Analysis and Prediction of the Queue Length for Non-Recurring Road Incidents

Banishree Ghosh
Nanyang Technological University
Singapore 639798
Email: banishre001@e.ntu.edu.sg

Justin Dauwels
Nanyang Technological University
Singapore 639798
Email: jdauwels@ntu.edu.sg

Ulrich Fastenrath
BMW Group
Parkring 19, 85748 Munich, Germany
Email: Ulrich.Fastenrath@bmw.de

***Abstract*—The queue length of an incident refers to the number of upstream links (i.e., links in the opposite direction of traffic flow) experiencing congestion due to the incident. In our work, we fuse incident records with traffic speed data from the expressways of Singapore for computing the queue length. Moreover, we propose a hybrid classification-regression model to predict the queue length of the incidents in real-time. At first, the model acts as a binary classifier. If the queue length of an incident is predicted to be higher than a predetermined threshold value, then the congestion is assumed to be impactful. Therefore, in the second step, the model performs regression analysis to predict the queue length of these incidents for fine-tuning. We also analyze the performance of different classification and regression methods in our work. In the classification step, all the methods perform almost equally well, and we can achieve $80\% - 96\%$ classification accuracy for different threshold values. However, in the regression step, Neural Network outperforms other methods. For the threshold value α = 250 m, the mean absolute percentage error is 65.78%, whereas for α = 1000 m the error value is 18.76%. Furthermore, we cluster the incidents to understand the underlying pattern of the external features and build separate regression models for each cluster of incidents.**

I. Introduction

Traffic incidents may have a significant adverse impact on human lives, either directly or indirectly. Moreover, traffic congestion caused by these incidents or crashes imposes enormous costs to the road users economically. Therefore, efficient traffic management is inevitable to mitigate the adverse effect of traffic incidents. Although the incident reports contain the details about the incidents, it is almost impossible to obtain the information about the impact of the mishaps from these reports. To this end, we perform a mathematical analysis to analyze the spread of congestion in the surrounding locality of the incidents. We compute the change in the speed-band to determine the impact on the neighboring links. Furthermore, the length of the queues due to the incidents vary widely from each other depending on several factors, such as day and time, types of roads, lanes affected, etc. In this paper, we implement a combination of classification and regression model to estimate the queue length based on these features.

A. Literature Review

Modeling and estimation of traffic congestion have been one of the most sought-after research areas in the field of urban transportation for the past years. In this subsection, we will briefly discuss related research works in this area.

The problem of traffic forecasting has been addressed from two different perspectives: (1) simulation (model) based and (2) data-driven based. Simulation models are built to simulate the network behavior and predict the future conditions [1]. These methods, however, suffer from exhaustive calibration and scalability [2]. On the contrary, data-driven models involve free-calibration. These methods include time-series analysis and machine learning techniques. A comprehensive overview of the data-driven methods is provided in [3]. In general, most of the studies consider the current and past traffic data from the particular link only where the incident happened to predict future values [4][5]. Additionally, some studies incorporate the information from neighboring links also [6], which improves the prediction performance of the models. Moreover, some of the literature aims to predict the impact qualitatively, whether it is severe or not. Hence, classification models have been implemented for prediction. In recent years several machine learning techniques, such as Decision Trees [7], Artificial Neural Networks (ANN) [8], etc. have been applied.

Chung *et al.* analyzed the spatiotemporal impact of incidents on the network using the time and location of the accident, and information from the loop detector. Therefore, this method is applicable for any road where accident data are available, and loop detectors are installed [9]. However, this approach has certain shortcomings. It can capture only the one-time impact of the incident. Usually, in real-time an incident is always followed by a growing/shrinking pattern of traffic. Hence, Pan *et al.* attempted to predict the impact of incidents in a time-varying spatial span [10]. Later, Yuye *et al.* predicted the impact of non-recurrent incidents from the city of Lyon, France [8]. However, they did not incorporate features such as lane number, time of the incident, whether shoulder is affected or not, etc. in their analysis, which we take into consideration in our work. Wang *et al.* applied a Naive Bayes (NB) classifier model to determine the probability of congestion and incidents in road networks [11]. However, they proposed a binary classifier which can not predict the exact queue length. Arthit *et al.* built a traffic congestion estimation model utilizing the data obtained from a single CCTV camera [12]. However, these cameras are set up at one end of the roads. Hence, it is not suitable for the long highways where velocity obtained from the CCTV camera may not be affected by the congestion at the other end. Therefore, there exist certain limitations in

the existing literature and we aim to improve these research gaps.

B. Our Contributions

Let us now briefly summarize our main contributions in this work:

1) We build a hybrid classification-regression model, which can provide an estimate of the congestion, both qualitatively as well as quantitatively.
2) We analyze the impact of the incidents for a maximum of 40 upstream neighboring links, where the length of each link varies in the range of $50 - 250$ m. Therefore, we take almost 4 km- 8 km in the upstream direction for each incident into consideration.
3) Also, we analyze the performance of our model for two different types of incidents, i.e., accidents and vehicle breakdowns and further provide a comparative analysis of the results.
4) Lastly, we compare the performance of different prediction algorithms in the classification-regression step.

