Accident Analysis and Prevention 145 (2020) 105696

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Real-time traffic accidents post-impact prediction: Based on crowdsourcing data

Yunduan Lin[a,b], Ruimin Li[a,*]

[a] Department of Civil Engineering, Tsinghua University, Beijing 100084, China
[b] Department of Civil and Environment Engineering, University of California, Berkeley, CA 94720, United States

ARTICLE INFO

Keywords:
Crowdsourcing data
Traffic accidents post-impact
Machine learning
Sequential prediction

ABSTRACT

Traffic accident management is a critical issue for advanced intelligent traffic management. The increasingly abundant crowdsourcing data and floating car data provide new support for improving traffic accident management. This paper investigates the methods to predict the complicated behavior of traffic flow evolution after traffic accidents using crowdsourcing data. Based on the available data source, the traffic condition is divided into four levels by congestion delay index: severely congested, congested, slow moving and uncongested. Four types of accidents are consequently defined based on the occurrence of each level. A hierarchical scheme is designed for identifying the most congested level and sequentially predicting duration of each level. The proposed model is validated using traffic accident data in 2017 from an anonymous source in Beijing, China by embedding three machine learning algorithms, random forest (RF), support vector machine (SVM) and neural network (NN), in the scheme. The results show NN outperforms the other two models when the assessment is conducted in absolute differences. Meanwhile, RF has a slightly better performance than SVM, especially when predicting the short-period congestion of severely congested level at the first time. By continuously updating the traffic condition information, significant improvement in accuracy can be acquired regardless of the exact model used. This study shows that emerging crowdsourcing data can be used in a real-time analysis of traffic accidents and the proposed model is effective to analyze such data.

1. Introduction

According to various factors, such as improper driving behavior, bad weather, and so on, traffic accidents become somewhat inevitable. They usually have a broad-scale impact on traffic conditions, especially in peak hour. Typically, accidents may generate congestion to some extent in part of the road network or even cause a chain breakdown in the entire system when the occurrence is overlapped with a road bottleneck. This type of congestion is characterized as nonrecurrent congestion along with congestion caused by large events, work zones and extreme weather. Nonrecurrent congestion, of which 72% is caused by accidents (Skabardonis et al., 2003), accounts for half to three-quarters of total congestion (Giuliano, 1989). To cope with this mobility challenge, accurately predicting Traffic Accidents Post-Impact (TAPI) is of great significance for better guiding traffic participants involved and more efficiently operating integrated transportation systems.

One key aspect of traffic accident analysis and prediction is to obtain both accident information and surrounding traffic conditions simultaneously in real time. Accident information in previous research was mainly single-sourcing from the Traffic Incident Management System (Li et al., 2018), which is created and operated by government departments or research institutions. Live traffic conditions can be acquired from various road sensors such as cameras, loop detectors and GPS-based on-board units. However, spatiotemporal sparsity issue occurs in the above conventional approaches due to the limited number of traffic detectors, especially when considering about the identification of accidents and recording surrounding traffic conditions simultaneously for real-time prediction. Due to the limited data availability, two major deficiencies exist in the previous studies. One is that the model is usually constructed only based on a small region, such as a specific highway (Zou et al., 2016; Al-Najada and Mahgoub, 2017) or urban freeway (Hojati et al., 2013). The other shortcoming is that the information of past accidents stored in the offline database is always complete while only a small part can be acquired in time for a real-time prediction. For example, whether there is fatality or injury in accident impacts accident duration. Usually, such information is collected afterwards which cannot be used in real-time analysis. Thus, more efforts need to be paid to retrieve timely information and conduct online

* Corresponding author.
E-mail address: lrmin@tsinghua.edu.cn (R. Li).

https://doi.org/10.1016/j.aap.2020.105696
Received 20 March 2020; Received in revised form 1 June 2020; Accepted 14 July 2020
Available online 21 July 2020
0001-4575/ © 2020 Elsevier Ltd. All rights reserved.

prediction.

In contrast, crowdsourcing data (CD) from mobile applications (APPs) has become an emerging data source for transportation systems due to its abundancy. Crowdsourcing, which converts all participants to potential supervisors of the transportation system, can have information from all the users spreading over the entire road network. Users on the platform can share information immediately once they observed any changes on roads while others who may be involved in future can make decision beforehand based on the consistently updating traffic conditions. Obviously, the exploitation of CD comes at the cost of infidelity and uncertainty when interpreting it. Supplemented with credit systems, crowdsourcing is becoming a comprehensive but cheap way to collect traffic related data (Yang et al., 2015; Hasan and Ukkusuri, 2014; Zhang et al., 2018). Rashidi et al. (2017) and Chaniotakis et al. (2016) provided a more detailed discussion about how CD could be utilized in studying transportation issues.

Recent studies about the application of CD in transportation fields mainly lie in the usage of social media, such as Twitter and Facebook. By extracting the traffic event from textual data, locating the event and associating with auxiliary data, traffic conditions can be estimated more accurately (Wang et al., 2013). However, we note that User-Generated Crowdsourcing Data (UGCD) which has a more direct connection with traffic conditions remains unexplored. Several navigation systems such as Google maps, Waze, Inrix as well as Autonavi and Baidu maps in China provide users with an interface for reporting various traffic incidents in real time, along with their navigation services. According to this feature, accident information as well as surrounding traffic conditions can be obtained simultaneously in real time. It is applicable to develop a novel method to analyze TAPI using UGCD, especially, to predict when the nearby road segment has resumed normal operation rather than just the clearance time of an accident.

In this study, we novally introduce how to use UGCD in real-time TAPI prediction and propose a hierarchical scheme to perform sequential prediction. Our work utilizes the power of open crowdsourcing data in traffic accident analysis, thus the public is able to capture more detailed perturbation of traffic conditions in a cost-efficient way without inquiring the government department. UGDC can also provide additional coverage to existing sources of the traffic management system (Amin-Naseri et al., 2018), which facilitates more traffic participants. Moreover, this work explores the potential of using advanced Artificial Intelligence model in comprehensively predicting accident impacts. The remainder of this paper is organized as follows: Section 2 is a brief literature review on traffic accident impact prediction and the emerging usage of crowdsourcing data in transportation field. Section 3 gives a detailed introduction and explanation of the data used. Section 4 presents the hierarchical model and embedded machine learning algorithms in this study, and Section 5 shows the numerical results with analysis. Finally, Section 6 summarizes the major findings of this study.

2. Literature review

2.1. Traffic accident impact prediction model

Traffic accident impact is usually measured by traffic accident duration, which is typically subdivided into 4 sections (Nam and Mannering, 2000), including detecting/reporting time, preparing/dispatching time, travel time and clearance time. Studies that focus on traffic accident duration, to a large degree, only consider specific components within overall duration, especially the reported time difference between occurrence and clearance.

Since the life circle of accident is a good indicator of its impact, probabilistic distribution analysis was used to characterize the evolution of accidents for decades solely based on accident information. Jones et al. (1991); Qi and Teng (2008) and Chung (2010) used a log-logistic distribution to fit the distribution of traffic accident duration. However, there were also studies such as Golob et al. (1987) and Chung and Yoon (2012), revealing that the distribution can be much better described by a log-normal distribution, while Hojati et al. (2013) and Alkaabi et al. (2011) found that a Weibull distribution fits best. However, the scalability of such analysis is greatly limited by the adopted dataset. Researchers then turned to use statistical models to describe the relationship between accident duration and other related factors. Among all the proposed statistical models, regression model is the most basic one, started by linear regression (Cohen and Nouveliere, 1997; Khattak et al., 2012) which considered the duration as a linear combination of different factors. Later, Wang et al., 2013; Agarwal et al. (2016) expanded regression models and combined the merits of several different models. The other representative class of models is the survival analysis/hazard-based model (Qi and Teng, 2008; Chung, 2010; Li et al., 2015): a parametric accelerated failure time (AFT) model that is widely used on different traffic duration time phases to figure out the impact factors; however, different results are achieved due to the differences in datasets and regions (Li et al., 2018). Zou et al. (2018) used copula approach to jointly analyze incident clearance and response time. The results showed that the proposed copula model can better estimate conditional survival probability of clearance time than AFT models.

