# Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements

Moinul Hossain, PhD
Associate Professor
Department of Civil and Environmental Engineering
Islamic University of Technology (IUT), Bangladesh

Mohamed Abdel-Aty, PhD, P.E.
Pegasus Professor
Department of Civil, Environmental, and Construction Engineering
University of Central Florida, USA

Mohammed A. Quddus, PhD*
Professor
School of Architecture, Building and Civil Engineering
Loughborough University
Ashby road, Loughborough, Leicestershire LE113TU, UK
Tel.: +44 (0)1509 228545
Email: m.a.quddus@lboro.ac.uk

Yasunori Muromachi, PhD
Associate Professor
Urban Design and Built Environment Graduate Major
Department of Civil and Environmental Engineering
School of Environment and Society
Tokyo Institute of Technology, Japan

Soumik Nafis Sadeek
Lecturer
Department of Civil Engineering
IUBAT-International University of Business Agriculture and Technology,
Bangladesh

*Corresponding Author

# Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements


## Abstract

Proactive traffic safety management systems can monitor traffic conditions in real-time, identify the formation of unsafe traffic dynamics, and implement suitable interventions to bring unsafe conditions back to normal traffic situations. Recent advancements in artificial intelligence, sensor fusion and algorithms have brought about the introduction of a proactive safety management system closer to reality. The basic prerequisite for developing such a system is to have a reliable crash prediction model that takes real-time traffic data as input and evaluates their association with crash risk. Since the early 21st century, several studies have focused on developing such models. Although the idea has considerably matured over time, the endeavours have been quite discrete and fragmented at best because the fundamental aspects of the overall modelling approach substantially vary. Therefore, a number of transitional challenges have to be identified and subsequently addressed before a ubiquitous proactive safety management system can be formulated, designed and implemented in real-world scenarios. This manuscript conducts a comprehensive review of existing real-time crash prediction models with the aim of illustrating the state-of-the-art and systematically synthesizing the thoughts presented in existing studies in order to facilitate its translation from an idea into a ready to use technology. Towards that journey, it conducts a systematic review by applying various text mining methods and topic modelling. Based on the findings, this paper ascertains the development pathways followed in various studies, formulates the ubiquitous design requirements of such models from existing studies and knowledge of similar systems. Finally, this study evaluates the universality and design compatibility of existing models. This paper is, therefore, expected to serve as a one stop knowledge source for facilitating a faster transition from the idea of real-time crash prediction models to a real-world operational proactive traffic safety management system.



**Keywords:** ITS, Real-time crash prediction model, design pathway, universal design requirements.

# Introduction

The concept of real-time crash prediction relates to the hypothesis that the probability of a crash occurring on a specific road section within a very short time window can be predicted using the instantaneous traffic dynamics (e.g. Lee et al., 2003a,b; Abdel-Aty et al., 2004; Pande and Abdel-Aty, 2005). The model built to serve the purpose is called a 'real-time crash prediction model' (RTCPM). This idea has potential to unlock the prospect of preventing some crashes that might have occurred otherwise. A number of studies have been conducted on this topic over the past one and a half decades and proposed models for predicting a traffic crash in real-time (e.g. Lee et al., 2003a,b,c; Abdel-Aty et al., 2004, 2006c; Abdel-Aty and Abdalla, 2004; Oh et al., 2005a,b; Dias et al., 2009; Hossain and Muromachi, 2012, 2013b; Xu et al.. 2013a,b,c; Yu and Abdel-Aty, 2013a,b; Roy and Muromachi, 2016; Roy et al., 2016; Sun and Sun, 2016; Katrakazas et al., 2016, 2017; Yang et al., 2018a,b; Roy et al., 2018b), identifying their types (Golob et al., 2004; Pande and Abdel-Aty, 2006a,b; Christoforou et al., 2011), understanding crash mechanism (Lee et al., 2003a,b,c; 2006a; Luo and Garber, 2006, Hossain and Muromachi, 2011, 2013a; Xu et al., 2012; Yeo et al., 2013), evaluating countermeasures through variable speed limits (Abdel-Aty et al., 2006a,b, 2008a; Lee and Abdel-Aty, 2008, Lee et al., 2004), ramp metering (Abdel-Aty and Gayah, 2010; Lee et al., 2006b), and variable message signs (Al-Ghamdi, 2007; Lee and Abdel-Aty, 2008). The recent trend has been focused on addressing the issues of transferability (Shew et al., 2013; Roy et al., 2018a), building them for specific road sections (e.g., weaving areas as shown by Wang et al., 2015), optimizing real-time safety and congestion in tandem (Park and Haghani, 2015), considering severity (Xu et al., 2013a) or simply, using more sophisticated modeling methods to improve accuracy (Xu et al., 2013b; Park and Haghani, 2015; Xu et al., 2016a, 2016b).

Although a substantial number of studies have been carried out in developing RTCPMs, the initiatives have been discrete. In addition, attempts to consolidate the existing knowledge with well-defined future guidelines in order to transform the idea into a system are still in their infancy. There have hitherto been five survey papers available concerning RTCPMs. Abdel-Aty and Pande (2007) were primarily engrossed in distinguishing between conventional crash prediction models (CPM) and RTCPMs postulating that the former identifies locations where ‘more crashes are likely to occur’, whereas the latter is concerned about locations where ‘a crash is more likely to occur’. Roshandel et al. (2015), on the contrary, conducted a brief systematic review coupled with a meta-analysis which had core interest in investigating the influence of traffic characteristics on crash occurrence. They identified several issues from existing studies: appropriateness of the variable selection, actual threat posed by the pre-defined crash precursors, trade-off between simple statistical models and data mining based approaches. They postulated that statistical methods, even though based on a strong theoretical basis, may not be capable of handling correlated variables whereas data mining-based approaches, which are capable of handling large data with correlated variables, may present outputs where the underlying mechanism is hard to comprehend. Their study argues the suitability of embracing the case-control approach which was common in most of the existing studies. This is because once the control is fixed, one can estimate the population of the control rather than opting for a subset, even though the data is large. Roshandel et al. (2015) was critical about the application of loop-detector based data as their location is fixed on the road and their distance from crash locations cannot be controlled, although 85% of the existing studies had their data collected through loop-detectors. In the end, the study provided a glimpse of the current knowledge and addressed some of the challenges and opportunities, however, left the readers with more questions than answers with respect to moving forward in developing and implementing RTCPM in real-world scenarios. Xu et al. (2015) also performed a meta-analysis with a quantification of the influence of traffic variables on crash risk. They applied three different Bayesian meta-analyses: fixed effect meta-analysis, random effect meta-analysis, and meta-regression. Later on, they developed a new RTCPM boosting their low sample size from Chinese expressways with results from the meta-analysis as informative priors. Their models constructed with meta-regression outperformed the models directly developed with limited data by 15%, which was further bolstered by 5% when they applied a Bayesian predictive density analysis to screen out the outliers in the limited data. Chu and Zhang (2017) conducted a literature review on RTCPMs based on studies published until 2015. Their study concentrated on four aspects of RTCPM building: data source, normal and pre-crash traffic conditions, variables space and predictive modeling methods where they discussed various approaches adopted in different studies for model construction. The conference paper is narrative, rather than systematic in nature and only touched base on development tendencies of RTCPM. Abdel-Aty et al. (2018) in their survey paper commenced with clearly distinguishing between traditional frequency-based road safety evaluation and real-time crash risk estimation and then progressed to summarize prominent studies dealing with the effects of near real-time traffic characteristics on crash occurrence. Their findings suggested that a number of traffic and weather-related parameters contribute to crash, most notably speed measured as the coefficient of variance of speed stood out to be the most significant. Their study concluded with several suggestions: (i) considering new vehicle-related variables, e.g., headway, for model construction; (ii) evaluating transferability of RTCPMs; (iii) testing various real-time interventions through traffic simulation; and (iv) taking the concept beyond safety estimation and amalgamating it with congestion pricing and alternate routing. Nonetheless, none of these reviewed studies had any major objective to present a systematic guideline on bridging the gap between an idea and a ready to use technology for RTCPMs.

This study fills that gap by summarizing and synthesizing the lessons learned from existing studies through a systematic review, identifying the adopted design pathways from the existing literature and formulating the universal requirements of real-time crash prediction models by combining the notions of existing studies and studies outlining similar technologies. Finally, it evaluates the universality of existing models to present the state-of-the-art, which will hopefully enable future researchers to transform the idea of real-time crash prediction into an actionable technology.