The remainder of this paper is organized as follows. In the next section, we describe our dataset. In Section III, we discuss the approaches to our analysis and estimation of queue length, whereas we analyze the performance of our model in Section IV. Finally, Section V provides concluding remarks and ideas for future research.

II. Description and Analysis of the Data

The dataset considered in this study is provided by the Land Transport Authority (LTA) of Singapore. It consists of traffic speed data and historical records of incidents.

The historical traffic incidents data contain the following attributes: Type of incident (vehicle breakdown or accident), position (road-segment id, latitude & longitude), time (start-time and end-time in terms of month, date, hour and minute), condition of the shoulder lane, number of lanes affected and their types, name of the expressway, and direction along which the incident happened. The lanes are numbered from right to left as lane 1, 2, 3 and so on. Therefore, this serial number represents the type of lane according to its position. Singapore's entire road network has 11 expressways, which are divided into 2156 road segments for analysis. We consider the records of incidents for six months (Aug 2014−Jan 2015) on those expressways. As the incidents data mainly comprise of vehicle breakdowns and accidents (they cover almost 90% of all incidents), we consider only these two types in our analysis. There were in total 6200 vehicle breakdowns and 2528 accidents recorded in this period.

In this study, we construct a matrix where each of the features is represented by one or more than one columns. We convert the categorical variables (such as the type of affected lane or the affected expressway etc.) into binary variables by one-hot assignment. The structure of the feature matrix is shown in Table I.

Furthermore, the average traffic speed in each road segment of the expressways is recorded at every 5 minute interval.

TABLE I: The features extracted from incidents data.

Attribute	Feature	Feature type
Type of Day	weekday/weekend	Categorical
Day of the week	Monday, Tuesday, Wednesday. etc.	Categorical
Time of the day	Peak-hour/off-peak	Categorical
Expressway	PIE, AYE, ECP, etc.	Categorical
Direction along the expressway	eastward, westward, northward, southward	Categorical
Condition of shoulder	not affected, affected	Categorical
Total number of lanes	1, 2, 3, 4, 5	Ordinal
Number of affected lanes	0, 1, 2	Ordinal
Type of affected lane	1st, 2nd, 3rd, etc. (from extreme right)	Categorical

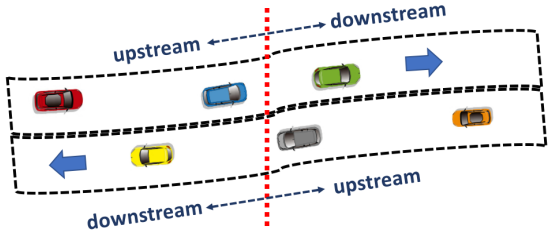

Fig. 1: Positions of upstream and downstream links.

The speed values, thus recorded, actually represent the average speed of all the vehicles that pass through that segment in this 5 minute span. In the datasets provided by LTA, the traffic speed values are discretized into ten speed-bands. While the first nine bands represent speeds up to 72 kmph (each spanning 8 kmph), the 10-th band corresponds to the speed greater than 72 kmph.

III. Approach

A. Determining the queue length

In this subsection, we describe the setup to analyze the spread of congestion due to the incidents. For this purpose, we first identify the upstream neighboring links for each incident. In Fig. 1 the positions of upstream links are shown for a bidirectional road. We take a maximum of 40 upstream links into consideration for each incident. Let's assume, the incident i happened on the third Monday in the month of Aug. Moreover, we suppose that the incident started at the time instant t_1 and ended at t_2 as recorded in the traffic incident report. For incident i, the set of upstream neighboring links is $u_i \in \{l_1, l_2, l_3, ...\}$, where l_1, l_2, l_3 etc. are individual links. From the dataset of traffic speed, we obtain the average speed in those upstream links for the span of $t_1 - t_2$ on the day of the incident. Furthermore, for every link in the set u_i, we compute the average traffic speed for the same time-span on the non-incident days. Let's assume, the average traffic speed in link l_1 at the time instant t_1 of the four Mondays in Aug is $s_{l_{1_{m_1}}}$, $s_{l_{1_{m_2}}}$, $s_{l_{1_{m_3}}}$, and $s_{l_{1_{m_4}}}$. We compute the average speed of the non-incident days (i.e. $s_{l_{1_{m_1}}}$, $s_{l_{1_{m_2}}}$, and $s_{l_{1_{m_4}}}$) and calculate the difference of this average and $s_{l_{1_{m_3}}}$ (as the incident happened on 3-rd Monday). Hence, for link l_1 at time instant t_1, the difference is

$$d_{l_{1_{t_1}}} = \frac{s_{l_{1_{m_1}}} + s_{l_{1_{m_2}}} + s_{l_{1_{m_4}}}}{3} - s_{l_{1_{m_3}}}, \quad (1)$$

