

# Dynamic Prediction of the Incident Duration Using Adaptive Feature-set

Banishree Ghosh<sup>1</sup>, Muhammad Tayyab Asif<sup>2</sup>, Justin Dauwels<sup>3</sup>, Ulrich Fastenrath<sup>4</sup>, Hongliang Guo<sup>5</sup>

**Abstract**—Non-recurring incidents such as accidents, vehicle breakdowns, etc. are leading causes of severe traffic congestions in large cities. Consequently, anticipating the duration of such events in advance can be highly useful in mitigating the resultant congestion. However, availability of partial information or ever-changing ground conditions makes the task of forecasting the duration particularly challenging. In this work, we propose an adaptive ensemble model that can provide reasonable forecasts even when a limited amount of information is available and further improves the prediction accuracy as more information becomes available during the course of the incidents. Furthermore, we consider the scenarios where the historical incident reports may not always contain accurate information about the duration of the incidents. To mitigate this issue, we first quantify the effective duration of the incidents by looking for the change points in traffic state and then utilize this information to predict the duration of the incidents. We compare the prediction performance of different traditional regression methods, and the experimental results show that the Treebagger outperforms other methods. For the incidents with duration in the range of 36–200 minutes, the mean absolute percentage error (MAPE) in predicting the duration is in the range of 25%–55%. Moreover, for the longer duration incidents (greater than 65 minutes), prediction improves significantly with time. For example, the MAPE value varies over time from 76% to 50% for the incidents having duration greater than 200 minutes. Finally, the overall MAPE value averaged over all incidents improves by 50% with elapsed time for prediction of reported as well as effective duration.

**Index Terms**—Incident duration prediction, traffic data, reported and effective duration, sequential prediction

## I. INTRODUCTION

Traffic incidents can severely disrupt the flow of traffic in already congested large metropolitans. The uncertainty in forecasting the impact of traffic incidents arises due to the challenges associated with predicting the duration of that particular event. Accurate forecast about the duration of such events can prove to be highly invaluable for traffic management authorities, logistics and taxi companies, as well as for motorists traveling in that area. The duration of these incidents turns out to be a major factor in determining their impact,

which in turn, depends on several factors such as day, time and location of the incident, the number of total and affected lanes, type of the incident, etc. Furthermore, we leverage traffic data (speed and flow) with these features to compute the predicted duration.

The overall duration of a traffic incident can be divided into the following different components: (1) reporting time ( $r_t$ ): the time taken to detect, verify and report the incident after its occurrence, (2) response time ( $s_t$ ): the time taken by the response team to arrive at the spot after reporting of the incident, (3) clearance time ( $c_t$ ): the time required by the team to clear the affected area, and (4) recovery time ( $v_t$ ): the time taken by the traffic condition to restore back to normal [1]. Since the response time and clearance time mostly depend on the timely actions of the response team, it is quite interesting to predict the span of these two stages from the perspective of the traffic management authorities [2]. Nonetheless, prediction of the entire incident duration is more useful to the drivers since the impact of the incident exists in the road network all over the four stages [3]. Therefore, considering the duration from the very beginning of the incident to its end is highly beneficial for driver advisory systems so that they can provide more robust guidance to the drivers based on the predictions. Hence, the total incident duration  $T$  considered in this work is the sum of these four stages [4]:

$$T = r_t + s_t + c_t + v_t. \quad (1)$$

In the following, we briefly review related studies on the forecasting of traffic incident duration.

### A. Literature Review

Incidents duration prediction has remained an active research problem for the last two decades in the area of transportation studies. We will now briefly discuss these studies, and elaborate on their strengths and limitations. In one of the earlier works, Golob *et al.* built log-normal models of incident duration by considering more than 9000 accidents caused by trucks in 2 years that happened on freeways in Los Angeles [5]. They studied a few features like collision type, lane closures, accident severity, etc. in the analysis. Later, Jones *et al.* incorporated the idea of conditional probability in incident duration prediction. They considered 2156 incidents from the Seattle metropolitan area and found that the length of incidents can be better approximated by a log-logistic distribution instead of a log-normal distribution [6]. Chung *et al.* also applied a log-logistic model to fit the incident data from Korean Freeways [7]. Hojati *et al.* implemented

<sup>1</sup>Banishree Ghosh is with the Energy Research Institute, Interdisciplinary Graduate School, Nanyang Technological University, Singapore.

<sup>2</sup>Muhammad Tayyab Asif is with Amazon. He was with BNY Mellon and NTU when this work was done.

<sup>3</sup>Justin Dauwels is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

<sup>4</sup>Ulrich Fastenrath is with the Traffic Information Management and Routing Optimization Department, BMW Group, Parkring 19, 85748 Munich, Germany.

<sup>5</sup>Hongliang Guo is with the Center for Robotics, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China.

Weibull and log-logistic distributions with fixed as well as random parameters. They incorporated random parameters in their analysis to consider the effect of unobserved or hidden features on incident duration prediction [3]. However, these studies considered individual predictive models. Moreover, these works mostly developed static models which did not incorporate real-time traffic conditions. Ruimin *et al.* fitted different statistical models for each stage of the traffic incident (e.g., dispatch time, clearance time, etc.) [8]. In another study, the same author built a mixture model combining the distributions, namely, generalized gamma, Weibull, and log-logistic for different incident clearance methods [9] and performed the incident duration prediction. However, in practice, the information about the distinct stages of traffic incidents may not always be available. They also did not present the results obtained at different time-instants by sequential prediction. Giuliano *et al.* analyzed various attributes associated with the frequency, pattern of occurrence and duration of the incidents. They showed that incident duration mostly depends on the type (injury accidents, non-injury accidents, and others), lane closures (no lane closed, one or more lanes closed), and time of the day (day, night). They used Analysis of Variance (ANOVA) [10] to estimate models of incidents duration based on those features. This particular method was selected because the independent variables are broadly categorical. They also performed Kruskal-Wallis test [11] to prove that overall differences between the categories are significant [12]. Nam *et al.* designed three separate statistical models for reporting time, response time and clearance time respectively using maximum likelihood estimation [13].

In the recent years, various data-driven modeling algorithms, such as Support Vector Regression (SVR) [14], Artificial Neural Networks (ANN) [15], etc. have also been applied for predicting the duration of traffic incidents. In this study, we apply various regression methods to analyze different types of incidents in Singapore. Pereira *et al.* also used several techniques to study the incidents in Singapore [16], however, they did not consider traffic data in their feature-set. He *et al.* studied the relationship of various external factors with the duration of the incidents [1]. They defined the total duration to consist of the first three phases, excluding the recovery time. However, they considered the average values of traffic data of two instants only; before and after the incident detection. Hence, the prediction has been performed only once at the beginning of the incident. Furthermore, Wu *et al.* analyzed incidents dataset from Utrecht in the Netherlands, however, they did not include important features like information about the blockage of lanes or carriageways in the feature-set [14]. Although, Yuye *et al.* analyzed real-time traffic data for predicting the impact of incidents, they considered traffic speed only as a categorical feature [15]. Moreover, they performed only one-time prediction of the occupancy in the roads. Lin *et al.* introduced a combination of MSP tree and hazard-based duration model for predicting incident durations [17]. The MSP builds a tree-based model and Hazard Based Duration Models (HBDM) construct the leaves of this tree. However, they did not include traffic data in their study.

Some studies have analyzed the impact of incidents by

incorporating information from traffic data albeit in a limited manner. Shefar *et al.* used the ratio of volume and capacity as a measurement of congestion [18]. Ceder *et al.* derived a relation between traffic congestion and hourly traffic volume [19]. In general, most of the studies assumed traffic flow to be the most important metric for estimating the traffic congestion. Golob *et al.* fitted linear regression models between traffic flow and the incidents in their analysis [20]. Skorput *et al.* built a mathematical model to detect the incidents based on the inward and outward traffic flow [21]. Khattak *et al.* introduced an online tool named iMiT which can dynamically predict incident durations, secondary incident occurrence and associated delays [22]. For this purpose, they worked with the traffic flow data available from the Hampton Roads Areas. However, the resolution of their flow data is 15 minutes, whereas we have the traffic flow values recorded at each 5-minute interval. Moreover, they relied on the incident report provided by HRTOC for their analysis. Therefore, the reported incident durations may incur a measurement error, since there might be a delay in reporting the incidents by HRTOC after they occurred. Recently, Ma *et al.* conducted a research on prediction of the freeway incidents clearance time in USA [2]. Although they obtained competitive results using the Gradient Boosting Decision Trees method, their approach has certain shortcomings. For instance, the resolution of the traffic volume data used in their study is too coarse (1 hour), which is not sufficient to understand the subtle changes in the impact of incidents accurately. Also, they did not perform the sequential prediction with time. In this study, we not only analyze the variations in upstream and downstream traffic flows but also consider speed as a covariate to model the impact of incidents in a more comprehensive manner. Last but not the least, Wei *et al.* performed the sequential prediction of incident duration using traffic data of fine resolution [23]. Although their feature set covers all important information pertaining to incident duration prediction, the size of their training and test dataset is small; their dataset comprises 24 incidents from a small part of an expressway in Taiwan, whereas we consider in this paper 11,278 incidents from all expressways along with 1,745 ramp incidents of the Singapore island. Moreover, the feature set in the approach of [23] is fixed, whereas we propose a method that can flexibly incorporate features, as incident and traffic data gradually becomes available. Besides, they did not report the error values for different classes of incidents separately, whereas we provide such information in this paper. Lastly, they used Artificial Neural Network model only, whereas we compare various regression methods in this paper.