## Methodology

The study is broadly divided into five parts: (i) systematic review, (ii) identification of design pathways, (iii) ascertaining the universal design requirements, (iv) determining the state-of-the-art by evaluating the existing studies against the universal requirements, and (v) providing a framework to construct RTCPMs fulfilling the universal design requirements. The final part also provides an informative discussion in light of the recent and anticipated future developments taking place in the emerging area of connected and autonomous vehicles (CAVs). The systematic review was conducted through topic modelling and text mining which are also known as Knowledge Discovery with Text (KDT). Correlation plot was prepared to identify the most followed design pathways. The overall process followed in this paper to achieve the objectives is illustrated in Figure 1.

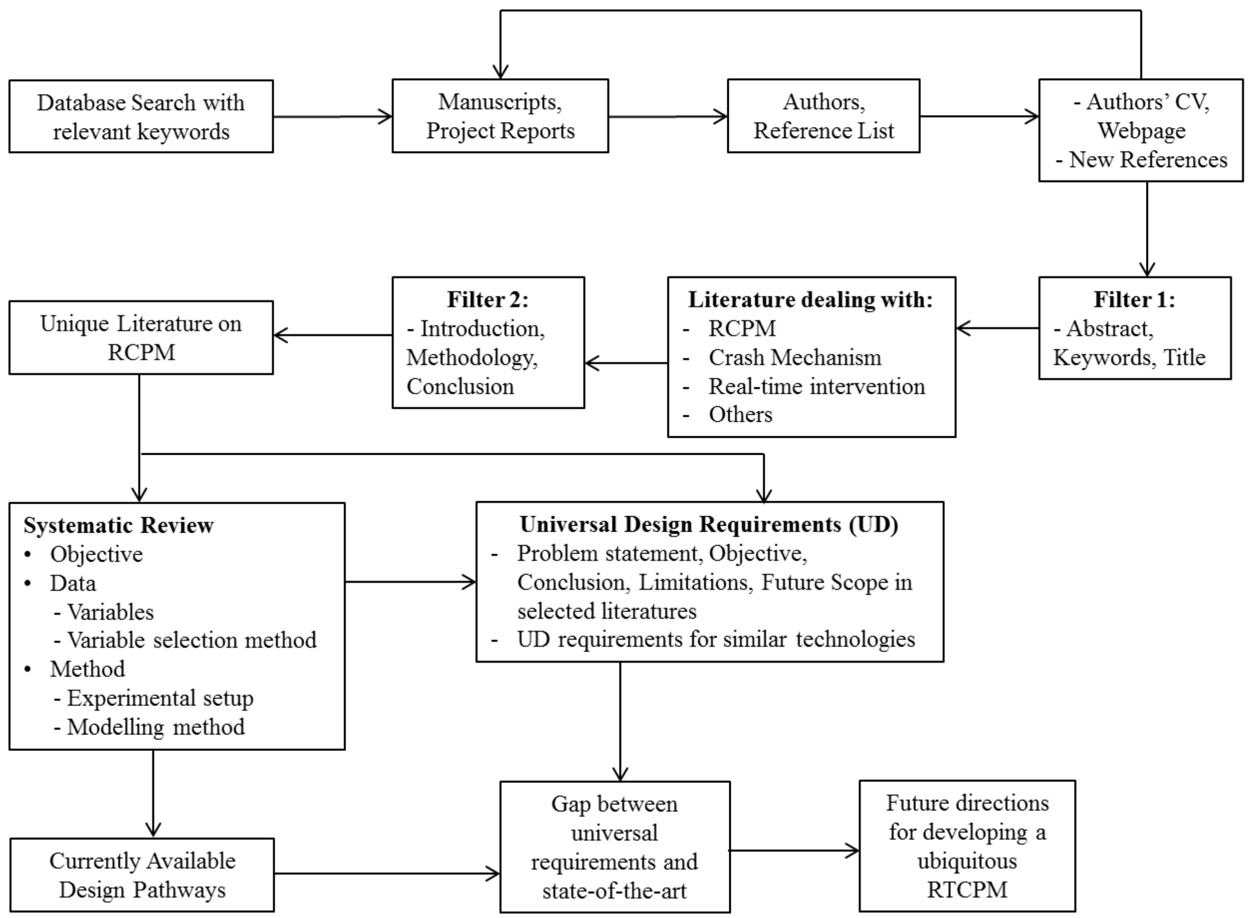



**Figure 1. Work flow diagram for a ubiquitous RTCPM**

The study commenced with conducting a comprehensive search in the Web of Science, Scopus, ProQuest, Google Scholar and society journal databases from North America, Europe and East Asia relating to transportation and/or safety using 'real-time crash prediction', 'real-time accident prediction', 'crash prediction model', 'accident prediction model', 'high resolution traffic data', 'traffic condition' and 'real-time intervention' as keywords to catalogue the relevant literature that mainly includes journal papers, conference papers, theses/dissertations and project reports. From the list, the authors were identified. Next, the detailed publication list of the authors was obtained from the internet (when available) and the reference list of the previously accumulated literature was inspected to source any literature that may be pertinent to real-time crash prediction. Afterwards, the title, keywords and abstract of each document was scrutinized to categorize them into four groups: real-time crash prediction (dealing with building RTCPMs), understanding crash mechanism (using high resolution detector data to understand the underlying determinants of crash), real-time intervention (methods to reduce crash hazards in real-time) and others (not pertaining to any of the aforementioned three groups). The studies falling into 'others' category were eventually truncated from the catalogue. Some of the studies dealing with understanding a crash mechanism or proposing a real-time intervention employed RTCPMs in order to explain the association between predictors and crash risk for the former case and appraised the real-time crash hazard after applying various interventions for the latter category. The RTCPMs applied in these studies were predominantly adopted from previous publications by the same author(s) where the sole objective was to construct a RTCPM. Through a rigorous exploration of the introduction, methodology and conclusion of these studies, duplicate RTCPMs were identified and subsequently removed from the catalogue. There were cases where the same literature was published in different forms in different times. In those cases, only the latest studies were considered. The final list consisted of 78 studies published between 2003 and 2018 and they are considered for a systematic review, the identification of design pathways and the evaluation of their universality. For ease of referencing, the studies are ordered chronologically as shown in Table 1. From here on, the studies will often be referred to as associated ID. For instance, Golob et al. (2004) is referred to as ID #6.

**Table 1. Study ID and Reference**

| Study ID | Authors Name | Study ID | Authors Name |
|---|---|---|---|
| 1 | Lee et al. (2003a) | 40 | Yu and Abdel-Aty (2013a) |
| 2 | Lee et al. (2003b) | 41 | Yu and Abdel-Aty (2013b) |
| 3 | Abdel-Aty and Abdalla (2004) | 42 | Yu et al. (2013) |
| 4 | Abdel-Aty and Pande (2004) | 43 | Paikari et al. (2014) |
| 5 | Abdel-Aty et al. (2004) | 44 | Xu et al. (2014a) |
| 6 | Golob et al. (2004) | 45 | Xu et al. (2014b) |
| 7 | Abdel-Aty and Pande (2005) | 46 | Xu et al. (2014c) |
| 8 | Abdel-Aty et al. (2005) | 47 | Lin et al. (2015) |
| 9 | Pande and Abdel-Aty (2005) | 48 | Shi and Abdel-Aty (2015) |
| 10 | Oh et al. (2005a) | 49 | Sun and Sun (2015) |
| 11 | Oh et al. (2005b) | 50 | Wang et al. (2015) |
| 12 | Pande et al. (2005) | 51 | Xu et al. (2015) |
| 13 | Abdel-Aty and Pande (2006) | 52 | Park and Haghani (2015) |
| 14 | Abdel-Aty and Pemmonaboina (2006) | 53 | Roshandel et al. (2015) |
| 15 | Abdel-Aty et al. (2006c) | 54 | Piradavani et al. (2015) |
| 16 | Hourdakis et al. (2006) | 55 | Xu et al. (2016a) |
| 17 | Lee et al. (2006a) | 56 | Fang et al. (2016) |
| 18 | Hellinga and Samimi (2007) | 57 | Xu et al. (2016b) |
| 19 | Lee et al. (2007) | 58 | Roy and Muromachi (2016) |
| 20 | Pande and Abdel-Aty (2007) | 59 | Katrakazas et al. (2016) |
| 21 | Abdel-Aty et al. (2008b) | 60 | Roy et al. (2016) |
| 22 | Zheng et al. (2010) | 61 | Sun and Sun (2016) |
| 23 | Jung et al. (2010) | 62 | Katrakazas et al. (2017) |
| 24 | Pham et al. (2010) | 63 | Abdel-Aty and Wang (2017) |
| 25 | Son et al. (2011) | 64 | Liu and Chen (2017) |
| 26 | Christoforou et al. (2011) | 65 | Wu et al. (2017) |
| 27 | Hossain and Muramachi (2011) | 66 | You et al. (2017) |
| 28 | Abdel-Aty et al. (2012) | 67 | Wang et al. (2017a) |
| 29 | Ahmed and Abdel-Aty (2012) | 68 | Wang et al. (2017b) |
| 30 | Hossain and Muramachi (2012) | 69 | Dimitriou et al. (2018) |
| 31 | Ahmed et al. (2012) | 70 | Park et al. (2018) |
| 32 | Qu et al. (2012b) | 71 | Roy et al. (2018a) |
| 33 | Hassan and Abdel-Aty (2013) | 72 | Wu et al. (2018) |

