Implications of Spatiotemporal Data Aggregation on Short-Term Traffic Prediction Using Machine Learning Algorithms

Weerasekera, Rivindu; Sridharan, Mohan; Ranjitkar, Prakash

DOI:
10.1155/2020/7057519
10.1155/2020/7057519

Citation for published version (Harvard):
Weerasekera, R, Sridharan, M & Ranjitkar, P 2020, 'Implications of Spatiotemporal Data Aggregation on Short-Term Traffic Prediction Using Machine Learning Algorithms', *Journal of Advanced Transportation*, vol. 2020, 7057519, pp. 1-21. https://doi.org/10.1155/2020/7057519, https://doi.org/10.1155/2020/7057519

Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 7057519, 21 pages
https://doi.org/10.1155/2020/7057519

Research Article

Implications of Spatiotemporal Data Aggregation on Short-Term Traffic Prediction Using Machine Learning Algorithms

Rivindu Weerasekera,[1] **Mohan Sridharan**,[2] **and Prakash Ranjitkar**[3]

[1]*Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand*
[2]*School of Computer Science, The University of Birmingham, Birmingham, UK*
[3]*Department of Civil and Environmental Engineering, The University of Auckland, Auckland, New Zealand*

Correspondence should be addressed to Rivindu Weerasekera; rwee015@aucklanduni.ac.nz

Received 23 July 2019; Revised 10 January 2020; Accepted 12 February 2020; Published 15 June 2020

Academic Editor: Shamsunnahar Yasmin

Short-term traffic prediction is a key component of Intelligent Transportation Systems. It uses historical data to construct models for reliably predicting traffic state at specific locations in road networks in the near future. Despite being a mature field, short-term traffic prediction still poses some open problems related to the choice of optimal data resolution, prediction of nonrecurring congestion, and the modelling of relevant spatiotemporal dependencies. As a step towards addressing these problems, this paper investigates the ability of Artificial Neural Networks, Random Forests, and Support Vector Regression algorithms to reliably model traffic flow at different data resolutions and respond to unexpected traffic incidents. We also explore different feature selection methods to identify and better understand the spatiotemporal attributes that most influence the reliability of these models. Experimental results indicate that data aggregation does not necessarily achieve good performance for multivariate spatiotemporal machine learning models. The models learned using high-resolution 30-second input data outperformed the corresponding baseline ARIMA models by 8%. Furthermore, feature selection based on Recursive Feature Elimination resulted in models that outperformed those based on linear correlation-based feature selection.

1. 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 [1]. 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 more important to make the 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 over 30 years old with early work utilizing Box-Jenkins ARIMA methods [2]. Recent approaches still use variations of the original ARIMA models, for example, seasonal ARIMA [3, 4], but there has been a shift towards using machine learning algorithms to address the traffic prediction challenges [5]. 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 [6]. These include building responsive algorithms that are able to predict nonrecurring congestion, determining the optimum data resolution, and identifying and modelling the important spatiotemporal dependencies in traffic data. The study described in this paper is a step towards addressing these challenges. We make the following key contributions:

(i) Explore the effect of the resolution of multivariate spatiotemporal input data on the accuracy of short-term traffic predictions models; we specifically consider models built using Artificial Neural Networks, Support Vector Regression, and Random Forests.

(ii) Evaluate the responsiveness of these predictive models to nonrecurring congestion events. Specifically, we study the reliability of the predictions provided by these models in the presence of unexpected events such as accidents.

(iii) Identify the spatiotemporal traffic attributes that most influence the performance of these models and their ability to model the complex 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 2. Next, Section 3 describes the dataset and methodology used to build and evaluate the predictive models, and Section 4 describes the machine learning algorithms used to build these models. Section 5 describes the hypotheses and measures used for experimental evaluation, and Section 6 analyzes the corresponding experimental results. Finally, Section 7 discusses the conclusions and directions for future work.

2. 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 spatiotemporal dynamics. We review some of the related work in this section.

Although studies in the existing literature predominantly use data aggregated over 5 min and 15 min 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 instance, Park et al. [7] investigated the effect of aggregation on travel time prediction and considered aggregation levels from 2 min to 60 min 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 [8] constructed a neural network model for making predictions and found that data aggregated over 5 min intervals gives better results than data aggregated over 1 min intervals. Vlahogianni and Karlaftis [9] 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; for example, Martin et al. [10] 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 spatiotemporal 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 [9]; however, these data smoothing techniques result in loss of information (and sensitivity) and make it difficult for the corresponding models to capture the spatiotemporal 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 spatiotemporal dynamics. For instance, Kamarianakis and Prastacos [11] used a Spatiotemporal 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 [12] found that vector autoregressive models that incorporate data from links neighbouring the link of interest perform better than ARIMA models that do not consider the data from the neighbouring links. Yang et al. [13] found that a sparse selection of neighbours chosen based on the level of correlation with the link of interest improves performance. Min and Wynter [14] showed that a multivariate spatiotemporal 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 [2–4, 11, 14], models based on machine learning and probabilistic estimation algorithms have also been explored because they are well-suited to model the complex spatiotemporal relationships in data. Popular approaches include Artificial Neural Networks (ANN) [15–19], Support Vector Machines (SVM) [20–24], k-Nearest Neighbours (kNN) [25–29], Kalman Filters [30–32], Bayesian Networks [33–35], and Random Forests [36, 37]. For instance, existing work has explored various ANN configurations. Wang et al. [19] 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. [38] used a binary neural network that incorporates spatiotemporal data for traffic prediction. Vlahogianni et al. [18] used a neural network model optimized with genetic algorithms and found that incorporating spatial and temporal data was helpful for multistep predictions. More recently, there have been efforts to use deep neural network architectures, including deep belief networks [39, 40] and stacked autoencoders [41].

There is no agreement in the literature regarding the number of upstream and downstream links (neighbouring any link of interest) that should be considered while building the predictive models. While some algorithms consider just one upstream or downstream link [24, 29], others consider a variable number of upstream and downstream links [38]. For an extensive review of spatiotemporal forecasting, please see Ermagun and Levinson [42]. As noted in Vlahogianni et al. [6], capturing spatial attributes in traffic data from a freeway is still an open problem.

Most existing work on short-term traffic prediction focuses on typical conditions [21]. 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 nonrecurring congestion such as accidents where periodic patterns do not hold. Of the studies that do not leave out nonrecurring congestion in their input data, a common approach is to create multiple models to deal with different conditions. For example, Dunne and Ghosh [43] used a model with nonlinear preprocessing in cases of congestion. Fusco et al. [44] reported good performance during nonrecurring congestion with a SARMA model, while a Bayesian Network performed better during recurring congestion. An online-SVR-based model was found to predict nonrecurring congestion accurately by Castro-Neto et al. [21]. Pan et al. [45] also highlight some of the challenges in capturing moving bottlenecks and nonrecurring congestion. See Vlahogianni et al. [6], Ermagun and Levinson [42], Oh et al. [46], and Oh et al. [47] for a more comprehensive overview of the existing literature in short-term traffic prediction.

In this study, we explore three machine learning algorithms that have demonstrated the ability to incorporate spatiotemporal 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. Please note that the primary objective of our study was not to introduce new algorithms. Instead, we make 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 nonrecurring congestion conditions. Third, we identify the spatiotemporal attributes that most influence the predictive accuracy of these models and their ability to model the complex dependencies in traffic data.

3. Methodology

This section introduces the study area and data and provides a mathematical formulation of the short-term traffic prediction problem (Section 3.1). This is followed by a description of the data preprocessing steps used in the proposed study (Section 3.2).

3.1. Study Area and Mathematical Formulation. This study was carried out in a 30 km 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 three lanes of roadway in each direction, and we only considered lanes going northbound in this study. The average length of a segment was 674 m, with the length varying between 52 m and 2252 m.

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% [48]. 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 model 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 [49]. It is not possible to accurately and comprehensively describe the current state of traffic using only information about flow. For example, if 200 vehicles pass over a detector during a 5 min 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. The combination of both volume and occupancy uniquely defines the current state of traffic. Unlike many existing studies that have only considered flow when making predictions, which does not define the traffic state uniquely, we consider both volume and occupancy because they each provide useful information. Together they help eliminate ambiguities, such as those described above.

For each predictive model, the input vector $X(s,t)$ is of the form:

$$X(s,t)=\begin{bmatrix} V_{t-T}^{1} & O_{t-T}^{1} & \cdots & V_{t-T}^{s} & O_{t-T}^{s} & \cdots & V_{t-T}^{S} & O_{t-T}^{S} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{t-1}^{1} & O_{t-1}^{1} & \cdots & V_{t-1}^{s} & O_{t-1}^{s} & \cdots & V_{t-1}^{S} & O_{t-1}^{S} \\ V_{t}^{1} & O_{t}^{1} & \cdots & V_{t}^{s} & O_{t}^{s} & \cdots & V_{t}^{S} & O_{t}^{S} \end{bmatrix}, \tag{1}$$

where V_t^s and O_t^s 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

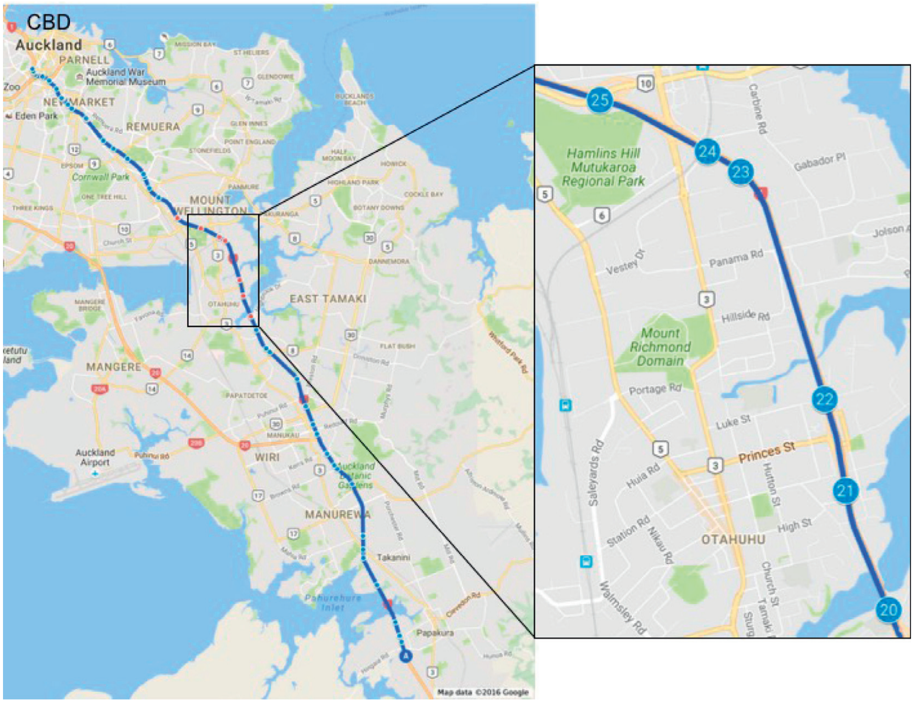


Figure 1: Study area with 45 road segments spread over 30 km of State Highway 1 (SH1) in Auckland. Volume and occupancy values from these road segments over a period of 30 days were used in this study.

interval for each specific segment s of interest. This output is a function of the input vector; for example, if traffic volume is to be predicted, the output of the models is $V^{s}_{t+5\,\mathrm{min}} = f(X(s,t))$. The goal of each machine learning algorithm is to build a model of this functional relationship between the inputs and outputs. The learned model can then be used to predict the output for any given input.