Apart from that, a number of studies applied the deep learning method for network traffic speed prediction of late. For example, Jia *et al.* built a Deep Belief Network (DBN) model in their work to perform the prediction [24]. Similarly, Ma *et al.* proposed a deep Convolutional Neural Network-based model (CNN) for traffic speed prediction [25]. However, these latest neural network models require a large amount of training data which may not be available for the incidents since these are comparatively rare events. In the previous studies, deep learning methods have been used for predicting traffic under normal conditions because large amount of data is

available in those cases. Moreover, CNN operates on the dense representation of data, such as network wide traffic, whereas typical accident data is highly sparse. Hence, we opted for traditional neural network approach in our study.

Adaptive network-based fuzzy inference system (ANFIS) have been considered to perform efficiently of late since it combines the efficiency of the fuzzy model in handling the uncertainty with the prediction accuracy of artificial neural network method using back-propagation [26]. For example, Tang *et al.* predicted the lane changing behaviour of the drivers using this model [27]. Besides, Eleni *et al.* proposed an approach to predict the incidents duration based on Fuzzy-Entropy Neural Network [28]. They considered the uncertainties arising from late clearance of the incidents or possibility of the occurrence of secondary incidents, therefore the fuzzy model is more suitable for them. However, they did not mention the error values of incidents duration prediction in their study.

In our earlier work [29], our primary goal was to group the incidents first through common latent similarities among them and then training the regression models for each cluster separately, in order to reduce the prediction error. However, we did not include traffic data in the feature set, and did not perform continuous prediction of the incident duration. Moreover, unlike the previous study, we do not apply Principal Component Analysis [30] for the purpose of feature selection in this work; instead we include temporal dynamics, as information about the traffic incident becomes available over time. Therefore, it is more suitable for practical implementation. Apart from that, we verify the incident durations reported by LTA, by analyzing the traffic speed and flow data from the upstream neighboring links. Next, we build separate regression models for predicting the reported duration and effective duration, whereas in the previous work we relied on the reported durations only. The previous studies related to incidents duration prediction are summarized in Table I.

TABLE I: Summary of the previous studies.

Literature	Use of traffic data	Prediction model
Khattak <i>et al.</i> [22]	Flow data with the resolution of 15 minutes	Dynamic
Qing <i>et al.</i> [1]	Traffic data before and after the incident detection	Static
Wu <i>et al.</i> [14]	Not used	Static
Pereira <i>et al.</i> [16]	Not used	Dynamic
Yuye <i>et al.</i> [15]	Traffic speed as a categorical feature	Static
Lin <i>et al.</i> [17]	Not used	Static
Ma <i>et al.</i> [2]	Hourly traffic volume	Static
Ghosh <i>et al.</i> [29]	Not used	Static
Wei <i>et al.</i> [23]	Speed and flow data of 5 minutes resolution	Dynamic
Our work	Speed and flow data of 5 minutes resolution	Dynamic

## B. Our Contributions

In this subsection, we summarize our main contributions which address the research gaps in the existing studies:

- 1) Most of the previous studies did not incorporate the continuous traffic speed and flow data in their analysis. Moreover, the proposed models did not include any mechanism to provide updated estimates based on real-time streaming data. Conversely, apart from the spatial, temporal and geographical features, we consider the traffic data of high

resolution in our analysis. Also, the prediction is performed sequentially with elapsed time over the entire span of the incident.

- 2) Our model is more befitting for practical implementation since in practice, our feature-set can be adjusted at different time instant based on the availability of the features. Moreover, our system is proposed to work even when multiple features are missing. It is likely that for some incidents important features are missing for the entire duration of the incidents. However, our model will still perform the prediction with the available information only.
- 3) The existing studies mostly relied on the incident reports provided by the external sources, which may not be 100% accurate always. Sometimes, these reports incur a delay in reporting the incidents after they occur. By contrast, we perform the prediction of both reported incident duration (according to the incident reports) and effective incident duration (computed based on the impact of the incidents in the neighboring upstream links) and provide a comparative analysis.

The remainder of this paper is organized as follows. In the next section, we describe our dataset. In Section III, we discuss the approaches and prediction methods applied in our analysis. Furthermore, we provide a comparative analysis of the effective and reported durations of the incidents in Section IV and discuss the prediction performance of different regression methods in Section V. In the following section, we demonstrate the variation in prediction performance of the ensemble model depending on different feature-sets. Moreover, we analyze the performance of our model in predicting the reported duration and effective duration more elaborately for different categories of incidents in Section VII and Section VIII, respectively. Furthermore, we consider the entrance/exit ramp incidents in our study and evaluate the prediction performance of our model for these incidents in Section IX. Next, in section X, we compare the performance of our prediction model with the previous studies. Finally, Section XI provides concluding remarks and ideas for future research.

## II. DESCRIPTION OF THE DATA

In this section, we describe our dataset which consists of the historical records of incidents and traffic data from the expressways and highway ramps of Singapore. The entire dataset is provided by the Land Transport Authority (LTA) of Singapore.

There were 11,278 incidents recorded on the expressways over the span of six months (Aug 2016–Jan 2017). Moreover, we also consider 1,745 incidents collected from 197 entrance/exit ramps of the expressways. We have a variety of features in our data, such as spatial features (road-segment or link id, latitude & longitude, the expressway and direction), temporal features (the time instants when the incident started and ended), incident features (type of incident) and geographical features (the status of the adjacent lanes including shoulder lane). In Singapore, the lanes are numbered from right to left as lane 1, 2, 3, etc. There are 10 expressways in the

entire network of Singapore comprising 2156 road segments or links. The median length of the links in Singapore is 100 meters. We show the histogram of length of all the links from the expressways of Singapore in Fig. 1 and find that the

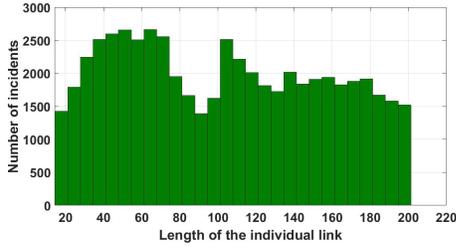


Fig. 1: Histogram of length of all the links from the expressways of Singapore.

maximum length is 202 meter. This high-resolution partition of road network allows us to avoid issues of multi-modal speed distributions within individual link [31].

In this study, we construct a matrix of the features which are either categorical or numerical. In our analysis, the categorical variables are nominal because no ordinal relation exists between different labels of the features. Therefore, if we replace the labels by different integers with a natural ordering, it may lead to poor performance of the model. Hence, we convert the categorical features (such as weekday/weekend, the direction, etc.) into binary ones using one-hot assignment method. The structure of the feature matrix is shown in Table V.

TABLE II: Features extracted from the incidents data.

Attribute	Feature	Feature type
Type of Day ( $d$ )	weekday/weekend	Categorical
Day of the week ( $w$ )	Monday, Tuesday, Wednesday, etc.	Categorical
Time of the day ( $m$ )	peak-hour/off-peak	Categorical
Expressway ( $e$ )	PIE, AYE, ECP, etc.	Categorical
Direction along the expressway ( $r$ )	eastward, westward, northward, southward	Categorical
Condition of shoulder ( $h$ )	not affected, affected	Categorical
Total number of lanes ( $n$ )	1, 2, 3, 4, 5	Ordinal
Number of affected lanes ( $a$ )	0, 1, 2	Ordinal
Type of affected lane ( $l$ )	1st, 2nd, 3rd, etc. (from extreme right)	Categorical
Type of incident ( $y$ )	accident, breakdown, obstruction, etc.	Categorical

The traffic data consists of the traffic speed and flow values from the expressways (including the ramps) of Singapore. The values are recorded at 5-minute interval in each of the 2156 links. The speed value, thus recorded, represents the average speed of all vehicles traversing the link during the 5-minute interval. On the other hand, the flow value indicates the total number of vehicles that pass through the link in this 5 minutes span. Although the relative distribution of different types of vehicles such as trucks, buses etc. may influence the duration of an incident, we do not consider them as separate features. This is because our models implicitly take those distributions into account. For instance, if a particular expressway is expected to experience a change in the speed flow pattern caused by massive truck movement during a particular time of the day, the model will try to learn the impact from the historical training set. We would re-train the model periodically to calibrate for the changes in traffic composition. In case the composition changes abruptly (say due to an

unforeseen circumstance), the performance of the model will naturally degrade. Although micro-simulation models may consider these factors explicitly, any simulation model will also require re-calibration for an abrupt change, making such models non-suitable for a real-time application.