| | | | |
|---|---|---|---|
| 34 | Hossain and Muramachi (2013a) | 73 | Yang et al. (2018a) |
| 35 | Ahmed and Abdel-Aty (2013) | 74 | Yuan et al. (2018) |
| 36 | Hossain and Muramachi (2013b) | 75 | Yang et al. (2018b) |
| 37 | Shew et al. (2013) | 76 | Yasmin et al. (2018) |
| 38 | Xu et al. (2013a) | 77 | Yuan and Abdel-Aty (2018) |
| 39 | Xu et al. (2013b) | 78 | Roy et al. (2018b) |

The systematic review has been conducted through identifying and discussing the basic intricate components of a RTCPM (e.g. variable space and their selection procedure, methodology, validation and evaluation), their chronological development, strength and limitations. The process commenced by performing topic modelling with the Latent Dirichlet Allocation (LDA) method so as to discover hidden semantic structures embedded in a study. Topic modeling is a method of automatically organizing and searching a large amount of textual data to discover the underlying theme in a document. LDA is an autonomous probabilistic model that applies bag-of-patterns representation to discover clusters of topics in unstructured corpus where topic is characterized by a distribution of words (Blei et al., 2003; Das et al., 2016). It is a generative statistical unsupervised model that requires no prior annotations of document. Rather, it auto-generates topics from the document by investigating the combination of document and word statistical data in relation to the topics. It represents documents as mixtures of topics that disclose words with certain probability. LDA is described with a plate diagram as illustrated in Figure 2.

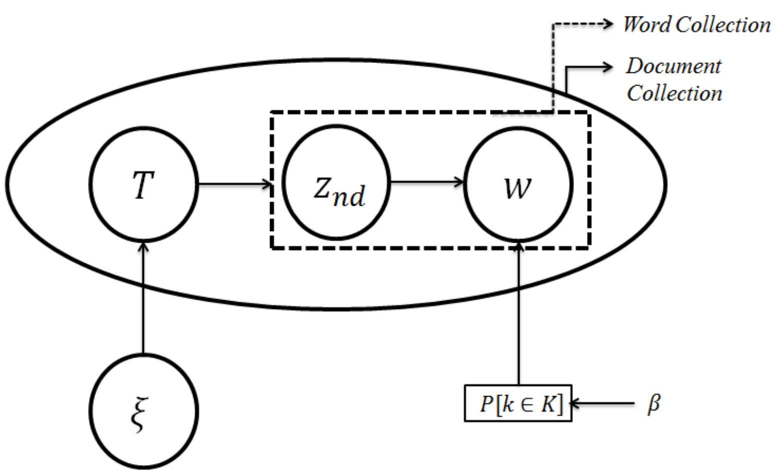



**Figure 2. Graphical representation of LDA for topic modeling**

In short, the algorithm is briefly discussed as follows:

1) The documents are produced with Q number of words, following Poisson distribution.
2) Then topic mixtures of fixed $k$ topics are chosen from these documents based on Dirichlet distribution, i.e. $T \sim Dir(\xi)$ where $\xi$ is prior on the per-document topic distribution and word distribution of each topic $k$ is determined by Dirichlet distribution also i.e., $P \sim Dir(\beta)$ where $\beta$ is prior to per topic word distribution.
3) LDA generates each word $w$

i) by picking up topics following multinomial distribution, i.e. topic $z_{nd} \sim$ multinomial (T)

ii) using the topic to generate the word (according to the topic's multinomial distribution), i.e. choosing a word $w$ from $P(w \mid z, \beta)$.

Here, T is the distribution of topics over document $d$, $z_{nd}$ is the topic for the n$^{th}$ word in the d$^{th}$ document, $\beta$ is the distribution over words over topics k. LDA inference can be done by variational expectation-maximization (VEM) algorithm or by Gibbs sampling (Grun and Hornik, 2011). In this research, the latter is applied for inferring document distribution T and topic-word distribution P. Here, $\xi$ and $\beta$ are the hyperparameters of LDA. Statistical inference from LDA algorithm depends heavily on the choice of hyperparameters to fit with the model. Although they are usually chosen in an ad-hoc manner (George and Doss, 2018) in this study, the proposed procedure suggested by Blei et al., (2003) has been followed.

Recently in academia a substantial number of systematic reviews (e.g., Das et al., 2016; Sun and Yin., 2017) have been conducted using the KDT to filter a large amount of literature to extract relevant information on a specific topic or to seek answers to questions that need to be addressed. Moreover, text analysis and topic modeling, aka KDT, are being used for real-time incident duration prediction by converging textual information into incident attributes (Pereira, 2013). KDT is a generic scientific branch of data mining which follows a process of identifying valid, important and interpretable patterns of unstructured textual data. It is founded on the assumption that the arrangements and occurrences of major words of a document hold its underlying messages. KDT methods commence with amassing a large structured set of texts known as 'corpus', whose noise is refined by removing redundant words, phrases, numbers and punctuations (Das et al., 2016). Their study constructed comparison word clouds and evaluated correlation of words as part of text mining.

Word clouds are used to determine the most frequent terms in a corpus. Let $p_{x,y}$ be the where the word $x$ occurs in document $y$, $p_y$ be the average rate across $n$ documents ($\Sigma_y p_{x,y}/n$). When comparing clouds, the size of each word is mapped to its maximum deviation $max_x(p_{x,y} - p_y)$. Its angular position is determined by the document in which that occurs the most (Das et al., 2016).

The systematic review directed to the recognition of design pathways followed in existing studies. Correlation was also conducted to identify the most followed design pathways. The formulation of universal requirements involved two steps. First, the problem statements (highlighted the limitations of the then models), objectives (presented the progresses made with that literature), conclusions, limitations and future scopes (stated what more to be expected from RTCPMs in future) outlined by the authors were extracted. This shed some lights on the shortcomings of the existing solutions and what qualities the researchers are expecting RTCPMs to possess. Afterwards, universal design requirements of similar systems were listed through literature review. A comprehensive list of universal design requirements was then compiled by combining the outcomes of these two steps. Then the universality of the existing literature was gauged. Finally, a framework was presented to develop a universal RTCPM.

This study employed various packages, such as, Open source statistical software "R" was used for text mining and topic modelling. The "topicmodels" package by Grun and Hornik (2011) was used for LDA. "Mallet" package (Mimno, 2015) to get the probability of topics in documents and probability of words in topics, "tm" for text mining (Feinerer and Hornik, 2015), "wordcloud" to visualize the clouds (Fellows, 2014) and "Rgraphviz" for correlation analysis plotting (Hansen et al., 2016) were employed.

## Systematic Review and Design Pathways

To abridge the RTCPM research information from a large archive of text, at the beginning of systematic review, topic modeling with the LDA method was performed on paper titles and abstracts. The generated topic along with the probabilities of topics and topic-words from the document groups, i.e. a combination of title and abstract, are outlined in Table 2.