In the similar way, we compute $d_{l_i} \in \{d_{l_{i_{t_1}}}, d_{l_{i_{t_1+1}}}, d_{l_{i_{t_1+2}}}, ...\}$ for the incident duration $t_1 - t_2$ for each individual upstream link. The average speed of the non-incident days is considered in this step to alleviate the impact of peak-hour or weekday congestion which may happen irrespective of any incident. Now, we assume the threshold of the difference in speed-band to be 1.5, i.e., if the average speed-band of the non-incident days exceeds the speed-band on the day of the incident by at least 1.5 (12 kmph), we consider that the link was congested because of the incident. Thus, we discard those links where the speed-band difference is below the threshold value and obtain the total number of affected links for each incident. Clearly, it is a subset u_{s_i} of u_i for incident i. Next, we sum up the lengths of the links of the subset u_{s_i} to obtain the queue length in meter for incident i. The queue lengths of the incidents, thus obtained, vary in the range of 50 m-3.5 km for the vehicle breakdowns and 40 m-6 km for the accidents. The histograms of the queue lengths are shown in Fig. 2 and Fig. 3 for the vehicle breakdowns and accidents, respectively. The mean and standard deviation of the queue lengths are 700 m and 501 m, respectively for vehicle breakdowns. On the other hand, the values for accidents are 1.67 km and 837 m, respectively.

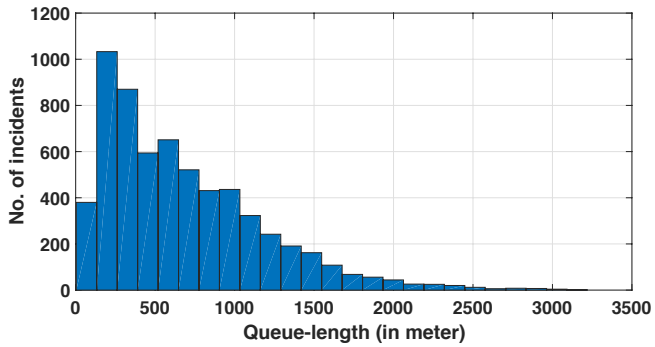

Fig. 2: Histogram of the queue length for vehicle breakdowns.

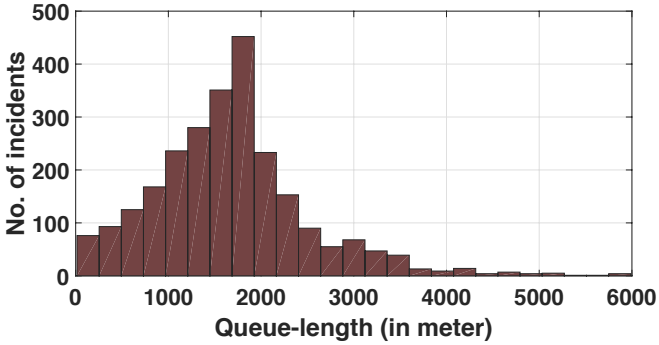

Fig. 3: Histogram of the queue length for accidents.

B. Prediction of the queue length

For prediction purpose, we introduce a combination of classification and regression model in our work. In the first step, we implement a fixed threshold based binary classifier, where if the queue length of an incident is higher than this predetermined threshold value, then the congestion is assumed to be impactful. Therefore, in the second step, we build regression models to predict the queue length of these incidents for more precise tuning. However, if the queue length is less than the threshold value, we may ignore those incidents. Therefore, in real-time at first, our system will classify the incident into one of the two classes depending upon the queue length. If it is forecast to have queue length higher than the threshold, the system will use regression models to estimate the exact value of queue length with greater accuracy.

For classification purpose, we employ three methods, such as Classification And Regression Tree (CART) [13], Support Vector Machine [5], and Treebagger [14]. Furthermore, we apply the regression methods, such as Classification And Regression Tree (CART) [13], Support Vector Regression [15], Linear Regression [16], Treebagger [14], Gaussian Process Regression (GPR) [17], and Neural Network [18] in the second step. The training and test datasets are selected by 3-fold cross-validation. We now briefly explain the methods below.

1) Classification And Regression Tree: It is a well-known method suitable for classification as well as regression. The trees are constructed by partitioning the dataset recurrently and fitting a model within each partition. Consequently, these partitions form a decision tree [19]. Each node of the tree corresponds a particular set of records T that is split by a specific test on a feature. For example, a split on a continuous attribute A can be induced by the test $A \leq x$. The set of records T is then partitioned into two subsets that lead to the left branch of the tree and the right one.

$$T_l = t \in T : t(A) \leq x \quad (2)$$

and

$$T_r = t \in T : t(A) > x. \quad (3)$$

Similarly, regression trees are constructed with numerical or categorical variables, and the output is always a continuous-valued function [19]. CART has certain advantages over other methods. It does not require feature selection beforehand. Also, it is suitable for both categorical as well as numerical variables. Last but not the least, this method is very robust to outliers [20].