The average speed values are indicated by ten discrete speed-bands. While the first nine bands represent the speeds up to 90 kmph (each spanning 10 kmph), the 10-th band extends over the values higher than 90 kmph. Conversely, the traffic flow data is continuous-valued. The statistical summary of the traffic data is depicted in Table III.

TABLE III: Statistics of the traffic data.

Traffic Data	Values
Minimum speed	0 kmph
Maximum speed	100 kmph
Mean speed	81.8 kmph
Median of the speeds	90 kmph
Mode of the speeds	100 kmph
Minimum flow	1 veh/hr
Maximum flow	9960 veh/hr
Mean traffic flow	2154.6 veh/hr
Median of the traffic flows	1653 veh/hr
Standard deviation of the traffic flows	1869.4 veh/hr

### III. APPROACH AND PREDICTION METHOD

In this section, we describe our approach to compute the effective incident duration using the traffic data. We also explain the prediction method and the performance metric we use in our analysis.

#### A. Approach to Compute the Effective Duration

The effective duration of an incident signifies the time span for which the congestion existed in the network because of the incident. In this subsection, we elaborate on the steps of computing the effective duration of the incidents using the traffic speed and flow data from the neighboring links. It is important to consider both speed and flow data as indicators of traffic congestion. From the fundamental speed-flow relationship [32], if both the speed and flow is lower than usual, the link is clearly congested. Consequently, considering two different traffic variables helps us avoid false flags raised due to sensor noise. Furthermore, rules based on a single traffic variable do not generalize well to diverse city scale networks. For example, the speed limit of the arterial roads is usually lower than that of expressways. Therefore, the speed will drop less during incidents at arterial roads compared to expressways. Hence, the change in flow is a better indicator of congestion for the incidents in arterial roads. Thus, considering both speed and flow allows us to design more robust decision rules for detecting congestion. We assume that we can access the real-time traffic data from the Land Transport Authority of Singapore directly, or perhaps through their website, which provides live updates of the traffic conditions on the Singapore road network [33]. Moreover, we assume that the traffic slowdown is caused solely by the incident.

*Step 1:* Let us assume that according to the report, the start and end times of the incident  $i$  which occurred at link  $\ell$

are  $T_{\text{rep\_start}}$  and  $T_{\text{rep\_end}}$ , respectively. Therefore, the reported duration of the incident is  $T_{\text{rep\_start}} - T_{\text{rep\_end}}$ . Since we have traffic data recorded at 5-minute interval, we divide the entire duration in discrete time instants and refer to each instant by  $t_j$ , where  $j \in \{0, 1, 2, \dots, L\}$ . Therefore, the reported start time  $t_0$  and end time  $t_L$  are defined by:

$$t_0 = \left\lfloor \frac{T_{\text{rep\_start}}}{5} \right\rfloor * 5, t_L = \left\lfloor \frac{T_{\text{rep\_end}}}{5} \right\rfloor * 5. \quad (2)$$

Furthermore,  $t_1, t_2, t_3, \dots$  indicate 5, 10, 15... minutes after the start time.

*Step 2:* From the traffic data, we obtain the traffic speeds  $s(u_\ell, t_j)$  and  $s(d_\ell, t_j)$  for each time instant  $t_j$ , where  $u_\ell$  and  $d_\ell$  are the upstream and downstream links of  $\ell$  respectively. The downstream link is considered for comparison with the upstream link in order to ensure that the traffic slowdown is caused due to the incident only. Moreover, we fetch the traffic flow values  $f(\text{inci}_\ell, t_j)$  and  $f(\text{non}_\ell, t_j)$  of link  $\ell$  for each time instant  $t_j$ , where  $f(\text{inci}_\ell, t_j)$  implies the flow value on the day of incident, and  $f(\text{non}_\ell, t_j)$  indicates the average flow of the non-incident same day of other weeks for the entire 6 months. For example, if a particular incident happened on a Monday, we compute the average flow of all other Mondays of 6 months except the day of incident to compare it with the flow of the incident-affected Monday. However, we do not compare the traffic flow of the upstream and downstream link because in practice, the average flow in the downstream link cannot exceed the average flow of the upstream link, unless there is an entrance to the expressway in between those two links. Moreover, even if there is an entrance, it is likely to be congested because of the incident. Besides, the maximum capacities of the upstream and downstream links might differ from each other.

Now, we obtain the values of  $s(u_\ell, t_j)$ ,  $s(d_\ell, t_j)$ ,  $f(\text{inci}_\ell, t_j)$ , and  $f(\text{non}_\ell, t_j)$  for each time instant  $t_j$ , where  $j \in \{-12, -11, \dots, 0, 1, 2, \dots, L-1, L, L+1, \dots, L+12\}$ . According to our notation,  $t_{-1}, t_{-2}, \dots$  indicate 5, 10, ... minutes respectively before the reported start time, and  $t_{L+1}, t_{L+2}, \dots$  correspond to 5, 10, ... minutes respectively after the reported end time of the incident. Therefore, for computing the effective duration we obtain the traffic speed and flow values for the entire duration from 60 minutes before the reported start time to 60 minutes after the reported end time. We consider a margin of 2 hours in our analysis because time is needed to detect, verify and report incidents (early start time) or the impact on the traffic may not immediately disappear as soon as the incident ends (late end time).

*Step 3:* Next, we compute the difference in speed  $d_s(\ell, t_j)$  and flow  $d_f(\ell, t_j)$ , where  $d_s(\ell, t_j)$  and  $d_f(\ell, t_j)$  are defined by:

$$d_s(\ell, t_j) = s(u_\ell, t_j) - s(d_\ell, t_j), \quad (3)$$

$$d_f(\ell, t_j) = f(\text{inci}_\ell, t_j) - f(\text{non}_\ell, t_j). \quad (4)$$

*Step 4:* We define  $T_{\text{eff\_start}}$  as:

$$T_{\text{eff\_start}} = (T_{\text{rep\_start}} + (t_j^* - t_0)) \ni t_j^* = \underset{t_j}{\operatorname{argmin}} \{t_j | (d_s(\ell, t_j) < 0, f(\text{inci}_\ell, t_j) < 0.8 * f(\text{non}_\ell, t_j))\}, \quad (5)$$

i.e., the time instant when the speed in the upstream link becomes less than that of downstream link and the average flow on the day of incident drops to less than 80% of the average flow of non-incident days is considered to be the first threshold point ( $T_{\text{eff\_start}}$ ) because the queue starts to grow in the upstream direction at this instant.

We now explain why we choose 80% to be the threshold ratio of the flow values of the incident day to the non-incident day. To this end, we determine the ratio of  $f(\text{inci}_\ell, t_j)$  and  $f(\text{non}_\ell, t_j)$  at different time instant  $t_j$  for different incidents (see Fig. 2).

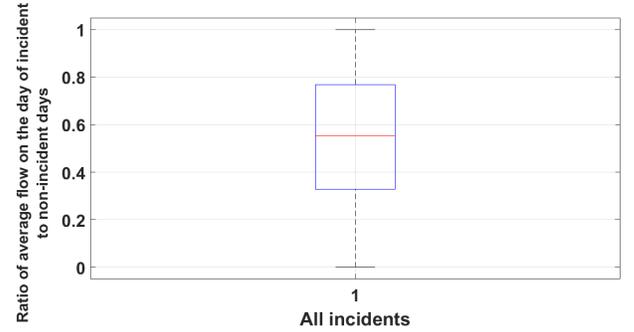


Fig. 2: Box plot of the ratios of traffic flow on the day of incident to that of non-incident days.

We observe that the ratio of 0.77 is at the 75-th percentile. Therefore, we choose the nearest round figure 0.8 (80%) to be the threshold ratio of the flow data.

*Step 5:* Similarly, we define  $T_{\text{eff\_end}}$  as:

$$T_{\text{eff\_end}} = (T_{\text{rep\_end}} + (t_j^* - t_0)) \ni t_j^* = \underset{t_j}{\operatorname{argmin}} \{t_j | (t_j > T_{\text{eff\_start}}, d_s(\ell, t_j) > 0, f(\text{inci}_\ell, t_j) > f(\text{non}_\ell, t_j))\}, \quad (6)$$

i.e., the time instant when the speed in the upstream link becomes greater than that of downstream link and the average flow of the day of incident is comparable to that of non-incident days is assumed to be the second threshold point ( $T_{\text{eff\_end}}$ ). We suppose that the traffic goes back to normal at this instant.

*Step 6:* Therefore, the reported incident duration is  $T_{\text{rep\_start}} - T_{\text{rep\_end}}$ , whereas the effective duration is  $T_{\text{eff\_start}} - T_{\text{eff\_end}}$ .