**Table 2. Top 6 topics from paper titles and abstracts**

| **Topic#** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| Words | Risk (0.061) | Freeway (0.051) | Mechanism (0.033) | Evaluate (0.031) | Predict (0.026) |
| | Realtime (0.057) | Realtime (0.047) | Realtime (0.029) | Realtime (0.028) | Freeway (0.021) |
| | Crash (0.048) | Crash (0.041) | Freeway (0.027) | Condition (0.028) | Model (0.23) |
| | Bayesian (0.045) | Traffic (0.035) | Crash (0.023) | Crash (0.021) | Crash (0.021) |
| | Predict (0.037) | Risk (0.031) | Data (0.021) | Freeway (0.018) | Traffic (0.021) |
| | Model (0.030) | Model (0.028) | Urban (0.018) | Traffic (0.013) | Realtime (0.020) |
| Prob. | 0.27 | 0.25 | 0.17 | 0.16 | 0.16 |
| **Topic#** | **6** | **7** | **8** | **9** | **10** |
| Words | Freeway (0.030) | Speed (0.026) | Segment (0.021) | Traffic (0.025) | Learning (0.031) |
| | Urban (0.030) | Threshold (0.021) | Character (0.021) | Weather (0.024) | Realtime (0.029) |
| | Data (0.024) | Traffic (0.021) | Realtime (0.020) | Character (0.019) | Traffic (0.015) |
| | Expressway (0.024) | Risk (0.019) | Data (0.019) | Crash (0.017) | Crash (0.014) |
| | Crash (0.021) | Character (0.018) | Crash (0.018) | Realtime (0.011) | Character (0.013) |

| | | | | | |
|---|---|---|---|---|---|
| | Predict (0.020) | Data (0.016) | Traffic (0.017) | Data (0.010) | Risk (0.013) |
| Prob. | 0.15 | 0.13 | 0.13 | 0.12 | 0.11 |
| **Topic#** | **11** | | | | |
| Words | Performance (0.028), Realtime (0.027), Character (0.017), Frequency (0.014), Crash (0.011), Risk (0.010) | | | | |
| Prob. | 0.10 | | | | |

The top eleven panels of topic with six tightly co-occurring terms from the paper title and abstract group combinedly can be observed from Table 2. Conditional probability of each of the topics over the word and document distribution is also given based on which ranking of the established topics. From Topic 1 to 11, the probability values for each cluster of the topics range between 0.10 and 0.27. Topic 1 includes: “crash”, “predict”, “model”, “risk”, “realtime”, and “Bayesian”. The probability of each word is presented within a parenthesis. The dominant words have found to be: risk, Bayesian, crash and real-time. Therefore, these words are skewed towards real-time crash risk prediction using Bayesian approaches. The same pattern of interpretation is followed for the other 10-topics. Topic 2 emphasizes the use of freeways (p=0.051) as study areas. Topic 3 focuses on revealing the crash mechanism (p=0.033) using real-time data from urban freeways. Topic 4 deals with evaluation (p=0.031) of traffic condition for crash risk with real-time traffic data. Topic 5 specifically focuses on real-time crash prediction (p=0.026) model building and Topic 6 narrows down the focus of study area within urban expressways and freeways (p=0.030). Topic 7 and Topic 8 indicate traffic characteristics the threshold (0.021) for speed (p=0.026) and real-time traffic data on road segment (p=0.021). Topic 9 focuses on weather data (0.024) and traffic characteristics (0.025). Topic 10 focuses on cutting edge learning methods (e.g. DNN, BN, DBN) (p=0.035) to evaluate crash risk. Finally, Topic 11 includes performance (p=0.028) measure of crash frequency relate to real-time crash characteristics. Combining the essences of the topics, it can be summarized that the selected manuscripts deal with crash prediction model building with real-time traffic data collected from urban expressways and freeways, some dealt with evaluation of traffic conditions and exploring the crash mechanisms and many adopted Bayesian as well as modern machine learning approaches for model construction. It is noteworthy to state here that although topic modeling (Topic 11) identified ‘frequency’ as a major keyword, the manuscripts dealing with frequency based crash risk analysis are not considered for further investigation in this study as RTCPMs deal with the estimation of crash risk at a given location at a given time whereas ‘frequency’ based crash prediction models deal with the identification of locations with high number of crashes. The distinctions are elaborately discussed by Abdel-Aty and Pande (2007).

On several occasions in this manuscript, various characteristics of the studies have been presented as ($XX{:}Y_1,\ldots,Y_n$) format where XX presents the total number of studies in the concerned category and $Y_i$ presents the corresponding Study IDs as listed in Table 1. The geographical distribution of the sources of 77 catalogued articles (excluding the review paper by Roshandel et al., 2015) is as follows: USA (45:3-17,20,22,23,25,28,29,31-333,35,37-42,44-48,50,52,55-

57,63,65,67,68,70,72,74,76,77), United Kingdom (2:59,62), Canada (4:1,2,18,43), China (7:49,51,61,64,66,73,75), Japan (8:27,30,34,36,58,60,71,78), Korea (1:19), Netherlands (1:21), France (1:26), Switzerland (1:24), Belgium (1:54) and Cyprus (1: 69). This suggests that most studies are coming from North America. The majority of the previous studies have been conducted on the interstate freeways in the USA/Canada (47:3-12,14,16-18,20-23,25-27,31,32,33,35,37-49,51,52,54-57,66,70,72) and some other study areas include: expressways (21:1,2,9,13,15,30,34,36,58,60,61,63-65,67,68,71,73,75,76,78), national roads (1:19), arterials(2:74,77), European motorways (3:24,59,62), North American state roads (3:28,29,50) and city streets in Cyprus (1:69). The chronology of the published studies based on their major objectives is presented in Figure 3. Also, Table 3 is included to identify the association of various studies with their major objectives. It is evident that the quest for an improved RTCPM is continuing. At times, they used such models to explore the underlying determinants of crashes; however, studies exploring to devise real-time countermeasures are quite scant.

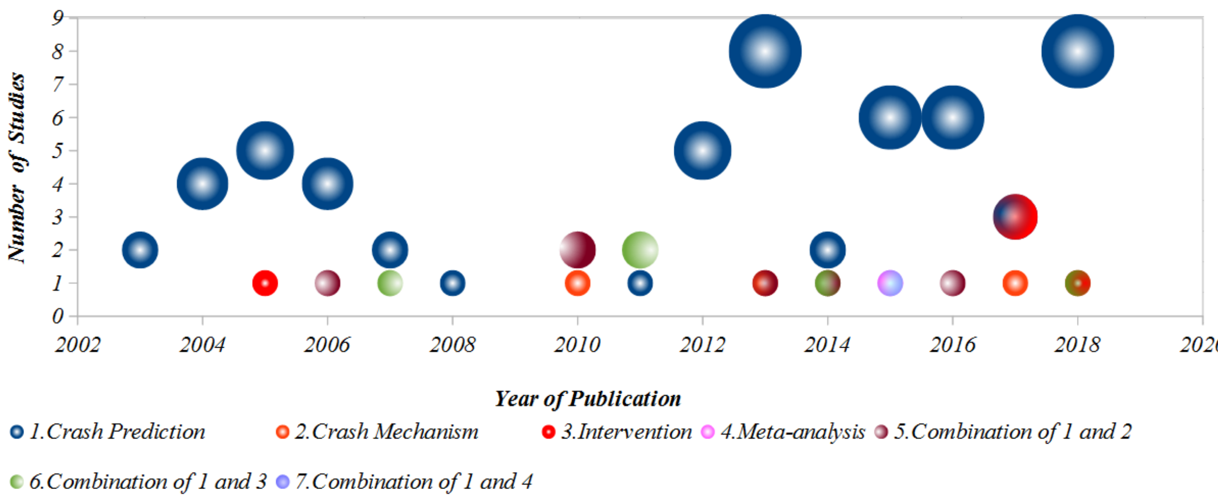



**Figure 3. RTCPMs constructed with various major objectives**

**Table 3. Studies with different objectives**

| Objective | Study ID |
|---|---|
| Crash Prediction | 57: 1-8, 10-16, 19-21, 25, 28-33, 35-41, 43-45, 47-50, 52, 54-56, 58-61, 64, 66, 67, 69, 71, 73-78 |
| Crash Mechanism | 3: 22, 42, 65 |
| Intervention | 5: 9, 62, 63, 68, 70 |
| Meta-Analysis | 1: 52 |
| Combination of 1 and 2 | 6:17, 23, 24, 34, 44, 57 |
| Combination of 1 and 3 | 5: 18, 26, 27, 46, 72 |
| Combination of 1 and 4 | 1: 51 |

Although the concept of RTCPM has evolved in course of time, the fundamental framework of their construct has remained mostly unchanged since Oh et al. (2000). The common modeling steps are as follows:

- selecting different descriptive statistics of the traffic flow parameters as variables;
- collecting data regarding these variables from one location or an array of longitudinal locations for each crash case;
- defining pre-crash and normal traffic conditions and separate traffic flow data into these two categories (with the exception from Xu et al. (2014a) where they divided the traffic states into four categories - free fluid traffic, bunched fluid traffic, bunched congested traffic, and standing congested traffic);
- treating the problem as a classification problem and use a suitable method to predict the crash probability, and finally
- evaluating the modeling performance.

