

Application of Machine Learning Techniques in Short-term Travel Time Prediction Using Multiple Data Sources

Homa Taghipour^a, Amir Bahador Parsa^{a,1}, Abolfazl (Kouros) Mohammadian^a

^aUniversity of Illinois at Chicago, 842 W Taylor Street, Chicago, IL 60607, USA

Abstract

Having access to accurate travel time is of great importance for both highway network users and traffic engineers. The travel time which is currently reported on several highways is estimated by employing naïve methods and using limited sources of data. This results in unreliable and inaccurate travel time prediction and could impose delay on travelers. Therefore, the main objective of this study is short-term prediction of travel time for highways using multiple data sources including loop detectors, probe vehicles, weather condition, network, accidents, road works, and special events in order to consider the effect of different factors on travel time. To this end, two machine learning methods, K-Nearest Neighbors and Random Forest, are employed. After applying data cleaning process on datasets and combining them, the models are trained to predict and compare short-term harmonic average speed as a representative of travel time for 5-minute prediction horizons in one hour ahead. The travel time is calculated as the ratio of the length of each link and the harmonic average speed for all reporting vehicles. Hence, a model is trained for each technique to predict travel time 5 minutes ahead, 10 minutes ahead, and all the way down to 60 minutes ahead. The results confirm satisfying performance of both models in short-term travel time prediction with slightly outperformance of Random Forest model. A feature importance and sensitivity analysis also applied for the Random Forest model, and traffic variables are found as the most effective variables in predicting the travel time.

Keywords: *Travel Time Prediction, Machine Learning, K-Nearest Neighbors, Random Forest, Feature Importance*

¹ Corresponding author. Tel: +1 (312) 996-9840.

Email addresses: htaghi2@uic.edu (H. Taghipour), aparsa2@uic.edu (A. B. Parsa), kouros@uic.edu (A. K. Mohammadian).

1 Introduction

2 Nowadays, having sufficient information about travel time plays a pivotal role for both road users
3 and traffic engineers. Hardly is it possible for road users to make decisions before and during their
4 trips, unless they have access to reliable information about travel time of the route they are passing
5 through. Having access to accurate travel time will also enable traffic engineers to increase the
6 efficiency and safety of traffic in road networks (Van Lint, 2008). Travel time can be defined as
7 the total time required for a vehicle to pass a route from one point (origin) to another point
8 (destination) with respect to all the delays which might be imposed on that vehicle (Zhu et al.,
9 2009). Currently, the travel time which is reported on Variable Message Signs (VMSs) of some
10 highways is estimated by employing naïve methods that use limited sources of data (Huisken and
11 van Berkum, 2003). This might cause inaccurate travel time prediction, and impose delay on
12 travelers, which consequently makes the travel time reports on the VMSs unreliable. Therefore,
13 the objective of this study is considering the effect of different factors such as weather condition,
14 accidents, road works, and special events, which are rarely considered in other studies, on travel
15 time of highways by combining these real-time sources of data and employing machine learning
16 methods.

17 Studies aimed at finding travel time could be divided into two main groups of estimating
18 and predicting travel time (Mori et al., 2015). Travel time estimation studies focus on calculating
19 travel times of trips that have already finished, with respect to the data captured during those trips
20 (Celikoglu, 2013; Li et al., 2013; Soriguera and Robuste, 2010). On the other hand, travel time
21 prediction studies, are concentrating on using the historical and real time data to forecast the travel
22 time for the future time intervals (Van Lint, 2004). There are several methods which are used for
23 travel time prediction, and the difference between methodologies stems from their complexity and
24 accuracy, that make each of them suitable for specific conditions.

25 Basically, travel time prediction models could be categorized into three main groups. The
26 first group are naïve models which generally consist of very simple methods (Huisken and van
27 Berkum, 2003). These models have several assumptions in their nature which make them simple
28 but less accurate (Van Lint, 2004). Thus, these models are generally used as the baseline for the
29 other methods or whenever just a simple prediction is required. One of the famous naïve models
30 is instantaneous predictor, which assumes the traffic condition remains consistent with the passage
31 of time (Van Lint and Van Hinsbergen, 2012). Based on this assumption, the most updated
32 estimation of travel time is assumed to be constant and equal to the travel time in the future. It is
33 obvious that expanding the prediction horizon will reduce the accuracy of this type of model.
34 Another kind of naïve predictor uses the historical travel time data to predict the future travel time
35 regarding the similarity of traffic condition in different periods of time (Schmitt and Julia, 2007).
36 This model works well when recurrent traffic conditions occur in a specific path (Van Lint and Van
37 Hinsbergen, 2012).

38 Second group is traffic-theory-based models. The basis of this group of models is on the
39 relation between traffic variables such as speed, density, and occupancy. Since the prediction refers
40 to travel time in future, the first step in these models is to simulate the traffic condition of a network
41 for a given time in the future and then travel time could be predicted based on traffic condition of
42 that network (Mori et al., 2015). Regarding this issue, three kinds of simulation comprised of
43 macroscopic (Papageorgiou et al., 2010), microscopic (Edara et al., 2017), and mesoscopic (Taylor,
44 2003) are applied in various studies.

