Comparative analysis of travel time prediction algorithms for urban arterials using Wi-Fi Sensor Data

Smit Thakkar
Research Scholar
School of Civil & Environmental Engineering
Queensland University of Technology
Brisbane, Australia
Smitbharat.thakkar@hdr.qut.edu.au

Shubham Sharma
Research Scholar
School of Civil & Environmental Engineering
Queensland University of Technology
Brisbane, Australia
shubham.sojls@gmail.com

Chintan Advani
Research Scholar
School of Civil & Environmental Engineering
Queensland University of Technology
Brisbane, Australia
chintan.advani@gmail.com

Shriniwas S Arkatkar
Associate Professor
Department of Civil Engineering
Sardar Vallabhbhai National Institute of Technology
Surat, India
sarkatkar@gmail.com

Ashish Bhaskar
Associate Professor
School of Civil and Environmental Engineering
Queensland University of Technology
Brisbane, Australia
ashish.bhaskar@qut.edu.au

Abstract— **Travel time is one of the elementary traffic stream parameters in both users' and transport planners' perspective. Conventional travel time estimation methods have performed out of sorts for Indian urban traffic conditions characterized by heterogeneity in transport modes and lack of lane discipline. Robust to these limitations, Media Access Control (MAC) matching is perceived to be a reliable alternative for travel time estimation. To assist with real-time traffic control strategies, this study aims at developing a reliable structure for forecasting travel time on Indian urban arterials using data from Wi-Fi/ Bluetooth sensors. The data collected on an urban arterial in Chennai has been used as a case study to explain the value of such data and to explore its applicability in implementing various prediction models. To this end, this study examines and compares three different machine learning algorithms k-Nearest Neighbour (k-NN), Random Forest (RDF), Naïve Bayes, and Kalman filtering technique for prediction. The performance of each model is evaluated to understand its suitability.**

Keywords— *Travel time prediction, Wi-Fi sensors,* Media Access Control, k-NN, Random Forest, Naïve Bayes, and Kalman filter

I. INTRODUCTION

Travel Time prediction is an important yet challenging task in Intelligent Transportation System Operation and Management. These predicted Travel Time/Speed profile serves as a primary input for short-term operational planning including traffic control design and adjustments, designing congestion calming measures, ramp metering, etc, and long-term strategic planning. In the last few decades, extensive work is reported on travel time prediction for freeways, urban arterials, and other signalized roads exploring a wide range of prediction tools. These prediction methodologies can be broadly categorized into model-based and data-driven approaches, exploring the application of traditional traffic flow theory, time series, and machine learning models. The model-based approaches use traffic dynamics, segment capacity, and demand estimation as model inputs, making predictions more complex and require maintenance over time. E.g. OLSM, TOPL (CTM), SBOTTP (CORSIM), DyanMIT-R, etc [1]. On the other hand, data-driven approaches [2] use the historical database and relate the existing traffic state to the most similar historic patterns of traffic parameters (traffic volume, flow, speed, etc.).Based on the pattern analysis, the historic database can be clustered into different groups like different hours of a day (peak hours and off-peak hours) and different types of days (weekdays, weekends, and special occasions). These approaches do not require traffic flow theories as inputs and make predictions using the database itself through machine learning and statistical tools. One of the major limitations of these data-driven approaches is their region-specific nature.

For urban arterials, several data-driven approaches have been used for the prediction over time like average speed techniques, linear regression, step-wise linear regression methods, linear and non-linear time-series analysis techniques, Autoregressive Moving Average Method (ARMA) [3], Autoregressive Integrated Moving Average Method (ARIMA) [3, 4], generalized autoregressive conditional heteroscedasticity model [5], and more. But they are not suitable for modelling non-linear relationships between traffic variables. Non-linear relationships can be modelled using other data-driven approaches, mostly machine learning techniques, such as Random Forest (RDF), Support Vector Machines (SVM) [6], Markov Chains, k Nearest Neighbors (k-NN), Naive Bayes classifiers, artificial neural networks (ANN) and are extensively explored in past. Another popular technique, the Kalman filtering approach and its variants, due to their ability to integrate the potential of data-driven and model-based approaches, have gained enormous attention for short-term real-time traffic state prediction. They estimate unknown variables by computing a joint probability distribution of the variables for each timeframe over a time series of measurements containing

statistical noise and tend to be accurate than a single measurement-based estimate.