The major variations in modeling have been found as follows:

- defining the scope of the model (i.e., high speed or low speed traffic conditions, different weather conditions, road geometry);
- defining pre-crash and normal traffic conditions;
- selecting the means (loop detector, video data, etc.) and methods (location and combination of detectors) of data extraction;
- selecting variable space, and
- deciding on the modeling method.
- considering study area: interstate freeways, expressways, recently arterials, arterial intersection, city streets etc.
- comparing model performance using various approaches and methods, e.g., Wang et al. (2017b) compared performance of combined real-time and frequency-based model against separately constructed frequency and real-time based models, Roy et al. (2018a) compared between Dynamic and Static Bayesian Networks, etc.

The following subsections discuss the major components of RTCPMs by presenting the state-of-the-art through a chronological narration. Some models considered crash severity, i.e., fatal, personal injury, property damage only, (12:1,6,14,28,38,39,40,44,50,52,57,70) or crash types, i.e., multi-vehicle, single vehicle, rear-end, side-swipe, collision/conflicts. (37:6,15,16,19,20,22,23,24-28,30,32,34,36,38–41,42,44,48–50,57-62,66,68-70,72,73) in their analysis.

## Type, spacing and arrangement of detector

The performance of RTCPMs vastly relies on the type, spacing and arrangement of the detectors that are selected with respect to the crash location to fathom crash potential. Out of the 77 studies chosen for review, 50 solely used loop-detectors to extract data on traffic flow variables (50:1-9, 12-18,20-27,30,32,34,36,38,39,43-47,49,50,54-58,60,61,66,69,71,73,75,78), five recent studies solely used Microwave Vehicle Detection System (MVDS) (5:19,48,63,67,76), two used Automated Vehicle Identification (AVI) (2:29,31), two used Bluetooth Detector (2:74,77), five studies used Remote Traffic Microwave Stations (RTMS) (5:40,41,42,51,72), and one study used probe vehicle (1:52). Rest of the studies used a combination of technologies, for example, loop detector & probe vehicle (5:10,11,52,59,62); loop detector & AVI (1:28); loop detector & radar (1:33); AVI & RTMS (1:35) and loop detector & MVDS (1:37), loop detector, RADAR & MVDS (1:65), MVDS & Video detectors (2:64, 68) to collect traffic flow data. All these sensors are capable of yielding count, speed and occupancy data, although the recent technologies have some advantages over loop detectors. For example, AVI system can provide measures about percentage of lane change per segment by comparing the unique tag ID for each individual vehicle at the beginning and end of the segment (Ahmed and Abdel-Aty, 2013). RTMS and AVI have similar capabilities except that the former captures time mean speed and the later senses space mean speed. However, both are low-cost and more scalable (Ahdi et al., 2012). MVDS uses radar detection technology which is cheap but sensitive to wind which may introduce error by swaying the poles on which they are mounted (Bugdol et al., 2014). Although the technology is an ideal source of Big Data (Shi and Abdel-Aty, 2015), it comes with associated high cost of installation and maintenance and cannot be therefore deployed on a large scale due to wiring and constant energy requirements (Ahdi et al., 2012). In two recent studies, Bluetooth data extracted from urban arterials (Yuan et al., 2018) and signalized arterial intersections (Yuan and Abdel-Aty, 2018) were employed to estimate the real-time crash risk.

There were 22 studies that did not mention anything about how the detectors were spaced whereas the remaining 55 studies reported detector spacing on the study area in various ways. Of which, the common ones are – average (29:3–8,12–15,17,18,20,23,38,39,45,48,55,57,58,60,62,64-66,68,69,78), minimum-average-maximum (4:21,38,48,53), average-median (1:6), minimum-maximum (8:1,2,22,28,38,48,55,61), average-standard deviation (2:45,48) and minimum-average-maximum-standard deviation (2:43,48). In general, most of the studies having loop-detectors reported an average detector spacing to be 0.8 km with the minimum value of 0.22 km and the maximum of 3.81 km. The average spacing was found to be 1.91 km for RTMS data. Ahmed and Abdel-Aty (2012) extracted data from AVI systems and reported the minimum (0.22–2.04 km), average (1.42–4.76 km), maximum (3.72–12.16 km) and standard deviation (0.88–3.60 km) values for both directions of all three road sections considered in their study. Shi and Abdel-Aty, using MVDS also provided minimum (0.16–32 km), average (0.73–1.6 km), maximum (1.60–5.90 km) and standard deviation (0.34–1.56 km) for both directions of three state roads considered in their study.

Like the detector technology and their spacing, the arrangement of detectors selected by various researchers to extract crash prone and normal traffic data also varied substantially. In many cases they have chosen the nearest detector from the crash location to extract pre-crash data (9:6,9,10,11,15,47,57,59,78). Other preferences were, nearest upstream (4:17,62,64,67), nearest downstream (1:25), one each in the upstream and downstream (7:2,38,39,55,58,60,73), one each in the upstream and downstream and the nearest from the crash site (7:31,40,50,51,69,71,75), one each in the upstream and downstream and the ramp (1:36), two each in the upstream and downstream (8:30,32,34,46,48,49,61,65), two each in the upstream and downstream and the nearest detector from crash (1:20), three each in the upstream and downstream (4:21,29,33,37), three each in the upstream and downstream and one in the nearest AVI location from crash (1:28), three and one each in the upstream and downstream respectively for loop detectors and AVI station (1:35), four in the upstream and two in the downstream (5:8,12-14,56), and five in the upstream and one in the downstream (4:3,4,5,7). Among the remaining studies, researchers modeled

RTCPM with microscopic data, hence, collected information from individual vehicle (1:52), reported to collect data from consecutive detectors but did not explain their locations (3:54, 72, 74), used probe vehicle data (1:52) and considered specific sections or influence area of upstream and downstream zone (2:68,76) rather than detector locations (2:1,68), or, did not report detector arrangements (11:5,18,22,23,25,27,42,53,63,66,70).

## Defining Pre-crash and Normal Traffic Conditions

Although researchers exhibited a wide variety of notions while defining a pre-crash condition, for the studies considering multiple time slices, a common approach was to extract the detector data for a 30-minute time period just before crash and divide it into six five-minute time slots (14:3,4,5,7,8,9,12,13,14,29,33, 64, 73, 75). One study used 6-minute prior time before a crash with three two-minute time slices (1:31) and another study used 20-minute prior time before crash using four five-minute time slices (1:77). However, there was an overwhelming motion towards defining a five-minute period, 5-10 minute before crash, as representative pre-crash time (44:3-5,7,8,9,12-14,17,20,21,27-34,36,37,39-42,44,46-51,54,55,57,59,61,62,63,66,67,68,76) and studies considering multiple time slices found this time period significant. As quite often the studies depended on crash time reported by various organizations - Department of Transport or expressway authorities (38:3,8,9,12–14,22–30,33–41,44–48,51,52,55-62), police report (8:1,4,5,7,17,23,25,54), traffic control center (3:2,18,19) – not mentioned if it had video data (1:1) – maintained surveillance camera, and various other sources, such as, CCTV footage (3:16,25,51), Bureau of Statistics, crash databases from centers or research laboratories, verbal interviews, etc., many authors were in favour of introducing a buffer time, 0-5 minute before crash, to compensate errors in reported crash time. Wang et al. (2015) postulated that 5-10 minute prior to crash provide accurate crash precursor condition as compared to that of 10-15 minute. Irrespective of their differences in defining pre-crash traffic, researchers unequivocally accepted the importance of accurately identifying the crash time for constructing RTCPMs. Only a few studies collected crash time from surveillance cameras on road (3:16,49,73). Most of the studies relied on the crash time that they obtained from authorities (33:3-5,7,8,11,15,17,23,26,29,30,31,33-37,40,41,47,58-63,67,68,71,76-78) or maintained reasonable buffer time between recorded crash time and pre-crash time (10:28,32,38,39,44-46,50,51,55). The attempts to determine the actual crash time included – detecting sudden drop in speed, often by plotting speed profile (6:1,2,22,24,31,54), identifying backward-forming shockwave upstream of the crash location (2:11,18), applying shock-wave and rule-based methods (3:9,13,14,), spotting speed and flow variation between adjacent lanes (1:27), drawing speed contour plots (2:52,57), estimating from the reported crash time by investigating upstream and downstream detectors' traffic flow variation for each crash (1:71). In an interesting recent study, the authors corrected crash time using information received from mobile phones along with video surveillance data (1:73).