45 The last group of models, which are data-based models, require a large amount of data and
46 use statistical methods to predict travel time. The more the data is available, the more accurate the
47 result will be. These models could be divided into two main groups of parametric and non-

parametric models (Mori et al., 2015). Parametric models have a pre-defined function and just the parameters of the model need to be estimated. There are several parametric models each of which has their own structure such as Linear Regression (Du et al., 2012), Bayesian Nets (Castillo et al., 2011), and Time Series models (Billings and Yang, 2006; Karimpour et al., 2017; Yang, 2005). Non-parametric models are another part of data-based models in which the structure of the function should be defined based on the data as well as number and typology of parameters. The most important goal of these methods is to utilize a large database and learn the relationships between variables, instead of using complex approaches, in order to predict travel time. One of the most famous group of models includes Artificial Neural Network models which have been widely used in transportation studies (Dharia and Adeli, 2003; Fan and Gurmu, 2015; Parsa et al., 2020a, 2020b, 2019a, 2019b). Specifically, K-Nearest Neighbors (Qiao et al., 2013; Yu et al., 2016; Zhao et al., 2018), Regression Trees (Nikovski et al., 2005), Local Regression (Simroth and Zahle, 2010), Support Vector Regression (Wu et al., 2004; Yildirimoglu and Ozbay, 2012), and Random Forest (Leshem and Ritov, 2007) are non-parametric models which are performing well for travel time prediction.

One of the most important and effective factors on choosing a method for estimating or predicting travel time is the available data. Although there are several kinds of methods each of which has its own advantages and disadvantages, none of them could be used unless adequate and related sources of data are available. Based on the classification done by Wu, et al. (Wu et al., 2004), there are several kinds of sensors using for obtaining traffic data and these sensors can be grouped with respect to their ability to obtain travel time directly. Accordingly, two main groups of sensors are link-based and point-based sensors. Link-based sensors measure the travel time directly for a route or link by passive probe vehicles, active test vehicles, or license-plate matching. In point-based sensors travel time is measured indirectly by devices such as loop detectors, laser detectors, and video cameras. The data which is collected from link-based sensors are more accurate, however point-based sensors are available more and could collect cost-effective real-time data.

Based on the characteristics of all different methods, since the main data source for this study is collected from loop detectors, the data-based methods are employed in order to predict the travel time. In general, to the best of the authors knowledge, most of studies in this area have considered only traffic related variables in order to develop a model for travel time prediction. However, other variables such as weather condition could have significant impact on the travel time (Qiao et al., 2012; Rakha et al., 2012), which is infrequently investigated in travel time prediction studies. Nookala showed that inclement weather condition could increase the traffic congestion and travel time, since the capacity of highway decreased in inclement weather, while the traffic demand does not change (Nookala, 2006). Therefore, the main goal of this study is to predict travel time for links of Eisenhower highway in the city of Chicago, by utilizing a combination of different sources of data including loop detectors, probe vehicles, weather condition, network, accident occurrence, road works, and special events. To this end, average speed of each link, and consequently travel time is predicted using KNN and Random Forest models.

In what follows, first the datasets and the process of cleaning and combining them, which are used in this study, are introduced. Also, the methods which are employed to predict the travel time are described. Then, the results of the models are shown and explained. In the end, the discussion of the results, conclusion and limitations of the study are discussed.

Data

Loop Detectors Data

In this study, different sources of data are used in order to develop models to predict travel time. The main part of dataset is the information captured by inductive loop detectors in the Chicago highways network which is collected by the Gateway Traveler Information System and provided by the Illinois Department of Transportation (IDOT). For this study, a 5-mile segment from west bound of the Eisenhower highway, consist of 14 loop detectors is selected. For each loop detector in each lane of the highway, number of vehicles, occupancy and average speed are reported every 20 second. These raw data points contain lots of missing and erroneous records which might be caused by detectors' malfunction, deterioration of pavement or any other reasons. Therefore, two types of single and combined thresholds are applied to the dataset, as a part of data cleaning procedure.

Single Threshold

There are several thresholds for each of the count, occupancy and speed values which are used in different studies. In this study, based on the characteristics of the selected highway, value of 3000 vehicles per hour per lane is set as the threshold for the count, which is almost equal to 17 vehicles per 20 second per lane. So, if the value of count in a lane does not lie within the domain of (0, 17), it is labeled as erroneous data point. Moreover, the extreme amounts for occupancy (i.e., more than 90% in 5 minutes) are assumed incorrect so that any data point with occupancy value outside the domain (0, 90) is labeled as erroneous data point. In addition, we set the value of 80 mph for the upper threshold of speed, and any point with more than 80 mph or less than 0 mph speed is considered as an outlier.

Combined Threshold

As mentioned earlier, in this study three variables of count, occupancy and speed are reported by loop detectors. The first threshold in this group is described as “only one zero variable out of three”, which means if only one of those three variables is zero whereas the other two are non-zero, that record is reasonably considered as an erroneous record. The second combined threshold is “only one non-zero variable out of three”. So that if two variables out of count, occupancy and speed are zero while the other one is non-zero, again this record of data is labeled as an error. Finally, the last combined threshold is “all zero variables”, in which a record is considered as an erroneous record if all the variables of count, occupancy, and speed are zero. This threshold filters out the data points that could impact the average speed of intervals incorrectly.

Data Imputation

The missing and erroneous data points could negatively affect the results of prediction models (Karimpour et al., 2019). Therefore, these points should be either eliminated from the dataset or imputed by new data points through one of the three following ways. First, “temporal estimation” is used when a detector has a missing data point, or it reports a value which exceeds threshold for an interval. In this case, an appropriate way for imputing the data point is taking the average of values for previous and next time intervals, especially when the time intervals are short (i.e., in this study it is 20 second). In another approach, which is “spatial estimation”, the missing or erroneous data points are imputed with the average of the values for previous and next loop detectors at the same time. This method is proper whenever loop detectors are close to each other, or using the temporal estimation is not applicable. Finally, “historical estimation” could be used whenever traffic condition is recurrent in a location in which missing or erroneous data is reported.

