Implications of spatio-temporal data aggregation on short-term traffic prediction using machine learning algorithms

Rivindu Weerasekera, Mohan Sridharan, and Prakash Ranjitkar

***Abstract*—Short-term traffic prediction, which uses historical data collected by traffic management agencies to construct models that can reliably predict the flow of traffic at specific locations in road networks, are a key component Intelligent Transportation Systems. Despite being a mature field, short-term traffic prediction still poses some open problems. For instance, it is not clear how the data resolution effects accuracy and responsiveness of models to non-recurring congestion, especially when considering spatio-temporal dependencies. In this paper, we evaluate the ability of Artifical Neural Networks, Random Forests and Support Vector Regression algorithms to reliably model traffic flow and their ability to be responsiveness to unexpected events such as accidents. We also look at different feature selection methods and examine the spatio-temporal attributes that most influence the reliability of these models. We find that aggregation is not necessary to achieve good performance for multivariate spatio-temporal models. We also find that feature selection based on Recursive Feature Elimination outperforms linear correlation based feature selection.**

I. Introduction

TRAFFIC congestion results in significant monetary losses in countries around the world, with the cost of traffic congestion in 2014 estimated to be $160 billion in the US alone [Schrank et al., 2015]. A significant amount of effort has been put into reducing congestion in cities. In many cities, it is becoming impractical to build new roads, or to expand existing roads, and it is becoming all the more important to make best use of the available resources. Intelligent Transportation Systems, Advanced Traffic Management Systems, and route guidance systems, use real-time data of traffic flow gathered from various sensors. In such systems, short-term traffic prediction, which helps make decisions based on predictions of traffic in the near-future, is more useful than just using the real-time data of traffic conditions. The field of short-term traffic prediction is not new. It is over 30 years old with early work utilizing Box-Jenkins ARIMA methods [Ahmed and Cook, 1979]. Recent approaches still use variations of the original ARIMA models (such as Seasonal ARIMA [Smith et al., 2002, Williams and Hoel, 2003]), but there has been a shift towards using machine learning algorithms to address the traffic prediction challenges [Karlaftis and Vlahogianni, 2011]. Although such models based on machine learning algorithms have been shown to be more reliable than the traditional ARIMA models, there are still many open problems [Vlahogianni et al., 2014]. These include building responsive algorithms that are able to predict non-recurring congestion, determining the optimum data-resolution, and identifying and modeling the important spatio-temporal dependencies in traffic data. The study described in this paper is a step towards addressing these challenges. We make the following key contributions:

- Explore the effect of resolution of multivariate spatio-temporal input data on the accuracy of the predictions made by the models built using three machine learning algorithms, Artificial Neural Networks, Support Vector Regression, and Random Forests.
- Evaluate the responsiveness of these predictive models to non-recurring congestion events. Specifically, we study the reliability of the predictions provided by these models in the presence of unexpected events such as accidents.
- Identify and examine the traffic attributes that most influence the performance of these models and their ability to model the complex, spatio-temporal dependencies in traffic data.

We illustrate these contributions using historical data of volume and occupancy measurements on a highway in Auckland (New Zealand). We first motivate the need for the proposed study by discussing related work in Section II. Next, Section III describes the dataset and methodology used to build and evaluate the predictive models, and Section IV describes the machine learning algorithms used to build these models. Section V describes the hypotheses and measures used for experimental evaluation, and Section VI analyzes the corresponding experimental results. Finally, Section VII discusses the conclusions and directions for future work.

II. Background

Many algorithms have been developed for short-term traffic prediction, which is a complex problem influenced by a variety of factors such as the resolution (i.e., the aggregation level) of the input and output data, and spatio-temporal dynamics. We review some of the related work in this section.

Although studies in the existing literature predominantly use data aggregated over 5min and 15min intervals, some prior studies have investigated the effect of data resolution on the reliability of the predictions provided by the corresponding models; the results have, however, been inconclusive. For

R. Weerasekera is with the Department of Electrical and Computer Engineering, The University of Auckland, New Zealand. e-mail: rwee015@aucklanduni.ac.nz

M. Sridharan is with the School of Computer Science, The University of Birmingham, UK. email: m.sridharan@bham.ac.uk

P. Ranjitkar is with Department of Civil and Environmental Engineering, The University of Auckland, New Zealand. email: p.ranjitkar@auckland.ac.nz

instance, Park et al. [2009] investigated the effect of aggregation on travel time prediction, and considered aggregation levels from 2min to 60min in the context of an ARIMA model. They concluded that higher levels of aggregation were required to forecast route travel time than when forecasting link travel times. Dougherty and Cobbett [1997] constructed a Neural Network model for making predictions, and found that data aggregated over 5min intervals gives better results than data aggregated over 1min intervals. Vlahogianni and Karlaftis [2011] looked at aggregation levels and, although they found that temporal aggregation may distort critical traffic flow information, they also concluded that further research was necessary to determine the optimum aggregation level(s).

The use of high-resolution data is challenging for multiple reasons. First, for some statistical models used for short-term traffic state prediction, it is necessary to ensure that the input data and the output data have the same aggregation level, but this constraint can be relaxed when machine learning algorithms are used to build predictive models. Second, while research shows that the high-resolution data (as expected) includes more accurate measurements, e.g., Martin et al. [2003] state that inductive loops are "one of the most accurate count and presence detectors", it also makes the noise in sensor measurements more distinct. Although data from these inductive loops can represent individual vehicles in the network, computational models developed to capture the flow of vehicles between segments or links in the network need to be robust to such noise and be able to capture spatio-temporal dynamics in order to exploit the information encoded in high-resolution data. Studies based on univariate time-series methods often perform aggregation to smooth out the variability in higher resolution data [Vlahogianni and Karlaftis, 2011], however these data smoothing techniques result in loss of information (and sensitivity) and make it difficult for the corresponding models to capture the spatio-temporal dynamics of traffic flow. In the study reported in this paper, we fixed the resolution of the output data (i.e., for the predictions being made) and examined the effect of different input data aggregation levels on the prediction accuracy.

There has been considerable research on analyzing the effects of spatio-temporal dynamics. For instance, Kamarianakis and Prastacos [2003] used a Spatio-Temporal Autoregressive Moving Average (STARIMA) model to incorporate data from links upstream to the link of interest in their prediction model, and Chandra and Al-Deek [2009] found that vector auto-regressive models that incorporate data from links neighboring the link of interest perform better than ARIMA models that do not consider the data from the neighboring links. Yang et al. [2015] found that a sparse selection of neighbors chosen based on the level of correlation with the link of interest improves performance. Min and Wynter [2011] showed that a multivariate spatio-temporal model with templates was able to provide very good prediction accuracy. However, these models depend on fixed correlations matrices that are modified infrequently. As a result it is difficult for these models to track changes or to capture sudden (or significant) changes between congested and free-flowing traffic conditions.

In addition to the approaches that build on the ARIMA models [Ahmed and Cook, 1979, Kamarianakis and Prastacos, 2003, Min and Wynter, 2011, Smith et al., 2002, Williams and Hoel, 2003], models based on machine learning and probabilistic estimation algorithms have also been explored because they are well-suited to model the complex spatio-temporal relationships in data. Popular approaches include Artificial Neural Networks (ANN) [Vlahogianni et al., 2005, Dunne and Ghosh, 2013, Sun et al., 2012, Ban et al., 2016, Wang et al., 2016], Support Vector Machines (SVM) [Castro-Neto et al., 2009, Jeong et al., 2013, Asif et al., 2014, Cheng et al., 2016, Yao et al., 2016], k-Nearest Neighbors (kNN) [Davis and Nihan, 1991, Chang et al., 2012, Oh et al., 2015b, Cai et al., 2016, Xia et al., 2016], Kalman Filters [Okutani and Stephanedes, 1984, Xie et al., 2007, Guo et al., 2014], Bayesian Networks [Ghosh et al., 2007, Horvitz et al., 2005, Pascale and Nicoli, 2011], and Random Forests [Zarei et al., 2013, Hamner, 2010]. For instance, existing work has explored various ANN configurations. Wang et al. [2016] developed a space-time delay neural network (STDNN) that included 22 links in central London and showed that this model outperforms a STARIMA model. Hodge et al. [2014] used a binary neural network that incorporates spatio-temporal data for traffic prediction. Vlahogianni et al. [2005] used a neural network model optimized with genetic algorithms and found that incorporating spatial and temporal data was helpful for multi-step predictions. More recently, there have been efforts to use deep neural network architectures including deep belief networks [Huang et al., 2014, Soua et al., 2016] and stacked auto-encoders [Lv et al., 2015].

There is no agreement in the literature regarding the number of upstream and downstream links (neighboring any link of interest) that should be considered while building the predictive models. While some algorithms consider just one upstream or downstream link [Xia et al., 2016, Yao et al., 2016], others consider a variable number of upstream and downstream links [Hodge et al., 2014]. For an extensive review of spatio-temporal forecasting, see Ermagun and Levinson [2018]. As noted in Vlahogianni et al. [2014], capturing spatial attributes in traffic data from a freeway is still an open problem.

Most existing work on short-term traffic prediction focus on typical conditions [Castro-Neto et al., 2009]. Traffic is (on average) inherently periodic with daily or weekly patterns and many studies exploit this periodicity in their algorithms. However, accurate predictions are arguably more useful in situations of non-recurring congestion such as accidents where periodic patterns do not hold. Of the studies that do not leave out non-recurring congestion in their input data, a common approach is to create multiple models to deal with different conditions. For example, Dunne and Ghosh [2011] used a model with nonlinear pre-processing in cases of congestion. Fusco et al. [2016] reported good performance during non-recurring congestion with a SARMA model, while a Bayesian Network performed better during recurring congestion. An online-SVR based model was found to accurately predict non-recurring congestion by Castro-Neto et al. [2009]. Pan et al. [2013] also highlight some of the challenges in capturing moving bottlenecks and non-recurring congestion. See Vlahogianni et al. [2014], Ermagun and Levinson [2018], Oh et al. [2015a,

2018] for a more comprehensive overview of the existing literature.

In this study, we explore three machine learning algorithms that have demonstrated the ability to incorporate spatio-temporal data in predictive models built for intelligent transportation and other applications. Specifically, we explore: (1) Artificial Neural Networks (ANN); (2) Support Vector Regression (SVR); and (3) Random Forests (RF). We chose ANN and SVR because they are the most widely used machine learning algorithms used to build predictive models in the literature. We chose Random Forests since it is an ensemble learning algorithm that requires a small number of parameters to be tuned. We would like to highlight that the aim of this study was not to introduce new algorithms. This study makes three key contributions. First, we examine how the predictive accuracy of models based on these algorithms changes as a function of the aggregation level of the input data. Second, we explore the ability of these models to respond accurately to non-recurring congestion conditions. Third, we identify the attributes that most influence the predictive accuracy of these models, to identify the important spatio-temporal dependencies in the traffic data and establish the ability of machine learning algorithms to model these dependencies.

III. Methodology

This section introduces the study area and data, and provides a mathematical formulation of the short-term traffic prediction problem (Section III-A). This is followed by a description of the data pre-processing steps used in the proposed study (Section III-B).

A. Study Area and Mathematical Formulation

This study was carried out in a $30km$ section of State Highway 1 (SH1) in Auckland, New Zealand. We considered data from 45 segments along SH1 from the suburb of Papakura towards Auckland City (see Figure 1). On average there are 3 lanes of roadway in each direction and we only considered lanes going northbound in this study. The average length of a segment was $674m$, with the length varying between $52m$ and $2252m$.