## B. Prediction Method and Performance Metric

In this subsection, we explain the steps for predicting the incident duration. In Singapore, the Land Transport Authority (LTA) gets notifications about incidents from sources such as traffic sensors, traffic cameras, police or motorists, and takes necessary response actions such as generating incident records, deploying quick response incident management teams and updating real-time traffic information on electronic signboards, website, mobile app and radio [34].

Now for each incident  $i$ , we extract the spatial, temporal and geographical features, such as type of day  $d(i)$ , day of week  $w(i)$ , time of the day  $m(i)$ , expressway  $e(i)$ , direction  $r(i)$ , total number of lanes  $n(i)$ , shoulder affected or not  $h(i)$ , number of lanes affected  $a(i)$ , type of affected lane  $l(i)$  and type of the incident  $y(i)$ . Moreover, since the duration of

the queue depends heavily on the traffic conditions of the feeding links, we incorporate the real-time traffic data as well for prediction purpose. The fact that the speed values are categorical, whereas the flow values are continuous, does not affect the prediction because the average speed and average flow are assumed as independent input features. We have the average speed and flow value  $s(l, t_j)$  and  $f(l, t_j)$  recorded during the interval  $(t_j - \delta_t, t_j)$ , where the sampling interval  $\delta_t$  is 5 minutes.

However, the features may not be available at the same time. Moreover, the values of the features may vary with elapsed time. Therefore, we construct different feature subsets based on the availability of the features in real life. The subsets of the features are demonstrated in Fig. 3. The first subset

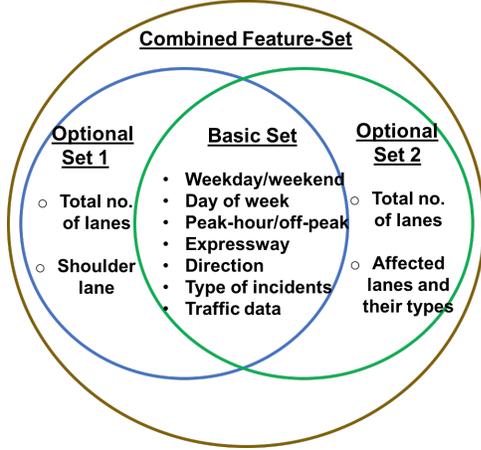


Fig. 3: Different sets of features.

contains the most generic features, which are usually available immediately after the incident happens. Hence, this subset of features is termed as basic set denoted by  $\mathcal{B}(i, t_j)$ . Some of the features in the basic set, such as traffic flow or speed values are updated at each instant. We can express the basic set mathematically as:

$$\begin{aligned} \mathcal{B}(i, t_j) = & (d(i), w(i), m(i), e(i), r(i), y(i), \\ & s(l, t_j), s(l, t_j - \delta_t), s(l, t_j - 2\delta_t), \dots, s(l, t_j - z\delta_t), \\ & f(l, t_j), f(l, t_j - \delta_t), f(l, t_j - 2\delta_t), \dots, f(l, t_j - z\delta_t)), \end{aligned} \quad (7)$$

and

$$\mathcal{B}(i, t_{j+1}) = a.\mathcal{B}(i, t_j). \quad (8)$$

Here  $z\delta_t$  represents the horizon of the past speed and flow values starting from the time instant when the incident started and  $a$  is the state transition vector. Therefore, the size of the feature vector increases with time.

Moreover, there are some optional feature subsets depending on how the information about the lanes are released by LTA. We train additional regression models which can incorporate these optional features if the required data is available. We denote the optional feature set (union of the optional feature subsets) by  $\mathcal{X}(i, t_j)$ . Although the total number of lanes remains constant, the condition of shoulder, number of affected lanes and type of affected lanes may change with time. As suggested by existing studies and our own data, it can take 5 – 10 minutes (1 – 2 prediction steps for our model) to

provide fine-grained information about the location of the incident [16]. Let us consider the following example. Suppose an accident happens on an expressway. One way in which an alert would be generated is if one of the drivers calls an ambulance, which will also alert the traffic management authority. At this point, the position of the accident might only be triangulated to the name of the road and nearby exit or landmark. Depending upon the resolution, different sections of a road may have different lanes and may or may not have shoulder lane. In this situation, our model will make a very generalized prediction based on the type of road, time of the day, etc. After a few minutes, the transportation authority might localize the spot using CCTV cameras and hence, provide a more specific location. Consequently, the information about the number of lanes and shoulder lane can be inferred. Further, the team reaches on-site and may provide (or even update information) about the number of affected lanes. In case CCTV camera is not available, the system would have to wait for on-site resource to update the ground conditions. We can write the optional set mathematically as:

$$\mathcal{X}(i, t_j) = (n(i), h(i, t_j), a(i, t_j), l(i, t_j)), \quad (9)$$

and

$$\mathcal{X}(i, t_{j+1}) = b.\mathcal{X}(i, t_j). \quad (10)$$

Here,  $b$  is the state transition vector. Our regression model performs the prediction using the basic set (i.e., the first feature subset as shown in Fig. 3), when  $\mathcal{X}(i, t_j) = 0$ . However, if the model is aware of the total number of adjacent lanes and the condition of shoulder lane while the current status of other lanes is still unknown to the model, the basic set and the optional feature set 1 (i.e., the second feature subset in Fig. 3) is considered for predicting the remaining duration. Similarly, if our predictive model has information about the closure of all other lanes but the shoulder lane, it performs the prediction with the basic set and the optional feature set 2 (i.e., the third subset as mentioned in Fig. 3).

Finally, our model takes the complete feature set (the combination of basic set and all optional sets in Fig. 3) into account, when all of these features are available. We denote the entire feature vector by  $\mathcal{F}(i, t_j)$ . Therefore, we can rewrite it as:

$$\mathcal{F}(i, t_j) = (\mathcal{B}(i, t_j), \mathcal{X}(i, t_j)), \quad (11)$$

i.e.,

$$\mathcal{F}(i, t_j) = (a.\mathcal{B}(i, t_{j-1}), b.\mathcal{X}(i, t_{j-1})). \quad (12)$$

Our goal is to find the relationship function  $\Phi$  between the feature set  $\mathcal{F}(i, t_j)$  and the remaining incident duration  $T(i, t_j)$  at the  $t_j$ -th instant:

$$T(i, t_j) = \Phi(\mathcal{F}(i, t_j)). \quad (13)$$

We build separate regression models using training data for all of these feature subsets at each 10 minute elapse from the starting point of the incident, because the remaining duration is predicted after 5 minutes, 15 minutes, 25 minutes, etc. until the incident ends. In our earlier work [29], we observed that the prediction performance can be improved by first grouping the

incidents through common latent similarities among them and then training data-driven predictors for each group. Therefore, our model incorporates clustering of incidents followed by applying the regression methods. We show the training and testing steps of the incident duration prediction model in Algorithm 1.

In this study, we consider various regression methods, i.e.,

---

**Algorithm 1** Incident Duration Prediction Algorithm

---

**Training step:**

- 1: **for**  $t = T_{\text{eff\_start}} : T_{\text{eff\_end}}$  by 10 min **do**
- 2:   Choose appropriate feature-set based on availability (from four sets of Fig. 3)
- 3:   **for**  $p = 1 : 4$  **do**
- 4:     Cluster the incidents based on features of the  $p$ -th set (K-means or GMM clustering)
- 5:     Build regression model for each cluster
- 6:   **end for**
- 7: **end for**

**Testing step:**

- 8: **while** the incident ends **do**
  - 9:   Find which feature-set is available
  - 10:   Choose nearest cluster based on Euclidean distance
  - 11:   Predict the duration using the appropriate regression model built by that cluster
  - 12: **end while**
- 

Classification And Regression Tree (CART) [35], Multi-Layer Perceptron (MLP) [36], Treebagger [37], Support Vector Regression (SVR) [38], Adaptive Fuzzy Neural Network [26], and Gaussian Mixture Regression (GMR) [39] for predicting the incident duration and compare their performances in Table IV. For this purpose, we compute the Mean Absolute Percentage Error (MAPE) values obtained by these regression methods:

$$\text{MAPE} = \frac{100}{N} \cdot \sum_{i=1}^N \left| \frac{e_i}{q_i} \right|, \quad (14)$$

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

$$e_i = q_i - \hat{q}_i. \quad (15)$$

We now consider an example of a potential real-world scenario. We examine the flow of events with time when an incident happens as shown in Fig. 4. Let us assume the traffic is normal at first at around 5:00 pm in an expressway. At 5:05 pm, a vehicle breakdown occurs and blocks the road. Therefore, traffic starts to slow down. Our model detects the change in the traffic data and speculates that an incident has happened. Although the incident reports are not available yet (since it takes some time to report the incidents to the management authority), the regression model performs the first prediction using the information it obtains from the traffic data. Meanwhile, the incident is reported approximately at 5:15 pm (10 minutes after the incident happened). The essential features are disseminated after the incident has been reported and therefore we can update our database (i.e., the location etc.)

according to the report. Our model performs the prediction again at 5:20 pm. Further, the details about the closure of the lanes are reported at 5:30 pm. Therefore, as all features are available by this time, the prediction gets more accurate. In the mean time, the response team reaches the location and clears the damaged vehicle from the road at 5:40 pm. However, traffic usually takes some time to go back to normal after the incident has been cleared. At 5:50 pm, the traffic conditions are recovered although the queue of the cars still exists. Finally, at 6:00 pm, the model finds the traffic data to be back to normal again and hence, stops the prediction.