So, in this case the erroneous data points are imputed using the reported data points from same location for same time of day and same day of week in the past. **Figure 1** shows a sample of weekly pattern of the average flow after applying data imputation process for the studied segments. This figure displays 7 daily traffic patterns, started from Monday, September 11, 2017 to Sunday, September 17, 2017. It is observed that weekdays have almost similar traffic patterns although the traffic pattern is different for Friday, Saturday, and Sunday.

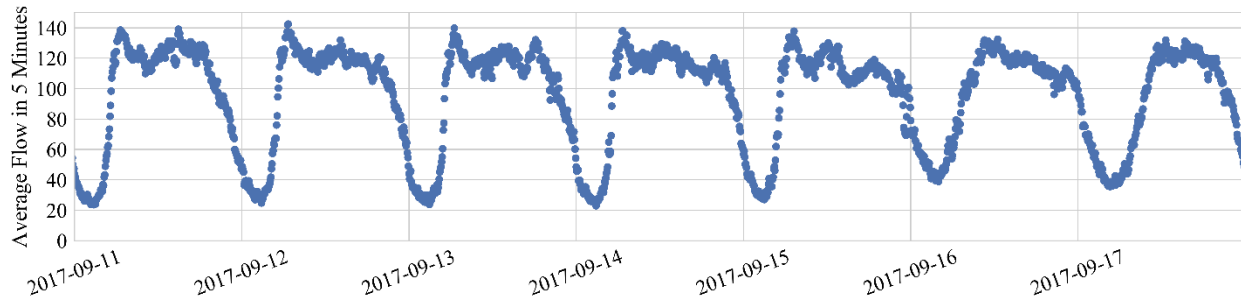


Figure 1 Weekly pattern of average flow from cleaned loop detectors dataset

In this study, the speed values in the loop detectors dataset are time mean speeds. The time mean speed is calculated by averaging the point speeds of all vehicles passing a loop detector at that detector's location. This speed could not reflect the travel time of a path segment accurately, since the speed of vehicles could vary across that segment. Therefore, in this study in order to consider the speed values another source of data which is collected by probe vehicles is utilized.

Probe Vehicles Data

This dataset is provided by the National Performance Management Research Data Set (NPMRDS) and includes the harmonic average speed, free flow speed, historical speed, travel time, and annual average daily traffic (AADT) for each link of highways. INRIX is creating the NPMRDS dataset of average speeds and travel times for specific road segments across the NPMRDS road network for every 5 minutes. INRIX uses its large data source includes millions of connected vehicles and trucks that are supplying location and movement data anonymously. They also implemented a path-processing algorithm to meet the NPMRDS requirements using its existing source data.

For the present study, value of travel time from this dataset is chosen to be the target variable. This variable is defined as the ratio between the segment length and the harmonic average speed of vehicles passing the segment. Since the travel time of each link is dependent to length of the link, the harmonic average speed is selected as the target, so that by predicting the average speed, value of travel time could be calculated using the length of each link. The missing values of this dataset are imputed by temporal estimation, spatial estimation, and historical estimation, as it is explained in the previous section. **Figure 2** shows histogram of the average speed from this dataset.

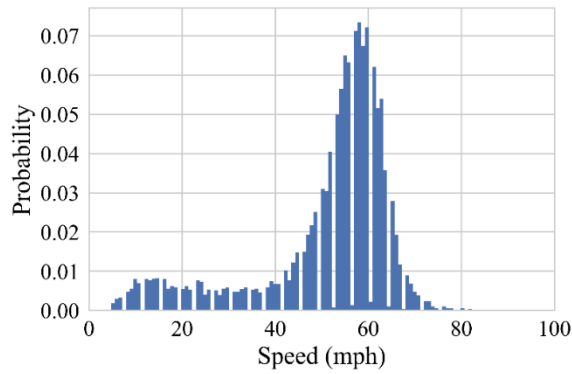


Figure 2 Average speed histogram from probe vehicles dataset

Weather Condition Data

The hourly weather condition data which is provided by the National Weather Service (NWS) is another source of data used in the present study. This data is available for airport stations and based on the location of selected segment, data from the Midway International Airport is utilized. In the dataset, weather condition is reported by 94 different states. In this study, all the states are grouped into four categories from fair to severe weather conditions. In the first category, weather conditions such as “Fair” and “Partly Cloudy” are grouped. The second category is comprised of weather conditions like “Fair with Light Haze” and “Light Haze Dust and Windy”. Unfavorable weather conditions such as “Light Drizzle” and “Light Rain Fog” constitute the third category. Finally, the fourth and the most severe category includes weather conditions like “Snow Fog” and “Thunderstorm Heavy Rain Fog”. In addition to the weather condition, surface temperature is also available in the dataset which is used in this study. Similar to other data sources there are plenty of missing values in this dataset which are imputed using the information from previous and next time intervals.

Accident Data

This dataset is also collected by the Gateway Traveler Information System. In the accident data details of accidents occurred in highways are reported. Time of accident, geographical location, and severity of accident are among the most important variables in this dataset.

Based on the accident data analysis and observing its effect on highway traffic at the studied area, when an accident occurs on a highway link, that link, the previous link, and the next link on the highway are affected on average for one hour after the accident occurrence. This effect is considered as binary variables to show the affected links and time intervals in the model. Two binary variables for previous link and next link are defined, since the effect of accident is different in upstream and downstream. Also, another variable in the model demonstrates the accident severity in three levels of minor, medium and major. Whenever the accident severity is not determined at accident detection time, the medium level, which is the most common one, is selected.