The strategy followed by various researchers in defining normal traffic condition has been to select a traffic condition from a crash eventless time period or a typical day, i.e., no crash or incidents took place during or near that time. Variations mainly introduced through how the studies negotiated with avoiding pre-crash conditions – by taking data at least 30 minutes earlier than the crash time from the same detectors (17:3–14,16,24,25,33,64), any typical 24-hr data when no crash took place (3:1,2,44), randomly chosen traffic data when no crash took place (9:9,20,37–

39,46,50,51,55), data extracted from the same detectors for same day and time of week but from other days when no crash took place within one hour from that time (20:27–32,34,36,40,41,45,47–49,54,58,60,61,73,75,78) and 2 hour (1:31), 3 hour (1:77) as well as 5 hour before-after that crash time (3:63,67,68).

## Variable space and selection method

Traffic flow variables have been at the core of the RTCPMs, the most common of those have been the subset of the average, standard deviation, coefficient of variation and other statistics or logarithmic transformations of speed, flow and occupancy aggregated at different upstream and downstream detector locations with respect to the crash location, and their differences in space, i.e., between longitudinally placed detector locations when data were extracted from multiple detectors, between laterally placed detectors (lane to lane difference) or, differences in various time slices. The data aggregation varies both in temporal and spatial scales, mainly due to the way the raw data were supplied. In a substantial number of studies, data were delivered aggregated for all lanes for every 20 seconds or 30 seconds which the studies further aggregated for one minute (5:16,56,62,71,78) or five minutes (33:1–5,7-9,12–17,28,33,37–41,55–57,44-49,62,64,65). In some studies, the supplied data were already aggregated for each 1 minute (6:21,29,54,58,60,63) and five minutes (12:22,24,27,29,30,34,36,60,68,69,72,75). Some studies aggregated their data to 15 minutes for simulation (1:63) and crash prediction (1:67)

Mostly these data were collected for the basic freeway segments, and some studies included traffic data from the ramps. Hossain and Muromachi (2013b) suggested that the conditions near ramp areas are substantially different from that of the basic freeway segments and separately built models for the ramp vicinities. Pande and Abdel-Aty (2007) included distance to the nearest ramp as an independent variable. Studies dated later 2017 started considering the traffic flow variables related to ramp areas along with the basic freeway segment (5:63,65-68). Some studies included density, queue length, exposure to traffic (Lee et al., 2003a), hazard ratio for average volume (Abdel-Aty and Pande, 2005), complex calculation of shockwaves (Yu and Abdel-Aty, 2005), safe stopping distance of individual vehicles (Son et al., 2008), average flow ratio calculated from the peak flow (Pande and Abdel-Aty, 2006b), congestion index (Dias et al., 2009; Hossain and Muromachi, 2012, 2013a; Shi and Abdel-Aty 2015; Roy and Muromachi, 2016; Roy et al., 2016), percentage of heavy vehicles (Pham et al., 2010; Wang et al., 2017b; Park et al., 2018), geometric mean of average flow ratios (Qu et al., 2012b), average journey time (Katrakazas et al., 2017) first order autocorrelation of count, speed and occupancy (Xu et al., 2014b), weaving volume ratio, speed difference between the beginning and end of weaving segment (Wang et al., 2015) as variables. Use of coarser data such as peak hour traffic data (Abdel-Aty et al., 2006c; Christoforou et al., 2011), 75th percentile of average, standard deviation and coefficient of variation of speed, 75th percentile of standard deviation and coefficient of variation of volume (Abdel-Aty et al., 2006c), or day of week (Xu et al., 2016b), mainly seen in conventional CPMs, were also practiced. RTCPMs built with microscopic traffic flow data also introduced traffic pressure, kinetic energy, coefficient of variation of time headway, mean velocity gradient and mean reaction time as variables (Hourdakis et al., 2006; Paikari et al., 2014). Abdel-Aty et al. (2012) represented speed as both time and space mean speeds. Although Xu et al. (2014b) did not estimate real-time crash risk in individual vehicle level; they utilized time and space headways as variables. Wang et al.

(2017b) introduced average daily standard deviation of speed which had a positive effect on crash frequency and Dimitriou et al. (2018) introduced lane of travel for each individual vehicle and location of loop detector in their model.

Road traffic crashes are attributed to various human, road geometry, vehicle and environment related factors. Traffic flow variables in RTCPMs can be considered as surrogate measures of human factors (62,17,23,25,69,72). Substantial number of studies have continued introducing geometric and environment related variables, such as the existence of curves (4:17,23,26,31), upstream and downstream on and off ramps, barrier, pavement condition (5:3,23,63,65,67), no. of lanes/lane changes/lanes blocked (7:23,38,67,69,70,72,76), median width (4:15,31,55,76), gradient (1:35), inner and outer shoulder width (5:23,39,55,68,76), pavement detail – surface condition (1:16), category and roughness (1:15), weather (8:2,6,23,31,42,50,66,76,77) – more specifically raining or not raining (2:14,74), amount of precipitation (4:31,35,76,77), lighting condition (3:2,6,76), visibility (clear or reduced) (6:31,35,42,47,72,77), sun position (night, cloudy, sun in back or side, sun in front) (1:16), etc., in their RTCPMs. Other interesting variables introduced include young neighbourhood and school hour and day of week (1:43), headway (2:69,72), congestion (1:70), length of road segment (2:68,76) and weaving influence length (3:63,67,76).

Crash is a rare event. Hence, the sample size containing crash data and their corresponding detector data are in most cases quite scant (only 30 studies having a sample size larger than 500). This induces a classical situation of large variable space and small sample size – requiring a suitable method to select the most important variables. Where some studies employed engineering judgment to choose the variables (2:1,32), most of the studies simply relied on the modeling method they applied to build the RTCPMs to cancel out the insignificant variables (29:2,4,5,7,8,9,12-14,16-18,20,22,26,28,38,41,46,50,52,55,57,69,72,73,74,76,77) and some did not report if they have followed any method to identify the most important variables (3:19,25,71). Others applied statistical methods such as t-statistics (4:3,10,11,76), standard error (1:15), p-value (5:50,63,65,67,68), nonlinear canonical correlation analysis (1:44), Pearson correlation (1:68), non-parametric Spearman's correlation test (1:54) and logistic regression (1:75). Recent studies that are based on Artificial Intelligence (AI) or data mining in constructing RTCPMs, mainly applied classification or pattern trees (6:35,37,40,47,64,70), random forest (13:21,24,27,29,33,39,43,45,47,48,60,61,66) or its variations such as (random multinomial logit models (3:30, 34,36,) to downsize the variable space. Some studies have also applied clustering (1:6), expectation maximization (EM) algorithm (2:49,78) or calculated Eigen values (1:56) to measure variable importance. To summarize, it can be concluded that the studies using statistical approaches to build RTCPMs mainly relied on the internal mechanisms of the models to drop insignificant variables, whereas, the studies applying AI and data mining approaches almost overwhelmingly applied either classification trees or random forest (random forest is considered as one of the latest and most efficient methods in evaluating and ranking variable importance (Harb et al., 2009) to identify the most important variables.