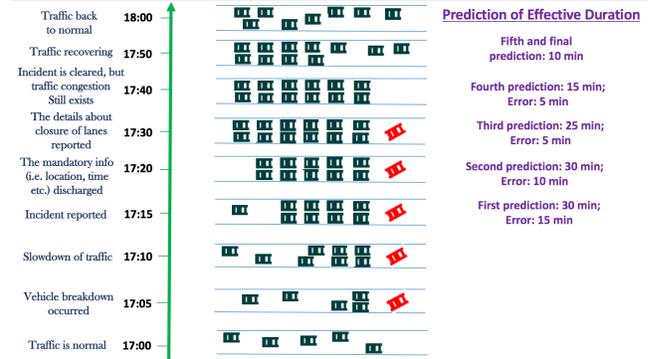


Fig. 4: A real-life scenario demonstrating how our algorithm works.

#### IV. COMPARISON OF EFFECTIVE AND REPORTED DURATION OF THE INCIDENTS

In this section, we illustrate a comparative analysis of the reported and effective duration of the incidents. As reported duration is obtained from the incident reports, whereas effective duration is computed using the traffic data, it is necessary to estimate if they are significantly different or not. In our work, there is a limitation in the approach to determine the effective duration of the incidents. The reported incident duration is a continuous variable, whereas the resolution of the traffic data is 5 minutes. We consider the reported duration to be the baseline and then determine the change in the start and end time of the incidents separately using the traffic data. Therefore, the difference in the reported and effective duration is always a multiple of 5.

We first show the scatter diagram of the reported duration and effective duration in Fig. 5. The correlation coefficient of the reported and effective duration is 0.89. We observe in

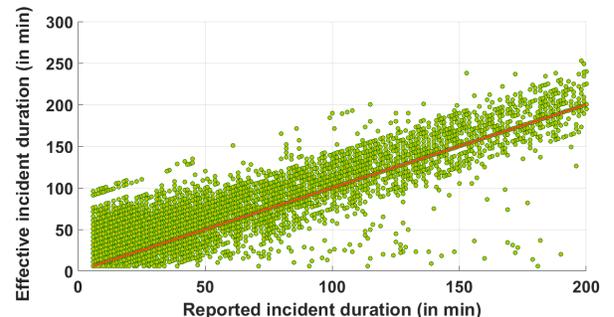


Fig. 5: Scatter diagram of effective duration (in minute) vs. reported duration (in minute).

Fig. 5 that the effective duration is greater than the reported duration for a large number of incidents (53.75% of all incidents). However, the effective duration can be less than the reported duration as well, though the percentage is small (13.01%). For the other incidents, the reported duration and effective duration are same. However, we observe that there are some outliers in the graph (the absolute difference of reported and effective duration is more than 60 minutes for 1.5% incidents). We discard them from our analysis.

To further analyze the differences in start time and end time separately, let's assume the difference between effective and reported start time is  $\Delta t_1$  and the difference between effective and reported end time is  $\Delta t_2$ . We show the cumulative distribution of the incidents with  $\Delta t_1$  and  $\Delta t_2$  in Fig. 6.  $\Delta t_1$

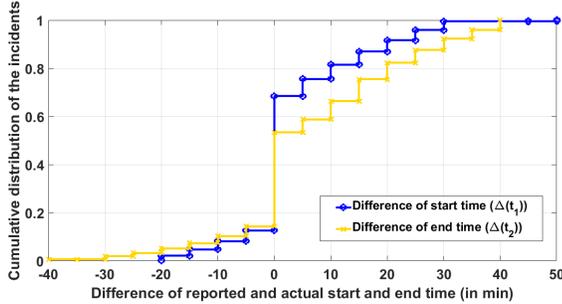


Fig. 6: Cumulative distribution of incidents with  $\Delta t_1$  and  $\Delta t_2$ .  $\Delta t_1$  is positive when time is needed to detect, verify and report the occurrence of an incident. Consequently, effective start time is before the reported start time for those incidents. Conversely,  $\Delta t_1$  is negative when effective start time is later than reported start time (i.e., traffic is normal although there is an incident). It may happen when the number of cars is much less than the capacity of the road at the time of incident. Furthermore,  $\Delta t_2$  is positive when there exists a time-gap between the clearance of the road and recovery of the traffic. Therefore, in these cases, the incident is reported to be ended although the traffic is not normal yet. By contrast,  $\Delta t_2$  is negative when effective end time is before reported end time (i.e., traffic goes back to normal before the incident is reported to end). It happens because sometimes incident vehicles are moved away from traffic lanes to road shoulder, hence traffic can resume to normal, even though incident is not yet over. Another reason could be that traffic flow is way below the capacity by the time the incident ends, hence there is no congestion and traffic can flow normally.

We observe in Fig. 6 that the steps in the positive side of  $\Delta t_1$  and  $\Delta t_2$  are steeper, which indicates that  $\Delta t_1$  and  $\Delta t_2$  are positive for a larger percentage of incidents. It also supports the fact that the effective duration is larger compared to reported duration for majority of incidents. Moreover, the cumulative frequency at  $x = 0$  of  $\Delta t_1$  is larger compared to that of  $\Delta t_2$ , which means that the reported start time overlaps with the effective one for many incidents, whereas the end times do not.

## V. PREDICTION PERFORMANCE OF DIFFERENT REGRESSION METHODS

In this section, we consider various traditional as well as recent methods, such as Classification And Regression Tree

(CART) [40], Multi-Layer Perceptron (MLP) [36], Treebagger [37], Support Vector Regression (SVR) [38], Adaptive Fuzzy Neural Network [26], and Gaussian Mixture Regression (GMR) [39] to model the relationship between various traffic factors and the incident duration.

For the first four regression methods, we divide the training data points into different clusters using  $K$ -means clustering algorithm [41] and build a regression model for each of them [29]. To determine the optimum number of clusters, we verify how the intra-cluster distance varies with the number of clusters  $K$ . The value of  $K$  where the largest drop occurs in the intra-cluster distance value is considered to be the optimum number of clusters. In the next step, we assign each testing data point to its nearest cluster and obtain a predicted value of the remaining duration by applying the model of that nearest cluster. The second step is repeated for all incidents in the test datasets. Moreover, we apply three-fold cross-validation to choose training and test datasets. Therefore, all the steps are iterated for each of the three training and test datasets.

We select the optimal parameters for each method through 10-fold cross-validation. We use an ensemble of five trees for Treebagger, and the MLP architecture with three hidden layers produced optimal results. Moreover, we observe that the SVR method with radial basis function as kernel provides the best result. For ANFIS method, we use Fuzzy  $c$ -means clustering (FCM) [42]. The number of clusters, maximum number of iterations and the minimum improvement in objective function between two consecutive iterations are set to 3, 200 and  $10^{-5}$ , respectively. Besides, for the Gaussian Mixture Regression method, we apply the Gaussian Mixture model [43] instead of  $K$ -means clustering prior to regression, where number of clusters is set to 2.

For comparison purpose, we provide the results for one-time prediction of the reported durations obtained by these methods. The results are mentioned in Table IV.

TABLE IV: MAPE values for one-time prediction of the incidents reported duration.

CART	Treebagger	MLP [15]	SVR	ANFIS [28]	GMR [44]
73.1	55.84	66.3	63.2	55.7	59.36

We find in Table IV that the Treebagger and ANFIS methods perform almost equivalently and these methods outperform others. However, since ANFIS method is computationally expensive compared to other methods, we do not prefer to use it further in our study. Since the Treebagger method assigns weights to the available features itself, we do not preset the weights ourselves and in the next sections, we consider the results obtained by the Treebagger method. For instance, the assigned weights to the features by our model for predicting the reported duration of all the incidents together are mentioned in Table V. We observe in Table V that the

TABLE V: The assigned weights to the features by the Treebagger method for predicting the reported duration of all the incidents together.

Weekday	Weekend	Day of week	Peak-hour	Off-peak	Express way	Direction	Shoulder lane	Total no. of lanes	Lane 1 or 2 affected	Other Lanes	Type of incident	Speed data	Flow data
0.26	0.09	0.1	0.33	0.15	0.31	0.2	0.68	0.38	0.5	0.17	0.5	0.34	0.6

status of the shoulder lane (affected or not), which main lane

is affected, the type of incident and the flow data are the most important features in predicting the duration.

