DEIM Forum 2019 H8-5

Forecasting of Pedestrian Counts at City Location Points

Sridhar BABU M †, Adam JATOWT†, Yihong ZHANG†, and Masatoshi YOSHIKAWA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshidahonmachi, Sakyo Ward, Kyoto, Kyoto Prefecture 606-8501, Japan
E-mail: †{sridharbabu1138,jatowt,yhzhang7}@gmail.com, ††yoshikawa@i.kyoto-u.ac.jp

Abstract The main purpose of this research work is to accurately analyze the pedestrian count flow in the city and predict the future count precisely. Walking is an important mode of transport, and is a non-polluting and an energy burning activity. With the increasing urban development and infalted fuel prices around the world more people prefer to use bicycles or walk than any other mode of transport. However, the flow of pedestrian varies across different location points in the same city. Existing research works only focus on the collective location points around the same area and are not very specific to particular locations. Our research focuses on each location point separately and makes use of the rich historical pedestrian count data collected from sensors placed around different locations in the city. In addition, it also incorporates other external factors associated with pedestrian movement to collectively model them and finally predict the future pedestrian count value. After pre-processing the input data and deploying the models, the recurrent neural network models have shown to outperform the time series models. The RMSE error measure for the best model is 17.23.

Key words Pedestrian Count, Time Series, Recurrent Neural Networks

1 INTRODUCTION

With millions of people travelling around the city every day, it becomes very vital for the urban planners and designers to come out with an optimized plan to accommodate the crowd. Increasingly, many city corporations and institutes for health and care excellence are collaborating to create new proposals to increase the amount of physical activity in peoples daily lives. As result, more and more number of people are regularly using the walking mode of transport and planners are asked to design more wide and sophisticated pavements including bumps and grooves with anti-glare surfaces for those with visual impairments.

So, the inference is rapid increase in the pedestrian walking. With the increase in pedestrian counts, there will be increase in number of business commodities in that particular locality. Particularly, if there are more number of people walking in the mid night time and early morning it can be regarded as a safe locality to travel. Safety index of the city is also directly related to the number of pedestrians moving in that particular location of city. Often crimes are happening in the secluded or less people moving areas. The pedestrian flow in a particular location point then provides us insights about the safety index of the location. If pedestrian flow is more high the probability of crime is less with the notion that crimes usually happen in remote or secluded streets. One more important and economical aspect of determining the pedestrian count is for urban planning and management. Either the case of opening a new shop, business in a locality or design of roads and pavements the planners need to do a detailed analysis on pedestrian flow in that locality since they are the potential customers.

The real question that needs to be answered is how to predict the future pedestrian counts at different location points in the city. There are two cases of scenarios for the above problem. One case assumes the availability of rich historical counts of pedestrian flow and requires developing temporal models to process and perform regression for future forecasting. The other case assumes complete lack of historical data (e.g. a city with no pedestrian count sensors installed). In this work, the focus is on the data rich location points i.e the city location points for which we have some historical pedestrian counts.

The following research work focuses on the case one, assuming the city has rich historical pedestrian counts. With the presence of these rich data, couple of time series and neural network models are proposed. These models are efficiently trained with these historical temporal data to forecast pedestrian counts for the future. In this work, the prime focus is only the pedestrian counts of next day in particular city location point.

The rest of the paper is organized as follows. Section 2 contains related works and literature survey. Section 3 is about the methodology. Section 4 presents the experiments

and datasets used for the project. Section 5 briefs about the conclusion of the work and future work that is in progress with the current research project.

2 RELATED WORKS

Pedestrian counting strategies are vital statistics for urban and city planning. Manual counting is however prone to errors and more cost expensive. In addition, higher pedestrian volumes are directly related with the roads or streets with high activity zones [1]. However, it is of prime importance to efficiently monitor and predict the future pedestrian demand zones for social welfare.

Predicting the future demand is useful for many real time applications. Once the passenger demand zone is identified, it will be a useful insight for the taxi, cabs to navigate to that particular destination. On the other hand, it is an additional information for navigation purposes. Once the most people moving area is known, that particular route can be avoided to reduce travel time. There are several works that analyze the pedestrian traffic congestion. One more reason for traffic congestion to happen is the case of planned special events happening in the city [5]. If the events details and previous pedestrian flow data are known these kind of chaotic situations could have been averted.