Traffic can be measured in different ways. The most common sensor used to collect traffic data is the Inductive Loop Detector, which comes in different forms. Dual loop detectors, which have two inductive loops placed a short-distance apart, are able to accurately capture the speed of a vehicle going over them, the volume (i.e., count of vehicles passing the detector), and occupancy (i.e., the amount of time a vehicle was over the detector). However, most of the loops in many cities (including Auckland) are single loop detectors, which can measure volume and occupancy, but can only estimate vehicle speed as a function of these measured values and the average effective vehicle length. Research shows that measuring speed with a constant effective vehicle length can lead to errors of up to 50% [Jia et al., 2001]. Using these derived speed estimates for making decisions can lead to misleading results—we thus did not use speed data in this study.

The fundamental diagram of traffic flow established by traffic engineers considers the relationship between three key traffic variables (1) flow (volume); (2) density; and (3) speed. Since density is difficult to measure directly, occupancy is frequently used as a substitute [Ryus et al., 2010]. Entirely describing the current state of traffic is not possible using only information about flow. For example, if 200 vehicles pass over a detector during a 5min interval, this could correspond to free-flow conditions during early mornings and evenings, but it could also correspond to highly congested conditions due to an accident during peak hours. Unlike many existing studies that have only considered flow variables when making predictions, we consider both volume and occupancy because both of these variables provide useful information.

For all the predictive models, the input vector $X(s,t)$ takes the form of:

$$X(s,t) = \begin{matrix}[V^1_{t-T}, & O^1_{t-T}, & \cdots, & V^s_{t-T}, & O^s_{t-T}, & \cdots, & V^S_{t-T}, & O^S_{t-T},\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V^1_{t-1}, & O^1_{t-1}, & \cdots, & V^s_{t-1}, & O^s_{t-1}, & \cdots, & V^S_{t-1}, & O^S_{t-1},\\ V^1_t, & O^1_t, & \cdots, & V^s_t, & O^s_t, & \cdots, & V^S_t, & O^S_t]\end{matrix} \quad (1)$$

where V^s_t and O^s_t denote volume and occupancy (respectively) of segment s at time-step t, S is the total number of segments, and T is the total number of historical time-steps considered. The output of each such model is the volume or occupancy aggregated over the subsequent five-minute interval for each specific segment s of interest. The goal of each algorithm used to build a predictive model is to find a functional relationship between the inputs and outputs. For instance, if traffic volume is to be predicted, the output V^s_{t+5min} of the models is given by:

$$V^s_{t+5min} = f(\, X(s,t)\,) \quad (2)$$

The output is thus a function of the input vector. The machine learning algorithms build models that approximate this function to predict the output for any given input.

B. Data Processing

Data from 30 days of April 2016 was collected for 45 segments ($S = 45$) on the motorway. In order to get segment level data from loop detectors, individual values were aggregated across the lanes (volume data was summed and occupancy was averaged) for each segment and at each point in time. We use the *volume and occupancy* values of all segments in the past 20 time-steps ($T = 20$), resulting in an input vector with 1800 attributes. To ensure that each segment has data from a reasonable number of upstream and downstream segments, predictions are only made for segments $20 - 25$ on the motorway (see Figure 1). Recall that volume and occupancy readings were reported every 30 seconds which correspond to **86400** time-steps. A naive aggregation would have resulted in smaller datasets of 8640 samples and 2880 samples for $5min$ and $15min$ aggregation respectively. To minimize the imbalance in the size of the datasets, a sliding window approach was used, resulting in a new sample being generated every 30 seconds for all the aggregation levels. The

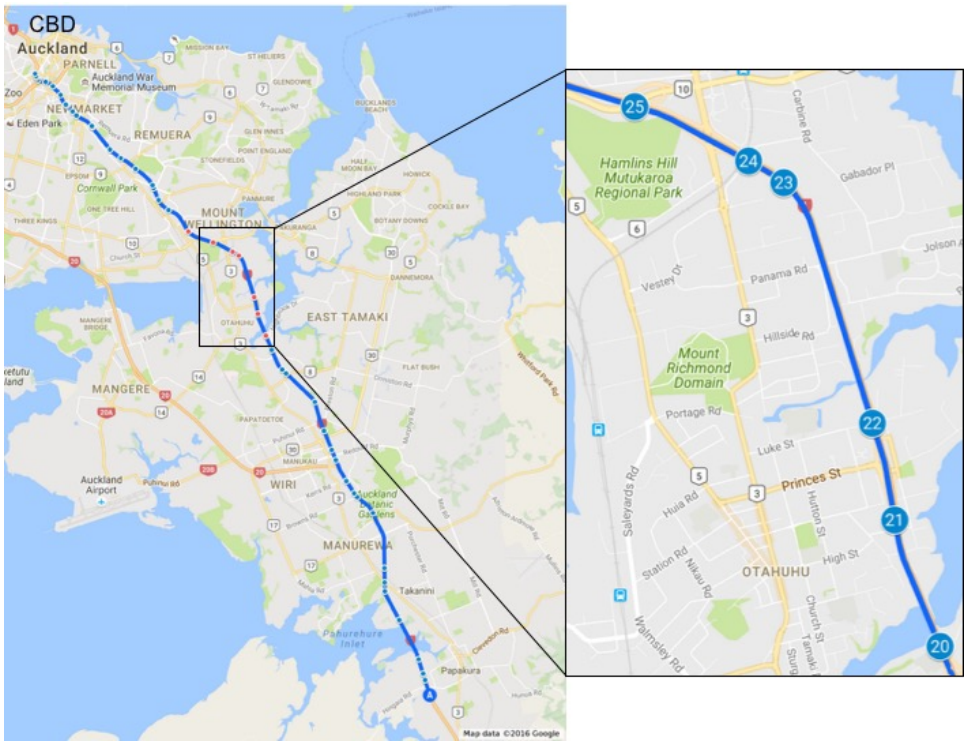


Fig. 1. Study area with 45 road segments on State Highway 1 in Auckland.

final size of the input dataset, with 20 time-steps included in each input sample, was thus 86370 samples for $30s$ resolution, 86190 for $5min$, and 85790 for $15min$ aggregation. Also, to ensure a fair comparison, the output is aggregated over the same time period for each model for all input time resolutions, i.e., the amount of time represented in the input depends on the resolution of the data, whereas in the output, all models will consider the aggregated values over the interval from when the final input reading was taken to five minutes past this time.

The dataset was pre-processed to remove some extreme values that were highly unlikely. First, we used winsorization [Ghosh and Vogt, 2012] to set the upper bound of the values in the dataset. Winsorization, a common approach for dealing with outliers, replaces all values above and below a certain percentile with the value of that percentile. In this paper, we set the upper percentile to 99.97% so that all values above this percentile are replaced by the value of this percentile. If a standard normal distribution is assumed, this choice of upper bound corresponds to clipping values that are ≥ 3.5 standard deviations from the mean. Figure 2 shows Segment 23 of the data before and after winsorization. Second, we scaled each attribute in the input data to lie $\in [0, 1]$—this scaling was especially important for producing stable results with Support Vector Regression and Artificial Neural Networks. Scaling was performed using the training data, and the corresponding scaling constants were applied to the test data. The occupancy values always stayed between 0% and 100% in the input and output, and no additional processing was needed to constrain the data to this range. Non-stationary time-series data is typically transformed to stationary data before applying time-series models. However, traffic data is

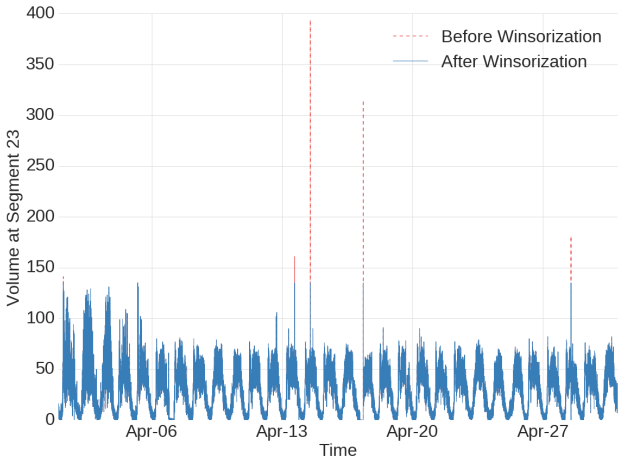


Fig. 2. Segment 23 before and after winsorization

considered to be cyclo-stationary and we model short-term traffic prediction as a multivariate pattern recognition problem with all data assumed to arise from the same underlying distribution. Thus, we did not perform any transformations to make the data stationary. Also, although the periodic nature of traffic can be exploited to improve the prediction accuracy of the learned models, doing so will make it difficult to reliably and quickly identify and respond to non-recurring congestion conditions.

Training of the models was accomplished using the first 20 days (57600 samples) of data, and the remaining 10 days of data were used for testing. The parameters of each model were tuned using the training dataset. Next, we briefly discuss the

algorithms that we used to build the models for short-term traffic prediction.

IV. Machine Learning Algorithms

In this section, we describe the three machine learning algorithms used to build the predictive models explored in this paper: Artificial Neural Networks (Section IV-A), Support Vector Regression (Section IV-B), and Random Forests (Section IV-C).

A. Artificial Neural Networks

Feedforward neural networks or multilayer perceptrons are the most common Artificial Neural Network (ANN) models. A neural network is composed of neurons arranged in layers with each layer containing one or more neurons. Each neuron is connected to all the neurons in its adjacent layers, and neurons within a layer are not connected. Each neuron takes a linear weighted sum of all its inputs x (from the layer before it) and passes it through a nonlinear activation function σ to produce the output y:

$$y = \sigma\left(\sum_{i=1}^{N}(w_i \cdot x_i)\right) \tag{3}$$

Each such output y is then used as an input to the next layer of neurons until the final (i.e., output) layer is reached. The weights associated with each neuron may be initialized randomly to enable each neuron to potentially learn a different function of its inputs.

The weights w_i associated with each neuron are the parameters defining the neural network model, and these parameters are estimated by minimizing a loss function that measures the difference between the output values estimated by the network and the ground truth values included in the training data. For regression problems, the squared error between the estimated and ground truth output values is generally used as the loss function. The back-propagation algorithm is then used to calculate the gradient of this error, and to propagate this gradient back through the network (towards the input layer) to update the weights of each neuron by gradient descent. Stochastic gradient descent algorithms are used widely to update the weights, and we used a stochastic gradient-based optimizer called *Adam* that is computationally efficient and is known to scale well to larger datasets [Kingma and Ba, 2014]. All parameters of this optimizer were set to their default values.

Although the nonlinear activation function in a neural network has traditionally been the sigmoid function, empirical results have indicated that the rectified linear unit (ReLU) activation function improves the ability to model complex relationships and reduces the time taken to train the model [Krizhevsky et al., 2012]. We thus used the ReLU activation function in a network with three hidden layers, each with 150 neurons. We performed 400 iterations of learning with mini-batches of data with 200 samples (each).

