WB Bladensburg to Kendall St NE	9.42	3.76
	WB 4th St NE to Capitol St NW	10.71	2.79
	WB Kendall St NE to 4th St NE	12.27	3.57
3:00pm to 7:30pm	WB Bladensburg to Kendall St NE	11.07	3.21
	WB 4th St NE to Capitol St NW	11.68	3.02
	WB Kendall St NE to 4th St NE	3.03	3.39

Figure 4 (a) and (b) show the Comparison of Waze-corrected speeds with VDS speeds on New York Ave NW Westbound from Kendall St NE to 4th St NE. It is noticed that the corrected speeds fit the ground truth value of travel speeds pretty well.

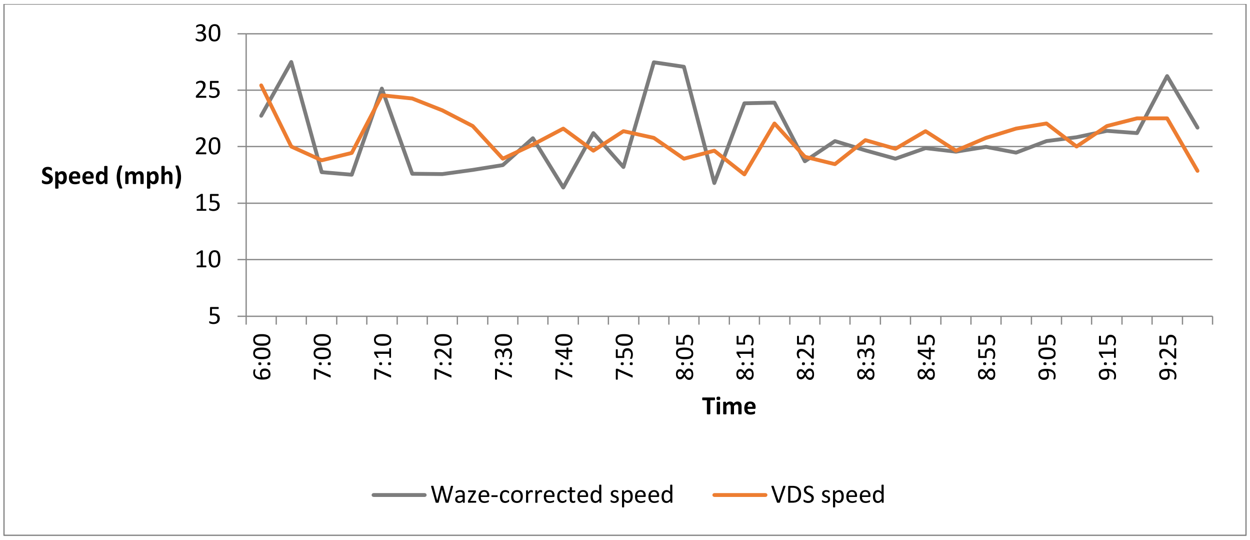

Figure 4 (a) Morning traffic patterns on New York Ave NW Westbound from Kendall St NE to 4th St NE