Road Work Data

This dataset which is collected by the Gateway Traveler Information System, is comprised of details of construction or maintenance operations in the highways. In this dataset variables such as start time, end time, geographical location, and severity are reported.

In order to consider the effect of this variable in the model a binary variable is defined.

When a roadwork project is executing on a link of highway, that link, the previous link, and the next link are affected from start time to end time of the project. Also, if one of the lanes of highway is blocked, variable of “number of lanes” considers its effect in the model.

Special Event Data

The special event dataset collected by the Gateway Traveler Information System provides detailed information about stadium events, parades, and road races. In this dataset several variables such as start time, end time, location, and type of events are reported.

When a special event is occurring close to highway links, those links are considered to be affected by that event for one hour before or one hour after that event. Considering the next or previous one-hour depends on the direction of extra traffic and the relative locations of the event and each link. For example, if a highway link is located in the path toward an event location, the extra traffic passes that link before the event in order to go to that location, so in this case we consider that link to be affected by the event one hour before start of the event. For the links which are located in the path from the event location toward other parts of the city, one hour after end of the event would be the affected time. Since these variables can considerably impact travel time of a route, we combine and use them in the model.

Network Data

This dataset includes several variables from the studied highway segment such as number of lanes, number of entrances, number of exits, and length for each of the highway links.

After applying data imputation process on each of the mentioned datasets, all of them are combined based on their location and timestamps, so that a link-based dataset which is aggregated in 5 minutes, from April 2017 to December 2017 is created for 15 links of west bound of the Eisenhower highway. **Table 1** demonstrates the variables of the final combined dataset which is used for developing travel time prediction models.

Table 1 Description of explanatory variables

Variable	Description
Time _i i= 1, 2, ..., 8	1: if time ID equals i*, and 0: otherwise
Friday	1: if the day of week is Friday, and 0: otherwise
Saturday	1: if the day of week is Saturday, and 0: otherwise
Sunday	1: if the day of week is Sunday, and 0: otherwise
WeatherID	Weather condition: 1: Fair or a few clouds 2: Cloudy with haze or light fog 3: light rain or drizzling 4: Heavy rain, snow, thunderstorm or fog
Temp_F	Temperature (Fahrenheit)

Occupancy	Aggregated occupancy at location of loop detector which is closest to midpoint of the link at the 5-minute time interval
Occupancy_p	Aggregated occupancy at location of previous loop detector on the road at the 5-minute time interval
Occupancy_n	Aggregated occupancy at location of next loop detector on the road at the 5-minute time interval
Count	Aggregated count of vehicles on loop detector which is closest to midpoint of the link at the 5-minute time interval
Count_p	Aggregated count of vehicles on previous loop detector on the road at the 5-minute time interval
Count_n	Aggregated count of vehicles on next loop detector on the road at the 5-minute time interval
AADT	Annual average daily traffic of the link
EntRamp	Number of entrance ramps on the link
ExtRamp	Number of exit ramps on the link
Miles	Length of the link in miles
Lanes	Number of lanes on the link
IsAcc_p	1: if an accident occurred on the link or previous link, at the 5-minute time interval, and next one-hour time intervals after accident occurrence, and 0: otherwise
IsAcc_n	1: if an accident occurred on the link or next link at the 5-minute time interval, and next one-hour time intervals after accident occurrence, and 0: otherwise
AccSeverityID	Accident severity: 1: Minor severity 2: Medium severity 3: Major severity
IsRW	1: if a roadwork project is executing on the link, previous link, or next link at the 5-minute time interval, and 0: otherwise
IsSE	1: if a special event is occurring close to the link at the 5-minute time interval, and either next one-hour, or previous one-hour time intervals**, and 0: otherwise
Speed (target)	Average speed of vehicles on a link at the 5-minute time interval

* Each 3-hour time interval of a day is considered as a binary variable. For instance, Time1 means from midnight to 3 am, and Time8 represents interval of 9 pm to 12 am.

** Considering the next or previous one-hour depends on the direction of extra traffic and the relative locations of event and link. For example, if a link is located in the path toward an event, the extra traffic passes this link before the event, so in this case we consider previous one-hour as the effective time.

1 Methods

2 Machine learning methods are compatible with the type of data that is available for this study, in
3 order to develop the travel time prediction model. Among different machine learning methods K-
4 Nearest Neighbors and Random Forest are selected based on their specific properties, which are
5 described as follows.

7 K-Nearest Neighbors

8 K-Nearest Neighbors (KNN) needs only a large volume of data points, and without developing a
9 mathematical model could predict the target variable. Also, there is no need to define parameters
10 of the model in advance, and the authenticity of the data is maintained since no smoothing
11 procedure is done to the data. Therefore, this method is an appropriate choice for the nonparametric
12 problem of travel time prediction (Yu et al., 2016).

13 This method is among the most famous supervised machine learning techniques, which are
14 widely used for regression and classification. As it is explained by Han et al. (Han et al., 2011),
15 the KNN method basically classifies a dataset by comparison to the similar records of a trained
16 dataset. In this study, the KNN regression algorithm is used, since the target is speed which is a
17 continuous value rather than a predefined class. In general, training dataset has n attributes, and
18 each of its records could be presented in a n -dimensional variable space by the value of the n
19 attributes. After that, to predict the target value for new records from test dataset, each new record
20 with its attribute values finds its location in the variable space, and the algorithm looks for the k
21 nearest neighbors that are closest to the new record among the trained data points. Therefore, in
22 this study the speed value for a new record would be predicted by taking average of the speed
23 values of k nearest neighbors to that record. In other words, the method searches for nearest
24 neighbors among all historical data points and uses them for the prediction.