## VI. OVERALL PREDICTION OF EFFECTIVE AND REPORTED DURATION WITH DIFFERENT FEATURE SETS

In this section, we evaluate the prediction performance of the regression models built on different feature sets according to their availability. We compute the results obtained by these regression models for both reported duration as well as effective duration prediction.

For comparison purposes, we provide the results only for the basic feature set and considering all features together. We compare the performance of our regression models in Table VI and Table VII for predicting the effective duration and reported duration respectively. The results are obtained by averaging over all incidents.

TABLE VI: The overall MAPE values (in percentage) obtained by the Treebagger model using basic feature set and all features in predicting reported durations.

Features	5 min	15 min	25 min	35 min	45 min	65 min	85 min	105 min	125 min	155 min
Basic	61.2	59.8	55.7	52.3	50.1	46.28	44.47	40.76	38.9	34.41
All	55.84	52.3	49.68	47.7	44.03	41.76	40.9	36.1	33.6	30.02

TABLE VII: The overall MAPE values (in percentage) obtained by the Treebagger model using basic feature set and all features in predicting effective durations.

Features	5 min	15 min	25 min	35 min	45 min	65 min	85 min	105 min	125 min	155 min
Basic	68.24	66.55	63.46	60	55.95	50.27	46.62	42.15	38.7	33.06
All	61	58.3	54.62	53.16	48.76	44.67	39.92	35.87	31.69	27.58

We find in Table VI and Table VII that the error values improve when we have all features compared to the basic feature set. At the beginning of prediction, the difference in MAPE values obtained by basic feature set and by all features is 5.36% for reported duration and 7.24% for effective duration prediction. However, these percentages are 4.39% and 5.48% respectively, at the end of prediction. In the previous section, we observe in Table V that the status of the shoulder lane (affected or not) and the type of affected lane have significant role in predicting the incident duration. Therefore, in general prediction improves by 5% – 10% if we consider all features instead of the basic feature set only.

## VII. PREDICTION OF REPORTED DURATION FOR DIFFERENT CLASSES OF INCIDENTS

In this section, we analyze the performance of our regression model in predicting the reported durations at different instants. At first, the histogram of the reported durations of the incidents is shown in Fig. 7. The mean, median, and mode of the reported incident durations are 50.05 minutes, 29 minutes, and 6 minutes, respectively. Next we categorize the incidents based on their total durations for comparing the error values. The MAPE values obtained by our Treebagger model using all features are mentioned for different categories of incidents in Table VIII. The rows represent the classes of the incidents based on the total reported durations, and the columns represent the elapsed time.

We observe in Table VIII that the MAPE values are very high for short-duration incidents (5 – 25 minute). Hence, the

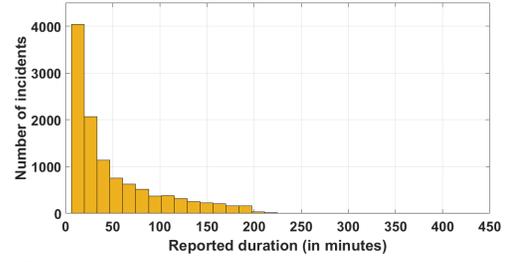


Fig. 7: Histogram of reported duration of the incidents.

TABLE VIII: Variation of MAPE values (in percentage) with elapsed time since the incidents start for reported duration prediction.

	5 min	15 min	25 min	35 min	45 min	55 min	65 min	85 min	105 min	125 min	155 min
5-15 min	98.24	–	–	–	–	–	–	–	–	–	–
16-25 min	87.84	93.7	–	–	–	–	–	–	–	–	–
26-35 min	66.20	70.58	64.98	–	–	–	–	–	–	–	–
36-45 min	56.68	54.66	51.02	48.32	–	–	–	–	–	–	–
46-55 min	47.51	43.65	38.45	33.93	37.94	–	–	–	–	–	–
56-65 min	39.23	36.22	30.78	27.2	29.94	35.62	–	–	–	–	–
66-85 min	33.37	32.35	24.18	22.61	20.23	21.02	24.97	–	–	–	–
86-105 min	29.4	25.86	21.07	21.87	20.44	18.93	16.07	21.24	–	–	–
106-125 min	23.21	24.66	21.74	19.64	17.72	15.16	14.29	18.14	23.02	–	–
126-155 min	43.15	38.36	37.64	34.36	31.72	28.10	25.67	21.08	20.34	23.36	–
156-200 min	52.57	49.15	47.01	44.63	42.67	38.84	36.60	32.77	30.06	28.11	25.02
Rest	75.32	69.36	61.64	58.4	60.22	57.55	58.82	46.37	40.21	50.45	49.63

prediction is not very reliable for these incidents. However, for the incidents in the middle range (46 – 125 minute), the MAPE values are 30% – 47% at first, which reduces to approximately 20% – 38% at the end of prediction. Consequently, we obtain a better prediction for longer durations. This is because the number of features corresponding to the traffic data increases with elapsed time, hence the prediction of remaining duration improves significantly as time elapses.

## VIII. PREDICTION OF EFFECTIVE DURATION FOR DIFFERENT CLASSES OF INCIDENTS

In this section, we study the performance of our proposed approach in predicting the effective duration of the incidents. The results are reported for the prediction performed by the Treebagger model. We have explained the method in Section III how we compute the effective duration of the incidents. The histogram of the effective durations is shown in Fig. 8. The mean, median, and mode of the incident durations are 60.75 minutes, 46 minutes, and 6 minutes, respectively. Similarly

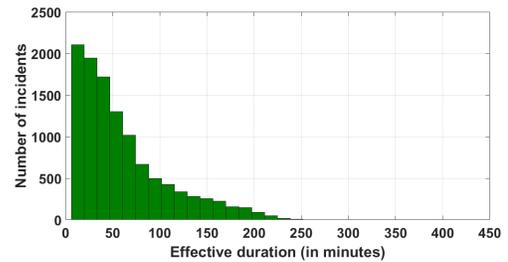


Fig. 8: Histogram of effective duration of the incidents.

as in the previous Section, the MAPE values obtained by our regression model taking all features into account are mentioned in Table IX. The rows indicate different classes of incidents based on total effective durations and the columns represent elapsed time.

TABLE IX: Variation of MAPE values (in percentage) with elapsed time since the incidents start for effective duration prediction.

	5 min	15 min	25 min	35 min	45 min	55 min	65 min	85 min	105 min	125 min	155 min
5-15 min	100.9	–	–	–	–	–	–	–	–	–	–
16-25 min	94.6	96.9	–	–	–	–	–	–	–	–	–
26-35 min	75.52	81.6	79.31	–	–	–	–	–	–	–	–
36-45 min	50.96	57.54	50.15	47.65	–	–	–	–	–	–	–
46-55 min	42.54	39.34	34.6	31.26	32.71	–	–	–	–	–	–
56-65 min	38.09	32.33	26.9	29.86	30.62	29.6	–	–	–	–	–
66-85 min	33.7	32.11	24.33	20.21	17.27	17	20.04	–	–	–	–
86-105 min	25.23	20.02	20.88	17.95	16.62	16.32	14.29	20.53	–	–	–
106-125 min	24.36	23.1	20.03	15.9	14.52	15.2	13.6	17.73	21.78	–	–
126-155 min	41.22	38.97	36.39	31.9	30.87	28	26.73	22.01	19.35	15.68	–
156-200 min	52.23	49.33	47.4	43.04	41.77	38.01	36.65	32.48	28.72	27.13	22.58
Rest	76	70.14	60.8	60.3	64.3	60.8	59.1	47.3	40.6	51.26	50

We find in Table IX that the error values are in general a bit higher for effective durations. In fact, the difference becomes smaller for longer duration incidents since effective duration is computed using traffic data and we can incorporate more features corresponding to the traffic data with elapsed time. Therefore, the performance of our model for predicting the effective duration is better for longer duration incidents ( $> 85$  minutes). Overall we achieve the similar pattern of the results as discussed in Section VII.

#### IX. PREDICTION OF REPORTED DURATION FOR THE RAMP INCIDENTS

In this section, we analyze the performance of our prediction model for incidents that occurred alongside different entrance/exit ramp locations. These incidents lasted for an average of 63 minutes with 50% of incidents lasting for about 36 minutes or less. Let us start by analyzing the temporal trend in performance of the model across all ramps. Table X shows the prediction errors in terms of MAPE by considering the reported duration of the incidents. At the start of the incident, our model had a mean error of 71.93%, which reduced to 54.46% after 35 minutes (25% improvement) and 45.26% after 65 minutes (35% improvement).

TABLE X: The overall MAPE values (in percentage) obtained by the Treebagger model using all features in predicting reported durations of the ramp incidents.