2) Support Vector Machine and Support Vector Regression: Support vector machine (SVM) and Support Vector Regression (SVR) are supervised learning algorithms often used in data-driven prediction [15]. It is useful for solving problems of classification, regression, pattern recognition, etc., where the output is expressed as a function of a linear combination of kernel functions based on a subset of the training data, which are termed as support vectors [21]. The basic idea is to find the optimal hyper-plane $\mathbf{w}$ such that the input vector $x_i \in \mathbb{R}^n$ is mapped from low-dimensional space to high dimensional

space by a nonlinear mapping function $\phi(x)$ to get a linear decision function [15]:

$$f(x) = w^T.\phi(x_i) + b : b \in \mathbb{R}. \quad (4)$$

3) Linear Regression: Linear regression is a statistical approach which is used to model the relationship between one dependent variable and one or more independent variables [16]. If $X_1, X_2, ..., X_n$ are independent variables and Y is the dependent variable, then the equation of the linear regression line is given by:

$$Y = c + b_1.X_1 + b_2X_2 + ...b_nX_n, \quad (5)$$

where $b_1, b_2, ..., b_n$ are regression coefficients and c is a constant. The advantage of this method is that it is much simpler compared to other models.

4) Treebagger: Treebagger is one of the most commonly used Ensemble Methods in machine learning. Ensemble models are constructed by melding a set of classifiers, and the new objects are classified by taking a weighted value of their predictions [22]. Bootstrap aggregating, also abbreviated as bagging, trains each classifier in the ensemble parallelly using a randomly drawn subset of the training set and weighs each classifier equally. As an example, the random forest algorithm combines a set of random decision trees to achieve very high classification accuracy [23]. The main advantage of this method is that it can handle both categorical and numerical features well. Moreover, it is suitable for very high-dimensional data as well as a large training set.

5) Gaussian Process Regression: It is a supervised learning method which implements Gaussian processes for regression or probabilistic classification problem. In this method, a Gaussian function is fitted to the data points, and the predicted values are obtained by interpolation from this function. The advantage of this technique is that the prediction is probabilistic; therefore the predicted value corresponds to a confidence interval which can provide an estimate of the precision [24].

6) Neural Network: The Neural Network is an information processing system having similar working principle as the biological nervous system. It comprises a large number of highly interconnected components (same as neurons) working in unison to solve the problems. The network learns by examples through a learning process.

The process of training a neural network involves tuning the values of the weights and biases to optimize the network performance. We apply Bayesian Regularization technique [25] for this purpose that updates the weight and bias values according to Levenberg-Marquardt optimization [25]. It optimizes the performance by minimizing the squared errors and weights of the network.

To analyze the performance of our predictive model, we compute three different metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N}}, \quad (6)$$

$$\text{MAE} = \frac{\sum_{i=1}^{N} |e_i|}{N}, \quad (7)$$

$$\text{MAPE} = \frac{100}{N}.\sum_{i=1}^{N} |\frac{e_i}{q_i}|. \quad (8)$$

where N is the total number of incidents, and e_i is the error between the actual and predicted queue length q_i and $\hat{q}_i$ respectively:

$$e_i = q_i - \hat{q}_i. \quad (9)$$

C. Clustering of the Incidents

We also investigate whether there are certain patterns in the external features which might be necessary for prediction of the queue lengths. Hence, we apply Gaussian Mixture Model [26] for clustering the incidents based on their features, and design regression models for each cluster separately.

The Gaussian Mixture Model is a probabilistic method which assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [26]. It implements the expectation-maximization algorithm [27] for fitting the distributions.

IV. Queue Length Prediction

In this section, we analyze the performance of our cascaded classification-regression model in predicting the queue length of the vehicle breakdowns and accidents. The flowchart of our proposed strategy is shown in Fig. 4.

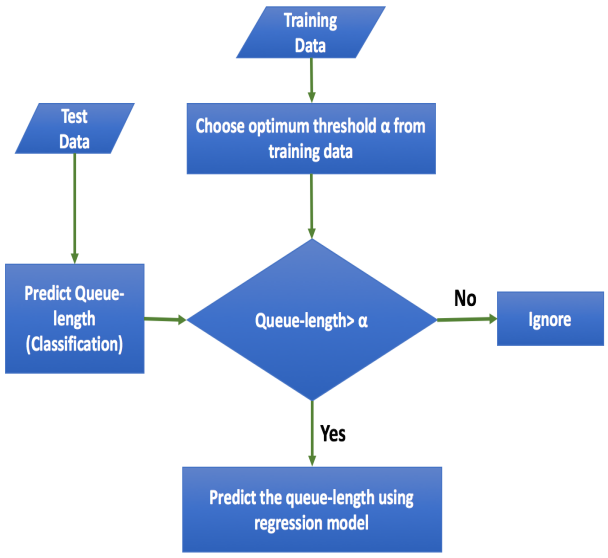

Fig. 4: Flow-chart of our prediction model.