B. Support Vector Regression

For classification problems, a Support Vector Machine computes a decision boundary that maximize the margin between this boundary and the closest data sample. Support Vector Regression (SVR) uses a similar approach for regression problems—errors corresponding to estimated values within an ε distance from the ground truth values are ignored. More specifically, given a set of training data, the objective is to find a function $f(x)$ that produces at most ε deviation from the actual target values y_i for the the training data, and is as flat as possible [Smola and Schölkopf, 2004]. For instance, a linear function $f(\mathbf{x}) = w^T\mathbf{x}+b$ is flat if it has a small w—this can be accomplished by minimizing $||w||^2$. Since a function that satisfies all the required constraints C may not exist, some slack variables (ξ, ξ^*) are introduced to allow for some errors. We then obtain the following formulation for SVR:

$$\begin{aligned} \text{minimize } \left\{\frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi+\xi^*)\right\} \text{ subject to} \quad & (4)\\ y_i - w^T\mathbf{x}_i - b \le \varepsilon + \xi_i & \\ w^T\mathbf{x}_i + b - y_i \le \varepsilon + \xi_i^* & \\ \xi_i, \xi_i^* \ge 0 & \end{aligned}$$

We can also incorporate nonlinear kernel functions to extend SVR to nonlinear problems. Popular kernels include linear kernel and the Radial Basis Function (RBF) kernel, which transform the input sample into a higher dimensional space that results in better separation (for classification) or estimation of values (for regression). We experimentally chose to use a linear kernel for SVR because it provided better results.

C. Random Forests

Random Forest (RF) [Breiman, 2001] is an ensemble method for building classification or regression models. Ensemble methods combine predictions from multiple models to improve accuracy. In an RF, the ensemble is a set of decision trees trained on B subsets of the full dataset. Each subset is selected by a technique known as bagging or bootstrap aggregation. If the training set is defined as input vectors $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}...$ and the corresponding (target) output values $Y = y_1, y_2, y_3...$, decision trees will be created as follows:

for b in $1...B$ **do**
  Pick N training samples randomly with replacement; call this subset $\{\mathbf{X}_b, Y_b\}$
  Train a decision tree Θ_b using $\{\mathbf{X}_b, Y_b\}$ where each split in a decision tree is based on a random subset of the attributes
end for

In other words, each subset is created by sampling from the training samples with replacement, and used to train a decision tree. The final prediction for a previously unseen input $\hat{\mathbf{x}}$ is computed as the average of the predictions from each trained decision tree:

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B}\Theta_b(\hat{\mathbf{x}}) \tag{5}$$

This approach ensures that individual trees are not highly correlated because of a small number of strong predictors. RF methods are popular because they provide some robustness to noisy data with outliers. They are also able to focus on attributes most useful to the regression or classification task under consideration, and ignore attributes that are less relevant. In our study, we used a RF with 100 trees.

V. Hypotheses and Measures

Once the machine learning algorithms described in Section IV are used to build models for short-term prediction of traffic volume, we experimentally evaluate the following hypotheses:

1) Predictive models based on machine learning algorithms are able to disregard the amplification of noise and variations in high-resolution data, and provide higher accuracy than models that do not use the high-resolution data.
2) The predictive models based on machine learning algorithms are responsive to non-recurring congestion events such as accidents, and this ability improves with the increase in the resolution of data.
3) The predictive models are able to capture complex relationships and the spatio-temporal evolution of traffic by assigning higher importance to volume and occupancy attributes extracted from segments near the segment of interest.

We experimentally evaluate these hypotheses using three measures (1) accuracy; (2) Root Mean Square Error (RMSE); and (3) Mean Absolute Error (MAE), defined as follows:

$$Accuracy = 1 - \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

where $\hat{y}_i$ is the predicted value and y_i is the ground truth value of the i^{th} data sample.

In addition, to quantify the responsiveness to non-recurring conditions, we computed these measures over samples that were representative of non-recurring conditions. Specifically, a sample $(\mathbf{x}_i, y_i)$ was considered if the difference between its output value and the weekly seasonal mean of the predicted variable was more than two standard deviations away from the mean of the distribution of output values:

$$|y_i - \hat{\mu}_i| > (2 * std) \tag{7}$$

$$std = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}{N-1}}$$

where std is the standard deviation and $\bar{\mu}_i$ is the mean of the values of the predicted variable during the corresponding time period for that day of the week.

VI. Experimental Results

This section discusses the results of experimentally evaluating the three hypotheses listed in Section V. We summarize the results in Sections VI-A, VI-B, and VI-D, and examine the computational efficiency of the proposed models in Section VI-C. Unlike results reported in many papers, the predictive models we built using the machine learning algorithms considered different traffic conditions such as peak and off-peak traffic at different times of the week, including weekends and public holidays. Recall that we explore different aggregation levels ranging from $30sec$ to $15min$ for the input data, but the output of each model is the volume or occupancy of vehicles (at a particular point in the highway) aggregated over a period of five minutes—see Section III-A for more details.

A. Using high-resolution data

As stated in Section III-A, the predictive models were constructed using the training set and evaluated on the test set. We repeated the trials to check that the performance of the models were stable using different random initializations. The standard deviation across different the segments are shown in parentheses.

The experimental results summarized in Table I show that all three machine learning algorithms performed better with $30sec$ aggregation level for input data in comparison with the $5min$ and $15min$ aggregation levels. While the increase in prediction accuracy with resolution may not be surprising, it is important to note that the increase in resolution also amplifies the noise and minor variations in the data.

Table I also shows results corresponding to two established methods for volume prediction in existing literature (ARIMA, historical average). For the ARIMA models, we applied a square-root transformation in addition to the first order difference and verified their stationarity. To compare the outputs from these methods with the outputs from the machine learning algorithms, we evaluated all models at the same output resolution of $5min$. For instance, for the $30sec$ aggregation level, the $5min$ aggregated output value was obtained by iterating and aggregating the output over 10 one-step ahead predictions. Also, results for the $15min$ input aggregation level were obtained by first applying the Stran-Wei temporal dis-aggregation [Stram and Wei, 1986] to extract $5min$ aggregated values from the $15min$ aggregated data. ARIMA(2,1,2) models were used for predicting volume at the $5min$ and $15min$ input aggregation levels, ARIMA(2,1,1) models were used for predicting occupancy at the $5min$ and $15min$ aggregation levels, and ARIMA(4,1,0) models were used for the $30sec$ input aggregation level. We used the Box-Jenkins method for selecting models and found the that the models identified above provided good performance. Note that the results in Table I include both recurring and non-recurring congestion events; we examine the non-recurring events in more detail in Section VI-B.

To further confirm the significance of these results, we conducted Diebold-Mariano (DM) tests for predictive accuracy [Diebold and Mariano, 1994]. The DM test compares

TABLE I
TRAFFIC VOLUME PREDICTION PERFORMANCE UNDER ALL CONDITIONS; STANDARD DEVIATION BETWEEN SEGMENTS ARE REPORTED IN PARENTHESES.

Model	Input Resolution (minutes)								
	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	0.906 (0.01)	34.5 (11.7)	23.8 (8.6)	0.889 (0.01)	44.5 (16.9)	30.1 (11.5)	0.865 (0.013)	53.6 (24.8)	37.3 (16.9)
RF	**0.910** (0.007)	**31.2** (11.7)	**22.2** (8.5)	0.904 (0.01)	34 (11.2)	23.8 (8.5)	0.89 (0.013)	39.9 (13.3)	28.1 (9.7)
SVR	0.905 (0.01)	34.7 (12.2)	24.4 (8.8)	0.894 (0.01)	39.5 (14.5)	27.9 (10.6)	0.882 (0.007)	43.7 (16.3)	30.9 (11.9)
Historical Avg	0.806 (0.01)	79.73 (35.7)	43.5 (17.4)	0.806 (0.01)	79.73 (35.7)	43.5 (17.4)	0.806 (0.01)	79.73 (35.7)	43.5 (17.4)
ARIMA	0.839 (0.02)	54.64 (18.3)	39.1 (13.2)	0.879 (0.01)	43.8 (15.6)	30.6 (11.4)	0.881 (0.011)	44.34 (16.29)	30.11 (11.36)

TABLE II
TRAFFIC **VOLUME** PREDICTION PERFORMANCE UNDER **NON-RECURRING** CONGESTION CONDITIONS; STANDARD DEVIATION BETWEEN SEGMENTS ARE REPORTED IN PARENTHESES.

Model	Input Resolution (minutes)								
	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	**0.913** (0.008)	**46.9** (16.4)	**33.2** (12.1)	0.88 (0.02)	66.5 (24.4)	46 (17)	0.84 (0.03)	80.2 (29.7)	59.3 (22.8)
RF	0.900 (0.012)	50.1 (17.4)	37.4 (13.2)	0.89 (0.01)	57.3 (19.6)	42 (15)	0.86 (0.02)	66.6 (21.1)	50.9 (16)
SVR	0.892 (0.015)	56 (18.9)	41 (14.1)	0.87 (0.02)	67.2 (21.4)	50 (16)	0.85 (0.03)	76.1 (22.9)	56.4 (17)
Historical Avg	0.139 (0.08)	232.9 (109)	192 (83.6)	0.139 (0.08)	232.9 (109)	192 (83.6)	0.139 (0.08)	232.9 (109)	192 (83.6)
ARIMA	0.851 (0.02)	73.8 (20.5)	54.2 (15.5)	0.67 (0.02)	176 (157)	126 (48)	0.86 (0.02)	77.7 (30.45)	51.6 (19.7)

TABLE III
TRAFFIC **OCCUPANCY** PREDICTION PERFORMANCE MEASURES UNDER ALL CONDITIONS; STANDARD DEVIATION BETWEEN SEGMENTS ARE REPORTED IN PARENTHESES.

Model	Input Resolution (minutes)								
	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	0.859 (0.019)	1.98 (0.64)	1.00 (0.37)	0.838 (0.01)	2.59 (0.74)	1.27 (0.44)	0.78 (0.027)	3.51 (0.89)	1.70 (0.57)
RF	**0.872** (0.006)	**1.83** (0.48)	**0.90** (0.30)	0.850 (0.01)	2.17 (0.55)	1.07 (0.35)	0.796 (0.026)	2.8 (0.7)	1.43 (0.47)
SVR	0.858 (0.011)	1.88 (0.46)	0.95 (0.30)	0.829 (0.01)	2.13 (0.52)	1.12 (0.33)	0.732 (0.04)	2.54 (0.59)	1.45 (0.34)
Historical Avg	0.433 (0.02)	7.49 (4.50)	3.56 (1.02)	0.433 (0.02)	7.49 (4.50)	3.56 (1.02)	0.433 (0.02)	7.49 (4.50)	3.56 (1.02)
ARIMA	0.689 (0.037)	20.53 (4.71)	10.12 (2.65)	0.833 (0.02)	2.37 (0.70)	1.17 (0.41)	0.834 (0.015)	2.59 (0.8)	1.22 (0.43)

TABLE IV
TRAFFIC **OCCUPANCY** PREDICTION PERFORMANCE UNDER **NON-RECURRING** CONGESTION CONDITIONS; STANDARD DEVIATION BETWEEN SEGMENTS ARE REPORTED IN PARENTHESES.