As the amount and variety of data generated in transportation systems grow explosively, not only the impact of each factor is appealing to researchers, but also the relationship among all variables as well as the structure of the model itself remains uninvestigated. To harness the massive data with unknown pattern, machine learning algorithms realize the data fusion and bridge the gap when corresponding the inputs to output without exogenous assumptions. Thus, lots of machine learning algorithms have been implemented to simulate human learning activities and solved a lot of problems with high accuracy. The typical machine learning methods used in traffic accident duration prediction include the following: (1) tree models: tree models characterize the nonlinear structure of model and output the average duration of accidents with similar characteristics, which can give a good accuracy; however, outliers in the input dataset will largely influence the results. Ma et al. (2017) proposed an efficient gradient-boosting decision tree model for prediction by using a threshold of 15 min. In comparison to traditional models and other methods including RF, SVM and back-propagation-neural network, this model is superior in both long-lasting and short-period incident prediction. (2) Artificial neural network (ANN): The ANN approach is a data-driven, self-adaptive and nonlinear methodology. Wei and Lee (2007) built a data fusion model using ANN techniques with 1 hidden layer and obtained MAPE under 40% mostly. Yu et al. (2016) compared the performance of ANN and SVM and concluded that SVM model has a comprehensively better performance despite of some long duration cases. (3) Hybrid models: deviation within these simple models activates researchers to employ a hybrid model for a more exhaustive prediction in recent years. Kim and Chang (2012) developed a hybrid model, which combines a tree model, a logit model and a Bayesian classifier together. Lin et al. (2016) proposed a combined M5P tree and hazard-based duration model. This hybrid model achieves a lower MAPE and identifies the significant variables much more easily than a single model. Shang et al. (2019) used Bayesian Optimization Algorithm to optimize parameters of RF while the relevant features are calculated by Neighborhood Components Analysis. Kuang et al. (2019) modeled the relationship of different features by a cost-sensitive Bayesian network and classified accidents according to its severity. A weighted K-nearest neighbor was then applied for duration prediction. Fu et al. (2019) considered traffic incident duration as a multi-task learning problem and proposed a spatiotemporal feature learning framework.

2.2. Real-time measurements and time series analysis used in duration prediction

As a traffic accident evolving, more information changes from unknown to known and thus could be added into prior consideration; that is, an up-to-date replenishment of information in model can result in a better prediction. To better encode the changing environment, either a real-time indicator or a time series model can be beneficial. Khattak et al. (1995) split the whole process into 10 phases. Only basic factors such as time, location and weather were obtained from reports in the first phase, which led to a blurry classification of incidents. Details about how one incident took place, how the operational conditions were around that spot and other descriptions could be added into the analysis in the following phases. Accuracy would increase after accumulation and correction of all data from different phases. Wei and Lee (2007) created two adaptive ANN-based models. The first one is used to forecast the duration at the first time of detection or report, while the other one includes multiperiod updates after the incident notification. Pereira et al. (2013) developed a sequential model which can consistently generate prediction updates whenever new text information is received while taking into account the elapsed time. Li et al. (2015) proposed a time-dependent mixture model which performs better than a model only with initial information. A reasonable structure in time series combined with timely detection of new information can make sequential prediction powerful. Shi and Abdel-Aty (2015) provided a real-time congestion measurement based on Big Data which could demonstrate the temporo-spatial change of congestion patterns. Both direct and indirect congestion indicators were found to have significant impact on rear-end crashes. Since different real-time indicators might be correlated with each other, Shi et al. (2016) took multicollinearity among independent variables into consideration and the performance of model was further improved by using Bayesian ridge regression to deal with the issue. Ghosh et al. (2018) provided updated prediction based on real-time streaming data by creating adaptive feature subsets based on the availability.

2.3. Crowdsourcing data used in transportation study

The data in most previous studies is from traffic accident reporting system in local traffic accident monitoring centers or emergency response agencies. Because the traffic conditions of different cities and regions have changed greatly over time, researchers cannot increase the sample size of the research objects by simply mixing a large amount of historical data with diverse background. In response to this situation, CD provides a significant advantage in that a large amount of data can be collected in a short time without specific gathering environment required. Rashidi et al. (2017) presented a bibliometric analysis with a focus on applications of social media data in modelling travel behavior, including travel demand modelling, mobility behavior, individuals' activity pattern, assessing public transport, traffic condition and incidents. They pointed out that the low acquisition cost and increasing amount made these data sources appealing. Nonetheless, special caution is required in using such data due to high extraction cost and sampling bias. Beheshti-Kashi et al. (2018) identified a list of textual sources in transportation which divides highly valuable user opinions into 3 categories, including social media based sources, traditional report sources and intra-organizational sources. Regarding to applications related with traffic accidents, Nguyen et al. (2016) carried out a detailed analysis of using Twitter to monitor traffic flow, which could be an ideal method of traffic incident detection. 5000 filtered tweets are labeled as either relevant or nonrelevant when training the machine learning models. Their model can not only advance the incident detecting time in comparison to Transport Management Centre (TMC) log time, but also discover some incidents that are not reported to TMC. Gu et al. (2016) primarily used natural language processing and several classifiers to extract and filter useful information from original text posted by users. Applications in different regions reveal great potential by covering existing reported incidents with just a small sample of tweets. UGCD has also been explored in accident analysis in recent years. Amin-Naseri et al. (2018) discovered that UGCD can provide additional coverage of accidents with low false alarm to conventional traffic management system. Timely reporting has also been found compared to probe-based alternative. Perez et al. (2018) extracted accident reports from Waze. They identified the repetitive reports according to road safety theory and obtained the patterns using clustering algorithm.

Comprehensively, the internet, associated with a social platform, can provide first-hand data without a chain of trivial reporting processes. Most of the previous work focuses on incident detection from social media; the subsequent impact, which extends to a sequential detection, is rarely mentioned, probably due to the lack of corresponding real-time traffic flow status data, or possibly for other reasons. To the best of our knowledge, traffic duration prediction has not been performed using UGCD up to now.

3. Data preprocessing and problem definition

3.1. Data description

The data used in this paper was collected from an anonymous navigation system. Compared to social media, such as Twitter, navigation system is highly related to the transportation system so that it provides a perfect interface for CD and traffic condition. In this study, real-time accident information and surrounding traffic conditions can be obtained by accident report and congestion level estimation functions of navigation system respectively.

The procedure of reporting accident to the navigation system is quite simple. Users can depict a traffic accident by selecting accident type, lane location from a preset framework while the time and location can be automatically detected. This kind of preset framework makes the report simple and efficient and can provide data in a unified format, but textual descriptions and photos can also be attached for details. The reported information will appear on the map as an icon, so that other users are informed about the change on road and can get further information by clicking on the icon. After a certain period, the icon will disappear from the map due to the expiration.

In this dataset, traffic conditions are obtained from the feedback provided by floating cars every 5 min to the granularity of road segment. However, relative congestion level usually counts more than absolute speed in reality when depicting the traffic status. In typical navigation systems, a 4-level traffic status defined by congestion delay index with color indicators is used. Congestion delay index I is given by Eq. (1):

$$I = \frac{v_{\text{free}}}{v} \tag{1}$$

where v is the current average speed of the investigated road segment, and v_{free} is the corresponding free flow speed. In this study, the boundaries for different levels are defined as,

Level 1: when $1 \le I < 1.5$, the segment is considered as uncongested, indicated by a green color (G);
Level 2: when $1.5 \le I < 2$, the segment is considered as slow moving, indicated by a yellow color (Y);
Level 3: when $2 \le I < 4$, the segment is considered as congested, indicated by a red color (R);
Level 4: when $I \ge 4$, the segment is considered as severely congested, indicated by a crimson color (C).