## Modeling method

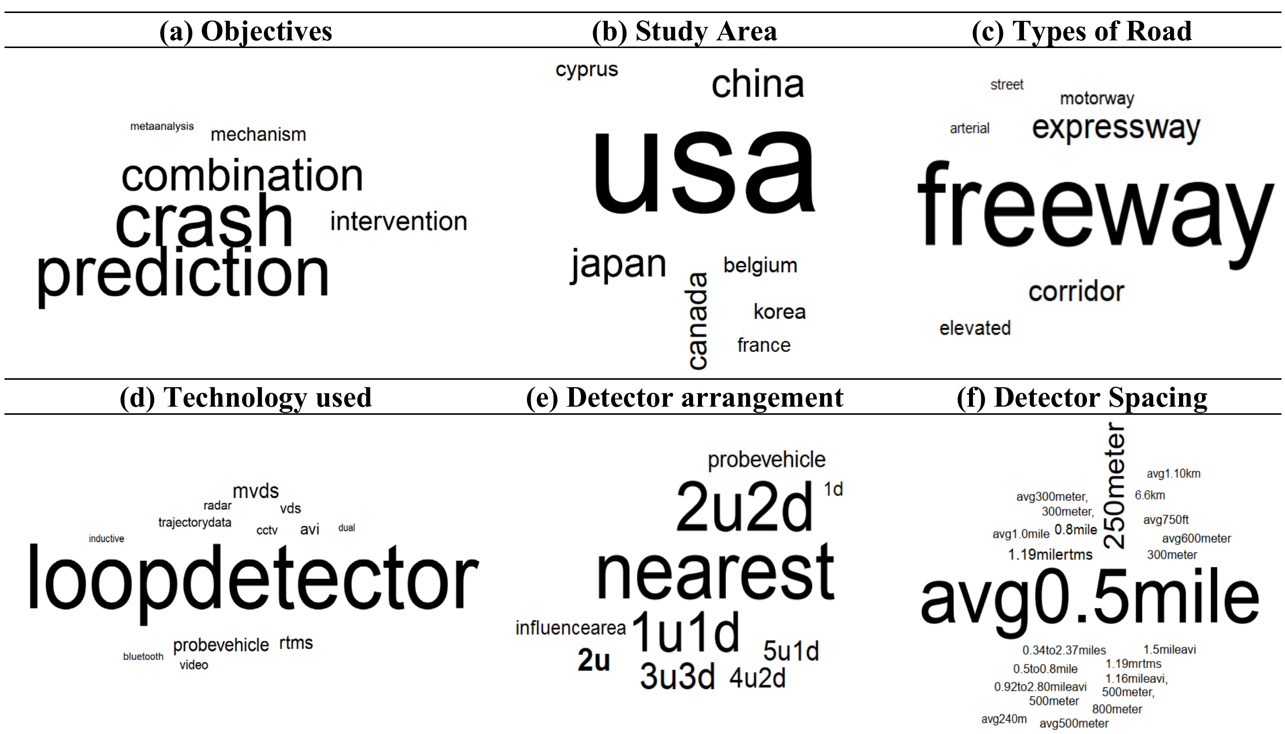
The fundamental modelling approach has been to collect data on various predictors as outlined in the previous section separately for pre-crash and normal traffic condition and then feed those into a modelling method suitable to predict dichotomous outcomes. However, in some cases, when severity or types of crashes were also predicted, methods allowing the dependent variables to have multi-classes were chosen. The typical modeling methods employed by researchers in developing RTCPMs so far can be broadly classified into two groups: statistical methods and artificial intelligence/data mining-based methods.

Among statistical methods, various forms of logit (40:5,8,9,12-14,16,17,20,22-24,28,29,31,33,38-41,44-46,48,50,55-57,63-65,67-69,72-75,76,77) and probit models (1:26) have been the primary choice. Some mixed generalized linear model e.g., Poisson-lognormal (2:41,68) and negative binomial model (3:15,25,72) was also preferred by some of the researchers. Among AI/data mining based methods, most of the proposed models applied various forms of neural networks (9:4,7,11,20,37,52,64,70,75), Bayesian networks (11:30,36,43,47,49,58,68,70,71,74,78) or classifying methods such as classification and regression trees (2:24,27), support vector machine, SVM (7:32,40,59,61,62,64,66), Principal Component Analysis (1:14) or simple rule based classifier (1:54). Some discrete attempts applied aggregated log-linear model (2:1,2), generalized estimating equations (1:3), Bayesian structural equation modeling (1:52), Bayesian classifiers (1:10), genetic algorithm (1:37), stochastic gradient boosting (2:35,70). Irrespective of modelling methods, the use of Bayesian approach in parameter estimation has been overwhelming among the recent studies (29:4,7,10,11,28,30,31,36,40,42,43-50,52,55,57,58,60,68,70,71,74,77,78). Xu et al. (2015) argued that RTCPMs directly developed with limited data may not capture the underlying relationships between the predictors and the outcome variables. They boosted the model performance by introducing informative priors where the predictors come with a distribution calculated through three different Bayesian meta-analyses - fixed effect meta-analysis, random effect meta-analysis, and meta-regression from existing studies. Finally, they developed a new RTCPM following Markov Chain Monte Carlo (MCMC) simulation-based Bayesian inference approach after refining the data for outliers by Bayesian predictive density analysis. Sun and Sun (2015) and Roy et al. (2016) compared Static Bayesian Network with Dynamic Bayesian Network to construct RTCPM with speed data and concluded that the latter method could capture the time dependency between different time slice data and hence could enhance the model performance. After model building and validation, the performances of the models build with SBN and DBN were compared by Roy et al. (2016). Their results demonstrated that the DBN model is able to predict 8.7% more crash conditions than that of the SBN. Katrakazas et al. (2016) examined the theory and application of a recently developed machine learning technique namely Relevance Vector Machines (RVMs) in the task of traffic conditions classification and found that RVMs could successfully be employed in real-time classification of traffic conditions. They rely on a fewer number of decision vectors, their training time could be reduced to the level of seconds and their classification rates are similar to those of SVMs. Katrakazas et al. (2017) also used two classifiers namely Support Vector Machines (SVMs) – a sophisticated classifier and k-Nearest Neighbors (k-NN) – a relatively simple classifier. The accuracy of both the SVM and k-NN classifiers was found to be consistent with recent studies on real-time collision prediction which used actual collision data along with the corresponding traffic data. To obtain higher accuracy, Roy et al. (2018a) and Yang et al. (2018b) applied Cell Transmission Model (CTM) with Dynamic

Bayesian Network and Deep Neural Network respectively. Roy et al. (2018a) argued that the detector spacing from one study area is highly likely to very from other study areas and demonstrated a CTM based model to transform any detector layout into a predefined detector layout and collected simulated traffic data to replace actual traffic data to construct RTCPM. They applied both BN and DBN and achieved accuracy of more than 84%. Interestingly, they did not find any significant difference between DBN and DN. Yang et al. (2018b) used full data set for RTCPM to overcome the limitation of matched-case control design and used a DNN to construct RTCPM yielding 96% accuracy – the highest accuracy rate so far for any existing RTCPMs.

Finally, most of the studies separated datasets for training and model evaluation. The evaluation process included calculating both accuracy of detection and false alarms. As most of the models yielded probabilities of crash, studies conducted a sensitivity analysis by introducing various threshold values to distinguish between crash and safe traffic conditions (57:4,5,7–11,13,14,16,21,23,24,28–33,35–39,43–55,57-68,70-75,77,78). Xu et al. (2016b) vividly presented the prediction performance of their RTCPMs using receiver operating characteristics (ROC) curve. Apart from these, Wang et al. (2017b) combined the frequency (Poisson log-normal) and the RTCPM (logistics regression) model to boost performance and studied if combing both models could provide better understanding of the crash mechanism. Moreover, they constructed a separate frequency-based model and an RTCPM as baseline models to compare performance. The results showed that the performance of integrated model was better than that of the individual models.

Figure 4 presents the comparison word clouds produced for various components of RTCPMs discussed above and presents at a glance the most frequently adopted approaches by various studies.