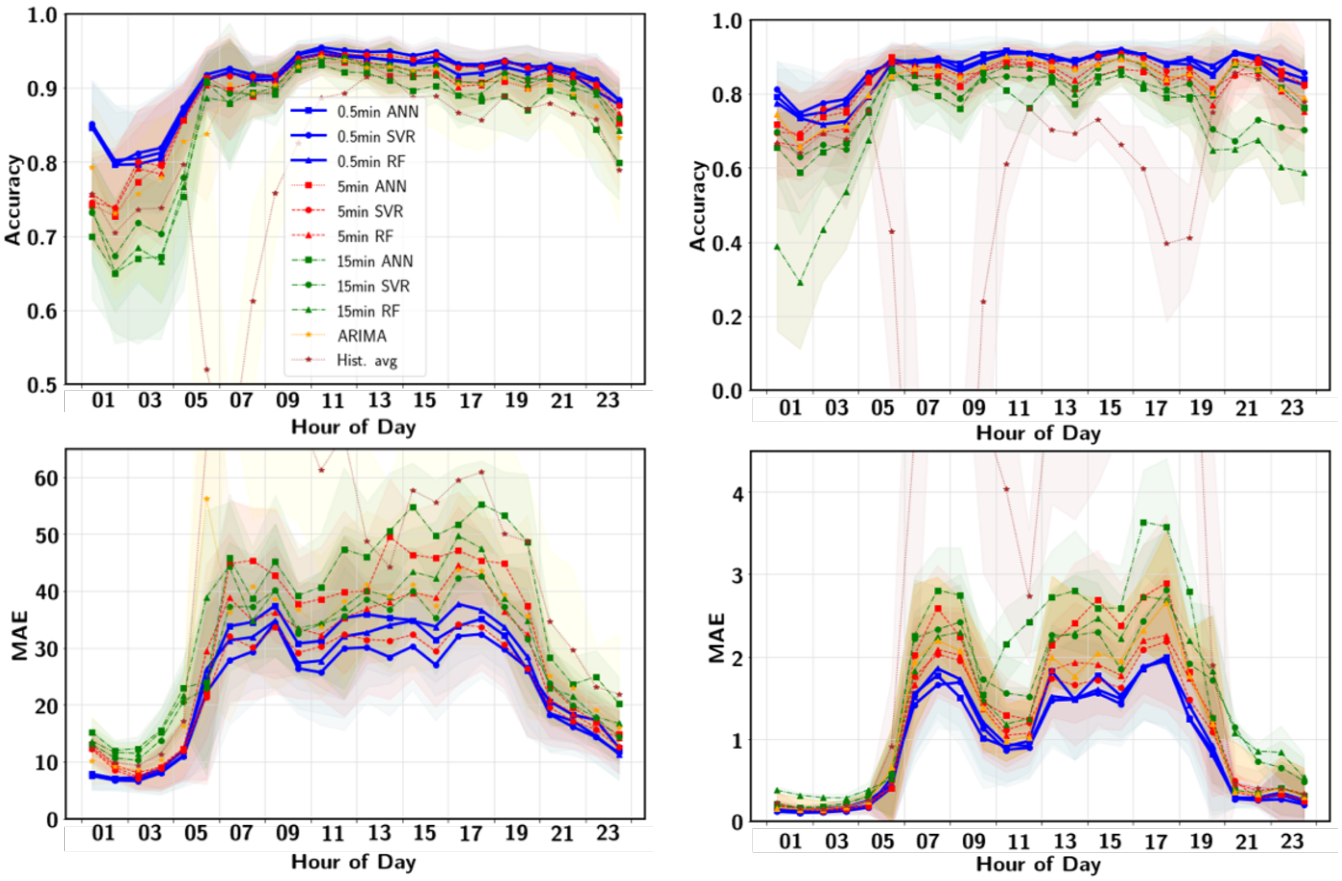
Model	Input Resolution (minutes)								
	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	0.869 (0.014)	1.93 (1.27)	0.938 (0.29)	0.837 (0.008)	2.77 (1.72)	1.316 (0.367)	0.80 (0.018)	3.503 (2.186)	1.628 (0.476)
RF	**0.873** (0.005)	**1.88** (1.22)	**0.905** (0.271)	0.851 (0.01)	2.21 (1.42)	1.072 (0.318)	0.796 (0.02)	2.845 (1.825)	1.418 (0.43)
SVR	0.858 (0.01)	1.923 (1.20)	0.954 (0.277)	0.828 (0.009)	2.18 (1.38)	1.13 (0.28)	0.73 (0.03)	2.58 (1.601)	1.44 (0.31)
Historical Avg	-1.57 (0.83)	18.0 (7.92)	16.44 (1.76)	-1.57 (0.83)	18.0 (7.92)	16.44 (1.76)	-1.57 (0.83)	18.0 (7.92)	16.44 (1.76)

the forecast accuracy of a pair of forecast methods. The null hypothesis of the DM test is that the two forecasts have the same accuracy. The null hypothesis will be rejected if the computed DM statistic falls outside the required significance level under a standard normal distribution. For a significance of 99%, the null hypothesis is rejected if the DM statistic falls outside -2.58 and 2.58. We used the mean squared error as the error metric. Table V shows the DM test statistic for each pair of models. Except for the 5min SVR and 15min RF models, all other models have significantly different levels of accuracy.

Table III, which summarizes the results of predicting occupancy, indicate similar trends. Although all three predictive models based on machine learning algorithms performed well, the model based on the Random Forest algorithm (Section IV-C) provided the highest accuracy. The average accuracy and MAE over different times of day for the three different data aggregation levels, are shown in Figure 3. We observe that, for each algorithm, the accuracy increases with the resolution. Overall, we observe that the performance of the predictive models based on machine learning algorithms improves significantly with the increase in resolution despite

TABLE V
DM TEST STATISTIC FOR EACH PAIR OF MODELS FOR PREDICTING VOLUME. CRITICAL VALUE: |2.58|

		0.5min			5min			15min		
		ANN	RF	SVR	ANN	RF	SVR	ANN	RF	SVR
0.5min	ANN	-	36.88	12.69	-54.59	23.10	-19.35	-71.81	-17.20	-38.30
	RF	-36.88	-	-27.87	-69.47	-11.95	-50.30	-94.20	-44.32	-63.09
	SVR	-12.69	27.87	-	-65.29	14.55	-52.00	-92.40	-31.63	-62.36
5min	ANN	54.59	69.47	65.29	-	71.18	48.64	-11.23	43.96	25.76
	RF	-23.10	11.95	-14.55	-71.18	-	-50.28	-100.06	-45.49	-66.02
	SVR	19.35	50.30	52.00	-48.64	50.28	-	-79.88	**0.49**	-57.11
15min	ANN	71.81	94.20	92.40	11.23	100.06	79.88	-	68.62	51.22
	RF	17.20	44.32	31.63	-43.96	45.49	**-0.49**	-68.62	-	-29.18
	SVR	38.30	63.09	62.36	-25.76	66.02	57.11	-51.22	29.18	-



(a) Volume Predictions

(b) Occupancy Predictions

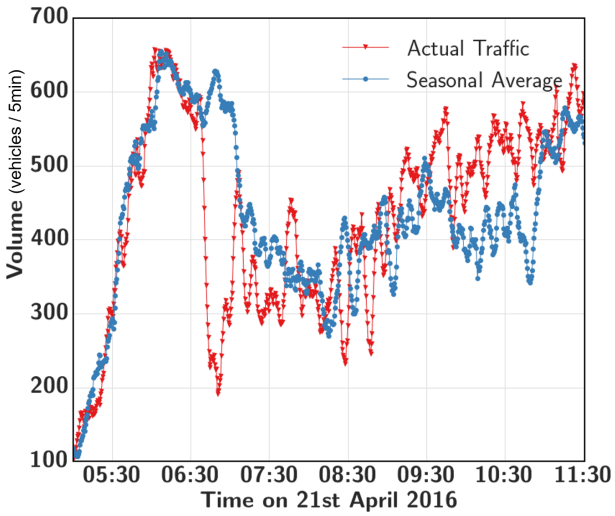
Fig. 3. Accuracy and MAE at different times of day. Shaded areas are the 95% confidence intervals across days in the test set.

the associated amplification of noise and minor variations in data. These results thus provide evidence in support of the first hypothesis, i.e., that predictive models based on machine learning algorithms are able to disregard the amplification of noise in high-resolution data, and provide higher accuracy than models that do not use the high-resolution data. The lower accuracy figures during overnight hours can be explained by the accuracy being a represented as percentage of vehicles and the average number of vehicles overnight being significantly lower (this is confirmed by the lower MAE values for the same period).

B. Non-recurring congestion

Next, we evaluate the second hypothesis (see Section V) by examining the responsiveness of the predictive models to non-recurring congestion events. We do so by only evaluating the trained predictive models on a subset of the test set; as described in Section V, this subset only included points that were significantly different from historical average values.

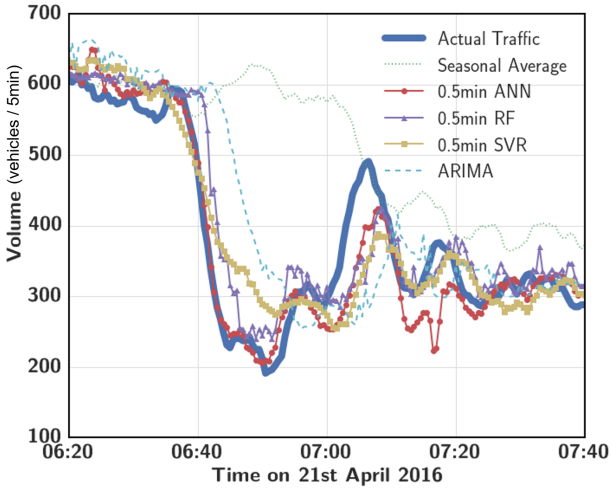
The results are summarized in Tables II & IV. We observe that the models built using input data at the $30sec$ aggregation level outperforms the models that use input data at $5min$ and $15min$ aggregation levels. Among the models based on



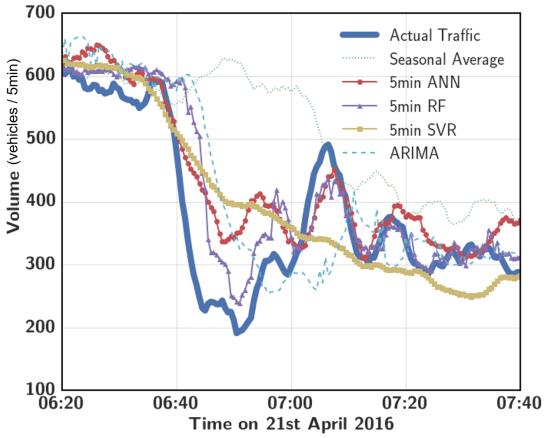
(a) Traffic volume during non-recurring congestion.

(b) NZTA tweet during congestion.

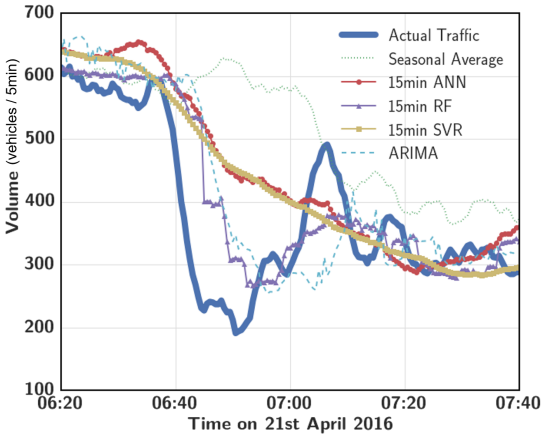
Fig. 4. (a) Traffic volume in Segment 23 on April 21, 2016 (Thursday) compared with weekly average; and (b) tweets from NZTA accessed from [@NZTA Akld & Nthlnd, 2015] on April 21, 2016.



(a) $30sec$ aggregation level.



(b) $5min$ aggregation level.



(c) $15min$ aggregation level.

Fig. 5. Traffic volume predictions in response to a non-recurring congestion event, for $30sec$, $5min$ and $15min$ input data aggregation levels; models using higher-resolution data respond better.

the machine learning algorithms, the model based on the ANN algorithm provides marginally better performance than that based on the RF algorithm for Volume predictions while the converse is true for Occupancy predictions. Furthermore, we observe that the multivariate predictive models based on machine learning algorithms provide better performance than the models based on historical average and ARIMA, which are established methods for short-term traffic prediction.