Since different users may report the same accident to the platform which is encouraged for credential purpose, repetitive records are stored. After filtering the information of all accident reports in Beijing,

including both urban and suburban areas during 2017, we finally screen out a list of 13,338 unique accident reports. Matching the reports with the traffic condition data by geocoding, we obtain a comprehensive description of an accident and the corresponding traffic conditions. To further capture other potential factors that may affect TAPI, we gathered more data about weather and air quality during the selected traffic accident period as well as the temporospatial properties of road segments.

3.2. A comprehensive definition for TAPI and accident types

Conventionally regarded as accident duration, TAPI, in fact, has a more sophisticated evolution. Not only the accident itself that matters, the entire recovering process should also be taken into account. More specifically, the congested level will show a unimodal trend, first experiencing a growing stage after an accident occurrence and then followed by a diminishing stage until traffic flow returns to normal. Accident may be cleared in either stage but TAPI will last much longer and cause lasting impact on roads. Moreover, congestion of different levels needs to be considered separately and differently. Based on the subdivision of traffic status, TAPI can be depicted comprehensively by the following 2 sets of value: (1) the most congested level that traffic condition will reach at the end of growing stage. (2) The duration from the reported start time to the end of each congestion level within the diminishing stage. Combining these two factors, the accidents can be classified into 4 types according to the number of levels reached consequently, as shown in Fig. 1. Type 4 is the worst case which means that after the accident happens, the traffic becomes increasingly congested until it reaches Level 4 and then congestion dissipates gradually. In contrast, Type 1 accident almost has no impact on the road conditions.

T_{start} stands for the start time of an accident. Ideally, congestion status may reach the peak after a period of fluctuation and slowly return to uncongested after the clearance of traffic accident. $N_4 = 1$, if the congestion reaches the level of severely congested, and $N_4 = 0$, otherwise. Similarly, $N_i (i = 1, 2, 3)$ indicates whether the traffic ever reaches level i. During the diminishing stage, $T_i (i = 1, 2, 3)$ is the first time back to level i. Although road conditions may fluctuate back and forth between different levels during the clearance of a traffic accident, we use the first occurrence of each congestion level in diminishing stage to represent the change of status in this study. It's because the first occurrence time shows the least time required for recovery while the following fluctuation has a higher probability to be caused by reasons other than accident. Therefore, we can define $T_{toi} = T_i - T_{start} (i = 1, 2, 3)$ to measure the exact duration of recovery to a specific level.

Taking the actual data as an example, all 4 types of accident congestion progress can be observed and examples are shown in Fig. 2.

4. Methodology

4.1. Hierarchical TAPI predicting scheme

Based on 4 accident types, we proposed a hierarchical TAPI predicting model by combining the prediction of the most congested condition and the duration of consequent congestion levels. At the very beginning of an accident, only basic information is known, a qualitative prediction about the most congested level is performed in the order of severity. When a level is predicted to happen, the duration prediction is activated for all levels below. As time goes by, more information about the accident and its consequent traffic conditions will be known. If the most congested level in reality is consistent with former prediction, new information such as when the former congestion level ends is added into the model, more accurate prediction for the following TAPI can be performed. Otherwise, the prediction of the most congested level is revised and the following prediction is performed consequently. The prediction process can be illustrated as Fig. 3.

Take the accident shown in Fig. 2d as an example, the hierarchical scheme will ideally make predictions in the following manner: (1) given the initial information of accident and associate environment factors, use a binary classification to predict whether Level 4 will occur; (2) Level 4 is predicted to happen. Thus, this accident is considered to be a type 4 accident and duration prediction for $T_{toi} (i = 1, 2, 3)$ is activated. Predicted value can be used as a reference for traffic guidance; (3) 65 min later, $N_4 = 1$ is detected which indicates a right classification at step 1. No additional adjustment is needed with a correct classification; (4) 85 min later, T_3 is detected. Add the known value T_{to3} into model and calibrate the predicted duration of lower levels $\widehat{T}_{toi} (i = 1, 2)$; (5) 110 min later, T_{to2} is detected. Further add the known value T_{to2} into model and calibrate the predicted duration $\widehat{T}_{to1}$; (6) 120 min later, T_{to1} is detected. Traffic flow returns to uncongested and the TAPI of this accident is fully depicted in the procedure.

4.2. Embedded algorithm

Since the scheme does not depend on any prior assumption or structure, it is totally data-driven and we can embed any algorithms for prediction. For comparison and illustration purpose, we use three common algorithms, random forest, support vector machine and neural network, in the following content.

RF is an ensemble model with each tree model catching part of the nonlinearity between TAPI and a subset of factors. Since the inner relationship among all variables and their different categories is difficult to interpret in real life, RF acts as a combination of multiple decision trees to find the most possible result without a bunch of assumptions. It is also effective in preventing overfitting to training data.

SVM constructs a hyperplane that can be used for classification. If linear inseparability occurs when dividing the space, segmentation

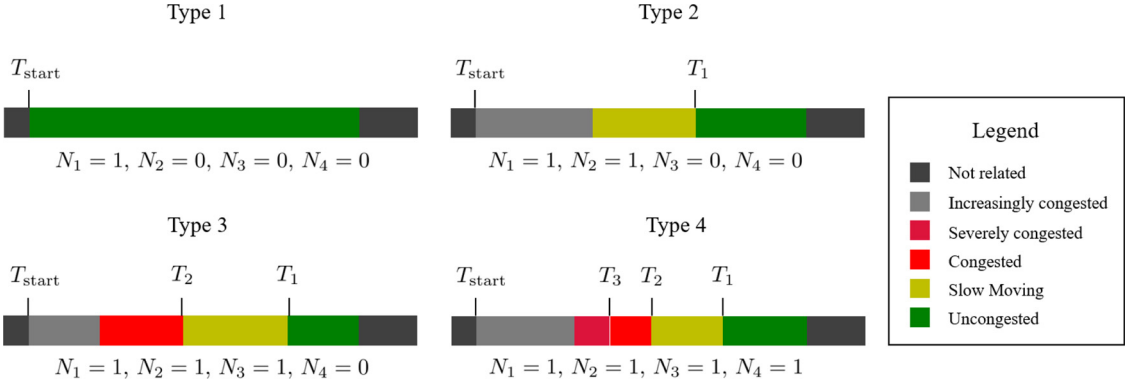
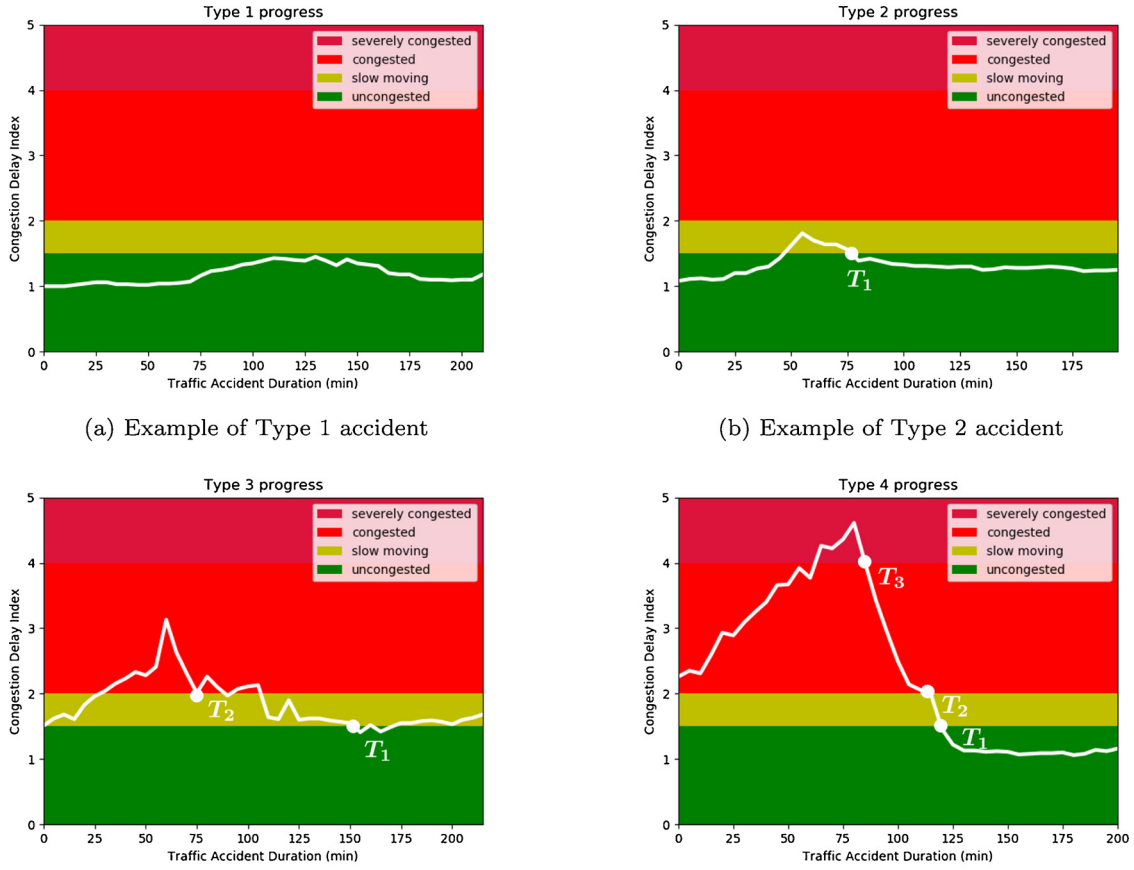