3.2. 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 5 min and 15 min 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 final size of the input dataset, with 20 time-steps included in each input sample, was thus 86370 samples for 30 s resolution, 86190 for 5 min, and 85790 for 15 min aggregation. Also, to ensure a fair comparison, the output is aggregated over the same time period for each model for all input time resolutions, that is, 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 preprocessed to remove some extreme values that were highly unlikely. First, we used winsorization [50] 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 volume values from segment 23 before and after winsorization.

Second, we scaled each attribute in the input data to lie $\in [0,1]$; this scaling was especially crucial 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. Nonstationary time-series data is typically transformed into stationary data before applying time-series models. However, traffic data is considered to be cyclostationary 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

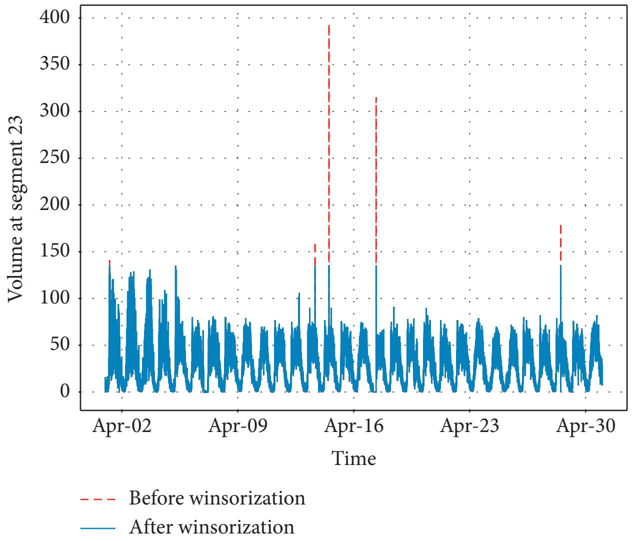


FIGURE 2: Volume values from segment 23 of study area before and after winsorization.

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 efficiently identify and respond to nonrecurring congestion conditions (also see Section 4.2).

Training of the models was accomplished using data from the first 20 days (57,600 samples), and data corresponding to the remaining ten days was 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.

4. 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 4.1), Support Vector Regression (Section 4.2), and Random Forests (Section 4.3).

4.1. Artificial Neural Network.

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:

$$y = \sigma\left(\sum_{i=1}^{N}(w_i \cdot x_i)\right). \tag{2}$$

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 backpropagation 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 [51]. 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 [52]. 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).

4.2. Support Vector Regression.

For classification problems, a Support Vector Machine computes a decision boundary that maximizes 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 training data and is as flat as possible [53]. 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:

$$\begin{aligned} \text{minimize} \quad & \left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi + \xi^*)\right\} \\ & y_i - w^T\mathbf{x}_i - b \le \varepsilon + \xi_i \\ \text{subject to} \quad & w^T\mathbf{x}_i + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0. \end{aligned} \tag{3}$$

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.