To further explore the responsiveness of the different models, we examine a known (i.e., reported) breakdown along the motorway in more detail. Figure 4(a) compares the average volume of traffic on Segment 23 of SH1 on Thursday with the traffic volume on a specific Thursday, April 21, 2016. The data corresponding to this date were in the test dataset, i.e., they were not used to train the predictive models. Figure 4(a) shows that there was a significant deviation from the average traffic around 6.40am on April 21, 2016. As reported on the social media site twitter, there was a breakdown near SH1 at $\approx 6.30am$ that day—see Figure 4(b). More specifically, the *Ellerslie on-ramp* mentioned in the tweet is near Segment 27 of SH1, which is $\approx 4Km$ from Segment 23 on SH1.

Figures 5(a)-5(c) show how the predictive models are able to accurately track the traffic volume corresponding to this event, as a function of the three different input data aggregation levels. For comparison, the figures also include the performance of the ARIMA approach. We observe in Figure 5(a) that using the high-resolution $30sec$ input data aggregation level enabled the machine learning models to predict the change in traffic volume at almost the same time-step when the non-recurring event occurred, whereas there is a lag when the other two aggregation levels are used. For additional examples on how the models predicted during non-recurring congestion, see Figure 9 in the Appendix. From these, it is possible to see that the $30s$ ANN model is able to respond to non-recurring congestion very quickly. It is also apparent that the SVR based models as well as the courses resolution models tend to smooth out shocks to traffic and are better at smoothing out noise in typical congestion conditions. The RF models tend to be in-between the ANN & SVR models and provide good performance overall.

Figure 6 shows that a model based on the ANN algorithm and input data at the $30s$ aggregation level accurately predicts traffic volume on a public holiday. Recall that this model had no information about the day of the week and the seasonal mean. Overall, these results provide evidence in support of the second hypothesis that the models based on machine learning algorithms and high-resolution data are more responsive to non-recurring congestion.

C. Computational Efficiency and Practical Scalability

Table VI summarizes the training time and testing time of the proposed models, when they are built and evaluated on an Intel Core $i7$ $3.4GHz$ desktop with $8GB$ of RAM. The time taken to generate a forecast was under 0.1 seconds for all models. The training time even in the most extreme case was under 20 minutes. Since the training process can easily be parallelized to create models for all segments on a network

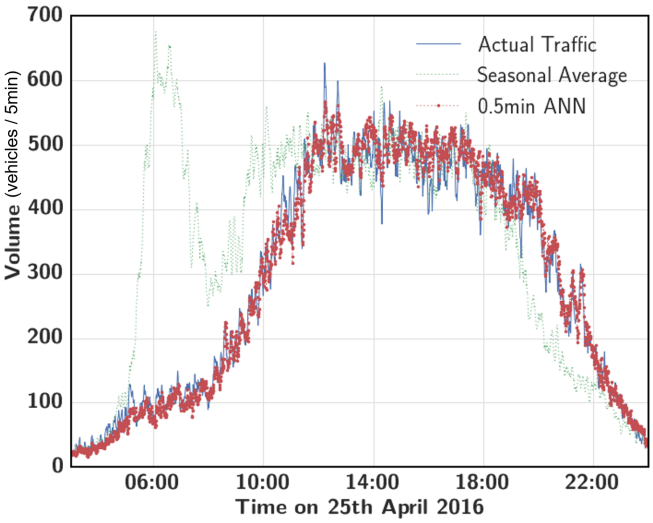


Fig. 6. Traffic volume prediction on April 25, 2016, a public holiday in New Zealand (ANZAC day).

and this can be done in an initial offline phase, we believe these methods can be easily implemented for forecasts over the entire traffic network.

We did not optimize our algorithms—performance could have been improved by using fewer training samples or tuning the algorithms' parameters, e.g., by using a smaller number of trees in the Random Forest or a smaller neural network. The different algorithms take different amounts of time for training and testing, e.g., models based on the (linear) SVR algorithm have the lowest training time and testing time—the nonlinear SVR models have a much longer training time ($\approx$ one hour for one model) but they did not perform as well as the linear model. The ANN-based models take longer to train but are fast during testing, whereas the RF-based ensemble models take longer to train and test.

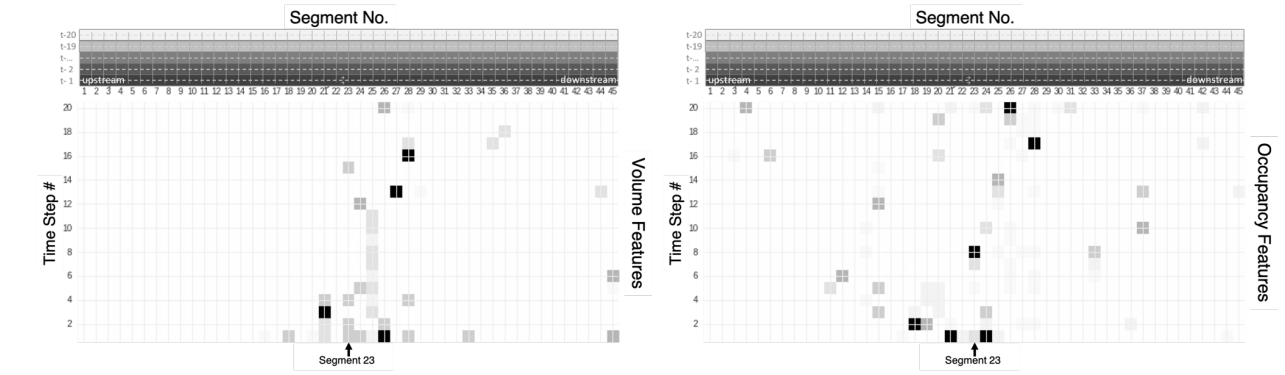
Overall, we believe our methods will easy scale to even large road networks. The re-training of the models can be undertaken as new data comes in over several weeks or months enabling the system to adapt to changes in the road network.

TABLE VI
TRAINING AND TESTING TIMEFOR EACH OF THE THREE PROPOSED MODELS FOR SHORT-TERM TRAFFIC PREDICTION. ALL MODELS WILL SCALE WELL FOR SHORT-TERM PREDICTIONS IN LARGE ROAD NETWORKS.

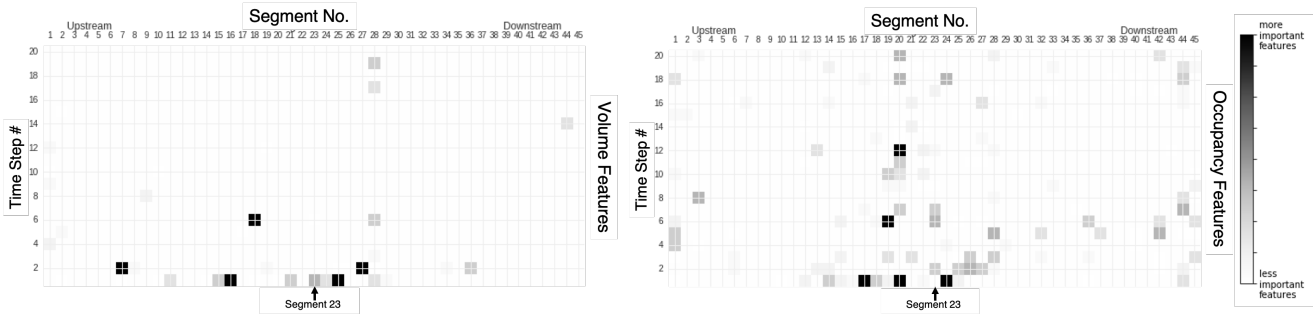
	Average training time for 57600 samples (seconds)	Average prediction time for one input sample (milliseconds)
ANN	283.8	0.16
RF	1154	82.08
SVR	4.743	0.0223

D. Attribute Selection

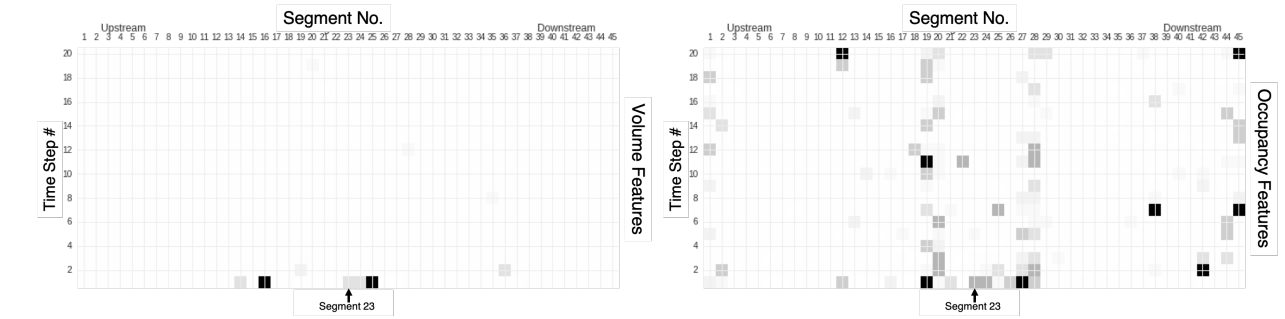
Next, we evaluate the third hypothesis regarding the ability to capture complex relationships and the spatio-temporal evolution of traffic. To do so, we first identify the attributes



(a) $30sec$ aggregation level.



(b) $5min$ aggregation level.



(c) $15min$ aggregation level.

Fig. 7. Ranking of attributes in terms of their relative importance to the performance of ANN models, for three different input data aggregation levels (Segment 23). Volume features on the left and Occupancy features on the right.

that most influence the performance of the proposed predictive models.

One of the most common approaches for identifying informative attributes is to compute the Pearson correlation coefficient between the target variable and each of the input attributes [Ermagun and Levinson, 2018]. However, the Pearson correlation coefficient is not able to capture nonlinear relationships that may exist between the input and output variables. We therefore used the Recursive Feature Elimination (RFE) approach to select the most relevant (i.e., informative) attributes [Guyon et al., 2002, Hastie et al., 2009]. RFE works by iteratively considering an increasingly smaller subset of attributes, dropping (in each iteration) the attributes considered to be the least relevant. In each iteration, we removed 10 attributes ranked lowest in terms of importance.

There are different ways to characterize the importance of attributes in RF-based models. Since any RF is a collection of decision trees, the *gini importance* of each attribute in all decision trees can be averaged, for instance, to arrive at the importance of the attribute. In the case of an ANN, the weights of the first layer of an ANN-based model can provide insight into the attributes that contributed significantly to making the predictions. In a similar manner, the weights assigned to each attribute of a linear SVM can be used to identify the relative importance of the attributes Chang and Lin [2008].

Figure 7 (and Figures 10-11 in the Appendix) visualize the relative ranking of each of the 1800 input attributes considered by the models for traffic prediction at Segment 23—darker shades represent the more informative attributes. For each figure, the plot on the left visualizes the volume attributes and the plot on the right visualizes the occupancy attributes. In each of these plots, the columns going from left to right along the x-axis represent the segments in spatial order along the motorway from the south to the north. Along the y-axis,

the first row is the most recent time-step and the top row is the oldest time-step, e.g., for the $30sec$ aggregation level for input data, row 20 corresponds to the data from 10 minutes before the current time-step. Overall, we observed that all three models provide higher rank to neighboring segments over a few time steps.