Fig. 1. Traffic conditions after accidents.



(a) Example of Type 1 accident

(b) Example of Type 2 accident

(c) Example of Type 3 accident

(d) Example of Type 4 accident

Fig. 2. Illustration of different types accident.

could be finished by mapping all points to a higher-dimensional plane. By using a kernel function to map data from lower-dimensional space to higher-dimensional space, simplified calculation can be operated directly in the mapped space, thus making the application of the algorithm feasible. In this study, a radial basis kernel function shown as Eq. (2) is chosen:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2} \tag{2}$$

NN is composed of input layer, hidden layer and output layer and weights between each layers are calculated. Generally, the output variable can be written as Eq. (3):

$$y = g_o(\mathbf{w}_o^T g_h(\mathbf{w}_h^T \mathbf{x} + \mathbf{b}_h) + \mathbf{b}_o) \tag{3}$$

where x is the input variables, w_h and b_h represent weights and bias between input layer and hidden layer, w_o and b_o represent weights and bias between hidden layer and output layer. g_h and g_o are activation functions which introduce nonlinearity into NN. More complexity can be captured by increasing the size or the depth of the layer.

5. Numerical results

5.1. Model inputs

We extract 3 kinds of crucial factors that may potentially impact TAPI, including temporospatial features, weather conditions and reported accident information. Furthermore, 8 independent variables could be split into 36 dummy variables, and the corresponding statistics of the happening probability and duration of each congestion level are shown as Table 1.

From Table 1, we can get some primary conclusions. (1) The probability to have high congestion levels increases as the road class becomes higher and severely congested after an accident occurs twice as likely on urban roads as rural roads. However, average duration is not much different in both cases. (2) Workday has significantly larger probability but similar duration to have congestion than holiday. (3) Peak hour increases the probability to have congestion while morning peak hour has further impact on the duration. (4) Middle lane has a higher probability to cause congestion compared to side lanes. (5) Foggy and sleet have positive impact on probability and duration respectively. (6) Congestion level before accident shows the largest variation among all factors.

5.2. Model evaluation

To evaluate our models, the whole dataset is split into 2 parts: 70% as the training set and 30% as the test set. We use a binary classifier based on RF and SVM respectively, to predict which level is the most congested level after a specific accident. To investigate the performance of these qualitative prediction models, two criteria: accuracy (ACC) and precision (PPV) are computed. The results are shown in Table 2.

ACC shown as in Eq. (4) is chosen to evaluate the overall predictive power of our models,

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \tag{4}$$

PPV shown as in Eq. (5) is chosen to evaluate the credibility of the predicted positive events which represent the occurrence of congestion,

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

On one hand, three models have similar performance with ACC over

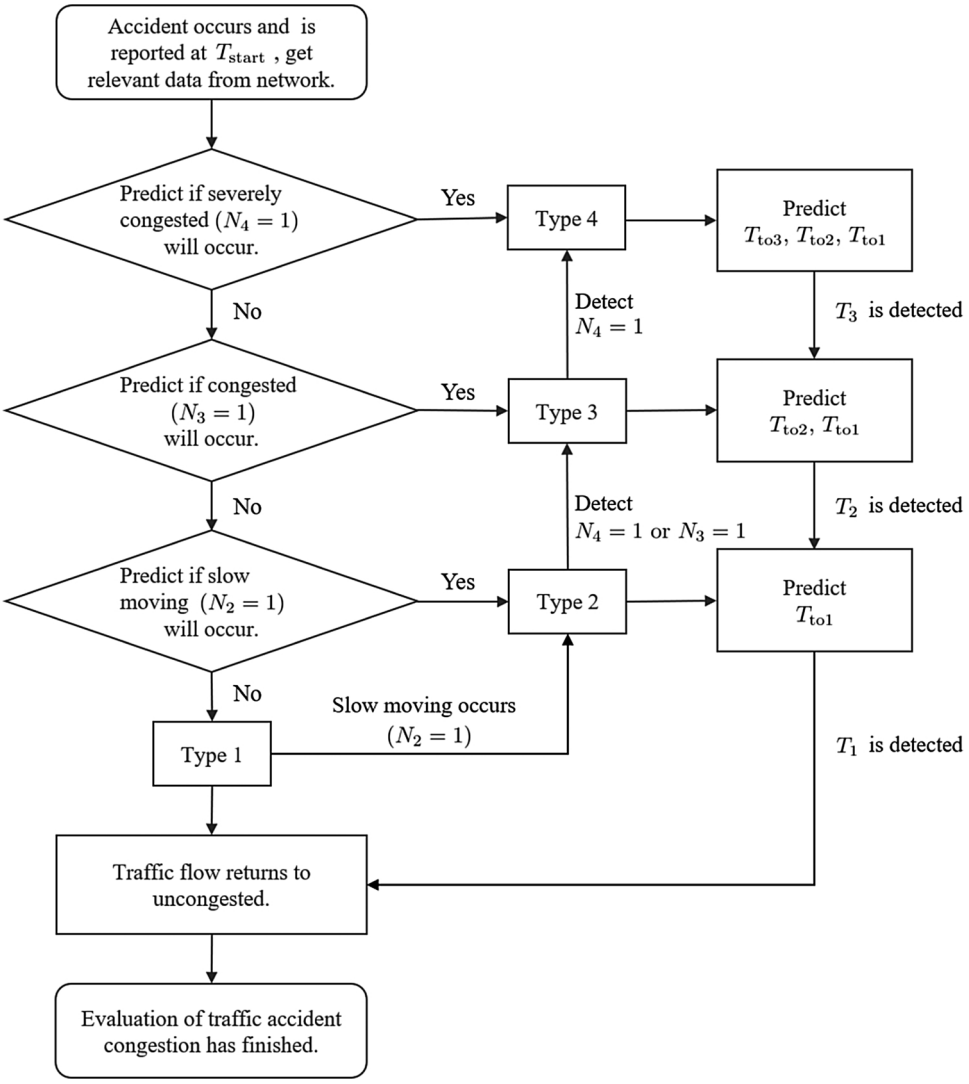


Fig. 3. A schematic diagram for hierarchical TAPI predicting model.