Another similar kind of problem is metro passenger flow forecasting in a city. To forecast the metro transit passenger flow two nearby events, namely, a tennis game and a baseball game are considered [3]. Social media is also widely used as a proxy to estimate the pedestrian flow in the locality. Presence of twitter data makes it an additional feature for regression modelling. Extraction of a number of unique user tweets were found to be directly proportional to passenger flow counts in the metro transit [3]. This gives the researchers information about using social media as a proxy for pedestrian count estimation.

In the above, both the physical and social media worlds are discussed. Similarly, there exists a cyber-social world that is predominantly gaining popularity in the research domain. Increasingly, number of people involved in the virtual gaming environment. As a result, the number of people playing pokemon-go in virtual cyber world were used to predict the pedestrian flow in the neighbouring locality [8]. However, they were not used as a single feature separately but their addition gives more useful insights and more conclusions were drawn.

Moreover, there are several factors that also influence people movement in the city. One such vital parameter is weather factor. Traffic flow characteristics, pedestrian walking environment, transportation safety are all having a major impact because of inclement weather[10]. Hence, the model may perform better if the external factors are also included in implementation.

Since the entire data is temporal, there are several types of time series analysis that could be done. For one kind of similar problem, the ARIMA model was developed to predict the water quality [6]. Previous works have also utilized the time series model, but on the raw sensor data without taking the seasonality into account [2]. Similarly, time series were used to predict the air quality but neural networks outperformed their accuracy [4]. In addition, the deep neural network seems to be promising in handling spatial temporal data with less error and more accuracy [7]. This gives us more information about using neural networks in our study and analyze the accuracy of models with respect to current problem.

3 METHODOLOGY

The flow of the system in fig(1a) depicts the entire process organization. After the raw data is extracted from the pedestrian counting system, the raw data and meta data are collaborated in the data processing before being sent for model building. After the data processing the below set of models are applied and each model separately performs the forecasting of future pedestrian counts. List of models used and their working details are mentioned below.

3 1 Auto Regressive Integrated Moving Average

Time series models are very resourceful since they can model both linear and non-linear patterns with accurate predictions. One such time series model which is useful for our current motivation is the Auto Regressive Integrated Moving Average model or ARIMA. It is a uni-variate time series models, basically used for modelling linear time series data and mainly for stationary time series. It is especially useful for short term forecasting of future values.

3 1. 1 AutoRegressive Model

This is one part of the ARIMA model, the AutoRegressive part of the model also represented as AR(p). AR(p) equation is

$$x_t = \phi 1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + e_t$$

3 1. 2 MovingAverage Model

Moving average model or MA(q) makes use of past forecast errors in regression like model. MA(q) equation is

$$y_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} + e_t$$

To model ARIMA time series, we need to implement couple of relation functions and determine the p,d,q parameters. The list of functions to be implemented includes, auto correlation function, partial auto correlation function. From the