A more careful examination of the results indicated that the predictive models based on SVR and RF assign higher importance to volume attributes than occupancy attributes when making decisions. Also, the same set of attributes do not contribute significantly to the performance of all three models. For all three models, the attributes that are considered important change when the resolution of the input data changes. For instance, for the models based on the $30sec$ aggregation level (i.e., highest resolution), the set of attributes considered to be important for decision making mostly included values (of volume and occupancy) from nearby spatial locations and time-steps. The number of attributes corresponding to downstream segments that are nearby is high for the higher resolution models, especially when predicting non-recurring congestion events. For the models based on the $5min$ and $15min$ aggregation levels, on the other hand, the set of attributes considered to be important also included values from more distant segments. These results add to the current knowledge about representing information for short-term traffic prediction. For instance, some recent research found that having more than one time-step of data from neighboring locations only provides minor improvements in performance [Yang et al., 2015]. Our results, on the other hand, indicate that volume and occupancy values from multiple neighboring locations and time steps may be important for accurate prediction of traffic depending on the resolution of the input data.

To further analyze the importance of the attributes, we considered the relative importance of different subsets of these ranked attributes. We observed that the performance, specifically accuracy, flattens out after including ≈ 100 attributes. Figure 13 shows the performance of the three models for the $30sec$ aggregation level, as a function of the number of attributes considered, with the attributes ordered in decreasing order of importance.A similar result was observed for the other two aggregation levels.

Finally, we compared the performance of the RFE approach for ranking attributes with the more common correlation-based approach and an approach that chose important attributes randomly—we considered the performance of the corresponding models under normal conditions and in the presence of non-recurring congestion events. Figures 8 and 12 indicate that the RFE approach outperforms the other two approaches for ranking attributes. In fact, in the case of non-recurring congestion, correlation-based attribute selection is closer in performance to the random selection of important attributes. These results provide evidence that correlation-based ranking of attributes is a poor choice for accurate traffic prediction under non-recurring congestion. The results also support the hypothesis that the predictive models based on the machine learning algorithms capture the complex spatio-temporal evolution of traffic by assigning higher importance to relevant attributes.

We believe that the reason for the poor performance of the Pearson correlation based feature selection is that the features that are most likely to be highly correlated to the output correspond to the closest neighbouring road segments. However in most cases, these features give redundant information. Segments further away may contain information about things like queues building up or a spike in traffic that are not necessarily correlated with the output but are quite informative for predictions. We believe the Recursive Feature Elimination approach is better able to capture these dependencies.

VII. Conclusions

Traffic congestion results in significant monetary losses in countries around the world. Although short-term traffic prediction helps make decisions based on predictions of traffic in the near-future and is more useful than just using the real-time data of traffic conditions, it also poses many open problems. For instance, existing approaches still find it difficult to (a) respond reliably and quickly to non-recurring congestion events; (b) accurately identify and model the spatio-temporal dependencies influencing traffic; or (c) reliably extract useful information from high-resolution traffic data. We have explored the construction and use of predictive models based on three established algorithms for addressing the aforementioned problems. Specifically, we investigated the use of Artificial Neural Network (ANN), Support Vector Regression (SVR) and Random Forest (RF), and evaluated the predictive performance of these models for three different aggregation levels of input data, $30sec$, $5min$ and $15min$. For each learned model, the output was a prediction (of volume or occupancy) over $5min$ period. However, the same methodology can be used to provide predictions over $10min$ or $15min$. Our experiments indicate that:

- Aggregation of high resolution data to a lower resolution is not required for accurate forecasting with machine learning algorithms. Aggregation may actually have a negative effect on accuracy for these multivariate models. Our results indicate that machine learning algorithms are able to extract useful information from high resolution data even though this data is highly variable with noise.
- By not exploiting periodic characteristics in traffic, the machine learning models studied here perform equally well under both recurring and non-recurring congestion without requiring any changes to the models. A significant difference between recurring and non-recurring congestion performance would have indicated that the model was not able to capture the underlying spatio-temporal evolution of traffic.
- We are able to visualise the importance of different input attributes and gain an understanding of the sophisticated spatio-temporal patterns captured by the machine learning models. With this we are able to see that the most recent data from neighbouring road segments have high predictive power.
- Our results indicate that feature selection based on the linear Pearson correlation coefficient analysis is not suitable for traffic forecasting models that aim to capture

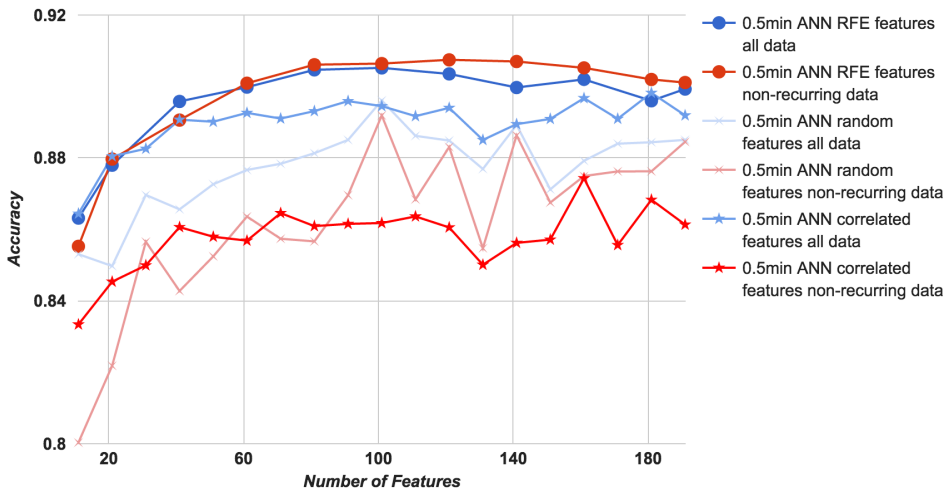


Fig. 8. Performance comparison of RFE, correlation-based and random-selection approaches for selecting important attributes; results correspond to an ANN model for the $30sec$ aggregation level.

non-recurring congestion. Even though this method is the most commonly used metric for feature selection [Ermagun and Levinson, 2018], its performance in this case is comparable to a random selection of features. Our experiments show that Recursive Feature Elimination provides a better ranking of attributes for feature selection.

One limitation of our study is that the analysis was done with a single dataset on one highway. Therefore, further analysis is required before these findings can be generalised.

These results however open up multiple directions for further research. First, we will encorporate these findings in more sophisticated machine learning algorithms for short-term traffic prediction. For instance, the complex, non-linear relationships influencing traffic flow may be modeled well using deep network architectures, especially when high-resolution input data is considered. Second, we will build on the indicated ability to track non-recurring congestion events to consider both accidents and weather conditions. This will require the underlying algorithms to model additional variables and their effect on traffic flow. Furthermore, we will explore network-wide traffic predictions towards the long-term objective of effective use of resources for smooth flow of traffic under a wide range of circumstances.

Data Availability

The terms of use of the data used to support the findings of this study does not allow the authors to distribute or publish the data. However, data can be accessed directly from NZTA through the following APIs: https://www.nzta.govt.nz/traffic-and-travel-information/infoconnect-section-page/.

Conflicts of Interest and Funding Statement

The authors declare that there are no conflicts of interests with regards to this publication and this research did not receive any specific funding.

Acknowledgments

The authors would like to thank Mike Duke from Auckland's Joint Transport Operations Centre (JTOC) for helping us with access to data to carry out the analysis.

References

Mohammed S Ahmed and Allen R Cook. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. *Transportation Research Record*, (722):116, 1979.

Muhammad Tayyab Asif, Justin Dauwels, Chong Yang Goh, Ali Oran, Esmail Fathi, Muye Xu, Menoth Mohan Dhanya, Nikola Mitrovic, Patrick Jaillet, Chong Yang Goh, Ali Oran, Esmail Fathi, Muye Xu, Menoth Mohan Dhanya, Nikola Mitrovic, and Patrick Jaillet. Spatiotemporal patterns in large-scale traffic speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):794–804, apr 2014.

Xiaojuan Ban, Chong Guo, and Guohui Li. Application of Extreme Learning Machine on Large Scale Traffic Congestion Prediction. In Jiuwen Cao, Kezhi Mao, Jonathan Wu, and Amaury Lendasse, editors, *Proceedings of ELM-2015 Volume 1: Theory, Algorithms and Applications (I)*, pages 293–305. Springer International Publishing, Cham, 2016. ISBN 978-3-319-28397-5.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Pinlong Cai, Yunpeng Wang, Guangquan Lu, Peng Chen, Chuan Ding, and Jianping Sun. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies*, 62:21–34, 2016.

Manoel Castro-Neto, Young Seon Jeong, Myong Kee Jeong, and Lee D. Han. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications*, 36(3 PART 2):6164–6173, 2009.

Srinivasa Ravi Chandra and Haitham Al-Deek. Predictions of Freeway Traffic Speeds and Volumes Using Vector Au-

toregressive Models. *Journal of Intelligent Transportation Systems*, 13(2):53–72, apr 2009.

H. Chang, Y. Lee, B. Yoon, and S. Baek. Dynamic near-term traffic flow prediction: system-oriented approach based on past experiences. *IET Intelligent Transport Systems*, 6(3): 292, 2012.

Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear svm. *Causation and Prediction Challenge, Challenges in Machine Learning*, (2):53–64, 2008.

Anyu Cheng, Xiao Jiang, Yongfu Li, Chao Zhang, and Hao Zhu. Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method. *Physica A: Statistical Mechanics and its Applications*, 466:422–434, 2016.

Gary A. Davis and Nancy L. Nihan. Nonparametric Regression and ShortTerm Freeway Traffic Forecasting. *Journal of Transportation Engineering*, 117(2):178–188, mar 1991.

Francis X. Diebold and Robert S. Mariano. Comparing Predictive Accuracy. NBER Technical Working Papers 0169, National Bureau of Economic Research, Inc, November 1994. URL https://ideas.repec.org/p/nbr/nberte/0169.html.

Mark S. Dougherty and Mark R. Cobbett. Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting*, 13(1):21–31, 1997.

Stephen Dunne and Bidisha Ghosh. Regime-based short-term multivariate traffic condition forecasting algorithm. *Journal of Transportation Engineering*, 138(4):455–466, 2011.

Stephen Dunne and Bidisha Ghosh. Using Neurowavelet Models. 14(1):370–379, 2013.

Alireza Ermagun and David Levinson. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, pages 1–29, 2018.

Gaetano Fusco, Chiara Colombaroni, and Natalia Isaenko. Short-term speed predictions exploiting big data on large urban road networks. *Transportation Research Part C: Emerging Technologies*, 73:183–201, 2016.

Bidisha Ghosh, Biswajit Basu, and Margaret O'Mahony. Bayesian time-series model for short-term traffic flow forecasting. *Journal of Transportation Engineering*, 133(3): 180–189, mar 2007.

Dhiren Ghosh and Andrew Vogt. Outliers: An Evaluation of Methodologies. *Joint Statistical Metings*, pages 3455–3460, 2012.

Jianhua Guo, Wei Huang, and Billy M. Williams. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C*, 43:50–64, 2014.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

Benjamin Hamner. Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1357–1359. IEEE, dec 2010.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

Victoria J Hodge, Rajesh Krishnan, Jim Austin, John Polak, and Tom Jackson. Short-term prediction of traffic flow using a binary neural network. *Neural Computing and Applications*, 25(7):1639–1655, 2014.

Eric Horvitz, Johnson Apacible, Raman Sarin, and Lin Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. *Conference on Uncertainty in Artificial Intelligence*, pages 1–10, jul 2005.