25 In this study, the Euclidean metric is used to calculate the distance between two points in
26 the n -dimensional variable space. Between two points of $X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $X_2 =$
27 $(x_{21}, x_{22}, \dots, x_{2n})$, the Euclidean distance is defined as **Equation 1**:

$$28 \text{ Euclidean distance } (X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad (1)$$

30 Before calculating the distance between points, values of attributes are normalized in order
31 to prevent attributes with larger values like occupancy outweighing attributes with smaller values
32 like binary attributes (Han et al., 2011).

35 Random Forest

36 Random Forest is an ensemble supervised machine learning method which could be used as
37 classifier or regressor for categorical or numerical datasets (Han et al., 2011). In this method a
38 combination of several random Decision Trees (DT) are utilized, so that each of the DTs votes to
39 predict value of the target variable. In order to combine DTs, Bootstrap Aggregation or bagging
40 that is a model averaging approach for combining machine learning methods to increase accuracy
41 is used. To develop a Random Forest model multiple random samples from training dataset are
42 selected with replacement in several iterations, and for each of them a DT is trained. Then, for a
43 new record from test dataset, each of the trained DTs returns value of the target variable. The final
44 result is calculated by taking average of all the predicted values for the target from DTs. Random
45 Forest is robust to noisy data and overfitting, also it is expected to have higher accuracy than
46 individual DT, since it decreases variance of the DTs (Han et al., 2011). Random Forest typically
47 works accurately and fast when a large dataset is available. Also, it could manage large number of

variables as the model inputs. Therefore, these characteristics make the Random Forest model a proper choice for predicting the travel time in this study (Fan et al., 2018).

Results

After applying data cleaning process on different data sources and combining them to create the final dataset, 80% of the data points are randomly selected for model training and the remaining 20% are used as the test data to evaluate the models. In addition, number of neighbors for the KNN model is assigned experimentally as 4, and for Random Forest model, 20 DTs in forest with max depth of 30 are set. Finally, KNN and Random Forest models are trained to predict and compare short-term average speed as a representative of travel time, as mentioned in previous sections, for 5-minute prediction horizons in one hour ahead. That is, a model is trained for each technique to predict travel time at the moment, 5 minutes ahead, 10 minutes ahead, 10 minutes ahead, and all the way down to 60 minutes ahead.

Figure 3 displays and compare the prediction accuracy (i.e., R-Squared Score) of KNN and Random Forest models. Random Forest model outperforms KNN model in all 13 prediction time intervals, however, the difference between their prediction accuracy is more negligible for very short-term prediction horizons. Both models perform extremely good at 0 minute ahead (i.e., at the prediction moment which would be travel time estimation), however, Random Forest with prediction accuracy of 95.3% slightly outperforms KNN with prediction accuracy of 92.9%. Then, prediction accuracy is decreasing by increasing prediction horizon in a way that prediction accuracy at 5 minutes and 10 minutes ahead for Random Forest model are 93.9%, and 91.6%, respectively, and those of KNN model are 91.1%, and 88.1%, respectively. Another point based on **Figure 3** is that prediction accuracy of both models is decreasing in a faster rate in the first 30 minutes ahead horizon than the horizon of 30 minutes to 60 minutes ahead. Despite the plateauing effect for prediction horizons above 30 minutes of both models, Random Forest model performance is more robust as it achieves prediction accuracy of 84% at 60 minutes ahead while that of KNN model is only 76.8%.

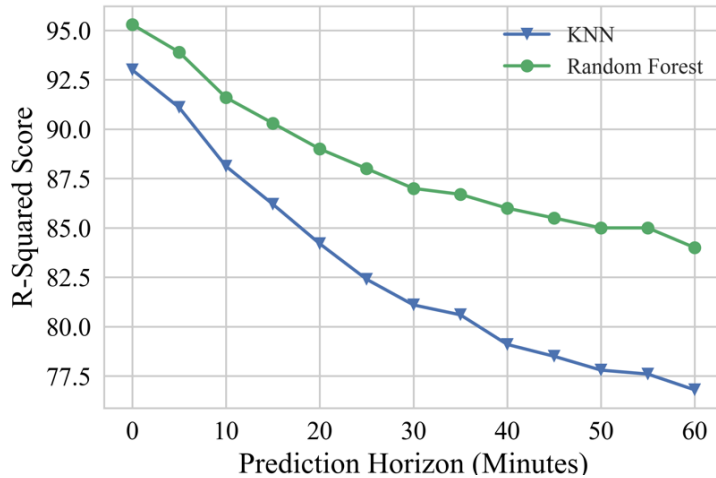


Figure 3 Comparison of KNN and Random Forest models' accuracy in different prediction horizons

Regarding **Figure 3**, it can be concluded that both models, especially Random Forest, are well-suited to predict travel time in a short-term horizon (i.e., 15 minutes ahead). Therefore, to better observe performance of these models in short-term prediction, true speed values versus predicted speed values of KNN, and Random Forest models are plotted in the **Figure 4**. This figure confirms that the shorter prediction horizon, the more prediction accuracy for both models. It is worth noting that, regarding training time, Random Forest model outperforms KNN model as well.

Accordingly, feature importance analysis is applied for the Random Forest model, and the result is shown in **Table 2**. As it is expected the traffic variables are the most important variables that affect the travel time. After traffic variables, temperature is found effective in the model which shows the importance of weather condition variable in traffic related predictions. Unlike our expectation, accident, special event, and road work incidents have relatively lower impact in the model. However, the reason for low importance of these variables might stem from small number of these incidents in the data. It is worth noting that since these incidents are correlated with traffic variables which are important inputs of the model, their effects are considered indirectly through the traffic variables in the model.