A. Variation of prediction error for different threshold values

At first, we show the results achieved solely by regression model in Table II, without classifying the incidents before the regression step (i.e., $\alpha = 0$). We mention the results obtained by the Treebagger method only.

Now, we vary the value of the threshold α, such as 250 m, 500 m, 750 m and 1000 m, and analyze the variation in

TABLE II: RMSE, MAE and MAPE values obtained by regression method only.

	RMSE (meter)	MAE (meter)	MAPE (%)
Breakdowns	647	503	150
Accidents	954	775.8	85.56

the errors. We show the distributions of the RMSE and MAE values obtained by Treebagger method in Fig. 5 for different values of α.

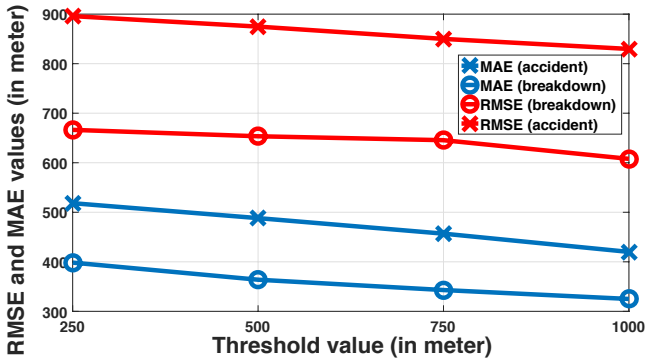

Fig. 5: Distribution of RMSE and MAE values obtained by Treebagger method for different thresholds.

We observe in Fig. 5 that in general, both RMSE and MAE values are higher for accidents. It is because the range of the queue lengths for accidents (see Fig. 3) is much higher compared to that of breakdowns (See Fig. 2). We show the distribution of absolute values of the errors with actual queue length in Fig. 6 and Fig. 7 to explain it more elaborately. We

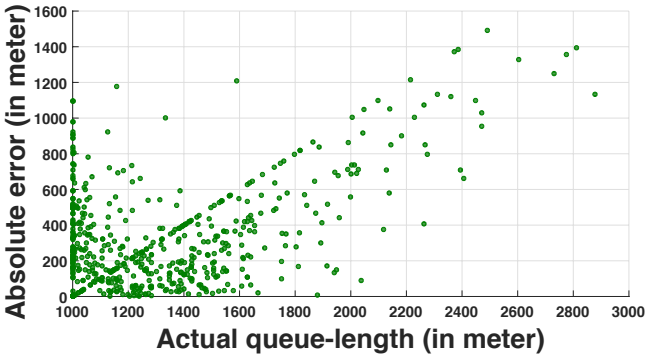

Fig. 6: Variation of absolute error values with queue length for vehicle breakdowns.

observe that the error increases with queue length in Fig. 6 and Fig. 7 which implies that the error is positively correlated with queue length. Therefore, it becomes intuitively evident why the maximum spread of errors is much higher (up to $3000 - 4000$ m) for the accidents due to their overall larger spread of queue lengths, while that of breakdowns is much smaller (up to 1500 m) due to their lower overall range of queue lengths.

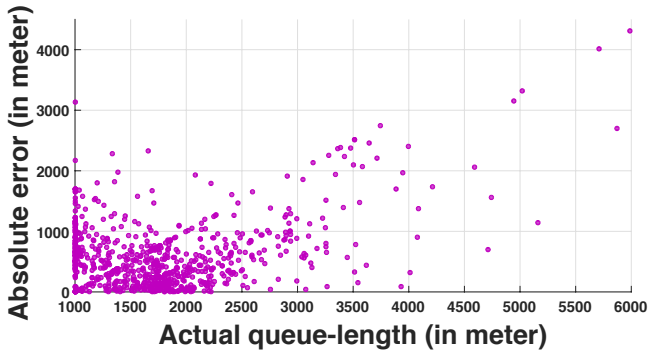

Fig. 7: Variation of absolute error values with queue length for accidents.