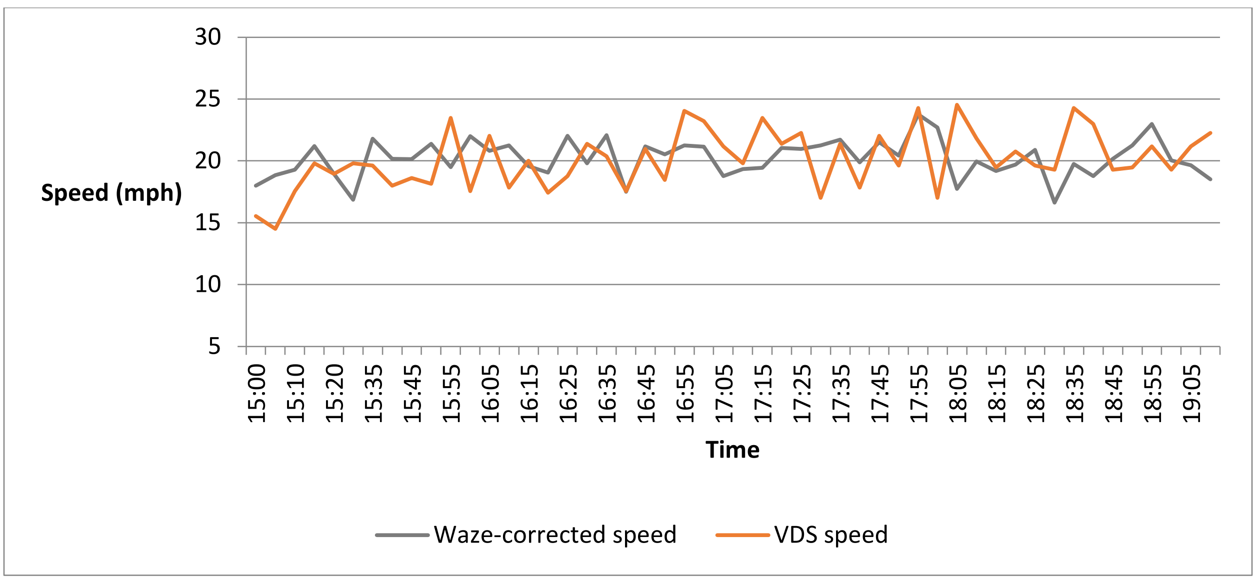

Figure 4 (b) Afternoon traffic patterns on New York Ave NW Westbound from Kendall St NE to 4th St NE

Analysis of Entire Corridors

In this section, the Waze-corrected speeds are compared with the ground truth data on the entire corridor. The morning rush hour and afternoon rush hour are considered separately and the corrected speeds are calculated. For each corridor, its correction factor b in the morning rush hour and afternoon rush hour is calculated respectively, and the Mean-Square-Error (MSE) between corrected speeds and ground truth speeds is also calculated, as shown in Table 4.

Table 4: Correction factor b and MSE on each corridor

Time Period	Road Segment	b	MSE
6:00am to 9:30am	New York Ave NW Westbound	5.38	1.33
	Rhode Island Ave NE Westbound	19.09	0.81
	Rhode Island Ave NE Eastbound	12.41	1.06
3:00pm to 7:30pm	New York Ave NW Westbound	9.67	2.19
	Rhode Island Ave NE Westbound	10.62	1.61
	Rhode Island Ave NE Eastbound	10.86	1.08

The comparison of the Waze speed data and the VDS data on New York Ave NW Westbound is shown in Figure 5 (a) and (b).