4.3. Random Forest. Random Forest (RF) [54] 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 $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \ldots$ and the corresponding (target) output values $Y = y_1, y_2, y_3 \ldots$, decision trees will be created as follows:

```
for $b$ in $1 \ldots B$ do
    Pick $N$ training samples randomly with replacement; call this subset $\{\mathbf{X}_b, Y_b\}$
    Train a decision tree $\Theta_b$ 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 created by sampling from the training set with replacement results in a decision tree. The prediction for any test input $\widehat{\mathbf{x}}$ is then the average of the predictions from each decision tree:

$$\widehat{y} = \frac{1}{B} \sum_{i=1}^{B} \Theta_b (\widehat{\mathbf{x}}). \tag{4}$$

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.

5. Hypotheses and Measures

We experimentally evaluated the following hypotheses regarding the predictive models learning using the machine learning algorithms:

(1) The learned models are able to disregard the amplification of noise and variations in high-resolution data and provide higher accuracy than models that do not use high-resolution data

(2) The learned models are responsive to nonrecurring congestion events such as accidents, and this ability improves with the increase in the resolution of data

(3) The learned models are able to capture the complex spatiotemporal evolution of traffic by assigning higher importance to volume and occupancy attributes extracted from segments near the segment of interest

As baselines for comparison, wherever appropriate, we used two established methods for volume prediction in existing literature (ARIMA, historical average). To experimentally evaluate the hypotheses, we used three measures: accuracy, root mean square error (RMSE), and mean absolute error (MAE), defined as follows:

$$\begin{aligned} \text{accuracy} &= 1 - \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \widehat{y}_i}{y_i} \right|, \\ \text{RMSE} &= \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \widehat{y}_i \right)^2}, \\ \text{MAE} &= \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \widehat{y}_i \right|, \end{aligned} \tag{5}$$

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

To quantify responsiveness to nonrecurring conditions, we computed these measures over samples that were representative of nonrecurring 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:

$$\begin{aligned} &\left| y_i - \widehat{\mu}_i \right| > (2 * \text{std}) \\ &\text{std} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - \widehat{\mu}_i \right)^2}{N - 1}}, \end{aligned} \tag{6}$$

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

6. Experimental Results

This section discusses the results of experimentally evaluating the three hypotheses listed in Section 5. We summarize the results in Sections 6.1, 6.2, and 6.4 and examine the computational efficiency of the proposed models in Section 6.3. Unlike results reported in many papers, our predictive models 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 30 sec to 15 min for the input data, but the output of each model is the volume or occupancy of vehicles (in a particular segment in the highway) aggregated over a period of five minutes—see Section 3.1 for more details.

6.1. Using High-Resolution Data. As stated in Section 3.1, 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 was stable using different random initializations.

The results summarized in Table 1 show that all three machine learning algorithms performed better with 30 sec aggregation level for input data in comparison with the 5 min and 15 min 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. As baselines for comparison, we considered two established methods for volume prediction in the 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 learned models, we evaluated all models at the same output resolution of 5 min. For instance, for the 30 sec aggregation level, the 5 min aggregated output value was obtained by iterating and aggregating the output over ten one-step-ahead predictions. Also, results for the 15 min input aggregation level were obtained by first applying the Stram-Wei temporal disaggregation [55] to extract 5 min aggregated values from the 15 min aggregated data. ARIMA (2, 1, 2) models were used for predicting volume at the 5 min and 15 min input aggregation levels, ARIMA (2, 1, 1) models were used for predicting occupancy at the 5 min and 15 min aggregation levels, and ARIMA (4, 1, 0) models were used for the 30 sec input aggregation level. These models were selected experimentally using the Box-Jenkins method.

The results in Table 1 indicate that the models corresponding to the 30 sec input aggregation level provide an average accuracy improvement of 8.1% over the ARIMA approach and an average 12.5% improvement over the historical average baseline. Note that these results include both recurring and nonrecurring congestion events; we examine the nonrecurring events in more detail in Section 6.2. To confirm the significance of these results, we conducted Diebold–Mariano (DM) tests for predictive accuracy [56]. The DM test compares the forecast accuracy of a pair of forecast methods. The test's null hypothesis 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 example, for a significance of 99%, the null hypothesis is rejected if the DM statistic $\notin [-2.58, 2.58]$. We used MSE as the error metric. Table 2 shows the DM test statistic for each pair of models. Except for the 5 min SVR and 15 min RF models, all other models have significantly different levels of accuracy.

Table 3, which summarizes the results of predicting occupancy, indicates similar trends. Although all three predictive models based on machine learning algorithms performed well, the model based on the Random Forest algorithm (Section 4.3) provided the highest accuracy.

Next, the average accuracy and MAE at different times of the day, for the three different data aggregation levels, are shown in Figure 3. For each algorithm, the accuracy increases with the resolution. Overall, we observe that the performance of the learned predictive models improves significantly with the increase in resolution despite the associated amplification of noise and minor variations in data.