Furthermore, the distribution of MAPE values is shown in Fig. 8 for the same values of α. We observe in Fig. 8 that

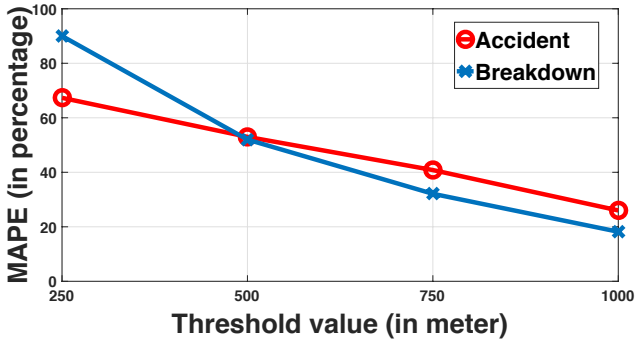

Fig. 8: Distribution of MAPE values obtained by Treebagger method for different thresholds.

the MAPE value is higher for breakdowns at lower α. This behavior of the MAPE curves can be explained by Fig. 2 and Fig. 3. We find in Fig. 2 that most of the breakdowns have queue length in the range of $300-400$ m. Therefore, when the value of α is 250 m, the higher density of the breakdowns in the vicinity of the decision boundary brings about an increased probability of misclassification. Consequently, the high MAPE value for the breakdowns comes from low classification accuracy for $\alpha = 250$ m. Conversely, the peak of the histogram in Fig. 3 exists farther from the classification threshold (in the vicinity of 2000 m). Therefore, it does not suffer from the similar problem of misclassification for $\alpha = 250$ m.

If we compare the results obtained by the cascaded model (Fig. 5 and Fig. 8) with that of the conventional regression-only model (Table II), we find that our proposed approach outperforms the traditional model regarding RMSE, MAE and MAPE values. It validates our choice of classification step before regression modeling.

Finally, we show the distributions of the RMSE, MAE and MAPE values averaged across all incidents obtained by Treebagger method in Fig. 9 for different values of α.

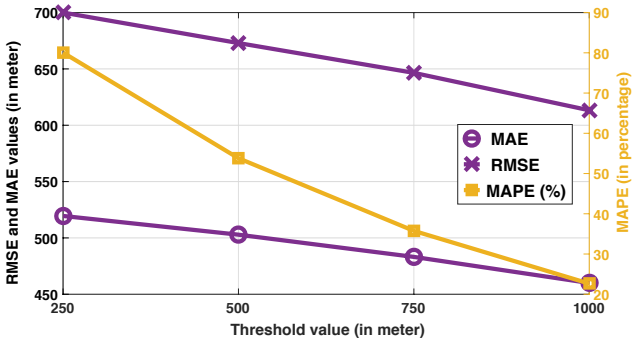

Fig. 9: Distribution of RMSE, MAE and MAPE values for all incidents obtained by Treebagger method for different thresholds.

Last but not the least, we see from Fig. 5 and Fig. 8 that the values of the errors decrease with increase in the threshold α. However, with increasing α the precision of the model is compromised because we discard those incidents for which the queue length is predicted to be less than α. Therefore, for higher α we fail to predict the exact queue length of a larger proportion of incidents. Hence, there is a clear trade-off between prediction accuracy and the range of precise operation of the model.

B. Performance of classification and regression methods

In this subsection, we analyze the performance of various algorithms in both classification and regression stages for predicting the queue length of the incidents. First, we show the classification accuracies obtained by three methods for different threshold values α in Table III.

TABLE III: Classification accuracies (in percentage) for all incidents in Singapore.

	CART	SVM	Treebagger
α = 250 m	96.09	95.13	96.78
α = 500 m	91.81	92.05	90.75
α = 750 m	84.29	85.17	85.03
α = 1000 m	79.56	76.24	78.15

We find that the methods achieve similar classification accuracy for a particular value of α. Also, classification accuracy decreases with increase in α. Moreover, in Table IV−VII we provide the prediction results obtained by six different regression methods averaged across all vehicle breakdowns and accidents together, for various threshold values. We

TABLE IV: RMSE, MAE and MAPE values for α = 250 m averaged across all incidents in Singapore.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	711.5	651.67	675.2	696.5	668	637
MAE (meter)	526.6	498	518.1	522.85	519	481.5
MAPE (%)	87.27	69.15	72.3	81.5	74.8	65.78

observe that Neural Network outperforms other methods.

TABLE V: RMSE, MAE and MAPE values for α = 500 m averaged across all incidents in Singapore.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	686	591.4	672	673	616.8	576
MAE (meter)	522.2	445.2	497.5	507	459	431
MAPE (%)	55.1	50.23	54.88	53.4	51.7	48.6

TABLE VI: RMSE, MAE and MAPE values for α = 750 m averaged across all incidents in Singapore.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	655.15	604.25	615.8	643	647.8	578.8
MAE (meter)	497.2	456.5	459.7	483.8	478.3	425.4
MAPE (%)	40.94	34.17	34.61	35.89	37.55	33.43

TABLE VII: RMSE, MAE and MAPE values for α = 1000 m averaged across all incidents in Singapore.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	633	590	590.9	613	600	548
MAE (meter)	482	430	460	454	447	395
MAPE (%)	25.3	19.8	22.35	21.9	22.64	18.76

Furthermore, the RMSE, MAE and MAPE values obtained by six regression methods for the threshold α = 1000 m are mentioned in Table VIII and Table IX for vehicle breakdowns and accidents, separately.