0.73 among all levels while NN slightly outperforms the other two ML models. However, the skewed fact should be noticed when predicting N_4 and N_2. According to the dataset, severely congested rarely happens. If we adopt a naïve algorithm to keep predicting 0 for all N_4, the model already achieves a high ACC. Similarly, always predicting 1 for N_2 also achieves a high ACC. But this kind of algorithm is not applicable since it does not serve the classification purpose. Thus, on the other hand, more indicators such as PPV are needed as a supplement. PPV is over 0.81 among all models where SVM has slightly better performance. The high PPV implies a high probability of congestion occurrence if the model gives a positive prediction and thus is quite informative to traffic participants. Note that, as the congestion level increases, PPV also increases and even approaches 1 when predicting the severely congested level. Overall, our model can have a good prediction with high ACC and PPV which makes a good start for TAPI prediction.

For the duration prediction of each level, we start with only relying on the initial information of accidents. The predicted results without updates are shown in Fig. 4.

All the models have a concentrated prediction, that is, more accurate prediction can be achieved at moderate duration while obvious deviation is shown at larger and smaller values. As the degree of congestion decreases, the predicted values show a positive bias. In addition, it is worth noting that NN model shows extremely high accuracy in T_{to3} prediction and a positive bias in other cases when congestion does not last long.

Considering the sequential prediction, we obtain more information as the accident evolves which can also be added into the model. Specifically, if T_{to3} is detected in reality, it can be considered as an independent variable when predicting T_{to2} and T_{to1} in Type 4 accidents. Likewise, actual T_{to2} can be utilized to predict T_{to1} in both Type 4 and Type 3 accidents. The results of models with updates are as shown in Fig. 5.

Comparing the results without updates in Fig. 4 and with updates in Fig. 5, we can see that sequential prediction can effectively improve all three models.

From the perspective of numerical criteria, mean absolute percentage error (MAPE) and root mean square error (RMSE) are used to measure the accuracy of these quantitative models as shown in Tables 3 and 4.

MAPE shown as in Eq. (6) is chosen to assess the overall performance of models,

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{t}_i - t_i}{t_i}\right| \times 100\% \tag{6}$$

RMSE shown as in Eq. (7) is chosen to understand the absolute value of deviation,

Table 1
Summary of occurrence probability and duration for each congestion level.

Category	Variable	Dummy variable	Average probability (%)			Average time (min)		
			N_4	N_3	N_2	T_{to3}	T_{to2}	T_{to1}
Temporo-spatial feature	Classification of roads	a_1 highway	18.59	60.22	80.67	55.60	71.60	80.30
		a_2 1st class road	12.26	51.42	79.25	58.27	74.86	77.20
		a_3 2^{ed} class road	9.58	39.59	68.32	50.77	63.42	70.50
		a_4 3rd class road	10.31	40.36	62.78	35.00	62.33	68.50
		a_5 4th class road	6.91	31.48	58.93	49.58	64.97	69.50
		a_6 expressway	29.79	73.65	88.73	50.53	68.95	78.19
		a_7 arterial	25.71	64.88	82.42	49.23	65.92	75.46
		a_8 2^{ed} trunk	21.34	57.10	76.57	51.05	68.08	75.46
		a_9 branch	17.11	53.22	74.71	54.87	68.91	76.66
		a_{10} street	18.32	52.01	75.46	41.70	61.44	68.52
	Workday or not	b_1 workday	23.92	61.71	80.32	50.64	67.98	76.08
		b_2 holiday	18.59	57.55	78.06	49.36	65.70	74.97
	Peak hour or not	c_1 morning peak	25.18	64.60	81.89	55.36	73.07	80.14
		c_2 evening peak	26.82	65.98	83.42	48.03	67.44	76.49
		c_3 off peak	20.10	57.16	77.57	48.78	64.69	73.55
	Lane location	d_1 left side	20.45	57.85	78.05	49.28	66.60	74.57
		d_2 middle	25.24	64.07	81.82	49.70	67.23	76.21
		d_3 right side	22.75	60.81	79.83	51.75	68.34	76.56
Weather condition	Weather	e_1 sunny/cloudy	22.27	60.79	79.61	50.35	67.17	75.66
		e_2 light rain	23.39	60.49	80.57	50.79	67.72	75.94
		e_3 moderate rain	24.00	60.00	80.00	47.33	70.42	76.13
		e_4 storm	23.72	63.46	77.56	50.14	68.69	80.25
		e_5 foggy	28.24	70.59	87.06	47.92	69.67	76.89
		e_6 sleet	36.36	50.91	63.64	58.50	83.57	83.71
	Pollution level	f_1 good	22.37	60.40	79.62	50.63	67.47	75.75
		f_2 lightly	24.06	62.05	80.50	49.29	67.21	76.09
		f_3 moderately	23.22	61.48	79.51	53.76	69.96	76.67
		f_4 severely	22.50	62.50	81.88	45.14	66.75	75.19
Accident information	Accident type	g_1 breakdown	23.11	60.71	79.97	49.03	69.01	77.26
		g_2 scratch	24.49	62.30	82.03	48.43	65.79	75.32
		g_3 pileup	21.24	60.08	78.98	52.27	67.96	75.83
		g_4 other	22.19	55.01	75.39	53.00	67.93	74.19
	Congestion level before accident	h_1 uncongested	10.83	38.57	63.07	58.80	68.11	70.19
		h_2 slow moving	14.04	68.21	99.71	55.15	63.17	70.61
		h_3 congested	35.90	99.71	99.89	46.73	65.96	85.78
		h_4 severely	99.03	99.46	99.68	45.39	77.77	87.64

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{t}_i - t_i)^2} \tag{7}$$

Same as the illustration in Figs. 4 and 5, RF and SVM models without updates has a MAPE over 27% besides the prediction of T_{to3} by RF model while NN models reaches below 20% in some cases. The accuracy is limited by two main reasons: (1) the granularity of detected traffic condition is 5 min which has restrictions on the original data accuracy and continuity. (2) MAPE is also influenced by the scale of the duration itself since it mainly describes the relative error to its absolute value. Small denominator will magnify errors between the actual value and predicted value. The prediction of T_{to3} by RF model reduces large MAPE from (2) by given good estimation on small duration. By adding the actual hitting time of previous congestion level can make the model reaches MAPE lower than 10%. In the version with updates, RF has better predicting power than SVM and NN.