Table 2 Feature importance results

Variable	Rank
Occupancy	1
Count_p	2
Occupancy_p	3
Count	4
Occupancy_n	5
Count_n	6
AADT	7
Temp_F	8
Miles	9
Time3 (Morning Peak-Hours)	10
Time7 (Evening Peak-Hours)	10
Sunday	11
Saturday	12
Friday	13
ExtRamp	14
Time _i (i = 1, 2, 4, 5, 6)	15
EntRamp	16
AccSeverityID	17
WeatherID	18
Lanes	19
Evets (IsAcc, IsRW, IsSE)	20

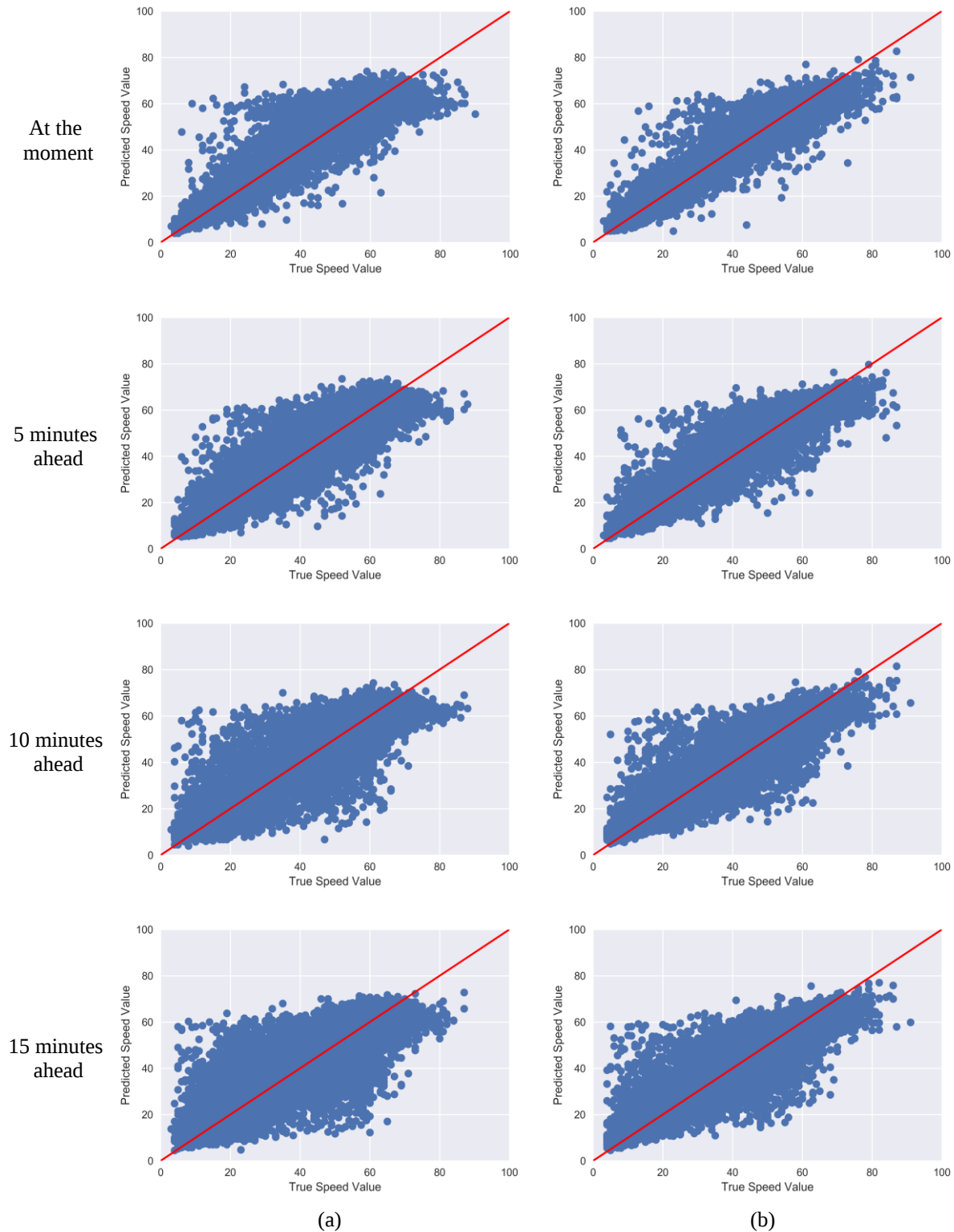


Figure 4 True speed values versus predicted speed values – (a) KNN, (b) Random Forest

The feature importance analysis could only relatively rank the importance of variables in the model, therefore in order to assess the potential impact of the most important variables on the

target variable a sensitivity analysis is also conducted. To this end, the top three important variables in the list which are occupancy (Occupancy, Occupancy_p, and Occupancy_n), count (Count, Count_p, and Count_n), and AADT are selected. Then, the final Random Forest model was run several times with change in each of these variables by $\pm 10\%$, $\pm 20\%$, $\pm 30\%$, $\pm 40\%$, and $\pm 50\%$ while other variables remain the same. **Figure 5** demonstrates the sensitivity of average speed for all the links and all the time intervals to the important variables. Based on the figure, it indicates that increasing the occupancy reduces the average speed, and it shows a linear relationship as it is expected. For the count variable, a nonlinear relationship is observed. As it is mentioned, the average speed value in here is for all the links and times in different conditions. Therefore, on average the speed is increasing by increasing the count from -25% to +50%, also by decreasing the count from -25% to -50%. The effect of AADT on the average speed is also nonlinear, that is by increasing the AADT the average speed increases, but there is no significant change in average speed by increasing the AADT from +10% to +50%, and from -50% to -10%.