TABLE VIII: RMSE, MAE and MAPE values for vehicle breakdowns at α = 1000 m.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	466	414.6	429	425	415	379.39
MAE (meter)	347	305.6	326.7	325.5	320.4	287.6
MAPE (%)	20.4	16.39	19.09	18.18	19	15.32

TABLE IX: RMSE, MAE and MAPE values for accidents at α = 1000 m.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	827	768.2	775.8	757	759	695.41
MAE (meter)	608.5	550	586.7	567	573	504.62
MAPE (%)	30.45	27.5	27.85	26.1	27.73	24.94

C. Performance of cluster-specific models

In this subsection, we aim to find the underlying patterns in the external features of the incidents. Hence, we group the incidents with common latent similarities together. Also, we analyze the performance of the cluster-specific regression models. We apply Gaussian Mixture Model to cluster the incidents, and we obtain that the optimum number of clusters are 2 and 3 for vehicle breakdowns and accidents, respectively. Now we implement Principal Component Analysis (PCA) [28] to see the important features. In Fig. 10 and Fig. 11 we visualize the clusters along two principal components obtained by PCA for vehicle breakdowns and accidents, respectively.

We observe that the clusters are well distant from each other. They are indicated by different colors in each case.

We find in Fig. 10 and Fig. 11 that the incidents are separated along the x-axis. Therefore, the features contributing

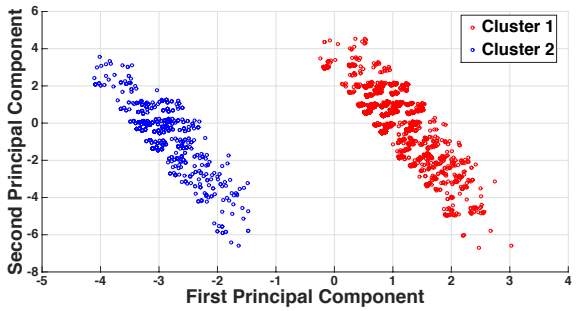

Fig. 10: Clusters of vehicle breakdowns in the two principal dimensions obtained by PCA.

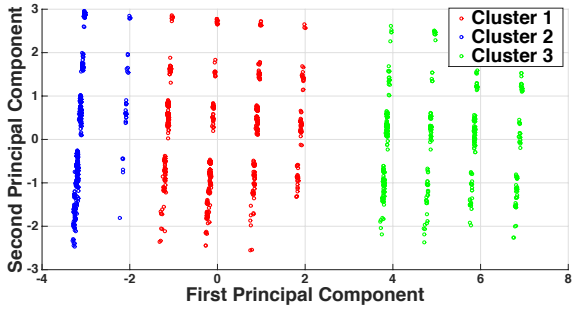

Fig. 11: Clusters of accidents in the two principal dimensions obtained by PCA.

the most to the first principal component are significant while clustering the incidents. The first principal components for vehicle breakdowns and accidents are tabulated in Table X and Table XI respectively, where the proportions of only five most significant features are specified.

TABLE X: The first Principal Component (PC) for the vehicle breakdowns.

	type of day	day of week	time of day	shoulder lane	number of lanes affected
PC1	0.54	0.26	0.11	0.1	0.09

TABLE XI: The first Principal Component (PC) for the accidents.

	type of day	time of day	number of lanes affected	total number of lanes	type of lanes affected
PC1	0.52	0.25	0.16	0.07	0.06

In Fig. 10, the left (blue) cluster contains the breakdowns happened on the weekdays. Conversely, the breakdowns in the right (red) cluster occurred on weekends.

In Fig. 11, the rightmost (green) cluster contains accidents occurred on the weekends and weekday off-peak hours. However, the red and blue clusters contain incidents of the weekday peak hours. The middle (red) cluster is comprised of the incidents having less than two lanes affected, whereas for the incidents in the leftmost (blue) cluster, on an average, four-five lanes are affected.

Furthermore, the RMSE, MAE and MAPE values obtained by clustering followed by regression methods are shown in Table XII and Table XIII for vehicle breakdowns and accidents, respectively. The overall RMSE, MAE and MAPE values have improved for the cluster-specific models, both for accidents and vehicle breakdowns. Therefore, we can infer that clustering helps to reduce the prediction error.

TABLE XII: RMSE, MAE and MAPE values for vehicle breakdowns at α = 1000 m.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	447.5	401	413.8	411.6	398.	365.5
MAE (meter)	328	291	317.3	313.8	313.2	274.5
MAPE (%)	19	15.32	17.68	17.65	18.2	15.1

TABLE XIII: RMSE, MAE and MAPE values for accidents at α = 1000 m.