The results discussed so far support the first hypothesis 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 values during overnight hours can be explained by the accuracy being represented as a 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.

6.2. Nonrecurring Congestion. Next, we evaluated the second hypothesis by examining the responsiveness of the predictive models to nonrecurring congestion events. We did so by only evaluating the trained predictive models on a subset of the test set comprising samples that were significantly different from historical average values. The results are summarized in Tables 4 and 5. We observe that the models built using input data at the 30 sec aggregation level outperform the models use input data at the 5 min and 15 min aggregation levels. Among the learned models, 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 learned predictive models 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 learned models, we examined 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 was in the test dataset, that is, not used to train the predictive models. Figure 4(a) shows that there was a significant deviation from the average traffic around 6.40 am on April 21, 2016. As reported on the social media site, Twitter, there was a breakdown near SH1 at ≈ 6.30 am that day (see Figure 4(b)). More specifically, the *Ellerslie on-ramp* mentioned in the tweet is near segment 27 of SH1, which is ≈ 4 km from segment 23 on SH1.

Figures 5(a)–5(c) show how the learned predictive models are able to track the traffic volume corresponding to this event, with each 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 30 sec input data aggregation level enabled the learned models to predict the change in traffic volume at almost the same time-step when the nonrecurring event occurred, whereas there is a lag when the other two aggregation levels are used; the performance is significantly worse with the baseline ARIMA model.

For additional examples of how the models predicted during nonrecurring congestion, see Figure 6. These plots indicate that the ANN model at the 30 sec input aggregation level responds very quickly to nonrecurring congestion. The

Table 1: Traffic volume prediction under all conditions.

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** (0.01)	**31.2** (11.7)	**22.2** (8.5)	0.904 (0.01)	34.0 (11.2)	23.8 (8.5)	0.890 (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)	79.7 (35.7)	43.5 (17.4)	0.806 (0.01)	79.7 (35.7)	43.5 (17.4)	0.806 (0.01)	79.7 (35.7)	43.5 (17.4)
ARIMA	0.839 (0.02)	54.6 (18.3)	39.1 (13.2)	0.879 (0.01)	43.8 (15.6)	30.6 (11.4)	0.881 (0.01)	44.3 (16.3)	30.1 (11.4)

Standard deviations across segments are reported in parentheses and numbers in boldface show the best results.

Table 2: Diebold–Mariano test statistic for each pair of models for predicting volume.

		0.5 min			5 min			15 min		
		ANN	RF	SVR	ANN	RF	SVR	ANN	RF	SVR
0.5 min	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
5 min	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	−100.1	−45.49	−66.02
	SVR	19.35	50.30	52.00	−48.64	50.28	—	−79.88	**0.49**	−57.11
15 min	ANN	71.81	94.20	92.40	11.23	100.1	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	—

Critical value: |2.58|; numbers in boldface indicate pairs of models that are not significantly different.

Table 3: Traffic occupancy prediction under all conditions.

Model	Input resolution (minutes)								
	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	0.859 (0.02)	1.98 (0.64)	1.00 (0.37)	0.838 (0.01)	2.59 (0.74)	1.27 (0.44)	0.780 (0.03)	3.51 (0.89)	1.70 (0.57)
RF	**0.872** (0.01)	**1.83** (0.48)	**0.90** (0.30)	0.850 (0.01)	2.17 (0.55)	1.07 (0.35)	0.80 (0.03)	2.80 (0.70)	1.43 (0.47)
SVR	0.858 (0.01)	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 (0.04)	20.5 (4.71)	10.1 (2.65)	0.833 (0.02)	2.37 (0.70)	1.17 (0.41)	0.834 (0.02)	2.59 (0.80)	1.22 (0.43)

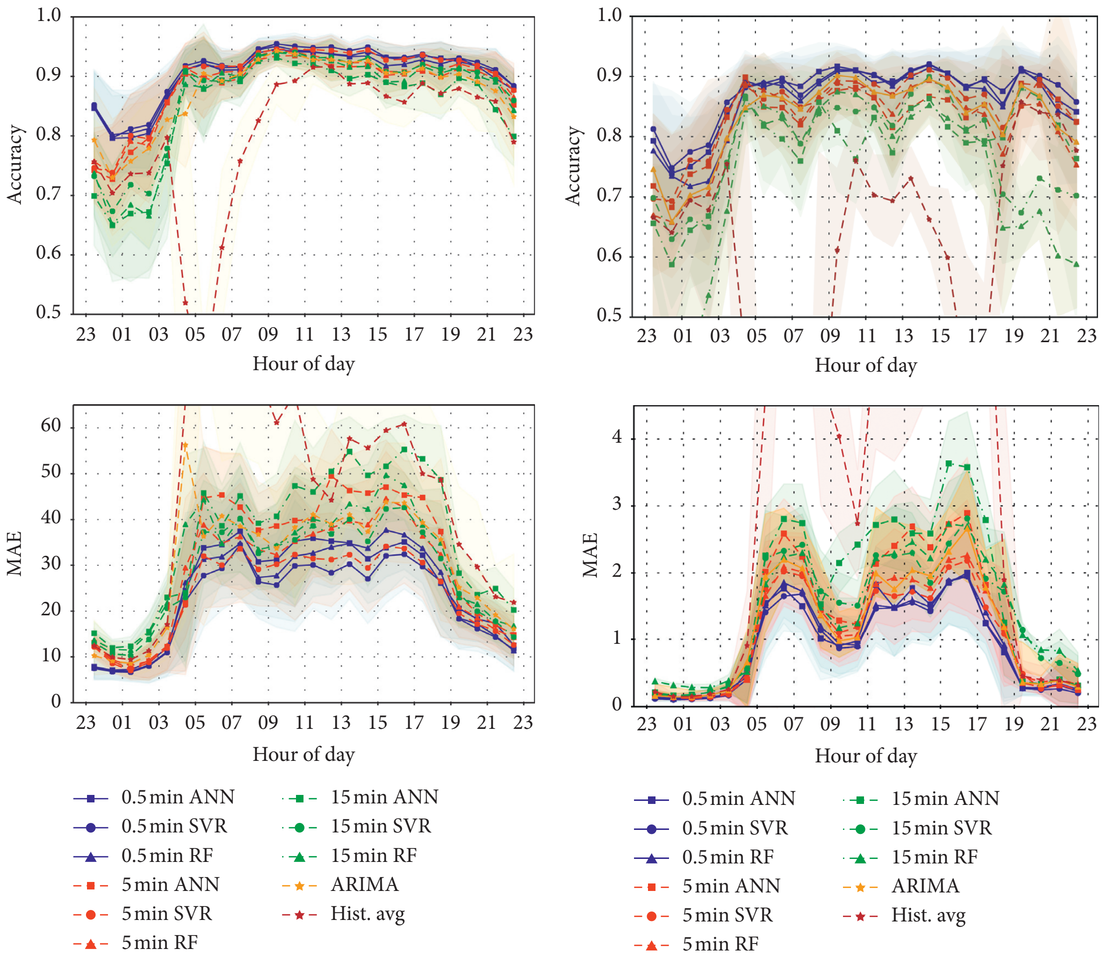
Standard deviations across segments are reported in parentheses and numbers in boldface show the best results.

SVR-based models and the coarser-resolution models tend to smooth out shocks to traffic and are better at smoothing out the noise in typical congestion conditions. The RF-based learned models tend to provide good overall performance that lies in between that provided by the ANN-based and SVR models.

Figure 7 shows that an ANN-based learned model at the 30 sec input data 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 support the second hypothesis that the models based on machine learning algorithms and high-resolution data are more responsive to nonrecurring congestion.

6.3. Computational Efficiency and Practical Scalability. Table 6 summarizes the training time and testing time of the proposed models, when they are built and evaluated on an Intel Core *i*73.4 *GHz* desktop with 8 GB 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 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, for example, 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; for example, 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 (≈ 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.



(a)

(b)

Figure 3: Accuracy and MAE at different times of the day. Shaded areas are the 95% confidence intervals across days in the test set. (a) Volume predictions. (b) Occupancy predictions.

Table 4: Traffic volume prediction under nonrecurring congestion conditions.

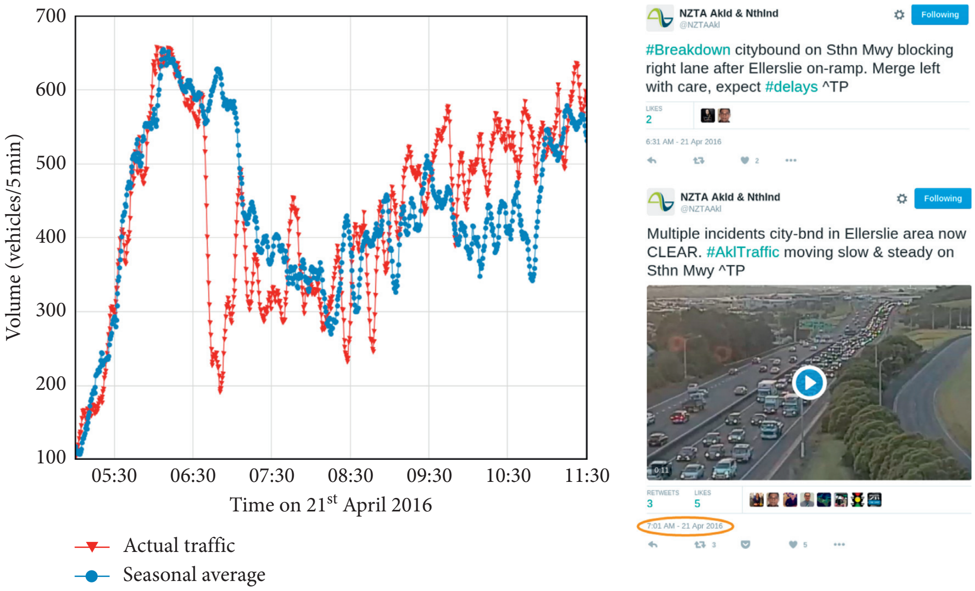
	Input resolution (minutes)								
Model	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	**0.913** (0.01)	**46.9** (16.4)	**33.2** (12.1)	0.880 (0.02)	66.5 (24.4)	46.2 (17.0)	0.840 (0.03)	80.2 (29.7)	59.3 (22.8)
RF	0.900 (0.01)	50.1 (17.4)	37.4 (13.2)	0.890 (0.01)	57.3 (19.6)	42.0 (15.2)	0.860 (0.02)	66.6 (21.1)	50.9 (16.0)
SVR	0.892 (0.02)	56.0 (18.9)	41.0 (14.1)	0.870 (0.02)	67.2 (21.4)	49.7 (16.1)	0.850 (0.03)	76.1 (22.9)	56.4 (17.0)
Historical avg.	0.139 (0.08)	232 (109)	192 (83.6)	0.139 (0.08)	232 (109)	192 (83.6)	0.139 (0.08)	232 (109)	192 (83.6)
ARIMA	0.851 (0.02)	73.8 (20.5)	54.2 (15.5)	0.670 (0.02)	176 (157)	126 (48)	0.860 (0.02)	77.7 (30.4)	51.6 (19.7)

Standard deviations across segments are reported in parentheses and numbers in boldface show the best results.

Table 5: Traffic occupancy prediction under nonrecurring congestion conditions.

	Input resolution (minutes)								
Model	0.5			5			15		
	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE	Accuracy	RMSE	MAE
ANN	0.869 (0.01)	1.93 (1.27)	0.94 (0.29)	0.837 (0.01)	2.77 (1.72)	1.32 (0.34)	0.80 (0.02)	3.50 (2.19)	1.63 (0.48)