Wenhao Huang, Guojie Song, Haikun Hong, and Kunqing Xie. Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2191–2201, oct 2014.

Young-Seon Jeong, Young-Ji Byon, Manoel Mendonca Castro-Neto, and Said M. Easa. Supervised Weighting-Online Learning Algorithm for Short-Term Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1700–1707, dec 2013.

Zhanfeng Jia, Chao Chen, Ben Coifman, Pravin Varaiya, Zhanfeng Jia, Chao Chen, Ben Coifman, and Pravin Varaiya. The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585)*, pages 536–541, 2001.

Yiannis Kamarianakis and Poulicos Prastacos. Forecasting Traffic Flow Conditions in an Urban Network: Comparison of Multivariate and Univariate Approaches. *Transportation Research Record*, 1857(1):74–84, 2003.

M. G. Karlaftis and E. I. Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399, 2011.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13, dec 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2): 865–873, April 2015.

P T Martin, Y Feng, and X Wang. Detector Technology Evaluation (MPC-03-154). page 140, 2003.

Wanli Min and Laura Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616, 2011.

@NZTA Akld & Nthlnd. @NZTA Akld & Nthlnd, 2015. URL https://twitter.com/NZTAAkl/status/722855139099426816.

Simon Oh, Young-Ji Byon, Kitae Jang, and Hwasoo Yeo. Short-term travel-time prediction on highway: a review of the data-driven approach. *Transport Reviews*, 35(1):4–32, 2015a.

Simon Oh, Young-Ji Byon, and Hwasoo Yeo. Improvement of search strategy with k-nearest neighbors approach for traffic state prediction. *IEEE Transactions on Intelligent*

Transportation Systems, 17(4):1146–1156, 2015b.

Simon Oh, Young-Ji Byon, Kitae Jang, and Hwasoo Yeo. Short-term travel-time prediction on highway: A review on model-based approach. *KSCE Journal of Civil Engineering*, 22(1):298–310, 2018.

Iwao Okutani and Yorgos J. Stephanedes. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B*, 18(1):1–11, 1984.

T. L. Pan, A. Sumalee, R. X. Zhong, and N. Indra-payoong. Short-term traffic state prediction based on temporal-spatial correlation. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1242–1254, Sept 2013.

Dongjoo Park, Laurence R. Rilett, Byron J. Gajewski, Clifford H. Spiegelman, and Changho Choi. Identifying optimal data aggregation interval sizes for link and corridor travel time estimation and forecasting. *Transportation*, 36(1):77–95, jan 2009.

Alessandra Pascale and Monica Nicoli. Adaptive bayesian network for traffic flow prediction. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 177–180. IEEE, 2011.

Paul Ryus, Mark Vandehey, Lily Elefteriadou, Richard Dowling G, and Barbara Ostrom K. *Highway Capacity Manual 2010*. Number 273. Transportation Research Board, 2010.

David Schrank, Bill Eisele, Tim Lomax, and Jim Bak. 2015 URBAN MOBILITY SCORECARD. Technical Report August, 2015.

Brian L Smith, Billy M Williams, and R Keith Oswald. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4):303–321, 2002.

a J Smola and B Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

Ridha Soua, Arief Koesdwiady, and Fakhri Karray. Big-Data-Generated Traffic Flow Prediction Using Deep Learning and Dempster-Shafer Theory. *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3195–3202, jul 2016.

Daniel O. Stram and William W. S. Wei. A methodological note on the disaggregation of time series totals. *Journal of Time Series Analysis*, 7(4), 1986.

Shiliang Sun, Rongqing Huang, and Ya Gao. Network-Scale Traffic Modeling and Forecasting with Graphical Lasso and Neural Networks. 138:1358–1367, 2012.

Eleni I. Vlahogianni and Matthew G. Karlaftis. Temporal aggregation in traffic data: implications for statistical characteristics and model choice. *Transportation Letters: The International Journal of Transportation Research*, 3(1):37–49, jan 2011.

Eleni I. Vlahogianni, Matthew G. Karlaftis, and John C. Golias. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, 13 (3):211–234, 2005.

Eleni I. Vlahogianni, Matthew G. Karlaftis, and John C. Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43:3–19, 2014.

Jiaqiu Wang, Ioannis Tsapakis, and Chen Zhong. A spacetime delay neural network model for travel time prediction. *Engineering Applications of Artificial Intelligence*, 52:145 – 160, 2016.

Billy M Williams and Lester A Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672, 2003.

Dawen Xia, Binfeng Wang, Huaqing Li, Yantao Li, and Zili Zhang. A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting. *Neurocomputing*, 179:246–26, 2016.

Yuanchang Xie, Yunlong Zhang, and Zhirui Ye. Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):326–334, jul 2007.

Su Yang, Shixiong Shi, Xiaobing Hu, and Minjie Wang. Spatiotemporal context awareness for urban traffic modeling and prediction: Sparse representation based variable selection. *PLoS ONE*, 10(10):1–22, oct 2015.

Baozhen Yao, Chao Chen, Qingda Cao, Lu Jin, Mingheng Zhang, Hanbing Zhu, and Bin Yu. Short-Term Traffic Speed Prediction for an Urban Corridor. *Computer-Aided Civil and Infrastructure Engineering*, 00:1–16, 2016.

Narjes Zarei, Mohammad Ali Ghayour, and Sattar Hashemi. Road traffic prediction using context-aware random forest based on volatility nature of traffic flows. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7802 LNAI, pages 196–205. Springer Berlin Heidelberg, 2013.

Mr Rivindu Weerasekera (BE(Hons)) is a doctoral candidate at the University of Auckland, New Zealand. He holds a first class honors degree in Electrical and Electronics Engineering from the University of Auckland. His research interests focus on the intersection of Intelligent Transportation Systems and Machine Learning.

Dr Mohan Sridharan (PhD) is a Senior Lecturer in the School of Computer Science at the University of Birmingham (UK). He was previously a Senior Lecturer in the Department of Electrical and Computer Engineering at The University of Auckland (NZ). Prior to his current appointment, he was a faculty member at Texas Tech University (USA), where he is currently an Adjunct Associate Professor of Mathematics and Statistics. He received his PhD in Electrical and Computer Engineering from The University of Texas at Austin (USA). Dr Sridharans primary research interests include knowledge representation and reasoning, machine learning, computational vision, and cognitive systems, in the context of autonomous robots and adaptive agents.

Dr Prakash Ranjitkar (PhD, MEng, BEng (Civil)) is a Senior Lecturer in Transportation Engineering in the Department of Civil and Environmental Engineering and a founding member of the Transportation Research Centre (TRC) at the University of Auckland, New Zealand.

Prakash has over 19 years of academic, research and consulting work experience in a range of transport and other infrastructure engineering projects. He has strong research interests in modelling and simulation of traffic, intelligent transportation system, traffic operations and management, traffic safety, human factors and applications of advanced technologies in transportation.

Prior to joining the University of Auckland in 2007, he worked for the University of Delaware in USA (2006-2007) and before that in Hokkaido University in Japan (2001-2006). Prakash is a member of IPENZ Transportation Group and Institute of Transportation Engineers (USA). He is an Editorial Board Member for the Open Transportation Journal and reviewer of Journal of Transportation Research Board, Journal of Eastern Asia Society for Transportation Studies, Journal of Intelligent Systems and IEEE Transactions of Intelligent Transportation Systems.

APPENDIX

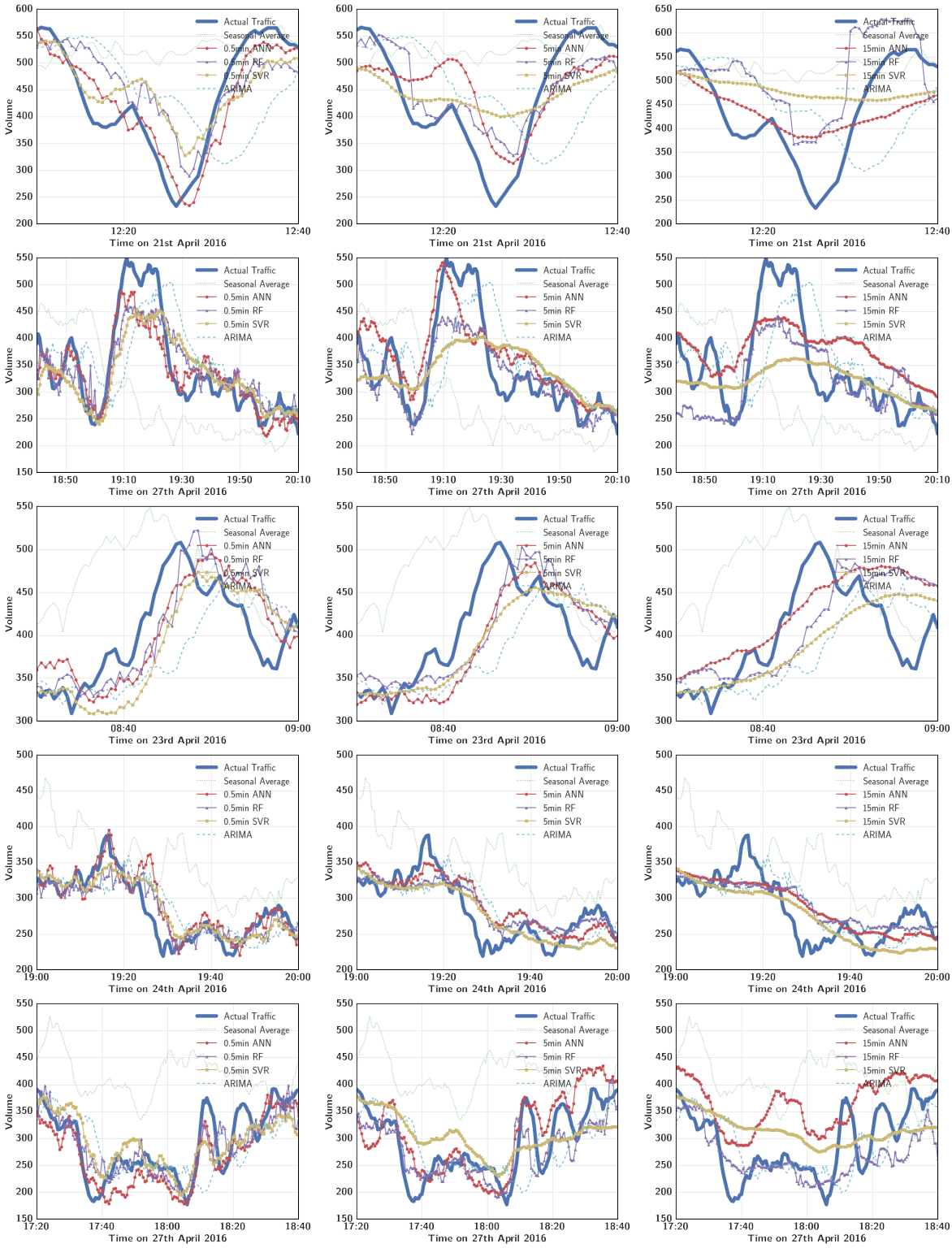
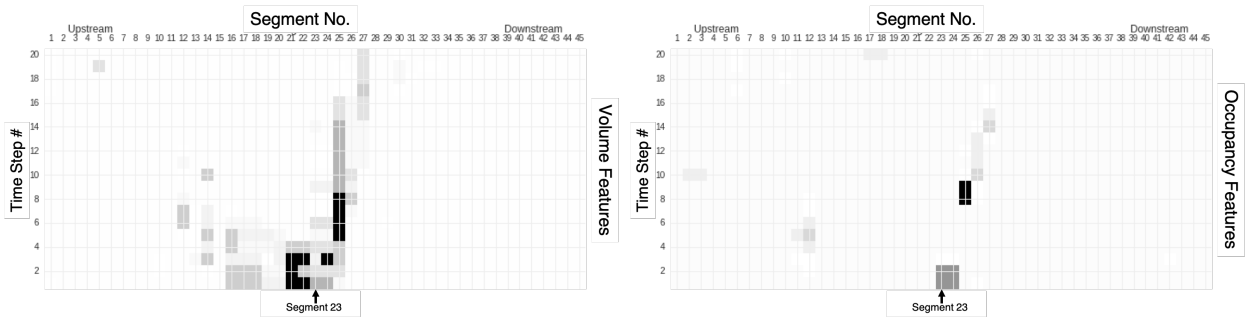
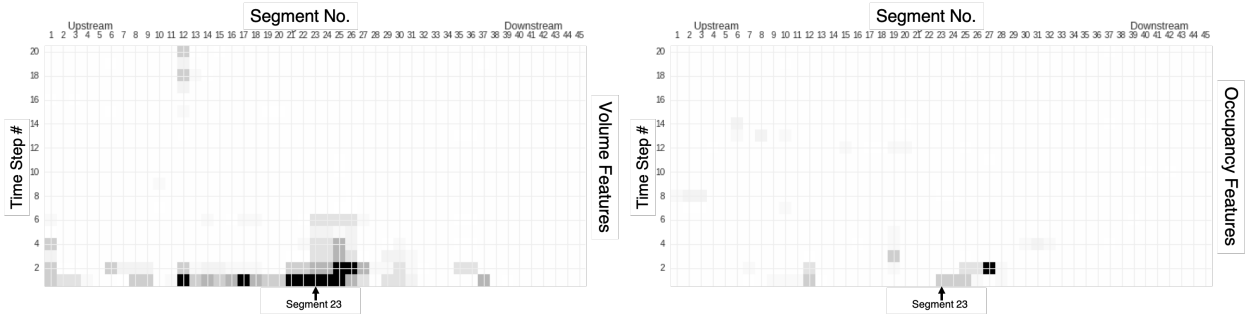