	CART	SVR	Linear Regression	Treebagger	Gaussian Process Regression	Neural Network
RMSE (meter)	796	745.2	759	734.2	741	690.3
MAE (meter)	600	532.2	575	556	560	498
MAPE (%)	27.68	25.31	25.7	24.2	25.98	23.89

V. Conclusion

In this paper, we aimed to compute the queue length of traffic incidents. To this end, we considered traffic speed data and incidents record from the expressways of Singapore. Moreover, we introduced a combined classification-regression model to predict the queue length of the incidents. We found that our hybrid model performs well in prediction (MAPE around 20% for α = 1000 m). However, our work has certain limitations as well. There exists a trade-off between prediction accuracy and precision range. Moreover, as we do not have continuous-valued speed data from the expressways of Singapore, the queue lengths of the incidents obtained using the speed-band data in our work may not be exactly accurate. For example, speed-band 4 corresponds to the traffic speed of $25-32$ kmph, whereas speed-band 5 covers the average speed of $33-40$ kmph. The difference in 25 kmph and 40 kmph is certainly not equivalent to the difference in 32 kmph and 33 kmph, even though both of them indicate the same difference in speed-band. Hence, the change in traffic speed due to the incidents may not be accurately reflected in the speed-band data. Therefore, in future, we plan to do the similar analysis for other cities, where exact values of speed are available in the traffic dataset.

Acknowledgment

The authors wish to thank the Land Transport Authority of Singapore for providing the incidents and traffic data from the expressways of Singapore.

REFERENCES

[1] M. P. Miska, *Microscopic online simulation for real-time traffic management*. TU Delft, Delft University of Technology, 2007.

[2] M. Bierlaire and F. Crittin, "An efficient algorithm for real-time estimation and prediction of dynamic od tables," *Operations Research*, vol. 52, no. 1, pp. 116–127, 2004.

[3] M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 2, pp. 871–882, 2013.

[4] C. Quek, M. Pasquier, and B. B. S. Lim, "Pop-traffic: a novel fuzzy neural approach to road traffic analysis and prediction," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 7, no. 2, pp. 133–146, 2006.

[5] L. Vanajakshi and L. R. Rilett, "A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed," in *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 2004, pp. 194–199.

[6] 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.

[7] L. Ruimin, Z. Xiaoqiang, Y. Xinxin, L. Junwei, C. Nan, and Z. Jie, "Incident duration model on urban freeways using three different algorithms of decision tree," in *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, vol. 2. IEEE, 2010, pp. 526–528.

[8] Y. He, S. Blandin, L. Wynter, and B. Trager, "Analysis and real-time prediction of local incident impact on transportation networks," in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 2014, pp. 158–166.

[9] Y. Chung and W. W. Recker, "A methodological approach for estimating temporal and spatial extent of delays caused by freeway accidents," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 3, pp. 1454–1461, 2012.

[10] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, "Forecasting spatiotemporal impact of traffic incidents on road networks," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 587–596.

[11] G. Wang and J. Kim, "The prediction of traffic congestion and incident on urban road networks using naive bayes classifier," in *Australasian Transport Research Forum (ATRF), 38th, 2016, Melbourne, Victoria, Australia*, 2016.

[12] A. Buranasing and C. Khemapatapan, "Traffic congestion estimation for short-highway in pre-timed systems," *International Journal of Modeling and Optimization*, vol. 3, no. 2, p. 158, 2013.

[13] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[14] H. L. Chang and T. P. Chang, "Prediction of freeway incident duration based on classification tree analysis," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 10, pp. 1964–1977, 2013.

[15] W. Wu, S.-y. Chen, and C.-j. Zheng, "Traffic incident duration prediction based on support vector regression," *Proceedings of the ICCTP*, pp. 2412–2421, 2011.

[16] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran, "Short term traffic forecasting using the local linear regression model," in *82nd Annual Meeting of the Transportation Research Board, Washington, DC*, 2003.

[17] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[18] W. H. Delashmit and M. T. Manry, "Recent developments in multilayer perceptron neural networks," in *Proceedings of the seventh Annual Memphis Area Engineering and Science Conference, MAESC*, 2005.

[19] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[20] R. Timofeev, "Classification and regression trees (cart) theory and applications," Ph.D. dissertation, Humboldt University, Berlin, 2004.

[21] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 131–159, 2002.

[22] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple classifier systems*. Springer, 2000, pp. 1–15.

[23] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[24] T. Idé and S. Kato, "Travel-time prediction using gaussian process regression: A trajectory-based approach," in *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM, 2009, pp. 1185–1196.

[25] M. Kayri, "Predictive abilities of bayesian regularization and levenberg–marquardt algorithms in artificial neural networks: A comparative empirical study on social data," *Mathematical and Computational Applications*, vol. 21, no. 2, p. 20, 2016.

[26] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 7, pp. 719–725, 2000.

[27] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[28] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*. Springer, 1986, pp. 115–128.