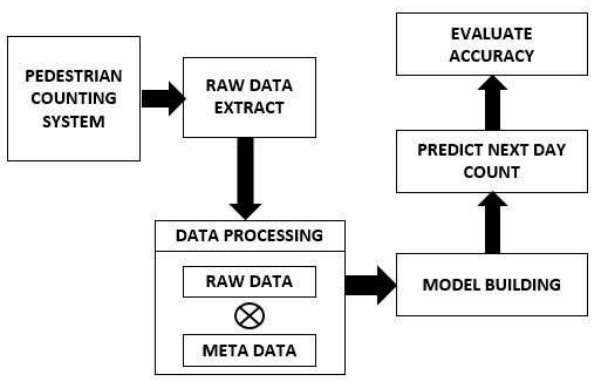

Figure 1 Architectural Flow of the System

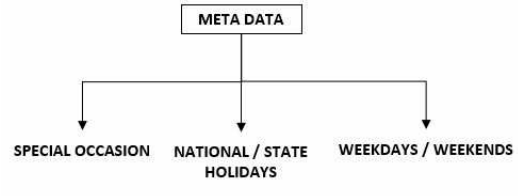

Figure 2 Metadata Analysis

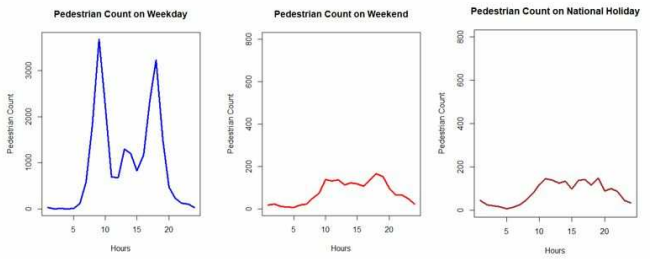

Figure 3 Difference of Pedestrian Count on different Days

implementation, we need to determine the parameter values for the model.

3 2 Vector Auto Regressive Model

The Above time series model is useful for modelling univariate time series data. Several cases we may have more features and more related set of time series. In such cases, Vector Autoregressive (VAR) models are very useful for modelling multivariate time series data. The VAR model is represented as follows:

$$y_t = d_t + A_t Y_{t-1} + A_2 Y_{t-2} + ... + A_p y_{t-p} + u_t$$

where $y_t = y_{1t}, ..., y_{kt}$ is a list of k observed time series at time t. In this case, k=2 (Pedestrian count and weather data). y_{t-1}, y_{t-2} are list of time series observed at time $t-1, t-2$ and so on.A_i is the coefficient matrix for the i^{th} value, u_t is the error and d_t is a constant.

Each set of variables in the time series is a linear function of past lags of itself and past lags of the other variables. Similar to ARIMA model VAR model also has a parameter value to be determined. The value can be determined with the help of auto correlation and cross correlation functions. Since we have the pedestrian count data, along with weather data time series we can use them to model the vector autoregressive model. By training simultaneous weather and pedestrian count time series we can use one of the two variable as a predictor variable to predict their future count value, for their corresponding weather series.

3 3 Recurrent Neural Network

Recurrent Neural Networks is one of the prominent type of neural networks used in the recent research works. The idea behind RNNs is to make use of sequential information. If we want to predict the next value in a sequence we need to know list of values came before in the sequence. The reason why RNNs are called recurrent is because they perform the same task for every element of a sequence, with the output being dependent on the previous computations.

Another way to think about RNNs is that they have a !Hmemory ! Iwhich captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but they suffer from vanishing gradients problem.

3 4 Long Short Term Memory

LSTM networks are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter and Schmidhubler[9] and were refined and popularized by many people in research domain. They work tremendously well on large variety of problems and are now widely used for variety of applications. LSTMs are explicitly designed to avoid a long-term dependency problem. Remembering information for long periods of time is practically their default behaviour, not something they struggle to learn.

All recurrent neural networks have the form of a chain of repeating modules of neural networks. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer. LSTM's also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.

4 EXPERIMENTS

4 1 DATASET DESCRIPTION

The dataset used for the model building consists of pedestrian count data collected from sensors placed around the different location points in the city of Melbourne. At present they have around 40 sensors placed in different streets of the city. These sensors with hourly granularity take the counting

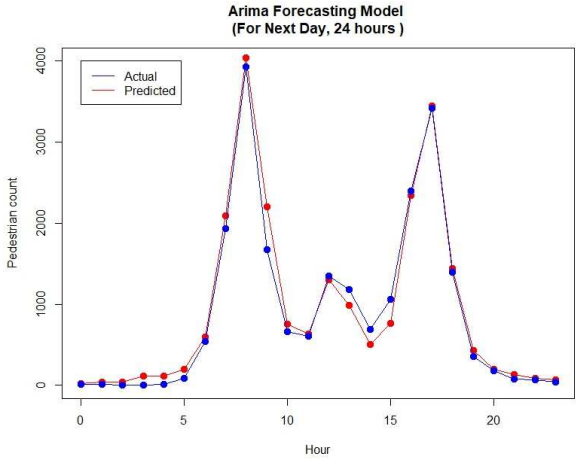

Figure 4 ARIMA Model forecasting for next day (24 hours)