RF	**0.873** (0.01)	**1.88** (1.22)	**0.91** (0.27)	0.850 (0.01)	2.21 (1.42)	1.07 (0.32)	0.796 (0.02)	2.85 (1.83)	1.42 (0.43)
SVR	0.858 (0.01)	1.92 (1.20)	0.95 (0.23)	0.828 (0.01)	2.18 (1.38)	1.13 (0.28)	0.73 (0.03)	2.58 (1.60)	1.44 (0.31)
Historical avg.	−1.57 (0.83)	18.0 (7.92)	16.4 (1.76)	−1.57 (0.83)	18.0 (7.92)	16.4 (1.76)	−1.57 (0.83)	18.0 (7.92)	16.4 (1.76)

Standard deviations across segments are reported in parentheses, and numbers in boldface show the best results.



NZTA Akld & Nthlnd
@NZTAAkl
Following

#Breakdown citybound on Sthn Mwy blocking right lane after Ellerslie on-ramp. Merge left with care, expect #delays ^TP

LIKES
2

6:31 AM - 21 Apr 2016

NZTA Akld & Nthlnd
@NZTAAkl
Following

Multiple incidents city-bnd in Ellerslie area now CLEAR. #AklTraffic moving slow & steady on Sthn Mwy ^TP

RETWEETS 3 LIKES 5

7:01 AM - 21 Apr 2016

(a) (b)

Figure 4: (a) Traffic volume in segment 23 on April 21, 2016, (Thursday) compared with the historical weekly average and (b) tweets from NZTA accessed from [57] on April 21, 2016.

Overall, we believe that models based on these machine learning methods will scale to large road networks. The retraining 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.

6.4. Attribute Selection. Next, we evaluate the third hypothesis regarding the ability to model the complex spatiotemporal evolution of traffic. To do so, we first identify the attributes that most influence the performance of the learned predictive models.

One common approach for identifying informative attributes is to compute the Pearson correlation coefficient between the target variable and each of the input attributes [42]. However, the Pearson correlation coefficient is not able to capture nonlinear relationships that may exist between the input attributes and thetarget variable . We, therefore, used the Recursive Feature Elimination (RFE) approach to select the most relevant (i.e., informative) attributes [58, 59]. 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 the 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 [60].

Figures 8(a), 9, and 10 visualize the relative ranking of each of the 1800 input attributes considered by the models for traffic prediction at a particular segment (segment 23 in these figures). The 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, for example, for the 30 sec 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 a higher rank to neighbouring 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 30 sec 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

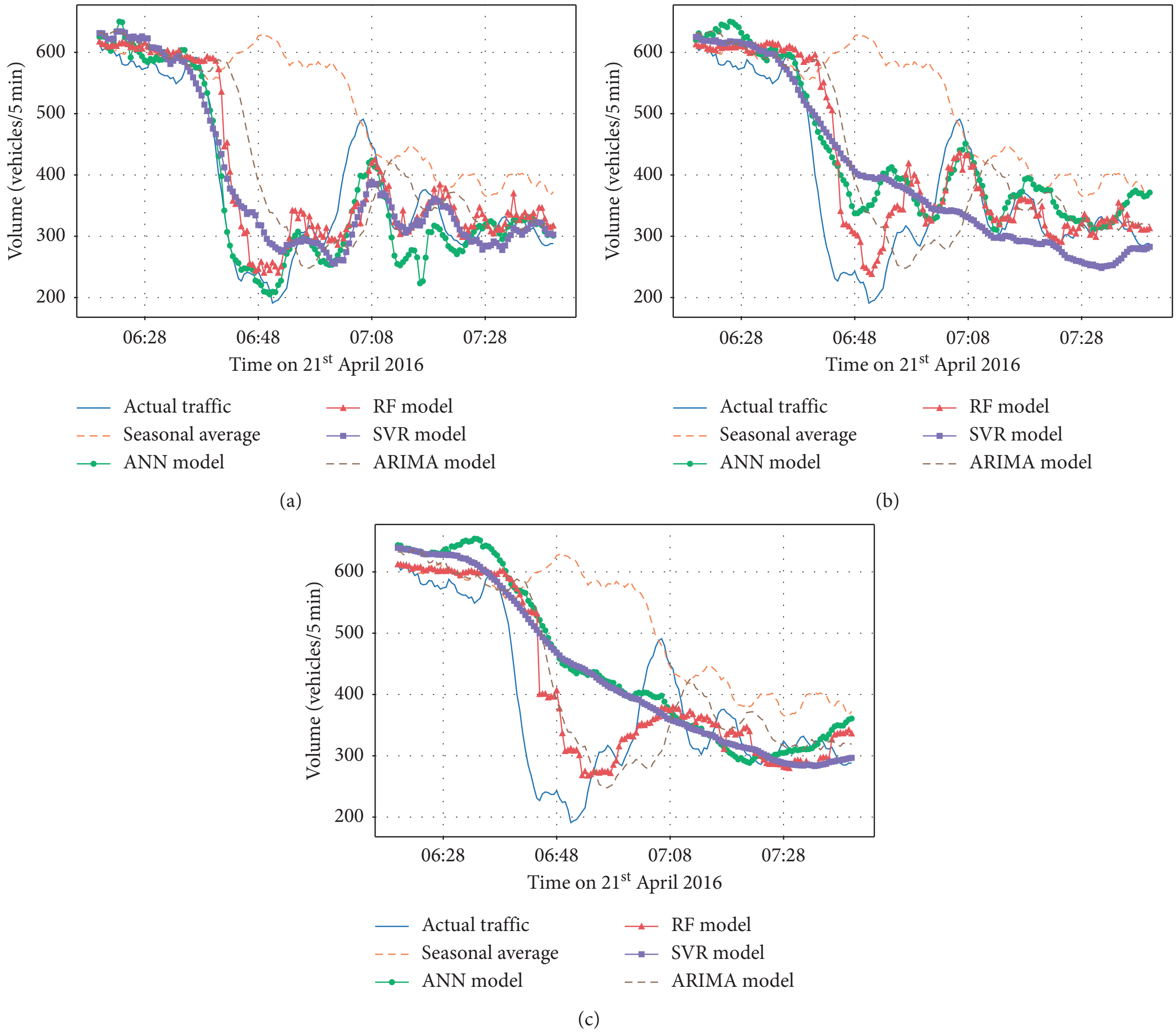


Figure 5: Traffic volume predictions in response to a nonrecurring congestion event for different input data aggregation levels (30 sec, 5 min, and 15 min); models using higher-resolution data respond better. (a) 30 sec aggregation level. (b) 5 min aggregation level. (c) 15 min aggregation level.

time-steps. The number of attributes corresponding to downstream segments that are nearby is high for the higher-resolution models, especially when predicting nonrecurring congestion events. For the models based on the 5 min and 15 min 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 neighbouring locations only provides minor improvements in performance [13]. Our results, on the other hand, indicate that volume and occupancy values from multiple neighbouring 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 11 shows the performance of the three models for the 30 sec 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 nonrecurring congestion events. Tables 7 and 8 as well as Figures 12 and 13 indicate that the

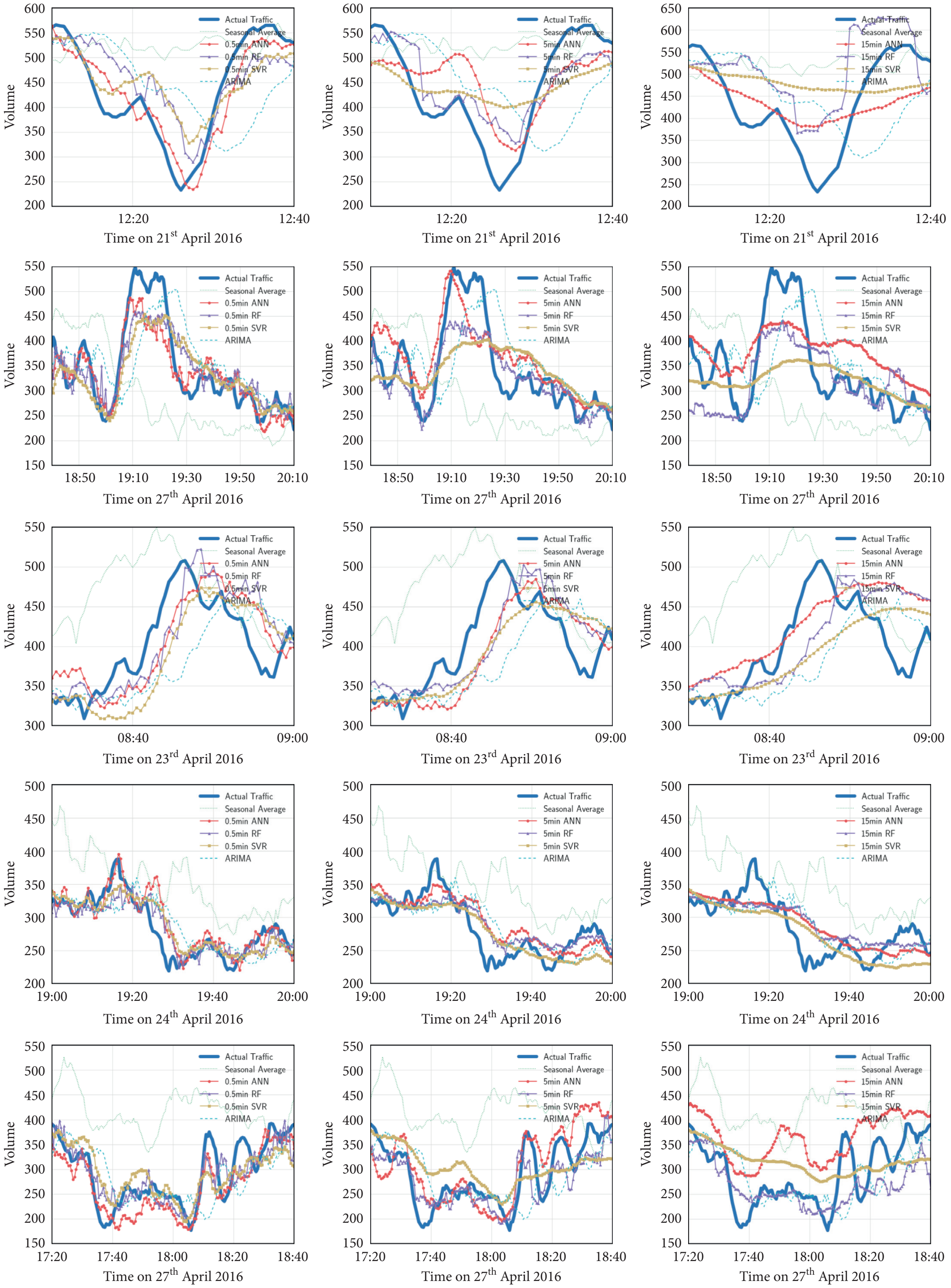


Figure 6: Additional examples of nonrecurring congestion events. Predictive models based on machine learning methods provide good tracking performance, especially at the high-resolution (30 sec) input data aggregation level.

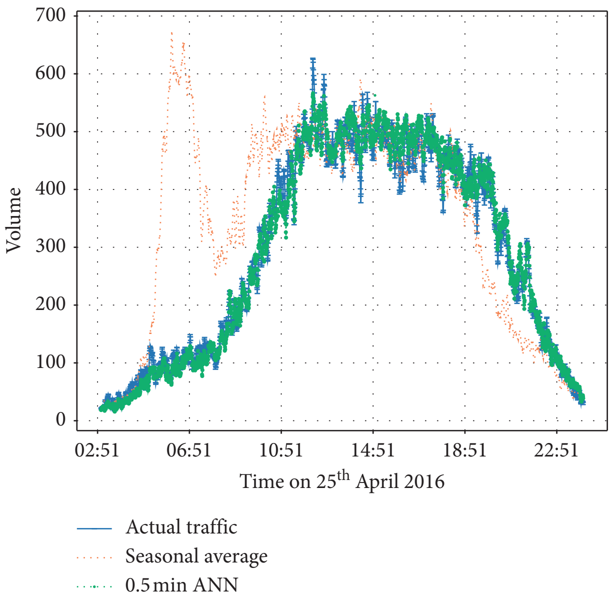
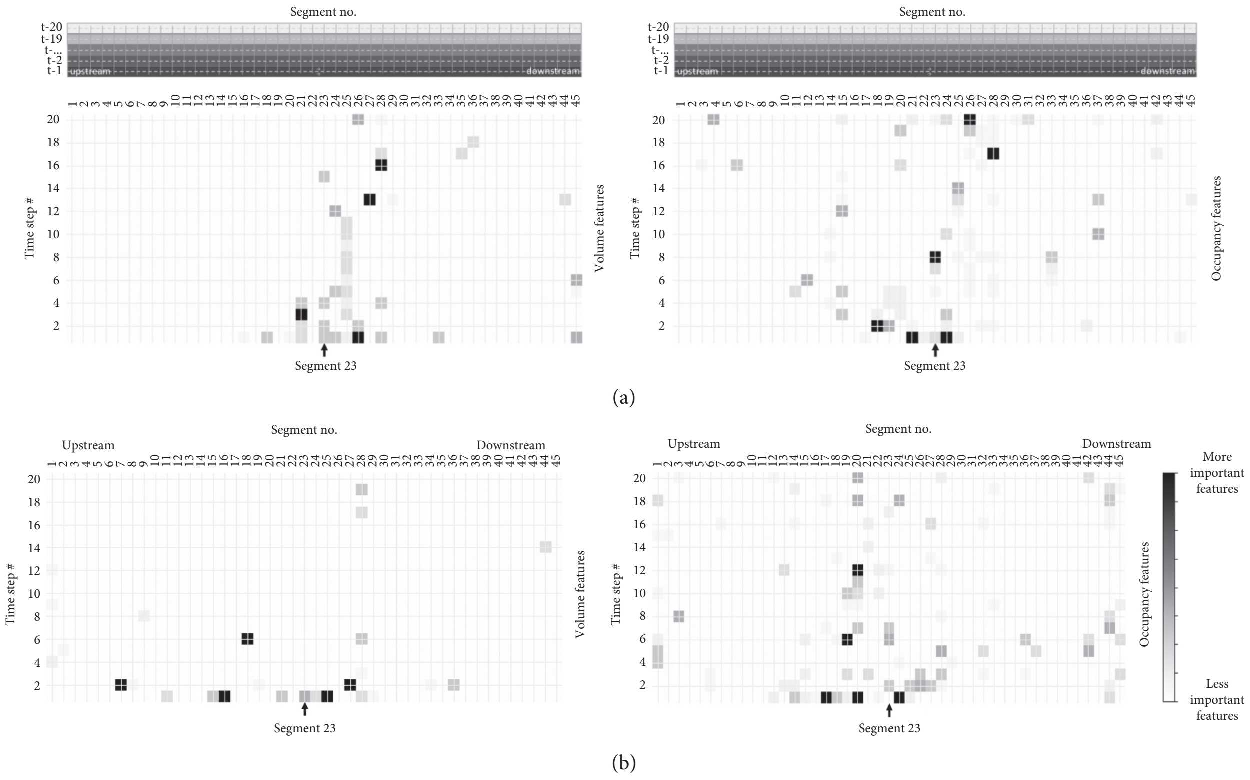


Figure 7: Traffic volume prediction on April 25, 2016, a public holiday in New Zealand (ANZAC day).

Table 6: Training and testing time for each of the three learned models for short-term traffic prediction. Results indicate that these learned models will scale well for short-term predictions in large road networks.

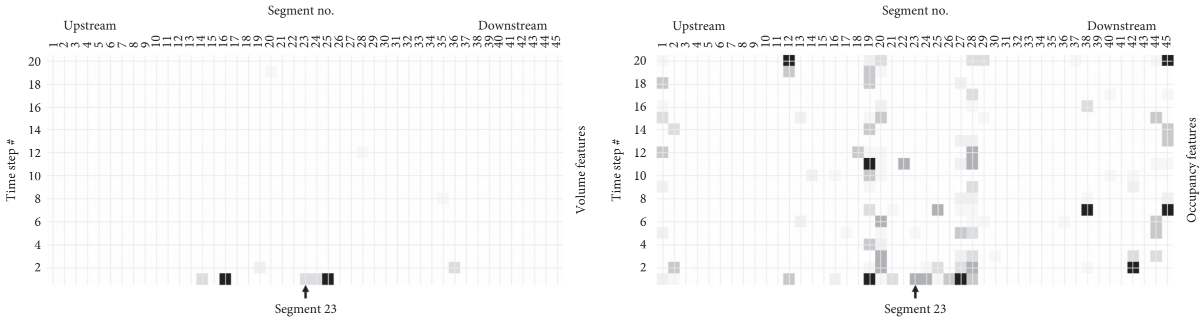
	Average training time for 57600 samples (seconds)	Average prediction time for one input sample (milliseconds)
ANN	283.8	0.16
RF	1154	82.08
SVR	4.743	0.0223



(a)

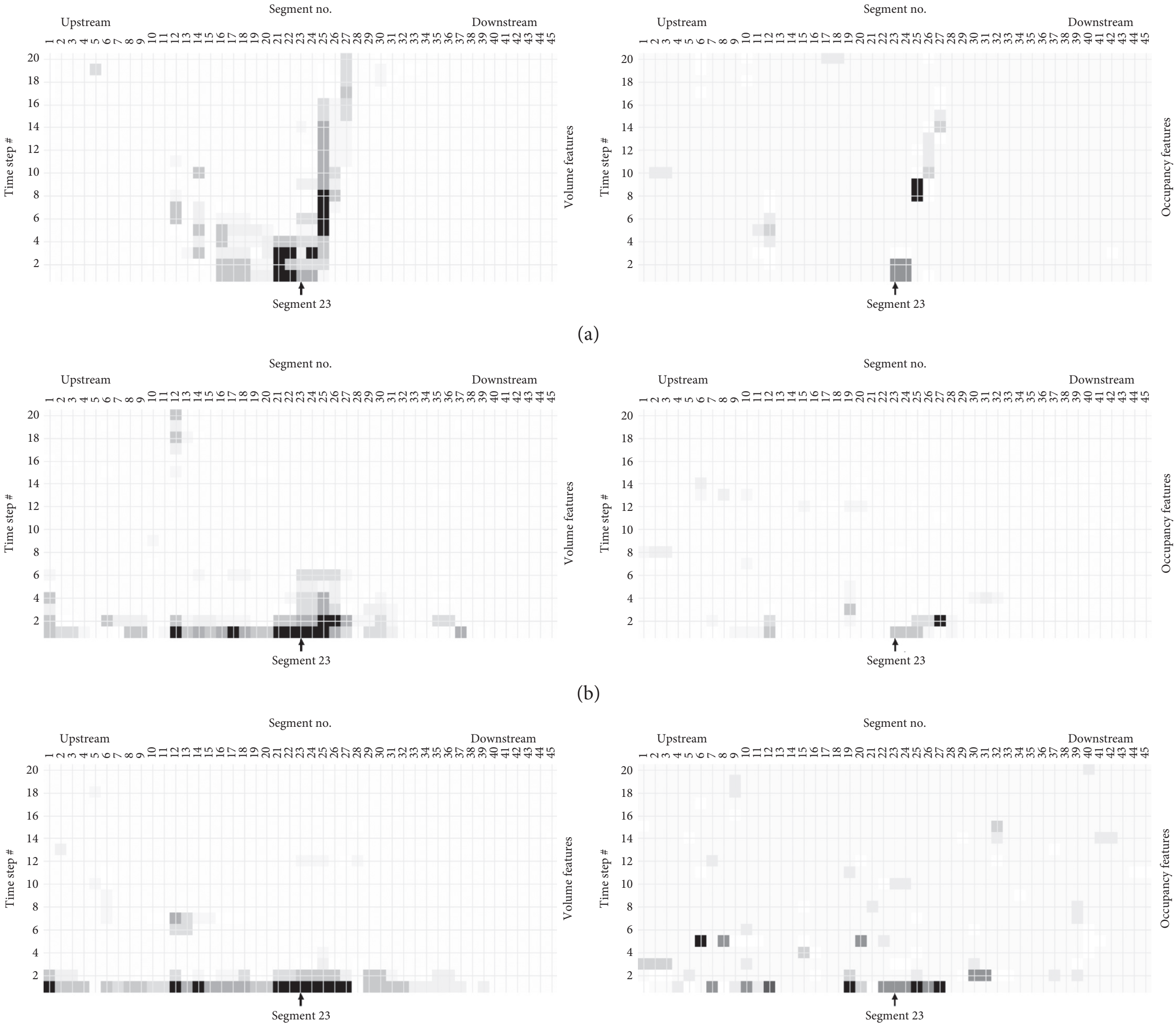
(b)

Figure 8: Continued.



(c)

Figure 8: Ranking of attributes in terms of their relative importance to the performance of ANN models, for each of the three different input data aggregation levels (segment 23). The plots for the volume features are on the left and those for the occupancy features are on the right. (a) 30 sec aggregation level. (b) 5 min aggregation level. (c) 15 min aggregation level.



(a)

(b)

(c)

Figure 9: Ranking of attributes in terms of their relative importance to the performance of SVR models, for three different input data aggregation levels (segment 23). The plots for the volume features are on the left and those for the occupancy features are on the right. (a) 30 sec aggregation level. (b) 5 min aggregation level. (c) 15 min aggregation level.

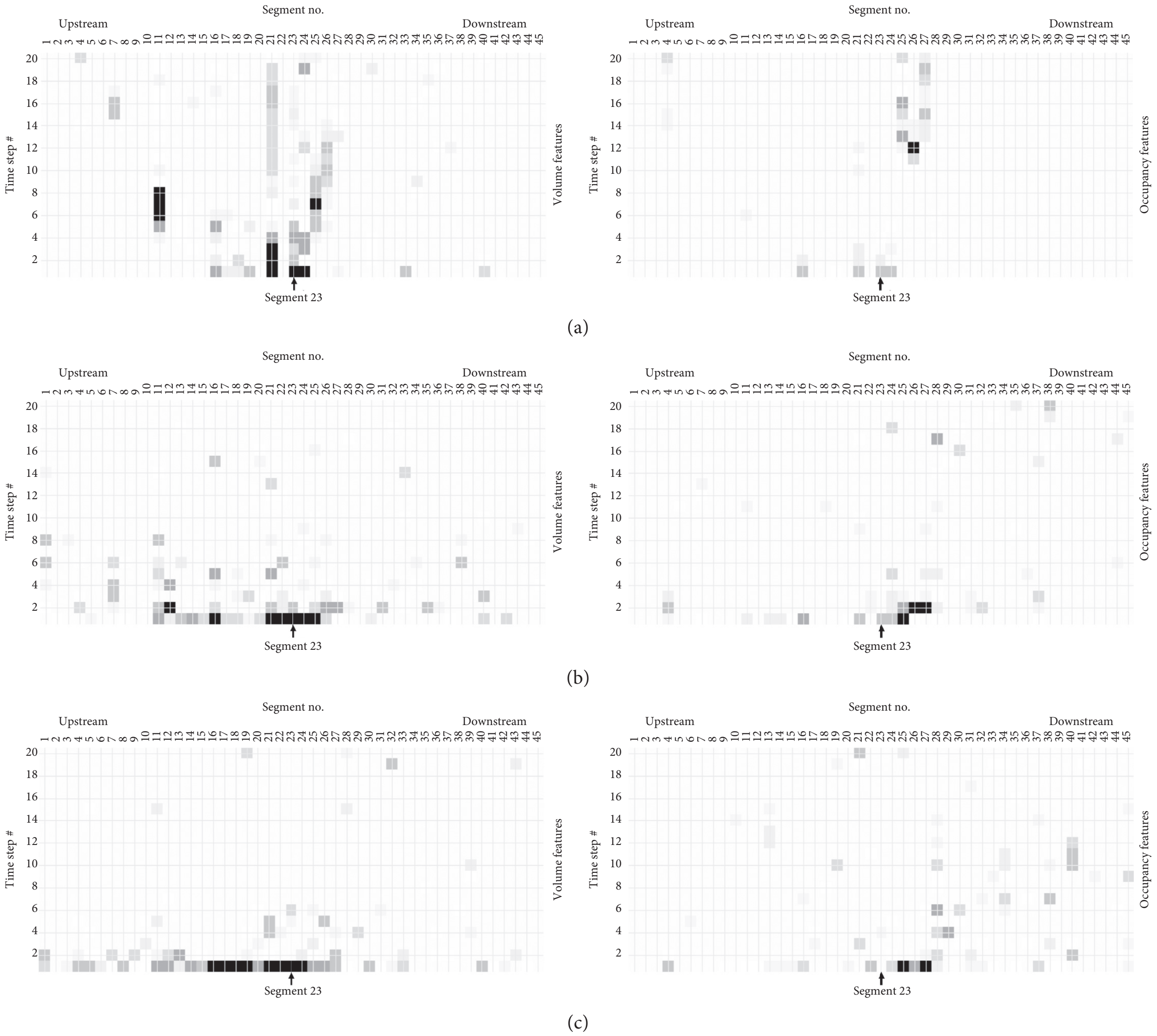


Figure 10: Ranking of attributes in terms of their relative importance to the performance of RF models, for three different input data aggregation levels (segment 23). The plots for the volume features are on the left and those for the occupancy features are on the right. (a) 30 sec aggregation level. (b) 5 min aggregation level. (c) 15 min aggregation level.

RFE approach outperforms the other two approaches for ranking attributes. In fact, in the case of nonrecurring congestion, the prediction accuracy using correlation-based attribute selection is similar to that with a random selection of the important attributes. One explanation for the poor performance provided by correlation-based feature selection is that the features that are most likely to be highly correlated to the output correspond to the road segments closest to the segment under consideration. However, in most cases, these features give redundant information. Segments further away may contain information about situations such as queues building up or a spike in traffic that is not necessarily correlated with the output but are quite informative for predictions. The RFE provides an opportunity to identify these dependencies, and the experimental results show that it is a much better choice for accurate traffic prediction, especially with nonrecurring congestion events. The experimental results also support the hypothesis that the predictive models based on the machine learning algorithms capture the complex spatiotemporal evolution of traffic by assigning higher importance to the attributes that are more relevant to the prediction task.