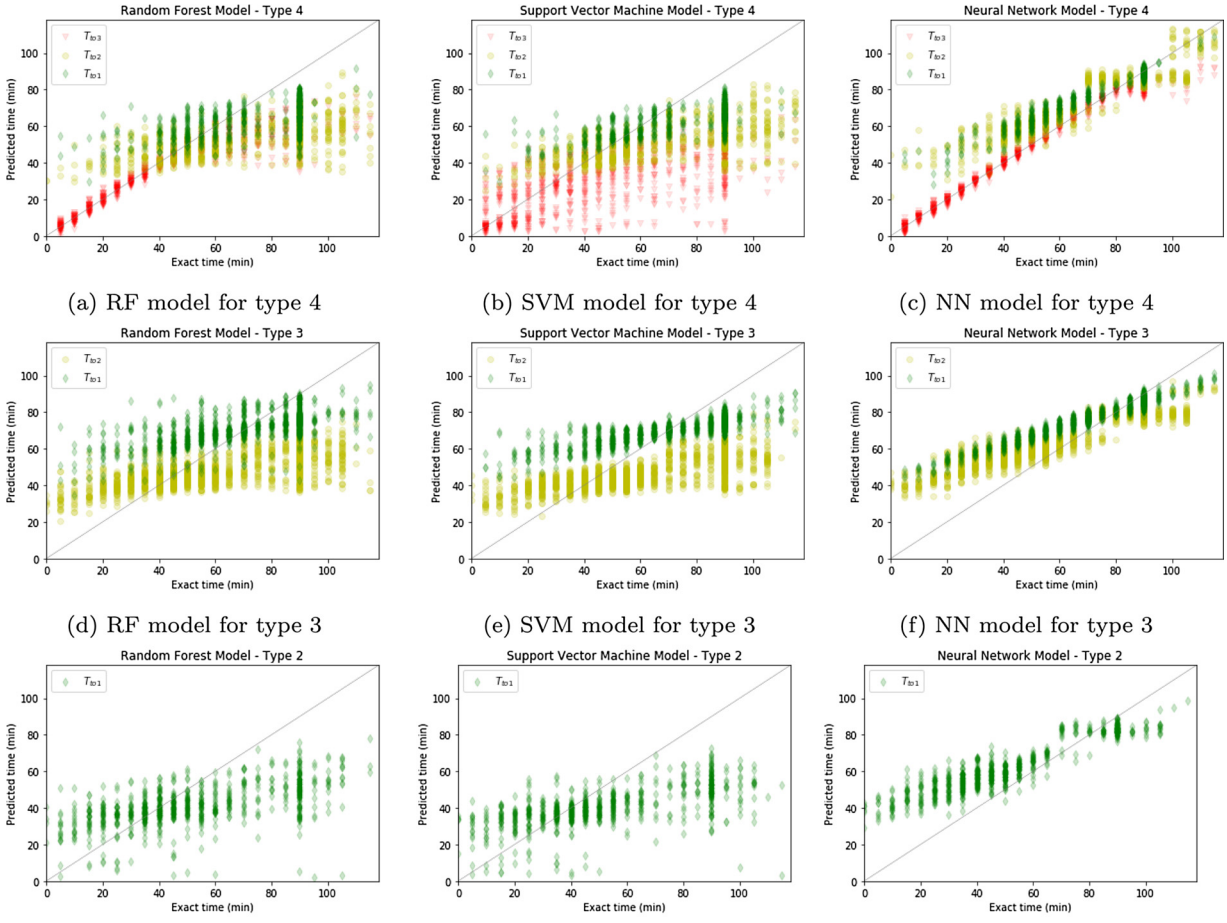
To further show the prediction performance on specific accident rather on the average level, we define the following assessment criteria on absolute prediction difference in Table 5, and corresponding proportion is shown in Fig. 6.

Despite the results of average criteria is not good enough, we find about 50% accidents can be highly accurate predicted even only with the initial information. More interestingly, NN shows great advantage over other two models when considering about the absolute difference. Meanwhile, Fig. 6a shows that inaccurate predicting also counts a lot which further implies the initial information about accident is not adequate to characterize the entire TAPI. Improvement from Fig. 6a and b comes from the increasing number of highly accurate predicting. Thus, it is reasonable to assume if we can get more detailed description of the accident and its surrounding traffic condition, the model can finally give an ideal prediction.

Since no previous study has used the same kind of data as our work,

Table 2
Results of predicting the most congested level.

	Actual value	RF				SVM				NN			
		0	1	ACC	PPV	0	1	ACC	PPV	0	1	ACC	PPV
N_4	0	3091	6	0.8391	0.9560	3095	2	0.8386	0.9924	3077	24	0.8423	0.9245
	1	638	267			644	261			607	294		
N_3	0	1226	335	0.7311	0.8354	1283	278	0.7309	0.8552	1128	409	0.7406	0.8178
	1	741	1700			799	1642			629	1836		
N_2	0	71	735	0.8051	0.8109	91	715	0.8048	0.8140	89	702	0.8096	0.8178
	1	45	3151			66	3130			60	3151		

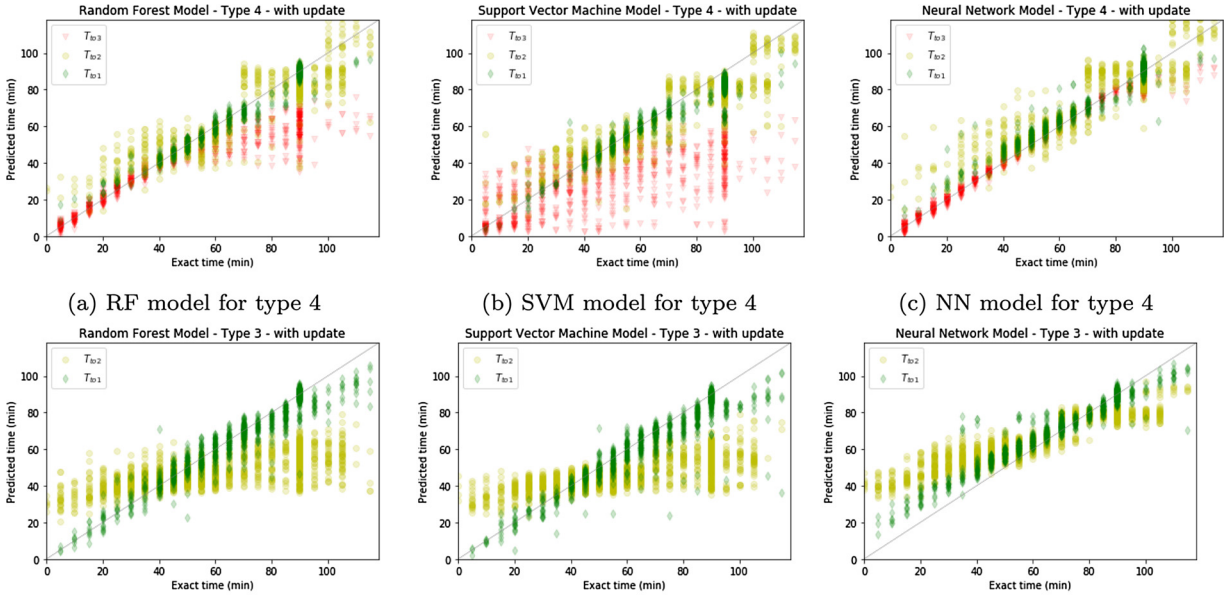


(a) RF model for type 4 (b) SVM model for type 4 (c) NN model for type 4

(d) RF model for type 3 (e) SVM model for type 3 (f) NN model for type 3

(g) RF model for type 2 (h) SVM model for type 2 (i) NN model for type 2

Fig. 4. Results of two models without updates.



(a) RF model for type 4 (b) SVM model for type 4 (c) NN model for type 4

(d) RF model for type 3 (e) SVM model for type 3 (f) NN model for type 3

Fig. 5. Results of two models with updates.

Table 3
Results of predicting congestion duration.

		RF		SVM		NN	
		MAPE (%)	RMSE (min)	MAPE (%)	RMSE (min)	MAPE (%)	RMSE (min)
Type 4	T_{to3}	13.43	21.06	48.77	33.53	3.10	8.09
	T_{to2}	32.78	34.06	33.04	34.74	17.40	10.86
	T_{to1}	30.51	30.59	30.29	30.93	13.59	10.04
Type 3	T_{to2}	39.82	25.26	39.27	25.56	42.86	14.85
	T_{to1}	27.95	21.41	27.48	20.97	21.35	13.42
Type 2	T_{to1}	39.22	24.08	36.44	24.23	53.80	15.68

Table 4
Results of predicting congestion duration with updating information.

		RF		SVM		NN	
		MAPE (%)	RMSE (min)	MAPE (%)	RMSE (min)	MAPE (%)	RMSE (min)
Type 4	T_{to2}	7.62	9.55	11.60	13.26	9.33	7.62
	T_{to1}	3.05	5.50	8.74	12.79	5.59	6.05
Type 3	T_{to1}	5.43	8.51	6.42	10.62	8.70	7.39

Table 5
Predicting assessment criteria.

Absolute difference (min)	Assessment
≤10	Highly accurate predicting
10–20	Good predicting
20–30	Reasonable predicting
> 30	Inaccurate predicting

we compare our model performance with some of the work which considered dynamic prediction in Table 6. The lower bound of MAPE has been greatly improved by our work which shows the great potential predictive power of CD in accident duration.