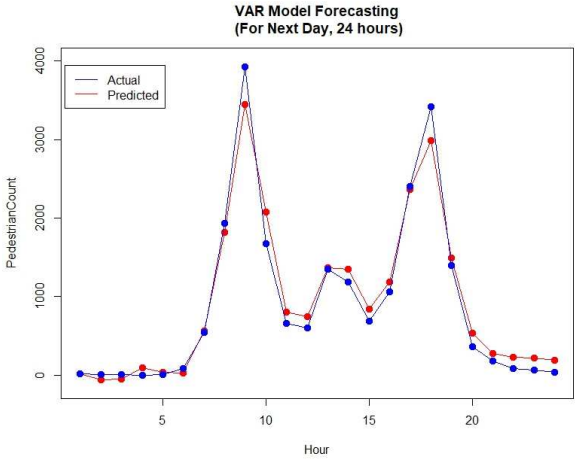

Figure 5 VAR Model forecasting for next day (24 hours)

of number of people walking or crossing that specific locality [1]. They send the updated counts to the central server, which publishes the data online. We collected the data from Jan 2016 to Dec 2018. Then the raw data is carefully pre processed before applying to the model building.

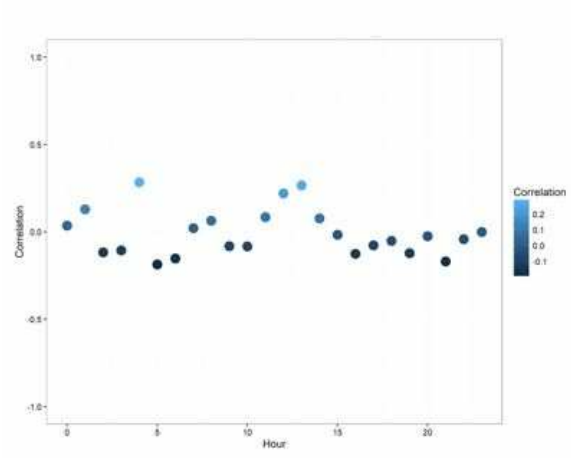

Figure 6 Weather vs Pedestrian Correlation

4 2 WEATHER DATA

In order to compute the pedestrian count and weather data correlation, we need to extract the weather details for that particular location. With the help of darksky API[2] interface, weather details for those different locations are collected in the same period of pedestrian count data. Then the correlation coefficient and in turn vector autoregressive model are computed.

Apart from the above two datasets, we also crawled the national holidays, state holidays and weekdays/weekends data. These are just the Meta data mentioned in the figure.

[1] http://www.pedestrian.melbourne.vic.gov.au/

[2] https://darksky.net/dev

4 3 WEATHER CORRELATION COMPUTATION

The weather data and its corresponding pedestrian count at same time are taken and for each hour the correlation coefficient is computed and then the average of the all the correlation coefficient is taken for overall results.

$$\forall Hour_1^{24} = \frac{1}{m}\sum_{i=0}^{m} pearson(w_i, p_i)$$

The correlation plot in fig(3) is average of all correlation obtained at all the sensor locations taken for the study. The correlation graph reveals that the weather at most of the hours has an influence on the pedestrian counts. At sometimes, the pedestrian counts may not be positively correlated with weather. For example,at hours 4,5,9 they are mostly negatively correlated. Nevertheless, they are well positively correlated at other hours, so we can include them for VAR modelling.

4 4 MODEL BUILDING

First one of the baseline model ARIMA model, is constructed. With the three to six months of data taken for model building, we predict the pedestrian count for the next day. With the help of ACF and PACF functions, we were able to retrieve optimum parameters for the model. As a result ARIMA gives a better result of forecasting.

With the same set of pedestrian count data, repeat the process along with the series of weather data obtained for the same time period. Then construct the VAR model by training the model and predicting the next day pedestrian count. It can be seen that VAR model slightly underperforms compared to ARIMA model.