The deployment of the data-driven prediction techniques also depends on the availability and quality of the data collection system. The technologies for collecting traffic data have evolved with time, from manual counting to smart sensor-based technologies like Magnetic loop detectors, Location-based GPS-tracking, image processing, etc. The literature available on travel time prediction for Indian-traffic conditions mostly uses data from GPS devices installed inside transit vehicles and hence, the data collected is specific to one mode and represents a small portion of the traffic stream [7-9].

Another passive and cost-effective data collection technique that has gained a lot of attention in recent times is tracking the Spatio-temporal movement of mobile devices via MAC-id reidentification using Bluetooth/Wi-Fi sensors [10]. Travel time measurements from these sensors do not depend on travel mode or the traffic lane discipline. This study uses Wi-Fi sensors for data collection on urban arterials. The motivation for using Wi-Fi sensors over Bluetooth has two reasons. Firstly, changes in the smartphone security features have reduced Bluetooth penetration because, in addition to turning Bluetooth on, they must also be put in the detectable mode. On the other hand, Wi-Fi only needs to be switched-on on mobile devices for allowing the corresponding access to the sensors.

Secondly, it is expected that the smart city mission of the Government of India, which promotes citywide Wi-Fi connectivity will encourage the use of Wi-Fi services and would help in achieving better matching rates *(terminology explained in Section V.A)* in the future. In this study, we find that the Wi-Fi sensors gives a slightly better sampling rate compared to Bluetooth sensors by comparing the penetration and matching rates of these two types of sensors on the same study section reported in a previous study [11, 12]. The high sampling rate and penetration into diverse travel modes also substantiate the choice of Wi-fi sensors over GPS-based travel time measurements for prediction studies.

The objectives of this paper can be summarized in three points. Firstly, to understand the Wi-Fi-based technology for the data collection on urban arterials from the viewpoint of heterogeneous traffic. Secondly, finding an optimal aggregation interval for estimating and predicting travel time and finally, analyzing the applicability of some popularly used data-driven travel time prediction algorithms on two different site conditions of different nature. The study concludes with detailed insight into the observed variation in the outcomes of each model and highlights the value of each model as opposed to the other.

II. Study Area, Data Collection

The study section selected is a segment of Rajiv Gandhi IT Expressway, Chennai, an urban arterial with service lanes on both sides, traced between First Foot-over Bridge (FOB1, 13.0039080º, 80.2474440º) and Tidel-Park Intersection (12.9876080º, 80.2513980º). The adjoining area to the study section is a hub of the IT sector and educational institutions. 2nd Avenue, a collector street meets the study section at (12.9945760º, 80.2499330º) at 1.1 km from FOB1. Except for this, no potential ingress or egress point is meeting the section directly. Wi-Fi sensors were deployed at a median of the road at three locations along the stretch as shown in Figure 1. Sensors used were off the shelf sensors used with 4 dBi antennas and were power operated (1A, 5V).

The study was carried out on two segments, the first being FOB 1 to 2nd Avenue (1.1 Km) and the second from 2nd Avenue to Tidel-Park (0.7 Km). Segment 1 has one end as midblock and the other end is a signalized intersection whereas segment 2 consists of signalized intersections at both ends, hence both routes are different in their configuration.

Wi-Fi data was collected for 40 days classified into 4 phases out of which the first three phases have been used to create a Historical Database. Each phase consists of data for the weekend (Saturday, Sunday) and weekdays. Using the MAC IDs matching technique, travel times were extracted by taking the difference of the first logged timestamp at both the upstream and downstream ends of the segment equipped with sensors.