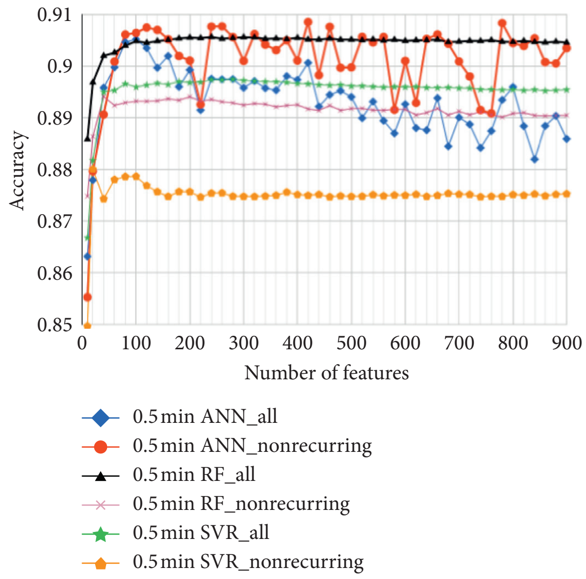


Figure 11: Accuracy of each of the three models for the 30 sec input data aggregation level, as a function of the number of attributes considered; attributes ranked in decreasing order of importance using RFE approach.

Table 7: Comparison of feature selection methods; traffic volume predictions under all conditions with 30 sec input data; given a constrained number of features, in most cases, the RFE method achieves better performance compared to random and correlation-based feature selection.

Model	Features	Accuracy			RMSE			MAE		
		Correlation	Random	RFE	Correlation	Random	RFE	Correlation	Random	RFE
ANN	10	0.864	0.853	0.863	40.5	47.0	42.2	27.9	32.7	29.1
	20	0.880	0.850	0.878	37.7	46.2	34.7	25.7	32.9	24.8
	40	0.891	0.866	0.896	35.4	39.5	31.2	24.3	28.0	21.7
	60	0.893	0.877	0.900	35.5	36.9	29.0	23.8	26.5	20.8
	80	0.893	0.881	0.905	35.2	36.6	**26.9**	23.7	25.5	**19.3**
	100	0.894	**0.896**	**0.905**	34.9	**30.8**	27.5	23.5	**21.9**	19.6
	120	0.894	0.885	0.904	34.6	33.8	29.0	23.3	23.9	20.4
	140	0.889	0.889	0.900	35.4	33.1	28.7	23.8	23.6	20.3
	160	0.897	0.879	0.902	33.1	37.6	29.6	22.5	26.3	20.5
	180	**0.898**	0.884	0.896	**32.7**	34.3	31.4	**22.3**	24.4	21.7
RF	10	0.871	0.841	0.886	39.9	49.6	34.6	27.5	34.5	24.2
	20	0.880	0.847	0.897	37.5	45.0	30.4	25.9	32.6	21.5
	40	0.886	0.875	0.902	36.1	36.2	28.3	24.7	25.9	20.1
	60	0.891	0.882	0.903	35.1	35.0	28.3	23.7	24.9	20.1
	80	0.892	0.866	0.904	35.0	36.0	28.0	23.5	25.5	19.9
	100	0.893	0.886	0.905	34.8	33.7	27.7	23.3	24.2	**19.7**
	120	0.894	0.885	0.905	34.7	34.6	27.8	23.2	24.5	19.8
	140	0.894	**0.890**	0.905	34.7	**32.8**	**27.7**	23.2	**23.3**	19.7
	160	0.894	0.883	0.905	34.7	33.7	27.7	23.2	24.2	19.7
	180	**0.895**	0.890	**0.905**	**34.6**	32.9	27.7	**23.1**	23.4	19.7
SVR	10	0.859	0.758	0.867	41.6	76.2	38.3	29.2	51.3	27.5
	20	0.870	0.843	0.882	40.1	55.5	33.7	27.8	36.4	24.4
	40	0.877	0.850	0.895	38.4	47.8	**30.9**	26.5	33.1	22.2
	60	0.881	0.854	0.895	37.8	49.7	32.1	25.7	34.2	22.7
	80	0.885	**0.879**	0.897	37.5	**38.2**	31.5	25.2	**26.5**	22.3
	100	0.886	0.876	0.896	**37.4**	38.2	32.1	**24.9**	26.9	22.6
	120	0.887	0.869	0.896	37.6	42.4	31.8	25.1	29.6	22.4
	140	0.887	0.878	0.897	37.7	38.9	31.6	25.1	26.9	22.3
	160	0.887	0.878	0.896	37.7	38.7	31.7	25.1	27.0	22.3
	180	**0.887**	0.878	**0.897**	37.8	38.8	31.4	25.1	27.1	**22.1**

Table 8: Comparison of feature selection methods.

Model	Features	Accuracy			RMSE			MAE		
		Correlation	Random	RFE	Correlation	Random	RFE	Correlation	Random	RFE
ANN	10	0.833	0.800	0.855	70.1	80.7	58.3	49.8	59.1	43.4
	20	0.845	0.822	0.880	66.1	69.4	47.8	46.5	53.6	35.8
	40	0.861	0.843	0.891	60.2	63.8	45.5	42.6	47.6	33.3
	60	0.857	0.864	0.901	62.8	55.6	41.1	43.8	41.6	30.1
	80	0.861	0.857	0.906	61.4	60.6	39.8	42.9	43.5	28.7
	100	0.862	**0.892**	0.906	60.4	**45.3**	**39.0**	42.3	**32.8**	**28.3**
	120	0.861	0.883	**0.907**	61.1	49.5	40.0	42.7	35.9	28.8
	140	0.856	0.886	0.907	63.5	48.2	40.0	44.1	35.0	28.4
	160	**0.874**	0.875	0.905	**55.6**	54.0	40.0	**39.1**	38.4	29.0
	180	0.868	0.876	0.902	57.2	52.1	43.9	40.2	38.0	31.0
RF	10	0.832	0.797	0.875	69.2	81.1	52.4	50.3	61.1	38.5
	20	0.842	0.836	0.886	65.9	62.9	48.0	47.8	48.9	34.7
	40	0.847	0.858	**0.894**	64.0	55.5	**43.9**	46.0	42.1	**32.2**
	60	0.852	0.856	0.892	**63.7**	58.7	44.6	45.1	43.7	32.8
	80	0.851	0.865	0.893	64.0	57.2	44.5	45.2	41.1	32.6
	100	0.852	0.868	0.893	63.8	54.2	44.2	45.0	40.4	32.5
	120	0.851	0.860	0.893	64.0	57.6	44.2	45.2	42.3	32.5
	140	0.852	0.868	0.893	63.8	54.0	44.0	45.0	40.4	32.4
	160	0.851	0.867	0.894	64.0	53.9	43.9	45.1	40.2	32.4
	180	**0.852**	**0.872**	0.893	63.9	**52.5**	44.1	**44.9**	**38.7**	32.4
SVR	10	0.825	0.722	0.850	72.6	114.1	59.3	53.0	84.4	45.2
	20	0.822	0.751	**0.880**	73.5	100.8	**47.2**	53.4	73.1	**36.5**
	40	0.834	0.796	0.874	**69.4**	81.6	50.0	50.1	60.4	37.7
	60	**0.834**	0.816	0.878	70.2	76.3	50.0	**50.0**	55.2	37.2
	80	0.832	0.838	0.879	71.5	66.0	49.9	50.5	48.7	36.9
	100	0.832	**0.842**	0.879	72.0	**63.8**	50.3	50.5	**47.5**	37.1
	120	0.829	0.828	0.877	72.7	70.0	50.8	51.2	51.8	37.5
	140	0.830	0.832	0.876	72.6	69.3	51.3	51.1	50.7	37.8
	160	0.829	0.839	0.875	72.7	65.9	51.7	51.1	48.3	38.0
	180	0.830	0.835	0.876	72.6	66.9	51.3	51.1	49.5	37.7

Traffic volume predictions under nonrecurring conditions with 30 sec input data; given a constrained number of features, in most cases, the RFE method achieves better performance compared to random and correlation-based feature selection.

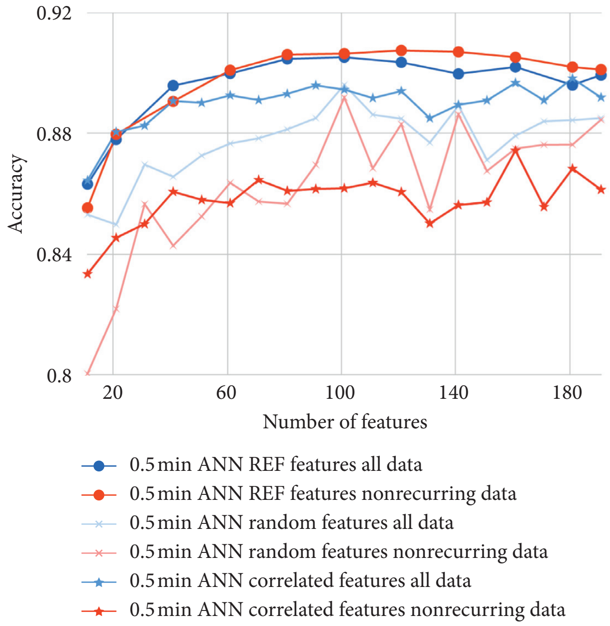


Figure 12: Performance comparison of RFE, correlation-based and random-selection approaches for selecting important attributes; results correspond to an ANN model for the 30 sec input data aggregation level.

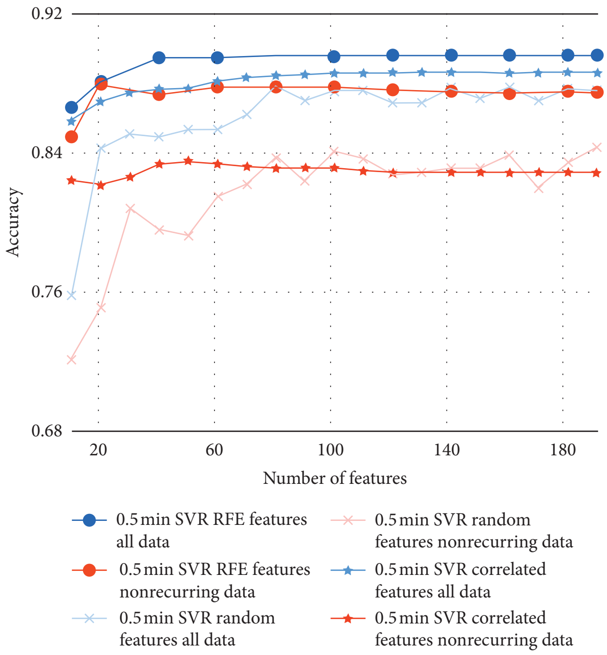


Figure 13: Performance comparison of RFE, correlation-based, and random-selection approaches for selecting important attributes; results correspond to an SVR model for the 30 sec aggregation level. RFE provides the highest accuracy.

7. Conclusions

Traffic congestion results in significant monetary losses in countries around the world. 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. Despite being a mature field, short-term traffic prediction poses many open problems such as the (a) choice of the optimal input data resolution; (b) reliable prediction and efficient tracking of nonrecurring congestion events; and (b) accurate modelling of the complex spatio-temporal dependencies influencing traffic estimation. We have explored the construction and use of predictive models based on three established machine learning 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 input data aggregation levels, 30 sec, 5 min, and 15 min. For each learned model, the output was a prediction (of volume or occupancy) over a 5 min period, although the same methodology can be used to provide predictions over 10 min or 15 min intervals as well. Our experiments indicate the following.