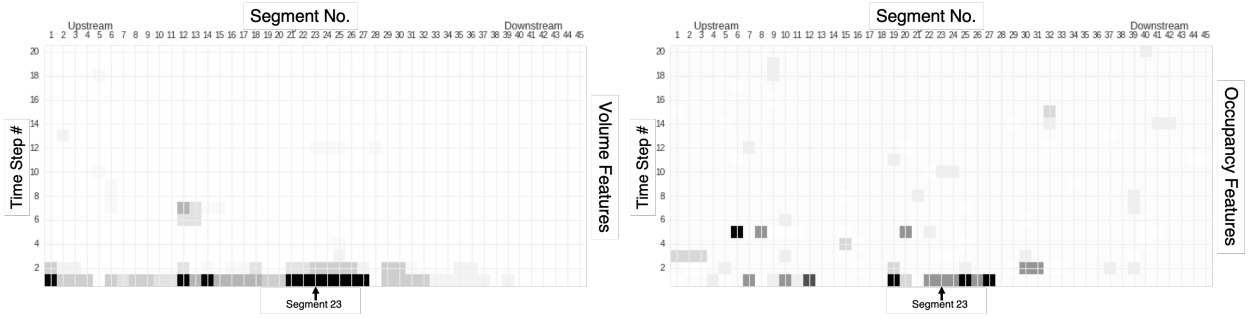
Fig. 9. More examples of non-recurring congestion



(a) $30sec$ aggregation level.

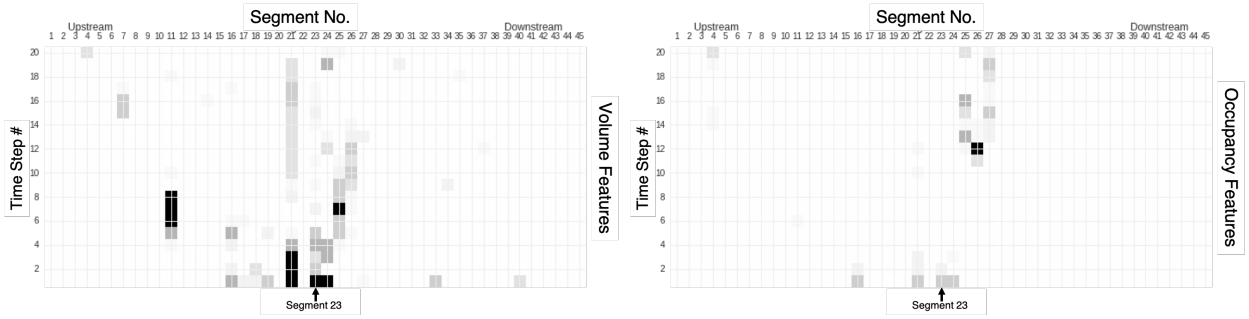


(b) $5min$ aggregation level.

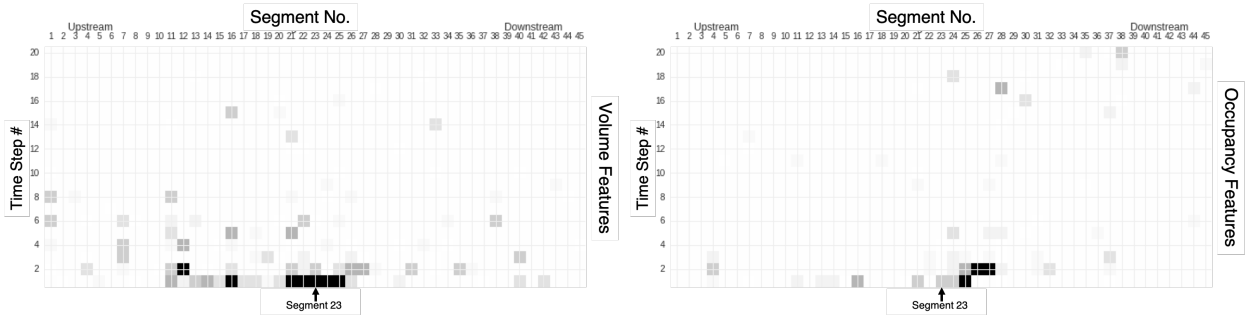


(c) $15min$ aggregation level.

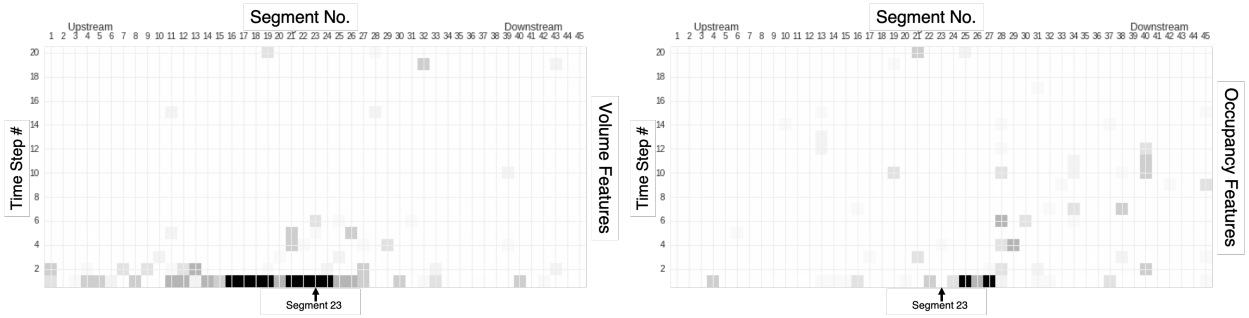
Fig. 10. Ranking of attributes in terms of their relative importance to the performance of SVR models, for three different input data aggregation levels (Segment 23). Volume features on the left and Occupancy features on the right.



(a) $30sec$ aggregation level.



(b) $5min$ aggregation level.



(c) $15min$ aggregation level.

Fig. 11. Ranking of attributes in terms of their relative importance to the performance of RF models, for three different input data aggregation levels (Segment 23). Volume features on the left and Occupancy features on the right.

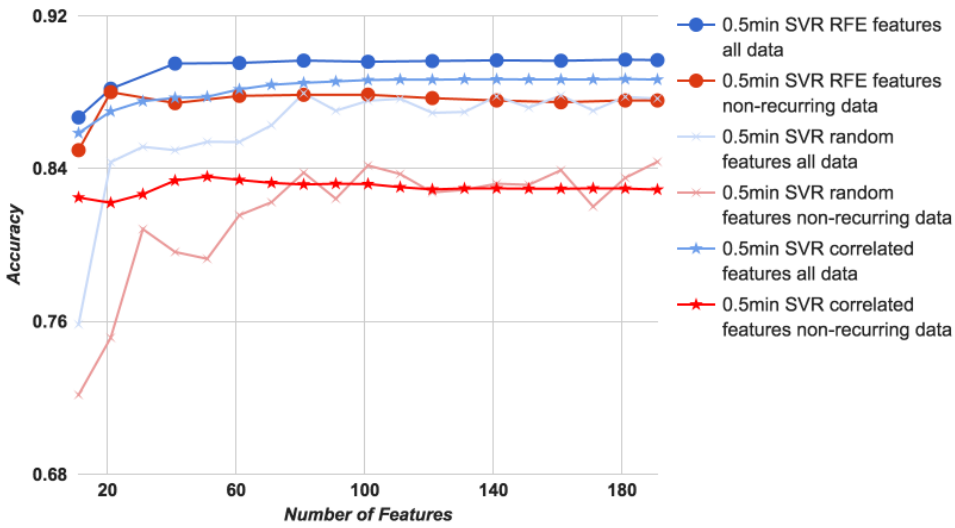


Fig. 12. Performance comparison of RFE, correlation-based and random-selection approaches for selecting important attributes; results correspond to an SVR model for the $30sec$ aggregation level.

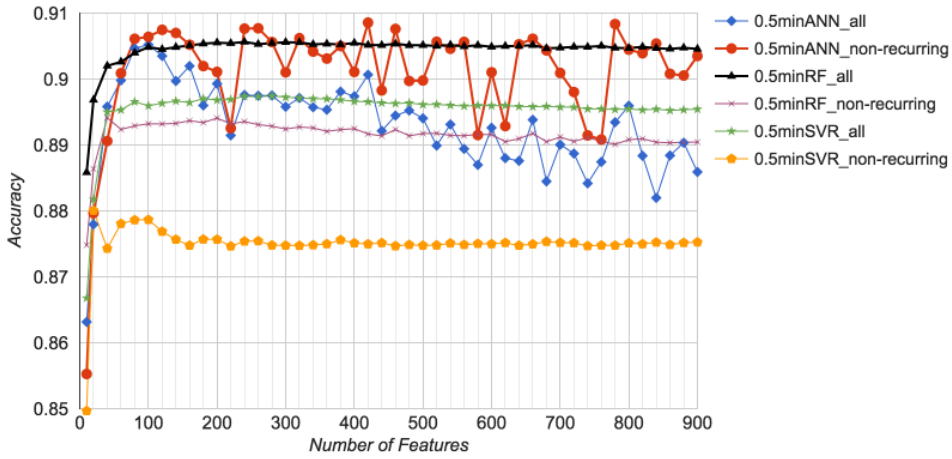


Fig. 13. Accuracy of each of the three models for the $30sec$ input data aggregation level, as a function of the number of attributes considered; attributes ranked in decreasing order of importance using RFE approach.