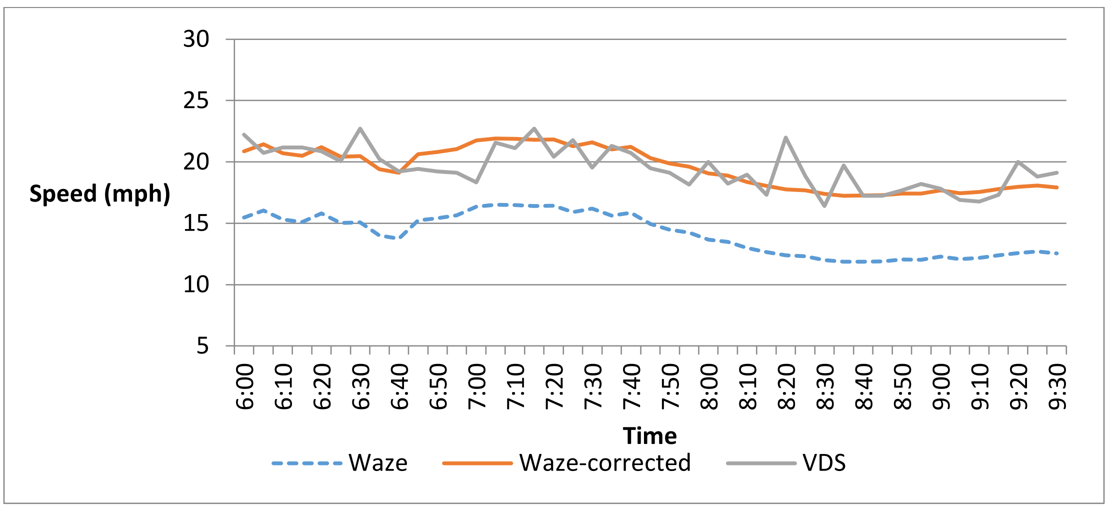

Figure 5 (a) Morning traffic patterns on New York Ave NW Westbound

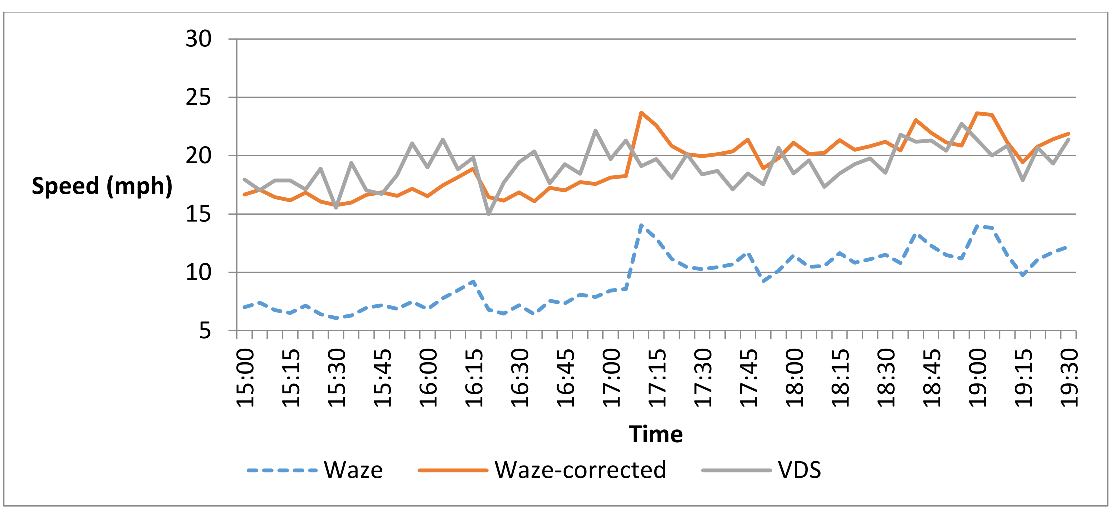

Figure 5 (b) Afternoon traffic patterns on New York Ave NW Westbound

Since the VDS data is not available on Rhode Island Avenue, the InRix data is used for comparison. The results are shown in Figures 6 (a) and (b):