5.3. Influencing factor analysis

Since different independent variables have different effects on the prediction, it is meaningful to identify the factors that have great impact for further research. Take RF model in the previous section for analysis, top 5 features and their corresponding importance scores are shown in Table 7.

For qualitative models, we find h congestion level before accident dominates the prediction. It is not surprising since the most congested level that the accident may reach is based on the former traffic condition. As a supplement, c_3 whether it's peak hour and a_6 whether the accident occurs on expressway also affect. h congestion level before accident is still important in duration prediction without updates, but other factors also have something to do with the prediction. It can be clearly seen in Table 7, a classification of roads, c peak hour or not, d lane location and g accident type have prominent contributions in the model while weather condition almost has no impact. When it comes to the sequential prediction, newly added variables will dominate the prediction instead.

Table 6
Performance comparison with previous study.

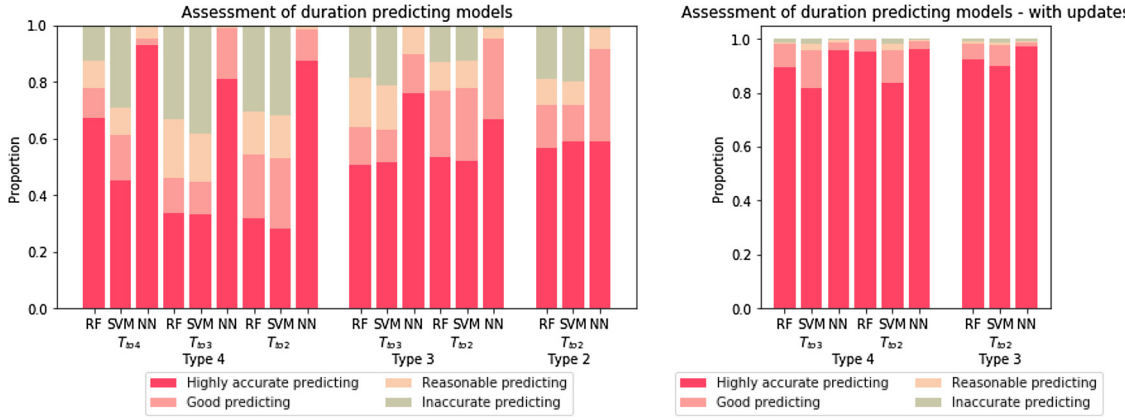
Literature	MAPE range
Wei and Lee (2007)	35–45%
Pereira et al. (2013)	40–100%
Li et al. (2015)	45.4–185.7%
Ghosh et al. (2018)	20–100.9%
Fu et al. (2019)	37.16–96.38%
Our work	5.5–53.8%

Table 7
Top 5 feature importance of different RF models.

Random forest model	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Type 4 – N_4	h_4 0.6733	h_3 0.1500	c_3 0.0235	b_2 0.0182	d_2 0.0147
Type 4 – N_3	h_3 0.5163	h_4 0.2253	c_3 0.1049	b_2 0.0311	d_2 0.0147
Type 4 – N_2	h_2 0.3777	h_3 0.2823	h_4 0.1764	a_6 0.0433	a_5 0.0237
Type 4 – T_{to3}	h_4 0.1269	c_3 0.1059	d_3 0.1035	g_2 0.0926	a_7 0.0815
Type 4 - T_{to2}	a_7 0.1014	h_3 0.0971	g_2 0.0967	c_3 0.0846	g_3 0.0775
Type 4 – T_{to2} – with updates	T_{to3} 0.2906	a_7 0.0730	a_6 0.0643	c_3 0.0582	c_2 0.0497
Type 4 – T_{to1}	h_4 0.2492	a_8 0.1593	a_7 0.1391	c_2 0.0913	h_3 0.0646
Type 4 – T_{to1} – with updates	T_{to2} 0.8939	T_{to3} 0.0902	c_3 0.0135	a_7 0.0024	
Type 3 – T_{to2}	c_3 0.1263	g_3 0.0906	d_2 0.0869	g_2 0.0866	h_3 0.0778
Type 3 – T_{to1}	h_3 0.1181	h_2 0.1047	g_3 0.0968	c_3 0.0951	b_2 0.0767
Type 3 – T_{to1} – with updates	T_{to2} 0.3613	d_3 0.0779	a_7 0.0615	c_3 0.0524	g_3 0.0497
Type 2 – T_{to1}	d_3 0.1034	c_3 0.1012	g_2 0.0865	d_2 0.0779	g_3 0.0759

6. Conclusion

Emerging crowdsourcing data provides a new source for the analysis of traffic accidents, and data fusion with real-time traffic condition information derived from floating car data provides a basis for this



(a) Assessment without updates (b) Assessment with updates

Fig. 6. Absolute difference assessment

work; this paper applies the machine learning method to predict the duration of different traffic condition states after traffic accidents and considers the updating effect by adding newly acquired data to the prediction. On a real-time basis, it ultimately aims to depict the evolution of surrounding traffic conditions after the occurrence of a traffic accident. On an offline basis, it can be used to assess the robustness of transportation system when traffic accidents happen. Specifically, we first extract UGCD from navigation APPs provided by the users. After filtering and calibrating UGCD into a set of traffic accidents, we can consecutively generate the features for every accident. Based on the geographical information of the accidents, we map each accident to the corresponding traffic flow data with the same spatiotemporal feature which can also be obtained from navigation apps. Additional related information can also be extracted in order to explore other factors that might affect TAPI. Preprocessing steps such as screening of raw data, selection of independent variables, definition of dependent variables, etc. are carried out to describe accident information and congestion. Using the congestion delay index, which is commonly used in current traffic guidance systems, the TAPI is divided into four levels: severely congested, congested, slow moving and uncongested. Based on the occurrence of each traffic condition level, the accident can be divided into four types. Predicting the recovering duration of each level for every type of accident can give an overall description of the persistence of congestion. RF, SVM and NN algorithms are applied to model and predict the congestion state and duration. NN model has a better performance in most cases, especially when considering the absolute difference between predicting duration and actual duration Moreover, the precision can become significantly improved over time with updated information involved regardless of the exact algorithm that is embedded. This result has great reference value for the prediction of real-time traffic conditions, for it can effectively guide the traffic participants to avoid relevant congestion sections.

There are several limitations in this study, which can be improved in the future: (1) the model in this study only considers a small subset of factors that may impact TAPI and thus has a limited degree of accuracy. The accident information recorded by crowdsourcing data is not as comprehensive as data obtained by traditional methods. For example, the specific types of accidents, the number of lanes and vehicles involved, which may affect the congestion time, are not normally recorded in a crowdsourcing dataset. In the future, it is possible to consider increasing the information of the network reporting system so that the state of the accident can be more accurately described. (2) Due to the large differences in traffic conditions at different times in different regions, models made by using a certain data sample are not applicable to all road conditions. Portability of the enhanced model can be considered in the future. From this study, we can see the feasibility of using crowdsourcing data to predict congestion caused by traffic accidents. With further development and enrichment in the future, crowdsourcing data can provide more accurate support for post-accident congestion prediction by integrating data from diverse sources.

Author statement

Yunduan Lin: conceptualization, methodology, software, investigation, writing-original draft preparation, writing-reviewing and editing. Ruimin Li: funding acquisition, supervision, investigation, writing-reviewing and editing, data curation, investigation, validation.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge assistance with accident data and float car data from concerned institution. The research reported in this paper is part of the Project supported by the National Natural Science Foundation of China (71871123). The financial support is highly appreciated.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.aap.2020.105696.