5 min	15 min	25 min	35 min	45 min	65 min	85 min	105 min	125 min	155 min
71.93	67.41	64.74	54.46	50.25	45.26	42.13	40.92	40.24	36.17

Table X shows the average performance of the model for ramps, which was aggregated over various kinds of incidents. In Table XI, we show the forecasting error for incidents with different durations (rows of the table). The table also shows how the accuracy changes with elapsed time (columns of the table). For ramps, we see similar temporal trends that we have seen for expressways as well (see Table VIII of the manuscript). As with expressways, the prediction accuracy improves as time progresses.

Since we considered a large heterogeneous urban network for analysis, it is natural to assume that the prediction performance of the model will not remain the same for different ramps at different locations. To analyze these trends, we group the ramps based on the expressways they serve. Moreover, Table XII shows the variations in the prediction performance

TABLE XI: Variation of MAPE values (in percentage) with elapsed time for ramp incidents.

	5 min	15 min	25 min	35 min	45 min	55 min	65 min	85 min	105 min	125 min	155 min
5-15 min	113	–	–	–	–	–	–	–	–	–	–
16-25 min	99.4	94.4	–	–	–	–	–	–	–	–	–
26-35 min	82.32	80.2	77	–	–	–	–	–	–	–	–
36-45 min	75.72	69.19	67.25	70.38	–	–	–	–	–	–	–
46-55 min	66.43	63.68	53.77	56.40	50.33	–	–	–	–	–	–
56-65 min	59.81	54.13	48.72	44.68	46.36	47.50	–	–	–	–	–
66-85 min	51.91	47.12	41.22	39.92	35.25	33.44	33.14	–	–	–	–
86-105 min	46.23	39.72	35.18	31.63	29.5	26.4	25.55	27.62	–	–	–
106-125 min	42.98	36.83	33.50	28.73	25.05	23.5	22.95	21.55	24.67	–	–
126-155 min	50.03	43.27	37.72	35.32	31.41	29.49	27.86	24.93	23.02	21.07	–
156-200 min	62.89	56.33	53.2	49.5	45.28	41.75	38.12	36.9	33.77	29.1	26.3
Rest	80.99	75.59	79.53	72.62	68.06	61.91	56.5	51.58	47.18	45.28	45.17

for incidents that occurred on ramps located at different expressways. For comparison purpose, we also show the prediction performance for the mainlines of the expressways connected to the ramps. Table XII also includes the median duration of the incidents along different locations. We observe that the median duration of the ramp incidents on KPE, MCE and, SLE are the least. These expressways are comparatively short in length. Moreover, SLE is located far from the center of the city (the busiest area of Singapore) and a significant part of MCE comprises an underground tunnel (3.5 km out of 5 km). Therefore, the ramps of these expressways may not be much incident-prone. The error values are also lower for these ramps. However, since our model performs better for the incidents in the middle range compared to the low-duration ones, the error value is higher for MCE ramps compared to those of KPE and SLE (the median duration of MCE ramp incidents is 5.5 minutes, whereas it is 10 – 15 minutes for KPE and SLE). On the other hand, PIE is the longest and busiest expressway in Singapore. Besides, although BKE is far from the center of the city, it connects two countries, Malaysia and Singapore. Therefore, the median durations of the ramp incidents on PIE and BKE are the highest, which leads to a significant prediction error. However, the median durations, as well as the MAPE values, are almost in the same range for mainline incidents of different expressways.

TABLE XII: The median duration and the MAPE values (in percentage) for the incidents on ramps and mainlines of different expressways.

	Expressways	AYE	BKE	CTE	ECP	KJE	KPE	MCE	PIE	SLE	TPE
Median duration (in minutes)	Mainlines	27	24	19	25	27	22	18	20	19	25
	Ramps	37	47	33.5	23	28	11	5.5	42	15	33
MAPE (in percentage)	Mainlines	60.92	55.68	52.01	58.82	50.17	46.44	42.91	54.54	52.49	49.68
	Ramps	71.62	88.28	74.12	63.6	67.93	49.55	55.5	88.37	53.19	74.31

#### X. COMPARISON OF OUR RESULTS WITH EXISTING LITERATURE

In this section, we compare our results to the existing literature. Araghi *et al.* provided a comparative analysis of k-NN and Hazard-based Models in their work [45]. The MAPE values obtained by the two methods were 41.1% and 43.7% respectively. We obtain the MAPE values 55.84% and 61% at first averaged over all incidents for reported and effective duration respectively, which improve to 30.02% and 27.58% after 155 minutes. Our regression model performs better for the incidents with duration larger than 55 minutes (See Table VIII and Table IX). On the other hand, Li *et al.*

classified the incidents based on the durations and determined the error values for different classes of incidents similar to us [8]. The MAPE value for short duration incidents (5 – 15 minutes) in their prediction was 184%, whereas we obtain the value as 100.9% (effective duration prediction). For long duration incidents (greater than 120 minutes), they obtained 74% MAPE in prediction whereas our error values are much lower as can be seen in Table VIII and Table IX for different categories of incidents. In another work of Li *et al.*, the MAPE value averaged over all incidents is 94.7% [9], whereas we obtain the MAPE value of 61% (effective duration). Khattak *et al.* developed an online incident management tool named iMiT which can predict the incident duration as well as identify the secondary incidents occurrence [22]. They found that the error values are higher for the extreme cases (incidents in the highest and lowest duration range). We also observe the similar behavior of our regression model. In the recent years, Lin *et al.* achieved better prediction accuracy using the M5P-HBDM model [17]. The MAPE value obtained by this model was 33.15% for all incidents. Qing *et al.* evaluated the error values of five different regression models for comparing the prediction performance [1]. The hybrid tree-based quantile regression method performs the best (MAPE value 49.1%). We obtain similar results for the incidents with duration greater than 35 minutes (See Table VIII and Table IX). Furthermore, Pereira *et al.* performed the sequential prediction of incident duration [16]. They also considered the successive release of features and therefore, predicted the duration for different feature-sets. However, they applied the text analysis approach for this purpose. Their model could generate reliable predictions after 15 min which gradually reduces the mean absolute percentage error from 100% to 40% with elapsed time. Our overall MAPE value varies over time from 61% to 27.58% considering all incidents together. In conclusion, we achieve competitive or better results for predicting traffic incidents duration compared to the state-of-the-art. We summarize the results obtained in the existing studies in Table XIII.

TABLE XIII: Comparison of our results with other studies.

Literature	MAPE values
Araghi <i>et al.</i> [45]	KNN: 41.1%, HBDM: 43.7%
Li <i>et al.</i> [8]	5 – 15 min: 184%, >120 min: 74%, overall: 56%
Ruimin <i>et al.</i> [9]	2 – 15 min: 185.7%, >15 min: 45.4%, overall: 94.7%
Khattak <i>et al.</i> [22]	5 – 15 min: 329%, >120 min: 80%, overall: 214%
Valenti <i>et al.</i> [46]	ANN: 44%, SVR: 36%, KNN: 36%
Qing <i>et al.</i> [1]	KNN: 59.2%, CART: 57.1%, Quantile Regression: 49.1%
Pereira <i>et al.</i> [16]	MAPE varies over time from 100% to 40%
Wei <i>et al.</i> [23]	MAPE in the range of 35% to 45%
	5 – 15 min: 100.9%, 16 – 35 min: 75% – 96%,
Our work	36 – 200 min: 20% – 50%, > 200 min: varies over time from 76% to 50%, overall MAPE varies over time from 61% to 27.58%

## XI. CONCLUSION

In this paper, we proposed a data-driven approach to forecast the duration of traffic incidents. We considered the duration reported by the authorities in addition to the effective duration, calculated from the traffic speed and flow data. Our proposed method can work with incomplete real-time traffic information and can also dynamically update the forecasts when new data

becomes available. To this end, we considered traffic data and incidents record from the expressways of Singapore. We built regression models for several combination of feature sets and performed the prediction at different time instants till the end of the incidents. We observed that for the incidents with duration in the range of 36 – 200 minutes, the mean absolute percentage error in predicting the effective as well as reported duration varies in the range of 20% – 50%. Moreover, for the longer duration incidents (greater than 200 minutes), in the beginning our model has a mean error of 76%, which reduces to 57.55% after 55 minutes (25% improvement) and 40.6% after 105 minutes (46.58% improvement). Overall, we achieve 55.84% prediction error for reported duration and 61% error for effective duration prediction averaged over all the incidents, which improve to 30.02% and 27.58% respectively, with elapsed time. In general, our regression model makes reliable prediction for all incidents except the lowest duration range.

There are certain limitations which we can address in our future work. For example, since the variable speed limit control strategy is not implemented for incident management in Singapore, we do not consider its effect in our study. Furthermore, we plan to extend our model for predicting the impact of planned events, such as music concerts, sports events, etc.

The clearance time is a significant part of entire incident duration. This study seems to suggest that the shoulder lane plays a significant role during the incidents in the expressways of Singapore. If the affected vehicle is small, it should be shifted to the shoulder lane as soon as possible so that the main lanes get cleared, thereby reducing the clearance time. Moreover, there are VMS (Variable Message Signs) displays installed on the expressways of Singapore to make the drivers aware of the incidents so that the overall congestion may be minimized.