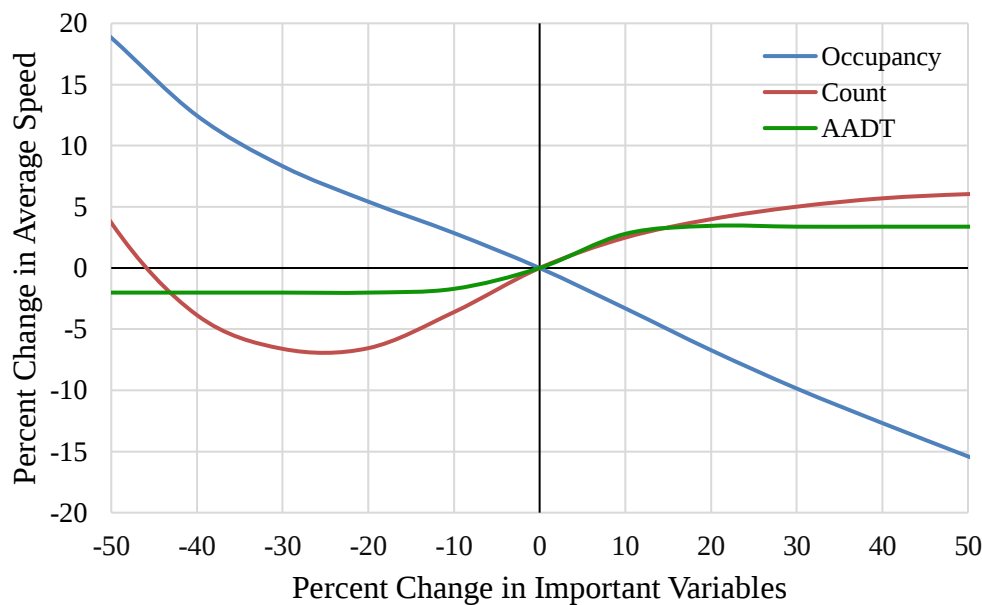


Figure 5 Sensitivity analysis of important variables in travel time prediction model

Summery and Conclusions

In this study multiple data sources including loop detectors, probe vehicles, weather condition, network, accidents, road works, and special events are combined, and two machine learning methods comprised of KNN and Random Forest are employed for short-term prediction of travel time for highways. Accordingly, a model is trained for each technique to predict travel time 5 minutes ahead, 10 minutes ahead, and all the way down to 60 minutes ahead.

Comparison of Random Forest and KNN models shows good performance of both models in short-term travel time prediction with slightly outperformance of Random Forest model for very short-term prediction horizons. Therefore, applying Random Forest model for 15 minutes or less prediction horizons is more reasonable although the prediction accuracy of longer prediction horizons is still acceptable.

Traffic related variables, especially *occupancy*, are found so effective in short-term travel time prediction using loop detector data. It suggests that among traffic variables collected by loop

1 detectors, occupancy can capture traffic condition better than the others. However, fusion of several
2 traffic data from different sources could increase prediction accuracy of the models.

3 Weather condition is an important variable in the studies related to traffic condition. In this
4 study, we utilized weather condition in our models and found it effective in predicting travel time.
5 It is suggested that to better capture impact of weather condition in the model, data collected from
6 more accurate sensors are required for the similar studies.

7 Finally, accident, special events, and road works are important variables which can have a
8 considerable impact on travel time prediction. Although in this study these variables have
9 relatively less impact on travel time prediction, for the future studies we suggest that rich amount
10 of incident data along with dynamic models could accurately capture the impact of these variables
11 on short-term travel time prediction.

12 **Acknowledgement**

13 Authors would like to show their gratitude to the Illinois Department of Transportation (IDOT) for
14 funding this project and providing the data sources.
15

References

- Billings, D., Yang, J.-S., 2006. Application of the ARIMA models to urban roadway travel time prediction-a case study, in: 2006 IEEE International Conference on Systems, Man and Cybernetics. IEEE, pp. 2529–2534.
- Castillo, E., Nogal, M., Menendez, J.M., Sanchez-Cambronero, S., Jimenez, P., 2011. Stochastic demand dynamic traffic models using generalized beta-Gaussian Bayesian networks. *IEEE Trans. Intell. Transp. Syst.* 13, 565–581.
- Celikoglu, H.B., 2013. Flow-based freeway travel-time estimation: A comparative evaluation within dynamic path loading. *IEEE Trans. Intell. Transp. Syst.* 14, 772–781.
- Dharia, A., Adeli, H., 2003. Neural network model for rapid forecasting of freeway link travel time. *Eng. Appl. Artif. Intell.* 16, 607–613.
- Du, L., Peeta, S., Kim, Y.H., 2012. An adaptive information fusion model to predict the short-term link travel time distribution in dynamic traffic networks. *Transp. Res. Part B Methodol.* 46, 235–252.
- Edara, P., Rahmani, R., Brown, H., Sun, C., 2017. Traffic Impact Assessment of Moving Work Zone Operations. Smart Work Zone Deployment Initiative.
- Fan, S.-K.S., Su, C.-J., Nien, H.-T., Tsai, P.-F., Cheng, C.-Y., 2018. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. *Soft Comput.* 22, 5707–5718.
- Fan, W., Gurmu, Z., 2015. Dynamic travel time prediction models for buses using only GPS data. *Int. J. Transp. Sci. Technol.* 4, 353–366.
- Han, J., Pei, J., Kamber, M., 2011. Data mining: concepts and techniques.
- Huisken, G., van Berkum, E.C., 2003. A comparative analysis of short-range travel time prediction methods, in: 82nd Annual Meeting of the Transportation Research Board, Washington, DC.
- Karimpour, A., Ariannezhad, A., Wu, Y.-J., 2019. Hybrid data-driven approach for truck travel time imputation. *IET Intell. Transp. Syst.*
- Karimpour, M., Karimpour, A., Kompany, K., Karimpour, Ali, 2017. Online Traffic Prediction Using Time Series: A Case study, in: *Integral Methods in Science and Engineering*, Volume 2. Springer, pp. 147–156.
- Leshem, G., Ritov, Y., 2007. Traffic flow prediction using adaboost algorithm with random forests as a weak learner, in: *Proceedings of World Academy of Science, Engineering and Technology*. Citeseer, pp. 193–198.
- Li, L., Chen, X., Li, Z., Zhang, L., 2013. Freeway travel-time estimation based on temporal-spatial queueing model. *IEEE Trans. Intell. Transp. Syst.* 14, 1536–1541.
- Mori, U., Mendiburu, A., Álvarez, M., Lozano, J.A., 2015. A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transp. A Transp. Sci.* 11, 119–157.
- Nikovski, D., Nishiuma, N., Goto, Y., Kumazawa, H., 2005. Univariate short-term prediction of