Apart from that, videography was carried for 1 hour at the mid-block section (FOB 1) and signalized intersection (2nd Avenue) to relate the sample size from MAC addresses to the actual traffic volume.

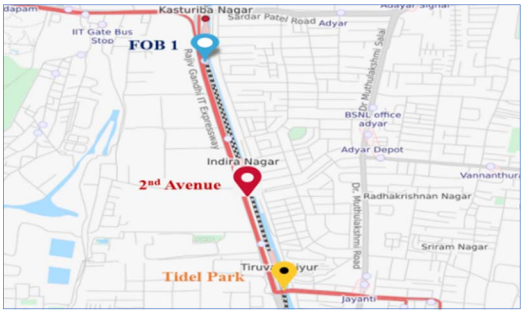

Figure 1: Study Sections

III. Data Preprocessing

Before analysis, collected data need to be treated, to eliminate the potential outliers due to devices associated with pedestrians, stationary devices, and other multiple matches. 3-stage filtering was adopted to eliminate outliers from the travel time dataset.

Stage 1- Minimum travel time based on the posted speed limit.

Stage 2- Maximum possible travel time based on pedestrian speed. The travel time value above this cannot be distinguished whether coming from a pedestrian crossing two sensor locations or a device associated with a vehicle.

Stage 3- Modified Z score filter [11], using threshold modified Z score value as 3.5.

The Modified Z score is given as follows:

$$M_i = \frac{0.6745(x - \widetilde{x})}{MAD}$$

Where:

MAD: Median Absolute Deviation & $\tilde{x}$: Median.

After eliminating outliers, the filtered travel time data was further analyzed to understand within-day and across-days patterns before training the prediction models.

IV. PREDICTION METHODOLOGIES USED

A. Kalman Filtering Technique

An optimum estimation algorithm predicts the parameter of interest in the presence of noisy measurements and has found a vast range of applications, mostly when the variables of interest can only be measured indirectly. The execution of the Kalman Filtering technique needs information concerning the system's dynamics, statistical data of system instabilities, and measurement error [13]. Let travel time evolution over intervals for a given segment is assumed to define system state S(t):

$$y(t) = \mathrm{A(t)}\, y(t-1) + w(t)$$

Here, y(t) represents travel time at time t, w(t) is a random error representative of the process error with a normal probability distribution, zero mean and covariance Q. A(t) is the Transition Function at time t. Measurement m(t)∈R is related to the state variable y(k) by:

$$\mathrm{m(t) = C\ y(t)+v(t)}$$

Where m(t) is the measured travel time for the segment under consideration at any time t and v(t) represents the Gaussian noise in the measurement with zero mean and covariance R. C was considered unity for single state prediction variable. The algorithm involves two datasets (D1 and D2) to execute the algorithm structure. One dataset (D1) out of two was used to compute the Transition function A(t) by time update equations, based on model hypothesis assuming transition function A(t) as the proportion of historic travel times conferring with the recognized patterns in the Historic Database:

$$A(t) = \frac{z\,(t)}{z(t-1)}$$

Where z(t) represents data in the historic database. The other dataset (D2) was utilized in the measurement equations to produce the final estimates. Following is the pseudo-code for the prediction process:

Step 1. Initializing process

Set t=0

A (0) =1

P (0)- = Var [z (0)]

N=Number of Time Intervals

Step 2. Estimating Travel Time and Measurement Error Covariance (P)

$\hat{y}(t)_- = \mathrm{A(t)} * \hat{y}(t-1)_+$

$P(t)_- = A(t-1)P(t-1)_+A^T(t-1) + Q(t)$

Step 3. Calculate Kalman Gain (K)

$K(t) = P(t)_-[P(t)_- + R(t)]^{-1}$

Step 4. Predicting Travel Time

$\hat{y}(t)_+ = \hat{y}(t)_- + K(t)\,[m(t) - \hat{y}(t)_-]$

Step 5. Updating P

$P(t)_+ = [1 - K(t)]\,P(t)_-$

Step 5. If t = N:

Break

Else:

t = t+1

Step 6. Repeat process from step 2 until convergence

B. k-Nearest Neighborhood (K-NN)

k-NN is a pattern recognition algorithm, which is non-parametric and is usually used to assign weights to the neighbouring contributors.[14] Initially, k-NN computes the Euclidean distance with increasing distances, and then the optimal number of k nearest neighbours can be identified by root Mean Square Error (RMSE)**.** Based on the nearest neighbours, the weighted average of the identified multivariate neighbours is computed. The target value is then predicted from the mode of the probable outcomes. Travel time historic data was taken as a training set whereas explanatory variables comprised time of day and no. of matched IDs. In this study, the k parameter considered for the k-NN classifier algorithm was the square root of the number of training patterns/samples.

C. Naive Bayes

Naive Bayes is a construction classifier, which assigns class labels to instances, grouped by a series of vectors to draw the label sets from the finite data sets. In Naive Bayes, the family of algorithms is trained by a certain common principle, known as Bayes' Theorem, and given as:

$$\mathrm{P(a|b) = P(b|a) \times P(a)/\ P(b)}$$

where:

P(a|b): Posterior probability; the probability of a given hypothesis b is true.

P(b): Prior probability of b; the probability of hypothesis a being true.

P(a): The probability of the "a" regardless of b.

The main point of interest in computing the posterior probability of P(a|b) from the prior probability p(a) with p(b) and P(b|a). The hypothesis with maximum probability is chosen after computing the posterior probability for different hypotheses and termed as maximum a posteriori (MAP**)** which is given as.

$$\mathrm{MAP\ (a) = max\{P(b|a) * P(a) / P(b)\}}$$

Based on the disclosed hypothesis with maximum probability, the probability of each class P(a) is back calculated and the class having maximum probability is predicted as the output for that certain dataset.

D. Random Forests (RDF)

Tree bagger (Random forest) is one of the machine learning technique, which operates by assembling the collection of decision trees by predicting the variable class or the mean based on the trained data set.

For example, let X = x_1, ..., x_n are the observations for the variables Y = y_1, ..., y_n, by repeated bagging (B times) for random samples with replacement; b = 1, ..., B. Based on the sampling with n trained observations, the variables are selected for training as X_b, Y_b. From that, training of trees can be carried out based on their nature, such as classification and regression. After training, predictions for a sample x' is made by taking the average predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

V. Analysis and Results

A. Preliminary Analysis

A preliminary analysis of the proportion of Wi-Fi devices detected by a Wi-Fi sensor (called penetration rate) concerning the traffic volume entering the influence zone of a sensor (radius = 35m) and the proportion of getting re-identified at two locations (called as the matching rate) was done to understand the representative sample size.[11] Travel time was extracted using the MAC IDs reidentification technique by taking the difference of the first logged timestamp at two ends of the segments for every MAC device. Outliers from the data were removed using a 3-Stage filter as discussed in Section III. Also, the pattern analysis of travel time over weekdays and weekend were performed to visualize historic patterns in the travel time database.

An average penetration rate of 69.77% was observed which does not, however, explicitly translate into the number of vehicles as this may also include more than one Wi-Fi devices from a single car and devices carried by pedestrians since the location is an intersection with pedestrian crossing facilities. Also, since the study area is a hub of IT industries, the penetration rate might be dominated by devices not associated with vehicles. The penetration rate is found to be fairly higher than the one observed in the Bluetooth sensor study carried on the same arterial in 2016 reporting a penetration of about 10% [12]. The matching rate was calculated by comparing the count of reidentified Wi-Fi devices with the actual number of vehicles using the road segment between two sensor locations. An average Matching Rate of 7-8% was observed in both sections. In a previous study on the same study segment using Bluetooth sensors [12], more varying matching rates in the range of 4-8% were observed.