(i) 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 despite the corresponding amplification of noise and variability in the sensor measurements.

(ii) By not explicitly exploiting the periodic characteristics in traffic, the machine learning models studied here perform equally well under both recurring and nonrecurring congestion without requiring any special changes to the models. The corresponding experimental results also indicate that these learned models are able to capture the underlying complex, spatiotemporal evolution of traffic.

(iii) Recursive Feature Elimination provides a good ranking of attributes for short-term traffic prediction. The more commonly used linear Pearson correlation coefficient-based feature selection [42] provides poor prediction accuracy similar to that with a random selection of features in the presence of nonrecurring congestion. Furthermore, feature selection enables us to visualize and better understand the spatiotemporal patterns modeled by the machine learning models.

These results open up multiple directions for further research. First, we will incorporate these findings in more sophisticated machine learning algorithms for short-term traffic prediction. For instance, the complex, nonlinear relationships influencing traffic flow may be modeled well using deep network architectures, especially when high-resolution input data is considered. We will also consider other datasets in order to generalize the findings reported in this paper based on data from a single highway. Second, we will build on the indicated ability to track nonrecurring congestion events in order 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 the smooth flow of traffic under a wide range of circumstances.

Data Availability

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

Conflicts of Interest

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 interest focus on the intersection of Intelligent Transportation Systems and Machine Learning. Dr. Mohan Sridharan (Ph.D.) 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), and a faculty member at Texas Tech University (USA) where he is currently an Adjunct Associate Professor of Mathematics and Statistics. He received his Ph.D. in Electrical and Computer Engineering from The University of Texas at Austin (USA). Dr Sridharan's primary research interests include knowledge representation and reasoning, interactive machine learning, cognitive systems, and computational vision, in the context of adaptive robots and agents. Dr. Prakash Ranjitkar (Ph.D., 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. He 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 interest 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, Prakash worked for the University of Delaware in USA (2006–2007) and before that in Hokkaido University in Japan (2001–2006). He 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.

Acknowledgments

The authors would like to thank Mike Duke from Auckland's Joint Transport Operations Centre (JTOC) for helping them obtain access to the data used for experimental evaluation in this paper.

References

[1] D. Schrank, B. Eisele, T. Lomax, and J. Bak, *Urban Mobility Scorecard*, International Transport Forum, Paris, France, 2015.

[2] M. S. Ahmed and A. R. Cook, "Analysis of freeway traffic time-series data by using Box-Jenkins techniques," *Transportation Research Record*, vol. 722, p. 116, 1979.

[3] B. L. Smith, B. M. Williams, and R. Keith Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, 2002.

[4] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal arima process: theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[5] 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*, vol. 19, no. 3, pp. 387–399, 2011.

[6] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: where we are and where we're going," *Transportation Research Part C: Emerging Technologies*, vol. 43, no. 3–19, 2014.

[7] D. Park, L. R. Rilett, B. J. Gajewski, C. H. Spiegelman, and C. Choi, "Identifying optimal data aggregation interval sizes for link and corridor travel time estimation and forecasting," *Transportation*, vol. 36, no. 1, pp. 77–95, 2009.

[8] M. S. Dougherty and M. R. Cobbett, "Short-term inter-urban traffic forecasts using neural networks," *International Journal of Forecasting*, vol. 13, no. 1, pp. 21–31, 1997.

[9] E. Vlahogianni and M. Karlaftis, "Temporal aggregation in traffic data: implications for statistical characteristics and model choice," *Transportation Letters*, vol. 3, no. 1, pp. 37–49, 2011.

[10] P. T. Martin, Y. Feng, and X. Wang, *Detector Technology Evaluation (MPC-03-154)*, International Transport Forum, Paris, France, 2003.

[11] Y. Kamarianakis and P. Prastacos, "Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1857, no. 1, pp. 74–84, 2003.

[12] S. R. Chandra and H. Al-Deek, "Predictions of freeway traffic speeds and volumes using vector autoregressive models," *Journal of Intelligent Transportation Systems*, vol. 13, no. 2, pp. 53–72, 2009.

[13] S. Yang, S. Shi, X. Hu, and M. Wang, "Spatiotemporal context awareness for urban traffic modeling and prediction: sparse representation based variable selection," *PLoS One*, vol. 10, no. 10, pp. 1–22, 2015.

[14] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[15] X. Ban, C. Guo, and G. Li, "Application of extreme learning machine on large scale traffic congestion prediction," in *Proceedings of ELM-2015 Volume 1: Theory, Algorithms and Applications (I)*, J. Cao, K. Mao, J. Wu, and A. Lendasse, Eds., pp. 293–305, Springer International Publishing, Cham, Switzerland, 2016.

[16] S. Dunne and B. Ghosh, "Weather adaptive traffic prediction using neurowavelet models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 370–379, 2013.

[17] S. Sun, R. Huang, and Y. Gao, “Network-scale traffic modeling and forecasting with graphical lasso and neural networks,” *Journal of Transportation Engineering*, vol. 138, no. 11, pp. 1358–1367, 2012.

[18] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, “Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach,” *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 211–234, 2005.

[19] J. Wang, I. Tsapakis, and C. Zhong, “A space-time delay neural network model for travel time prediction,” *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 145–160, 2016.

[20] M. T. Asif, J. Dauwels, C. Y. Goh et al., “Spatiotemporal patterns in large-scale traffic speed prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 794–804, 2014.