## ACKNOWLEDGMENT

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

## REFERENCES

- [1] Q. He, Y. Kamarianakis, K. Jintanukul, and L. Wynter, "Incident duration prediction with hybrid tree-based quantile regression," in *Advances in Dynamic Network Modeling in Complex Transportation Systems*. Springer, 2013, pp. 287–305.
- [2] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method," *IEEE Transactions on Intelligent Transportation Systems*, 2017.
- [3] A. T. Hojati, L. Ferreira, S. Washington, P. Charles, and A. Shobeirinejad, "Modelling total duration of traffic incidents including incident detection and recovery time," *Accident Analysis & Prevention*, vol. 71, pp. 296–305, 2014.
- [4] D. Chimba, B. Kutela, G. Ogletree, F. Horne, and M. Tugwell, "Impact of abandoned and disabled vehicles on freeway incident duration," *Journal of transportation engineering*, vol. 140, no. 3, p. 04013013, 2013.
- [5] T. F. Golob, W. W. Recker, and J. D. Leonard, "An analysis of the severity and incident duration of truck-involved freeway accidents," *Accident Analysis & Prevention*, vol. 19, no. 5, pp. 375–395, 1987.

- [6] B. Jones, L. Janssen, and F. Mannering, "Analysis of the frequency and duration of freeway accidents in seattle," *Accident Analysis & Prevention*, vol. 23, no. 4, pp. 239–255, 1991.
- [7] Y. Chung, "Development of an accident duration prediction model on the korean freeway systems," *Accident Analysis & Prevention*, vol. 42, no. 1, pp. 282–289, 2010.
- [8] R. Li, "Traffic incident duration analysis and prediction models based on the survival analysis approach," *IET intelligent transport systems*, vol. 9, no. 4, pp. 351–358, 2014.
- [9] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Competing risks mixture model for traffic incident duration prediction," *Accident Analysis & Prevention*, vol. 75, pp. 192–201, 2015.
- [10] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.
- [11] N. Breslow, "A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship," *Biometrika*, vol. 57, no. 3, pp. 579–594, 1970.
- [12] G. Giuliano, "Incident characteristics, frequency, and duration on a high volume urban freeway," *Transportation Research Part A: General*, vol. 23, no. 5, pp. 387–396, 1989.
- [13] D. Nam and F. Mannering, "An exploratory hazard-based analysis of highway incident duration," *Transportation Research Part A: Policy and Practice*, vol. 34, no. 2, pp. 85–102, 2000.
- [14] 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.
- [15] 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.
- [16] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Text analysis in incident duration prediction," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 177–192, 2013.
- [17] L. Lin, Q. Wang, and A. W. Sadek, "A combined m5p tree and hazard-based duration model for predicting urban freeway traffic accident durations," *Accident Analysis & Prevention*, vol. 91, pp. 114–126, 2016.
- [18] D. Shefer, "Congestion, air pollution, and road fatalities in urban areas," *Accident Analysis & Prevention*, vol. 26, no. 4, pp. 501–509, 1994.
- [19] A. Ceder and M. Livneh, "Relationships between road accidents and hourly traffic flow: analyses and interpretation," *Accident Analysis & Prevention*, vol. 14, no. 1, pp. 19–34, 1982.
- [20] T. F. Golob and W. W. Recker, "Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions," *Journal of transportation engineering*, vol. 129, no. 4, pp. 342–353, 2003.
- [21] P. Škorput, S. Mandžuka, and N. Jelušić, "Real-time detection of road traffic incidents," *PROMET-Traffic&Transportation*, vol. 22, no. 4, pp. 273–283, 2010.
- [22] A. Khattak, X. Wang, and H. Zhang, "Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays," *Intelligent Transport Systems, IET*, vol. 6, no. 2, pp. 204–214, 2012.
- [23] C.-H. Wei and Y. Lee, "Sequential forecast of incident duration using artificial neural network models," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 944–954, 2007.
- [24] Y. Jia, J. Wu, and Y. Du, "Traffic speed prediction using deep learning method," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 1217–1222.
- [25] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.
- [26] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.
- [27] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Systems with Applications*, vol. 91, pp. 452–463, 2018.
- [28] E. I. Vlahogianni and M. G. Karlaftis, "Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties," *Computer-Aided Civil and Infrastructure Engineering*, vol. 28, no. 6, pp. 420–433, 2013.
- [29] B. Ghosh, M. T. Asif, J. Dauwels, W. Cai, H. Guo, and U. Fastenrath, "Predicting the duration of non-recurring road incidents by cluster-specific models," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 1522–1527.
- [30] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.
- [31] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, "Estimating arterial traffic conditions using sparse probe data," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 2010, pp. 929–936.
- [32] T. V. Mathew and K. K. Rao, "Fundamental relations of traffic flow," *Lecture Notes in Transportation Systems Engineering. Bombay, India: Department of Civil Engineering Indian Institute of Technology Bombay*. Google Scholar, 2017.
- [33] "Datamall," <https://www.mytransport.sg/content/mytransport/home/dataMall.html>, 2017.
- [34] "One.motoring - traffic news," [https://www.onemotoring.com.sg/content/onemotoring/en/on\\_the\\_roads/traffic\\_news.html](https://www.onemotoring.com.sg/content/onemotoring/en/on_the_roads/traffic_news.html), 2017.
- [35] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [36] M. Riedmiller, "Advanced supervised learning in multi-layer perceptrons from backpropagation to adaptive learning algorithms," *Computer Standards & Interfaces*, vol. 16, no. 3, pp. 265–278, 1994.
- [37] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [38] 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.
- [39] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 286–298, 2007.
- [40] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
- [41] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.
- [42] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.
- [43] 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.
- [44] X. Liu, L. Pan, and X. Sun, "Real-time traffic status classification based on gaussian mixture model," in *Data Science in Cyberspace (DSC), IEEE International Conference on*. IEEE, 2016, pp. 573–578.
- [45] B. N. Araghi, S. Hu, R. Krishnan, M. Bell, and W. Ochieng, "A comparative study of k-nn and hazard-based models for incident duration prediction," in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp. 1608–1613.
- [46] G. Valenti, M. Lelli, and D. Cucina, "A comparative study of models for the incident duration prediction," *European Transport Research Review*, vol. 2, no. 2, pp. 103–111, 2010.



**Banishree Ghosh** received the Bachelors degree in Electronics and Telecommunication Engineering from Jadavpur University, Kolkata, India. She is currently pursuing her Ph.D. in the Interdisciplinary Graduate School, Nanyang Technological University, Singapore. Her research interests include traffic modeling, urban informatics, and predictive analytics.



**Muhammad Tayyab Asif** (S'12, M'16) received the Ph.D. degree in Electrical and Electronic Engineering from Nanyang Technological University, Singapore. Currently, he is working as an Applied Scientist in Amazon. Before that, he was a Research Scientist in IBM Research and BNY Mellon. His research interests include intelligent transportation systems, sensor fusion, connected vehicles, network optimization, and modeling of large-scale networks.



**Dr. Justin Dauwels** is an Associate Professor with the School of Electrical and Electronic Engineering at the Nanyang Technological University (NTU) in Singapore, and Deputy Director of the ST Engineering NTU corporate lab. His research interests are in data analytics with applications to intelligent transportation systems, autonomous systems, and analysis of human behavior & physiology. He obtained the PhD degree in electrical engineering at the Swiss Polytechnical Institute of Technology (ETH) in Zurich in December 2005. He was a postdoctoral fellow at the RIKEN Brain Science Institute (2006-2007) and a research scientist at the Massachusetts Institute of Technology (2008-2010). He has been a JSPS postdoctoral fellow (2007), a BAEF fellow (2008), a Henri-Benedictus Fellow of the King Baudouin Foundation (2008), and a JSPS invited fellow (2010, 2011). His research on intelligent transportation systems has been featured by the BBC, Straits Times, Channel 5, and various other media outlets. His research on diagnosis of Alzheimers disease is featured at a 5-year exposition at the Science Center in Singapore. His research team has won several best paper awards at international conferences. He has filed 10+ US patents related to data analytics.



**Ulrich Fastenrath** is currently the head of traffic information management and routing optimization at BMW Group, Germany. He joined the company in May 2012 and is responsible for research and development activities in traffic and routing space.

He received Ph.D. in theoretical solid state physics at the University of Cologne in Germany and was also a two-year post-doctoral fellow at the Technical University of Munich.



**Hongliang Guo** received his Bachelor of Engineering in Dynamic Engineering and a Master of Engineering in Dynamic Control at the Beijing Institute of Technology, China. He received PhD in Electrical and Computer Engineering from the Stevens Institute of Technology, USA. He later joined Alameda in Rotterdam, the Netherlands as a postdoc research in 2011. His research interests include self-organizing systems and agent-based technologies. He is now with the Center for Robotics, University of Electronic Science and Technology of China.