- road travel times, in: Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005. IEEE, pp. 1074–1079.
- Nookala, L.S., 2006. Weather impact on traffic conditions and travel time prediction.
- Papageorgiou, M., Papamichail, I., Messmer, A., Wang, Y., 2010. Traffic simulation with METANET, in: Fundamentals of Traffic Simulation. Springer, pp. 399–430.
- Parsa, A.B., Chauhan, R.S., Taghipour, H., Derrible, S., Mohammadian, A., 2019a. Applying Deep Learning to Detect Traffic Accidents in Real Time Using Spatiotemporal Sequential Data. arXiv Prepr. arXiv:1912.
- Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020a. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accid. Anal. Prev.*
- Parsa, A.B., Shabanpour, R., Mohammadian, A., Auld, J., Stephens, T., 2020b. A Data-Driven Approach to Characterize the Impact of Connected and Autonomous Vehicles on Traffic Flow.
- Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A. (Kouros), 2019b. Real-time accident detection: Coping with imbalanced data. *Accid. Anal. Prev.* 129, 202–210. <https://doi.org/10.1016/j.aap.2019.05.014>
- Qiao, W., Haghani, A., Hamed, M., 2013. A nonparametric model for short-term travel time prediction using bluetooth data. *J. Intell. Transp. Syst.* 17, 165–175.
- Qiao, W., Haghani, A., Hamed, M., 2012. Short-term travel time prediction considering the effects of weather. *Transp. Res. Rec.* 2308, 61–72.
- Rakha, H., Arafah, M., Park, S., 2012. Modeling inclement weather impacts on traffic stream behavior. *Int. J. Transp. Sci. Technol.* 1, 25–47.
- Schmitt, E.J., Julia, H., 2007. On the limitations of linear models in predicting travel times, in: 2007 IEEE Intelligent Transportation Systems Conference. IEEE, pp. 830–835.
- Simroth, A., Zahle, H., 2010. Travel time prediction using floating car data applied to logistics planning. *IEEE Trans. Intell. Transp. Syst.* 12, 243–253.
- Soriguera, F., Robuste, F., 2010. Requiem for freeway travel time estimation methods based on blind speed interpolations between point measurements. *IEEE Trans. Intell. Transp. Syst.* 12, 291–297.
- Taylor, N.B., 2003. The CONTRAM dynamic traffic assignment model. *Networks Spat. Econ.* 3, 297–322.
- Van Lint, J.W.C., 2008. Online learning solutions for freeway travel time prediction. *IEEE Trans. Intell. Transp. Syst.* 9, 38–47.
- Van Lint, J.W.C., 2004. Reliable travel time prediction for freeways. Netherlands TRAIL Research School.
- Van Lint, J.W.C., Van Hinsbergen, C., 2012. Short-term traffic and travel time prediction models. *Artif. Intell. Appl. to Crit. Transp. Issues* 22, 22–41.
- Wu, C.-H., Ho, J.-M., Lee, D.-T., 2004. Travel-time prediction with support vector regression.

- IEEE Trans. Intell. Transp. Syst. 5, 276–281.
- Yang, J.-S., 2005. A study of travel time modeling via time series analysis, in: Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005. IEEE, pp. 855–860.
- Yildirimoglu, M., Ozbay, K., 2012. Comparative Evaluation of Probe-Based Travel Time Prediction Techniques Under Varying Traffic Conditions. Transp. Res. Board 91st Annu. Meet. Washingt. DC.
- Yu, B., Song, X., Guan, F., Yang, Z., Yao, B., 2016. k-Nearest neighbor model for multiple-time-step prediction of short-term traffic condition. J. Transp. Eng. 142, 4016018.
- Zhao, J., Gao, Y., Tang, J., Zhu, L., Ma, J., 2018. Highway Travel Time Prediction Using Sparse Tensor Completion Tactics and-Nearest Neighbor Pattern Matching Method. J. Adv. Transp. 2018.
- Zhu, T., Kong, X., Lv, W., 2009. Large-scale travel time prediction for urban arterial roads based on Kalman filter, in: 2009 International Conference on Computational Intelligence and Software Engineering. IEEE, pp. 1–5.