Since the time series models are having quite a long series of data, it becomes little difficult to capture the dependencies. These traditional models have difficulties capturing them and hence the methodology to try out neural networks

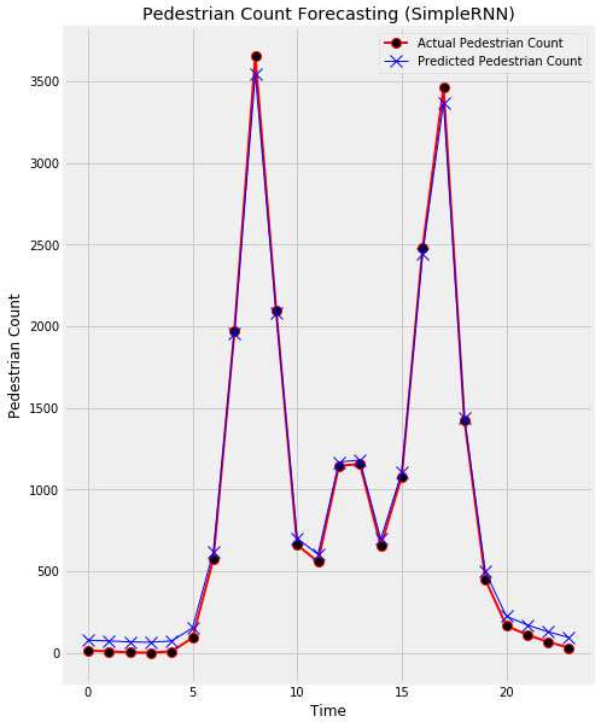

Figure 7 SimpleRNN forecasting for next day (24 hours)

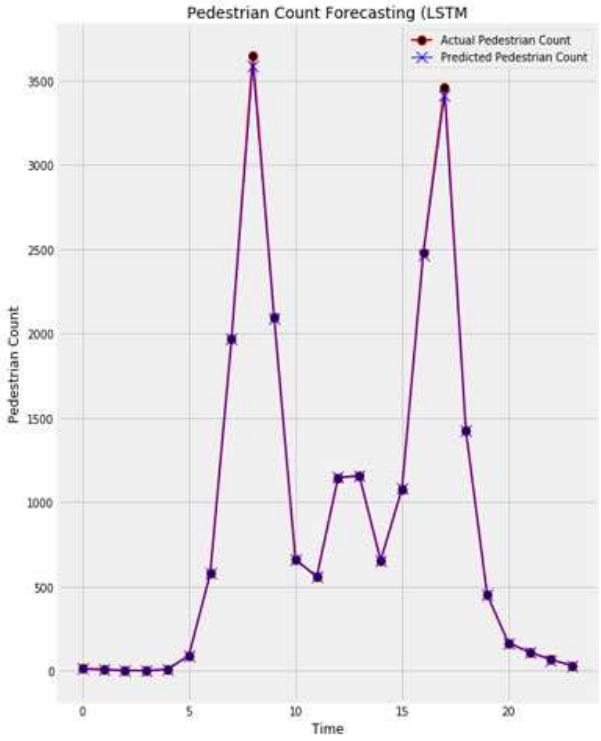

Figure 8 LSTM forecasting for next day (24 hours)

was proposed.

Recurrent neural networks a type of neural networks particularly suited for sequential data is taken to handle this problem. The experiment is repeated with the recurrent neural network for learning the model. Model training and testing was done with the same set of data as it was done with time series models. RNN performed way better than the traditional time series models.

With the same set of training data, each day of twenty four hours are sent as input to the model and alternatively remaining set of days are sent as input once the previous day is processed. As a result a set of sequential data for recurrent neural network model is developed. Finally, once the model training and testing is done it predicts the pedestrian count for the next day.

Since recurrent neural network suffers from vanishing gradient problems, special kind of neural network and one of the prominent ones for handling sequential time series data LSTM is used. Exactly the same procedure, that is followed for RNN is repeated for LSTM modelling. Same set of data with same training and testing time are taken. LSTM model outperforms all the previous models. Though the difference in accuracy measures are less between neural network models, but we can clearly see the difference in computation of LSTM and traditional time series models.

5 RESULTS

All the mentioned time series and neural network models are implemented for the given problem.

Model	RMSE
ARIMA	55.64
VAR	57.26
BasicRNN	25.23
LSTM	17.23

Table 1 Model and Error Accuracy Measures

fig(3) is the results of time series forecasting for the future pedestrian count. It trains and predicts the pedestrian count for the next day. Fig(3.a) represents ARIMA model and Fig(3.b) represents VAR model.