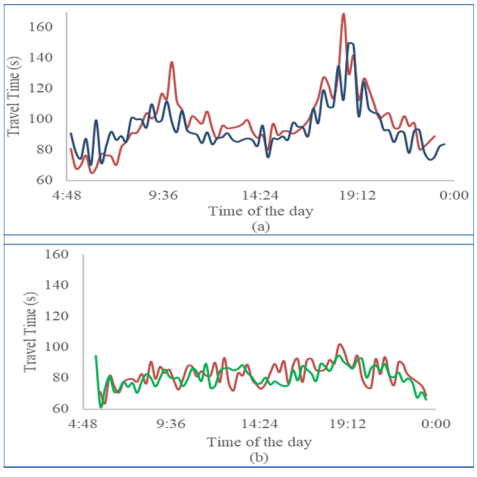

Figure 2: Comparison of Travel time pattern for two (a) Mondays and (b) Sundays for segment 1

Once the travel time values were measured and processed for outlier removal, within-day, and across days travel time trends were extracted and studied. The two weekday and weekend travel time trends for Segment-1 are compared in Figure 2. From the figure, it is evident that two Mondays' travel time profiles follow an overall recurring trend with non-recurring minor fluctuations. Similar across-day patterns were observed for Sundays too. However, within-day trends were found to be different for weekdays and weekends with a dual peaked trend on weekdays and a flatter curve on weekdays with small nonrecurring fluctuations. Similar dynamics were observed for Segment-2 but with higher travel time and sharper peaks due to the presence of signal controls on both ends.

Since the weekday–weekend variation and segment-wise variation are identified to be the most noteworthy, it was decided to investigate each segment for weekday and weekend distinctly. The patterns recognized through this analysis are the basis for identifying historical data which, as discussed in the following section, was utilized as an input for travel time forecasting models.

B. Optimum Aggregation interval

The choice of aggregation interval is usually based on the prediction application, variability within the dataset, and ability of the prediction algorithm to model and handle the within dataset variability.

For real-time traffic control, a smaller aggregation interval is needed. However, this choice is hindered by the small sample size of aggregated data, high susceptibility of getting impacted by outliers. To understand it better, the travel time dataset was examined for within dataset variance using Mean Square Error of Estimate (MSEE) for travel time. Travel time data from sensors were analyzed for different days. Data was aggregated for different time intervals ranging from one minute to one hour for both segment-1 and segment 2, and MSEE values were computed by using the equation:

$$MSEE = \sum_{i=1}^{n} \big(X(h) - xi(h)\big)^2 / V(h)$$

where,

V(h)= Total number of samples in the aggregated dataset,

X(h)= Mean Travel time,

xi(h)= i^{th} Mac Id Travel time.

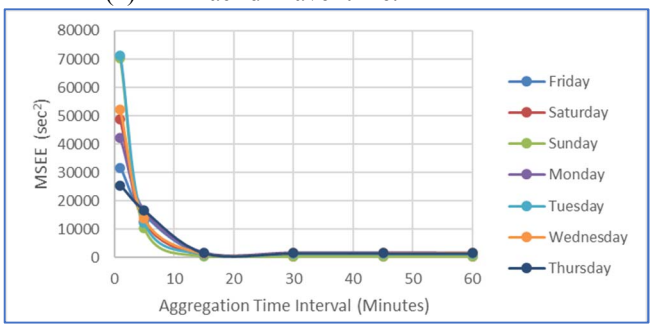

Figure 3:The plot of MSEE vs Aggregation Time Interval for segment 1

The weighted average of MSEE concerning sample size was calculated for different aggregation intervals. Figure 3 and Figure 4 shows the plot of MSEE (sec^2) vs aggregation interval (minutes) for Segment 1 and Segment 2, respectively. From the differences in the curves for Segment-1 and Segment-2, it is

evident that variance in travel time is dependent on the highway geometry and traffic controls employed and significantly affects the selection of optimum travel time aggregation for travel time estimation and prediction. Furthermore, MSEE values for 15-minute aggregation intervals were found to be less than 5 minutes' travel time aggregation owing to the larger sample size. It is apparent that for a smaller aggregation interval of 1-minute, large values of MSEE are recorded. This is due to the relatively smaller sample size and hence results in being highly affected by the presence of outliers. A steep descent is observed as the aggregation interval size is increased and following that, the curve flattens. We used the "Elbow method", a heuristic used mostly in clustering analysis to obtain the optimum aggregation interval in which the knee/elbow of the curve is used as the optimum cluster number. For most of the curves in Figure 3 and Figure 4, the elbow lies between 10 to 15 minutes.