[21] M. Castro-Neto, Y.-S. Jeong, L. D. Han, D. Lee, and Han, “Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 6164–6173, 2009.

[22] A. Cheng, X. Jiang, Y. Li, C. Zhang, and H. 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*, vol. 466, pp. 422–434, 2016.

[23] Y.-S. Jeong, Y.-J. Byon, M. M. Castro-Neto, and S. M. Easa, “Supervised weighting-online learning algorithm for short-term traffic flow prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1700–1707, 2013.

[24] B. Yao, C. Chen, Q. Cao et al., “Short-term traffic speed prediction for an urban corridor,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 2, 2016.

[25] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, “A spatiotemporal correlative *k*-nearest neighbor model for short-term traffic multistep forecasting,” *Transportation Research Part C: Emerging Technologies*, vol. 62, pp. 21–34, 2016.

[26] 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*, vol. 6, no. 3, p. 292, 2012.

[27] G. A. Davis and N. L. Nihan, “Nonparametric regression and short-term freeway traffic forecasting,” *Journal of Transportation Engineering*, vol. 117, no. 2, pp. 178–188, 1991.

[28] S. Oh, Y.-J. Byon, and H. Yeo, “Improvement of search strategy with k-nearest neighbors approach for traffic state prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1146–1156, 2015.

[29] D. Xia, B. Wang, H. Li, Y. Li, and Z. Zhang, “A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting,” *Neurocomputing*, vol. 179, pp. 246–263, 2016.

[30] J. Guo, W. Huang, B. M. Williams, and Williams, “Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[31] I. Okutani and Y. J. Stephanedes, “Dynamic prediction of traffic volume through Kalman filtering theory,” *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.

[32] Y. Xie, Y. Zhang, and Z. Ye, “Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 326–334, 2007.

[33] B. Ghosh, B. Basu, and M. O’Mahony, “Bayesian time-series model for short-term traffic flow forecasting,” *Journal of Transportation Engineering*, vol. 133, no. 3, pp. 180–189, 2007.

[34] E. Horvitz, A. Johnson, S. Raman, and L. Liao, “Prediction, expectation, and surprise: methods, designs, and study of a deployed traffic forecasting service,” in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 1–10, Edinburgh, Scotland, July 2005.

[35] A. Pascale and M. Nicoli, “Adaptive bayesian network for traffic flow prediction,” in *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 177–180, IEEE, Nice, France, June 2011.

[36] B. Hamner, “Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow,” in *Proceedings of the IEEE International Conference On Data Mining, ICDM*, pp. 1357–1359, IEEE, Sydney, Australia, December 2010.

[37] N. Zarei, M. A. Ghayour, and S. Hashemi, “Road traffic prediction using context-aware random forest based on volatility nature of traffic flows,” in *Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 7802, pp. 196–205, Springer, Lecture Notes in Computer Science, Springer, 2013.

[38] V. J. Hodge, R. Krishnan, J. Austin, J. Polak, and T. Jackson, “Short-term prediction of traffic flow using a binary neural network,” *Neural Computing and Applications*, vol. 25, no. 7–8, pp. 1639–1655, 2014.

[39] W. Huang, G. Song, H. Hong, and K. Xie, “Deep architecture for traffic flow prediction: deep belief networks with multitask learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.

[40] R. Soua, A. Koesdwiady, and F. Karray, “Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory,” in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3195–3202, Vancouver, Canada, July 2016.

[41] 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*, vol. 16, no. 2, pp. 865–873, 2015.

[42] A. Ermagun and D. Levinson, “Spatiotemporal traffic forecasting: review and proposed directions,” *Transport Reviews*, vol. 38, no. 6, pp. 786–814, 2018.

[43] S. Dunne and B. Ghosh, “Regime-based short-term multivariate traffic condition forecasting algorithm,” *Journal of Transportation Engineering*, vol. 138, no. 4, pp. 455–466, 2011.

[44] G. Fusco, C. Colombaroni, and N. Isaenko, “Short-term speed predictions exploiting big data on large urban road networks,” *Transportation Research Part C: Emerging Technologies*, vol. 73, pp. 183–201, 2016.

[45] 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*, vol. 14, no. 3, pp. 1242–1254, 2013.

[46] S. Oh, Y.-J. Byon, K. Jang, and H. Yeo, “Short-term travel-time prediction on highway: a review of the data-driven approach,” *Transport Reviews*, vol. 35, no. 1, pp. 4–32, 2015.

[47] S. Oh, Y.-J. Byon, K. Jang, and H. Yeo, “Short-term travel-time prediction on highway: a review on model-based approach,” *KSCE Journal of Civil Engineering*, vol. 22, no. 1, pp. 298–310, 2018.

[48] Z. Jia, C. Chen, C. Ben, and P. Varaiya, “The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors,” in *Proceedings of the ITSC 2001. 2001*

IEEE intelligent transportation systems. Proceedings (Cat. No.01TH8585), pp. 536–541, Oakland, CA, USA, August 2001.

[49] P. Ryus, M. Vandehey, L. Elefteriadou, G. Richard Dowling, and K. Barbara Ostrom, *Highway Capacity Manual 2010: Number 273*, Transportation Research Board, Washington, DC, USA, 2010.

[50] D. Ghosh and A. Vogt, "Outliers: an evaluation of methodologies," *Joint Statistical Metings*, vol. 2012, pp. 3455–3460, 2012.

[51] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, pp. 1–13, Banff, Canada, April 2014.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1–9, New York, NY, USA, (NIPS 2012).

[53] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[54] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[55] D. O. Stram and W. W. S. Wei, "A methodological note on the disaggregation of time series totals," *Journal of Time Series Analysis*, vol. 7, no. 4, pp. 293–302, 1986.

[56] F. X. Diebold and R. S. Mariano, *Comparing Predictive Accuracy*, National Bureau of Economic Research Inc., Cambridge, MA, USA, 1994, https://ideas.repec.org/p/nbr/nberte/0169.html.

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

[58] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[59] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, Springer Series in Statistics, 2009.

[60] Y.-W. Chang and C.-J. Lin, "Feature ranking using linear SVM: causation and prediction challenge," in *Proceedings of the 3rd JMLR Workshop on Causality and Conference WCCI2008*, no. 2, pp. 53–64, London, UK, 2008.