References

Agarwal, S., Kachroo, P., Regentova, E., 2016. A hybrid model using logistic regression and wavelet transformation to detect traffic incidents. IATSS Res. 40 (1), 56–63.
Alkaabi, A.M.S., Dissanayake, D., Bird, R., 2011. Analyzing clearance time of urban traffic accidents in Abu Dhabi, United Arab Emirates, with hazard-based duration modeling method. Transp. Res. Rec. (2229), 46–54.
Al-Najada, H., Mahgoub, I., 2017. Real-time incident clearance time prediction using traffic data from internet of mobility sensors. Proceedings of the 2017 IEEE 15th Intl. Conf. on Dependable, Autonomic and Secure Computing, 15th Intl. Conf. on Pervasive Intelligence and Computing, 3rd Intl. Conf. on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) 728–735.
Amin-Naseri, M., Chakraborty, P., Sharma, A., Gilbert, S.B., Hong, M., 2018. Evaluating the reliability, coverage, and added value of crowdsourced traffic incident reports from Waze. Transp. Res. Rec. 2672 (43), 34–43.
Beheshti-Kashi, S., Buch, R., Lachaize, M., Kinra, A., 2018. Big textual data in transportation: an exploration of relevant text sources. Proceedings of the International Conference on Dynamics in Logistics 395–399.
Chaniotakis, E., Antoniou, C., Pereira, F., 2016. Mapping social media for transportation studies. IEEE Intell. Syst. 31 (6), 64–70.
Chung, Y., 2010. Development of an accident duration prediction model on the Korean freeway systems. Accid. Anal. Prev. 42 (1), 282–289.
Chung, Y., Yoon, B.J., 2012. Analytical method to estimate accident duration using archived speed profile and its statistical analysis. KSCE J. Civil Eng. 16 (6), 1064–1070.
Cohen, S., Nouveliere, C., 1997. Modelling incident duration on an urban expressway. In: Papageorgiou, M., Pouliezos, A. (Eds.), Transportation Systems 1997, Vols. 1–3. Pergamon Press Ltd, Oxford, pp. 297–301.
Fu, K., Ji, T., Zhao, L., Lu, C.-T., 2019. Titan: a spatiotemporal feature learning framework for traffic incident duration prediction. Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 329–338.
Ghosh, B., Asif, M.T., Dauwels, J., Fastenrath, U., Guo, H., 2018. Dynamic prediction of the incident duration using adaptive feature set. IEEE Trans. Intell. Transp. Syst. 20 (11), 4019–4031.
Giuliano, G., 1989. Incident characteristics, frequency, and duration on a high volume urban freeway. Transp. Res. Part A – Policy Pract. 23 (5), 387–396.
Golob, T.F., Recker, W.W., Leonard, J.D., 1987. An analysis of the severity and incident duration of truck-involved freeway accidents. Accid. Anal. Prev. 19 (5), 375–395.
Gu, Y.M., Qian, Z., Chen, F., 2016. From twitter to detector: real-time traffic incident detection using social media data. Transp. Res. Part C Emerg. Technol. 67, 321–342.
Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C Emerg. Technol. 44, 363–381.
Hojati, A.T., Ferreira, L., Washington, S., Charles, P., 2013. Hazard based models for freeway traffic incident duration. Accid. Anal. Prev. 52, 171–181.
Jones, B., Janssen, L., Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accid. Anal. Prev. 23 (4), 239–255.
Khattak, A., Wang, X., Zhang, H., 2012. Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays. IET Intell. Transp. Syst. 6 (2), 204–214.
Khattak, A.J., Schofer, J.L., Wang, M.H., 1995. A simple time-sequential procedure for predicting freeway incident duration. IVHS J. 2 (2), 113–138.
Kim, W., Chang, G.L., 2012. Development of a hybrid prediction model for freeway incident duration: a case study in Maryland. Int. J. Intell. Transp. Syst. Res. 10 (1), 22–33.
Kuang, L., Yan, H., Zhu, Y., Tu, S., Fan, X., 2019. Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted k-nearest neighbor. J. Intell. Transp. Syst. 23 (2), 161–174.
Li, R.M., Pereira, F.C., Ben-Akiva, M.E., 2015. Competing risks mixture model for traffic incident duration prediction. Accid. Anal. Prev. 75, 192–201.
Li, R.M., Pereira, F.C., Ben-Akiva, M.E., 2018. Overview of traffic incident duration analysis and prediction. Eur. Transp. Res. Rev. 10 (2), 13.
Lin, L., Wang, Q., Sadek, A.W., 2016. A combined m5p tree and hazard-based duration model for predicting urban freeway traffic accident durations. Accid. Anal. Prev. 91, 114–126.
Ma, X.L., Ding, C., Luan, S., Wang, Y., Wang, Y.P., 2017. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans. Intell. Transp. Syst. 18 (9), 2303–2310.
Nam, D., Mannering, F., 2000. An exploratory hazard-based analysis of highway incident

duration. Transp. Res. Part A Policy Pract. 34 (2), 85–102.
Nguyen, H., Liu, W., Rivera, P., Chen, F., 2016. Trafficwatch: real-time traffic incident detection and monitoring using social media. In: Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J.Z., Wang, R. (Eds.), Advances in Knowledge Discovery and Data Mining, PAKDD 2016, Pt I. Springer-Verlag Berlin, Berlin, pp. 540–551.
Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2013. Text analysis in incident duration prediction. Transp. Res. Part C Emerg. Technol. 37, 177–192.
Perez, G.V.A., Lopez, J.C., Cabello, A.L.R., Grajales, E.B., Espinosa, A.P., Fabian, J.L.Q., 2018. Road traffic accidents analysis in Mexico city through crowdsourcing data and data mining techniques. Int. J. Comput. Inform. Eng. 12 (8), 604–608.
Qi, Y., Teng, H.L., 2008. An information-based time sequential approach to online incident duration prediction. J. Intell. Transp. Syst. 12 (1), 1–12.
Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp. Res. Part C Emerg. Technol. 75, 197–211.
Shang, Q., Tan, D., Gao, S., Feng, L., 2019. A hybrid method for traffic incident duration prediction using boa-optimized random forest combined with neighborhood components analysis. J. Adv. Transp. 2019.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C: Emerg. Technol. 58, 380–394.
Shi, Q., Abdel-Aty, M., Lee, J., 2016. A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accid. Anal. Prev. 88, 124–137.
Skabardonis, A., Varaiya, P., Petty, K.F., Trb, 2003. Measuring Recurrent and Nonrecurrent Traffic Congestion. Freeways, High-Occupancy Vehicle Systems, and Traffic Signal Systems 2003: Highway Operations, Capacity, and Traffic Control. Transportation Research Board Natl Research Council, Washington, pp. 118–124.
Wang, S., He, L., Stenneth, L., Philip, S.Y., Li, Z., Huang, Z., 2013. Estimating urban traffic congestions with multi-sourced data. Proceedings of the 2016 17th IEEE International Conference on Mobile Data Management (MDM) 82–91.
Wei, C.H., Lee, Y., 2007. Sequential forecast of incident duration using artificial neural network models. Accid. Anal. Prev. 39 (5), 944–954.
Yang, F., Jin, P.J., Cheng, Y., Zhang, J., Ran, B., 2015. Origin-destination estimation for non-commuting trips using location-based social networking data. Int. J. Sustain. Transp. 9 (8), 551–564.
Yu, B., Wang, Y.T., Yao, J.B., Wang, J.Y., 2016. A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. Neural Network World 26 (3), 271–287.
Zhang, Z.H., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transp. Res. Part C Emerg. Technol. 86, 580–596.
Zou, Y., Henrickson, K., Lord, D., Wang, Y., Xu, K., 2016. Application of finite mixture models for analysing freeway incident clearance time. Transportmetrica A: Transp. Sci. 12 (2), 99–115.
Zou, Y., Ye, X., Henrickson, K., Tang, J., Wang, Y., 2018. Jointly analyzing freeway traffic incident clearance and response time using a copula-based approach. Transp. Res. Part C Emerg. Technol. 86, 171–182.