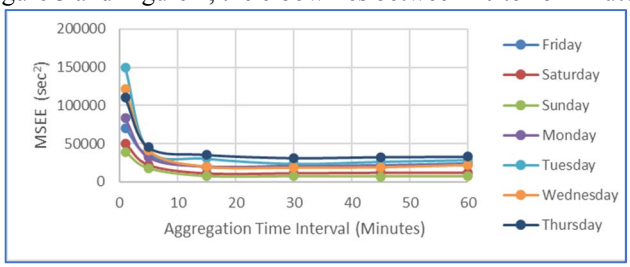

Figure 4: The plot of MSEE vs Aggregation Time Interval for segment 2

It should be noted that the aggregation interval size of more than 15 minutes was observed to have no significant relative difference in MSEE values. Also, for real-time traffic state information, planning, and conventional highway design practices, time intervals more than 15-minutes are usually not used. To emphasize the impact of within-dataset variance and aggregation interval on the performance of the prediction algorithms, this study used 5-minutes and 15-minutes aggregation intervals.

C. Travel Time Prediction

Mean Absolute Percentage Error (MAPE) and Theil's Inequality Coefficient are used as performance metrics for measuring the performance of prediction algorithms and are shown in the Equations below. The fewer the errors, the more reliable the predicted travel times are, and thus superior the prediction system. The acceptable limit of Theil's inequality coefficient based on the literature review is found to be 0.2 [26]. The performance assessment is done at two stages: aggregation interval wise and segment-wise to understand the applicability and robustness of each model under different traffic dynamics and within-dataset variations.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\ Observed\ Travel\ time - Predicted\ Travel\ time\ |}{Observed\ Travel\ time}\text{x}100$$

$$Theil's\ inequality\ coefficient\ = \frac{\sqrt{\frac{1}{n}\Sigma(\hat{y}_i - y_i)^2}}{\sqrt{\frac{1}{n}\Sigma(\hat{y}_i)^2} + \sqrt{\frac{1}{n}\Sigma(y_i)^2}}$$

The dataset was divided into 3 sets: training, test, and validation. The validation set consisted of 3 days of Wi-Fi data observed from the field and was compared with the prediction results. Figure 5 shows the comparison of the predicted travel times with the observed travel time profile for 5 minutes (a) and 15 minutes time bin (b) for Segment-1 for a weekday. Figure 5 shows the comparison of the MAPE and Theil's inequality coefficient respectively for the two study segments over two aggregation intervals.

Table 1 Performance evaluation metrics for Prediction Algorithms

Algorithm	MAPE (%)				Theil's Inequality coefficient			
	5 min		15 min		5 min		15 min	
	Segment-1	Segment-2	Segment-1	Segment-2	Segment-1	Segment-2	Segment-1	Segment-2
RDF	7.30	13.31	4.75	10.45	0.05	0.08	0.03	0.06
Naïve Bayes	12.53	22.67	7.59	20.66	0.08	0.16	0.05	0.11
k-NN	7.59	17.89	5.49	14.87	0.05	0.12	0.03	0.1
Kalman Filter	11.96	20.28	5.54	8.20	0.07	0.14	0.03	0.06