Fig(4) is the future forecasting graph obtained by the neural network models. Fig(4.a) and fig(4.b) represent the SimpleRNN based and LSTM based forecasting results respectively. In all the results of forecasting, the graphs are plotted for pedestrian count against the corresponding hours of the day. Each graph has both the actual value of pedestrian count observed and the predicted value by the model curves plotted. Time series model tend to have a quite a huge margin of errors in forecasting compared to the neural networks.

Table.1 shows the list of models tried in the experiment and their corresponding error measures. The accuracy measure used here to determine the model efficiency is root mean square error (RMSE). In the Table.1 each model with their

RMSE are mentioned. The model and accuracy results provide a clear cut insight that LSTM performs bettern than simple RNN and in more brief the neural network outshine the time series models in learning short term forecasting of pedestrian counts.

6 CONCLUSION and FUTURE WORK

In this paper, the pedestrian count data along with external parameters like weather are included for the modelling and have been successfully implemented. The time series models like ARIMA with only pedestrian count data, VAR with both people count and weather data are implemented. Then neural network models for sequential data learning like RNN, LSTM are also implemented successfully. The overall work is more of analysis on the prediction of pedestrian count data for the next day, taking into account the rich historical data in the past. From the analysis, conclusion is LSTM performs better in terms of accuracy compared to the traditional time series models for forecasting the future pedestrian counts.

For future work, there is more scope in improving the performance of model building for the current type of problem. In the current work, there are lot of models constructed for each location point and for different days. In future, there will be considerable need to develop less number of models by incorporating information about holidays, weekdays and weekends into the model rather than being done at preprocessing.

Secondly, the most crucial and challenging work will be to predict the pedestrian counts at location points where there are no sensors deployed and there is complete absence of historical data. Forecasting the count in data scarce locations in the city by utilizing the values in the data rich city will be significant and essential for research community and urban development managers.

References

[1] Schneider, Robert J., Todd Henry, Meghan F. Mitman, Laura Stonehill, and Jesse Koehler. *"Development and application of the San Francisco pedestrian intersection volume model."* (2013).

[2] Wang, Xianjing, Jonathan Liono, Will Mcintosh, and Flora D. Salim. *"Predicting the city foot traffic with pedestrian sensor data."* (2017).

[3] Ni, Ming, Qing He, and Jing Gao.*"Forecasting the subway passenger flow under event occurrences with social media."* IEEE Transactions on Intelligent Transportation Systems 18, no. 6 (2017): 1623-1632.

[4] Freeman, Brian S., Graham Taylor, Bahram Gharabaghi, and Jesse Th. *"Forecasting air quality time series using deep learning."* Journal of the Air and Waste Management Association (2018): 1-21.

[5] Kwoczek, Simon, Sergio Di Martino, and Wolfgang Nejdl. *"Predicting and visualizing traffic congestion in the presence of planned special events."* Journal of Visual Languages and Computing 25, no. 6 (2014): 973-980.

[6] Faruk, Durdu mer. "A hybrid neural network and ARIMA model for water quality time series prediction." Engineering Applications of Artificial Intelligence 23, no. 4 (2010): 586-594.

[7] Zhang, Junbo, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. *"DNN-based prediction model for spatio-temporal data."* In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 92. ACM, 2016.

[8] Wang, Derek, Tingmin Wu, Sheng Wen, Donghai Liu, Yang Xiang, Wanlei Zhou, Houcine Hassan, and Abdulhameed Alelaiwi. *"Pokmon GO in Melbourne CBD: A case study of the cyber-physical symbiotic social networks."* Journal of computational science 26 (2018): 456-467.

[9] Pedestrian Counting System - Government of Melbourne initiative *http://www.pedestrian.melbourne.vic.gov.au/*

[10] Lin, Lei, Ming Ni, Qing He, Jing Gao, and Adel W. Sadek. *"Modeling the impacts of inclement weather on freeway traffic speed: exploratory study with social media data."* Transportation Research Record: Journal of the Transportation Research Board 2482 (2015): 82-89.