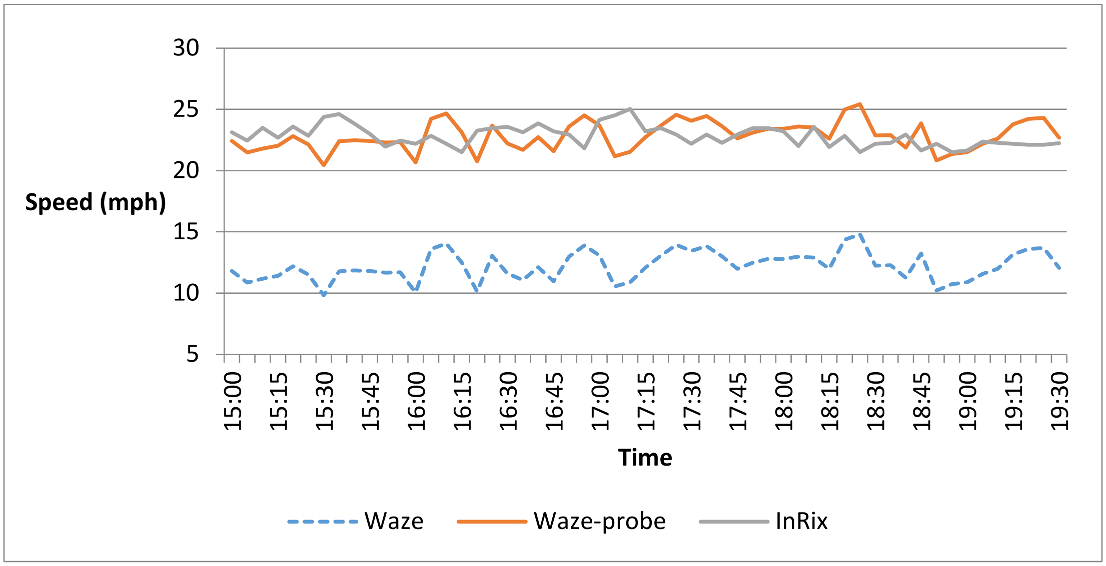

Figure 6 (a) Afternoon traffic patterns on Westbound Rhode Island Avenue NE

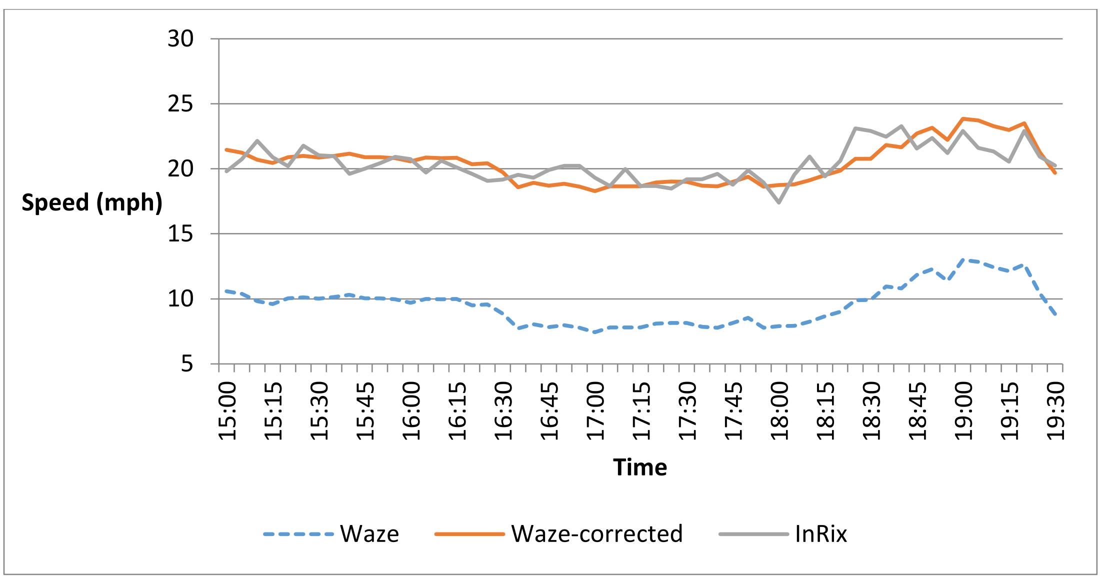

Figure 6 (b) Afternoon traffic patterns on Eastbound Rhode Island Avenue NE

Figure 5 and 6 show that the corrected average speeds fit the ground truth value of traffic speeds pretty well. The results of analysis show that the derived speed data from the Waze application is able to provide additional layer for understanding the traffic patterns and evaluating the effectiveness of the arterials.