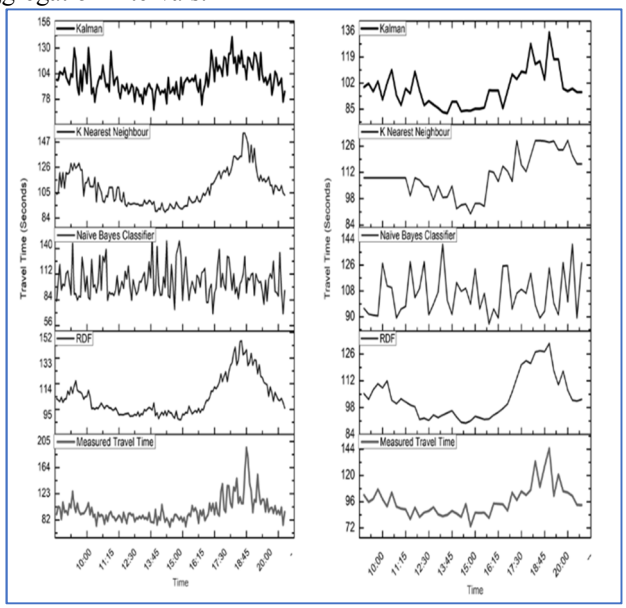

Figure 5: A sample comparison of weekdays predicted travel time and measured travel time pattern of Segment-1 at (a) 5 Minutes and (b) 15 Minutes aggregations

In Figure 5, The bottommost curves represent the actual travel time profiles and other curves are the predicted profiles for the labelled algorithms. The primary observation is that the Naïve Bayes Algorithm did not capture the recurring trend at all for both the aggregation time bins. The parameters used in the model training are the number of matched Wi-Fi devices between two sensors, and the travel time. One of the reasons for the poor performance of the algorithm is the inherent assumption of the independent predictors which is not true for this case. Comparatively, the MAPE and Theil's coefficient for RDF and Kalman filter and k-NN methods are lower than the

Naïve Bayes approach with all of them able to capture the evening peak period.

The RDF algorithm creates multiple decision trees and integrates them to make stable predictions. Compared to the rest of the methodologies, smooth predicted profiles can be observed (Figure 5) for both the observation bins. RDF is a good predictor for predicting recurrent trends but fails to capture the stochastic fluctuations. Kalman filter was the only approach to predict the two small sub-peaks in the evening peak period for 15 minutes observation bin as it tries to combine the predicted state and the noisy measurements to generate optimal estimates of the predicted states. k-NN also captured the overall trend but can be sensitive to the irrelevant features and the scale of the data. With big datasets, k-NN can be a computationally expensive option as it stores all the training dataset to make classifications or predictions.

Overall, the following two observations are noteworthy. Firstly, because of the lower sample size and high within dataset variance for 5-minute aggregation, all algorithms have better prediction performance for 15 minutes aggregation bin as the data is more disaggregated in the prior case along with significantly lower sample size. Secondly, All the algorithms performed better in case of Segment 1 when compared with Segment 2. Segment 2 has intersections at both ends with a signal operating with a cycle length of 5-10 minutes at the tail end in peak hours. Segment 1 has a minor intersection at one end and a mid-block at the other. Because of the highly diverse configuration, the Segment-2 is characterized by comparatively high variance in the travel time of different vehicle modes evident from Figure 3 and Figure 4. This within dataset variance has a very significant impact on the performances of all the four models. Overall, from the results, it was observed that the RDF and Kalman Filter algorithm ranks better in replicating the measured travel time, followed by K-NN, and lastly Naïve Bayes.

VI. Conclusion

The matching percentage between a pair of sensors was observed in the range of 7-8% of the total volume of vehicles which highlights the huge potential of Wi-Fi technology-based traffic data collection. This study focuses on evaluating the performance of four different travel-time prediction algorithms namely, Random Forest, Naive Bayes, k-NN, and Kalman Filters. Towards this purpose, travel time data using Wi-Fi-based sensors were collected and added as an input to the aforementioned algorithms to comprehend suitability and hence, the credibility of different models for travel time prediction under mixed traffic conditions. The entire analysis is done using a single source of traffic-data-collection method and hence, the potential application of the Wi-Fi sensor-based traffic-data-collection method for its use in travel time prediction is substantiated. Random forest and Kalman filter technique owing to the high range sample size have performed consistently well in capturing the temporal fluctuations in travel time, whereas, for k-NN, the high degree of unpredictability was observed in a few scenarios. The Naïve Bayes model due to conditional independence assumption results in loss of accuracy in prediction, and hence fails to deliver reliable real-time predictions. K-NN classifier being robust to noisy training data is a good tool for real-time prediction of travel times, but effective only when the training dataset is sufficiently large but as the data size grows, prediction becomes computationally expensive. As a future scope, non–linear, modified adaptive Kalman filtering algorithm can be modelled under an asymmetrical environment for Wi-Fi captured data, where all variances of the zero-mean Gaussian white noises are unknown, which is expected to be a more effective, and appropriate model for real-time applications. The study concludes by highlighting the necessity of understanding the nature of the dataset as prediction accuracy is a function of characteristics of the study segments, traffic dynamics, aggregation interval, and prediction methodology.

Acknowledgment

The authors acknowledge the support of Centre of Excellence in Urban Transport (CoEUT) at Indian Institute of Technology Madras for data collection and extraction efforts.

References

[1] S. Oh, Y.-J. Byon, K. Jang, and H. Yeo, "Short-term travel-time prediction on highway: A review on model-based approach," *KSCE Journal of Civil Engineering,* vol. 22, pp. 298-310, 2018.

[2] S. Oh, Y.-J. Byon, K. Jang, and H. Yeo, "Short-term travel-time prediction on highway: a review of the data-driven approach," *Transport Reviews,* vol. 35, pp. 4-32, 2015.

[3] D. Billings and J.-S. Yang, "Application of the ARIMA models to urban roadway travel time prediction-a case study," pp. 2529-2534.

[4] A. Guin, "Travel time prediction using a seasonal autoregressive integrated moving average time series model," pp. 493-498.

[5] M. Yang, Y. Liu, and Z. You, "The reliability of travel time forecasting," *IEEE Transactions on Intelligent Transportation Systems,* vol. 11, pp. 162-171, 2009.

[6] P. Gao, J. Hu, H. Zhou, and Y. Zhang, "Travel time prediction with immune genetic algorithm and support vector regression," pp. 987-992.

[7] A. Khadhir, B. Anil Kumar, and L. D. Vanajakshi, "Analysis of global positioning system based bus travel time data and its use for advanced public transportation system applications," *Journal of Intelligent Transportation Systems,* pp. 1-19, 2020.

[8] H. Bahuleyan and L. D. Vanajakshi, "Arterial path-level travel-time estimation using machine-learning techniques," *Journal of Computing in Civil Engineering,* vol. 31, p. 04016070, 2017.

[9] R. Jairam, B. A. Kumar, S. S. Arkatkar, and L. Vanajakshi, "Performance comparison of bus travel time prediction models across Indian cities," 2018.

[10] A. Bhaskar and E. Chung, "Fundamental understanding on the use of Bluetooth scanner as a complementary transport data," *Transportation Research Part C: Emerging Technologies,* vol. 37, pp. 42-72, 2013.

[11] S. Sharma, H. Maripini, A. Khadhir, S. S. Arkatkar, and L. Vanajakshi, "Analysis and Use of Wi-Fi data for Signal State Identification," *Transportation Research Procedia,* vol. 48, pp. 1008-1021, 2020.

[12] J. K. Mathew, V. L. Devi, D. M. Bullock, and A. Sharma, "Investigation of the use of Bluetooth sensors for travel time studies under Indian conditions," *Transportation Research Procedia,* vol. 17, pp. 213-222, 2016.

[13] J. Barceló Bugeda, L. Montero Mercadé, M. Bullejos, O. Serch, and C. Carmona Bautista, "A kalman filter approach for the estimation of time dependent od matrices exploiting bluetooth traffic data collection," pp. 1-16.

[14] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, "An improved k-nearest neighbor model for short-term traffic flow prediction," *Procedia-Social and Behavioral Sciences,* vol. 96, pp. 653-662, 2013.