CONCLUSIONS

Applying the social traffic data to evaluate the effectiveness of roadways is a new concept in traffic management. The paper explored the possibility of using the Waze data to evaluate the traffic conditions on roadways in Washington DC. A data-mining approach has been developed to derive average travel speeds from the crowdsourced Waze database. The results from the approach were compared with the data from the VDS system and the InRix system on the same roadways. The analysis has shown the Waze data presents a promising low-cost data source for understanding the traffic patterns and evaluating roadway effectiveness. Since the average speed data derived from the Waze application is usually a biased estimate of the true speeds, further research is needed to improve the accuracy of the estimation. Even though the analysis is performed on two corridors, the developed approach can be extended to other critical corridors. As more and more drivers use the Waze application, it is expected the amount of traffic information from the Waze users will exponentially increases in the coming years. The social traffic data will have more application in the field of traffic operations and incident management.

REFERENCES:

[1] "Traffic Data Collection and its Standardization", Springer, 2010, ISBN 978-1-4419-6070-2

[2] U.S. Department of Transportation, Federal Highway Administration, Office of Highway Policy Information, Highway *Performance Monitoring System Field Manual*. March 2012.

[3] Cesar A Quiroga, "Performance measures and data requirements for congestion management systems", Transportation Research Part C: Emerging Technologies, Volume 8, Issues 1–6, February–December 2000, Pages 287–306.

[4] Yim, Y., B., Y., Cayford, R., Investigation of Vehicles as Probes Using Global Positioning System and Cellular Phone Tracking: Field Operational Tests, California Partners for Advanced Transit and Highways (PATH), University of California, Berkeley, 2001.

[5] Danilo Valerio, Witek T, Ricciato F., Pilz R. and Wiedermann W. "Road traffic estimation from cellular network monitoring: a hands-on investigation". In: IEEE Personal Indoor Mobile Radio Communication Simposium, IEEE PIMRC 2009, September 2009

[6] Dowling, Richard. "NCHRP Report 616: Multimodal Level of Service Analysis for Urban Street", Transportation Research Board, January 2012.

[7] Highway Capacity Manual (2010). Fifth Edition, Transportation Research Board of National Academics, Washington, D.C., 2010

[8] Transportation Research Board,"NCHRP SYNTHESIS 311 Performance Measures of Operational Effectiveness for Highway Segments and Systems: A Synthesis of Highway Practice", 2003.

[9] Waze Website, https://www.waze.com/

[10] Michael Fire, D. Kagan, R. Puzis, L. Rokach and Y. Elovici, "Data Mining Opportunities in Geosocial Networks for Improving Road Safety," IEEE 27th Convention of Electrical and Electronics Engineers in Israel, 2012.

[11] Barb Darrow, "Boston is using big data to solve traffic jams", in the Fortune website: http://fortune.com/2015/05/21/boston-is-using-big-data-to-solve-traffic-jams/, May 21, 2015.

[12] Silva, Thiago H., et al. "Traffic condition is more than colored lines on a map: characterization of waze alerts." International Conference on Social Informatics. Springer International Publishing, 2013.

[13] Iovanovici, Alexandru, Lucian Prodan, and Mircea Vladutiu. "Collaborative environment for road traffic monitoring." ITS Telecommunications (ITST), 2013 13th International Conference on. IEEE, 2013.

[14] Martelaro, Nikolas, David Sirkin, and Wendy Ju. "DAZE: a real-time situation awareness measurement tool for driving." Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, 2015.

[15] Jeske, Tobias. "Floating car data from smartphones: What google and waze know about you and how hackers can control traffic." Proc. of the BlackHat Europe (2013): 1-12.

[16] Brown, Joshua WS, Olga Ohrimenko, and Roberto Tamassia. "Haze: privacy-preserving real-time traffic statistics." Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013.

[17] MongoDB Website, https://www.mongodb.com/

[18] InRix System Website, http://inrix.com/