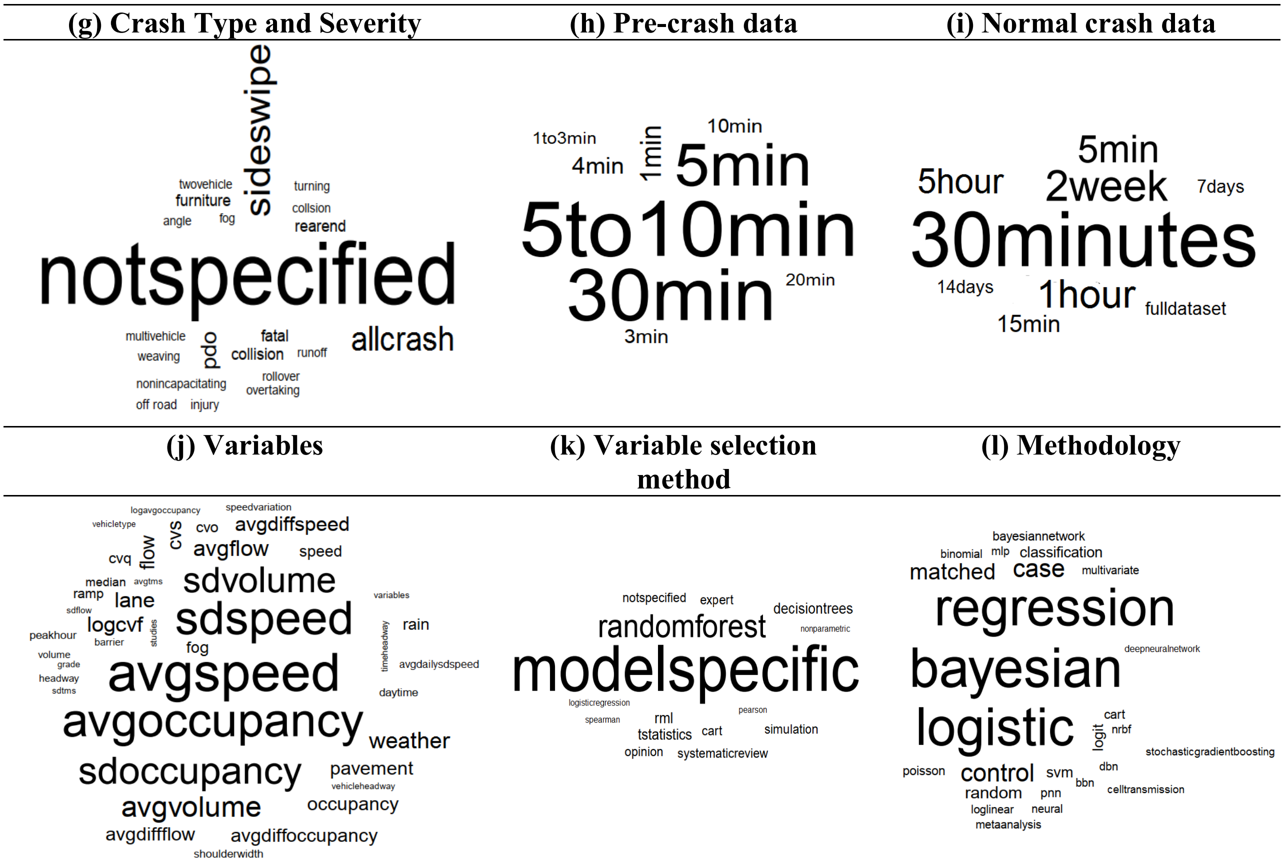



**Figure 4. Comparison word clouds for various components of RTCPMs**

## Design pathway

Summarizing the discussion of the previous subsections, the various major components and subcomponents of RTCPM construction are identified in Figure 4. Based on that, the design pathways followed by various studies have been presented in Table 5. To elaborate, the study with ID 13, i.e., Abdel-Aty and Pande (2006) has been coded as "A1-B2-C2-D1aviii-E1e-F2-G3-H1:15-I2a". Matching the characters with Table 4, it is understood that their main objective was to develop a crash prediction model (A1), they collected traffic data using loop detectors (B2), variable selection method is model specific (C2), they used multivariate logistic regression as modelling method (D1aviii), choose a detector layout of 4 in upstream and 2 in downstream from which data were extracted (E1e), ramp was not considered (F2), did not check whether their proposed model is transferrable to another location (G3), road geometry and time of the day is regarded as non-traffic variables (H1:15) and the outcome variable is the severity of the crash (I2a). It should be noted that in cases where multiple options were present, i.e., here, for non-traffic variables (H), both geometry (1) and time of day were considered (15), the numbers are separated with a colon, i.e., "H1:15".

**Table 4. Taxonomy of components of RTCPMs**

| | | |
|---|---|---|
| **A.** Main Objective | 1. Crash prediction<br>2. Crash mechanism<br>3. Intervention<br>4. Meta-analysis | 5. Combination of 1 and 2<br>6. Combination of 1and 3<br>7. Combination of 1, 2 and 3<br>8. Combination of 1 and 4 |
| **B.** Source of Traffic Data | 1. Probe vehicle<br>2. Loop detector<br>3. Bluetooth | 4. RTMS<br>5. MVDS<br>6. Others (AVI/ Video, RADAR)<br>7. Systematic review |
| **C.** Variable selection method | 1. Specifically mentioned name of the method<br><br>a. t-statistics, b. random forest, c. random multinomial logit, d. Classification tree, e. Simulation, f. common variables in several studies, g. Frequent pattern tree, h. nonparametric Spearman's correlation test, i. p-value and sign of the estimator, j. clustering, k. standard error, l. NLCCA, m. EM algorithm, n. Eigen values, o. Pearson Correlation, p. Logistic Regression<br><br>2. Model specific<br>3. Not specified<br>4. Expert opinion | |
| **D.** Modeling method | 1. Statistical approach<br>a. logistic regression<br>i. matched case control, ii. simple, iii. conditional, iv. sequential, v. Bayesian conditional parameter, vi. Bayesian random parameter, vii. Bayesian, viii. Multivariate, ix. Bayesian matched case-control, x. Multilevel, xi. Multilevel Bayesian, xii. Random parameter, xiii. Mixed, xiv. ordinal<br>b. Aggregated log linear model<br>c. Multivariate Probit<br>d. Bayesian classifier<br>e. Generalized estimating equations (GEE)<br>f. Non-linear Canonical Correlation Analysis<br>g. Bayesian Statistics<br>h. Seemingly unrelated negative binomial<br>i. Poisson, Negative binomial, Zero-hurdle Poisson, Zero hurdle negative binomial<br>j. Bayesian Structural Equation Modelling<br>k. Binary response logit model<br>l. NRBF,<br>m. Binary Logit,<br>n. Bayesian Bivariate Poisson-lognormal model<br>o. UFC<br>p. Bayesian Hierarchical Poisson Model<br>q. Poisson log-normal Model<br>r. Multinomial Logit Model<br>s. Random Parameter Negative Binomial | 2. AI/Data mining<br>a. Neural network<br>i. Simple, ii. Probabilistic, iii. Bayesian, iv. Deep, v. Others<br><br>b. Bayesian Network<br>i. Static, ii. Dynamic<br>c. Classification trees<br>i. CART, ii. SVM, iii. Rule based classifier, iv. RVM<br>d. Genetic algorithm<br>e. Stochastic Gradient Boosting<br>f. k-NN<br>g. PCA<br>3. Others<br>a. Heuristic ad hoc method, and Near-optimal method, b. Fixed effect, Random effect and meta-regression + MCMC simulation-based Bayesian inference, c. Cell Transmission Model, d. ALNEA Ramp Algorithm, e. Surrogate Safety Assessment Model, f. No details provided |
| **E.** Detector layout | 1. Provided with respect to crash<br>a. nearest, b. each in upstream and downstream (1U-1D), c. 2 in both upstream and downstream (2U-2D), d. 3 in both upstream and downstream (3U-3D), e. 4 in upstream and 2 in downstream (4U-2D), f. 5 in upstream and 1in downstream (5U-1D) g. others<br>2. Not provided<br>3. Provided but not on relation to the crash point rather than in the unit of length | |

| | |
|---|---|
| **F.** Ramp consideration | 1. Yes, and modeled separately<br>2. No<br>3. Considered as variable or any other way |
| **G.** Transferability | 1. Checked<br>2. Suggested<br>3. Not checked |
| **H.** non-traffic variables | 1. Geometry<br>2. Pavement<br>3. Weather<br>4. Lighting<br>5. Combination of 1 and 2<br>6. Combination of 1 and 3<br>7. Combination of 1 and 4<br>8. Combination of 2 and 3<br>9. Combination of 2 and 4<br>10. Combination of 3 and 4<br>11. Combination of 1, 2 and 3<br>12. Combination of 1, 2 and 4<br>13. Combination of 1, 3 and 4<br>14. Combination of 2, 3 and 4<br>15. Combination of 1-4<br>16. Time of the day<br>17. not specified<br>18. Traffic Signal |
| **I.** Dependent/Outcome variable | 1. Crash, No crash<br>2. Multiclass<br>a. Crash with severity, No crash<br>b. Crash with type, No crash |

Finally, the correlation plot is presented in Figure 5 to highlight the most commonly undertaken design pathways in existing studies. For proper understanding of the terms used such as 1U-1D, readers are referred to Table 4. The variables for which the correlation values were less than 0.1 were excluded from the diagram. It can be observed that the predominant practice for constructing RTCPMs have been to use loop detectors to collect traffic data, use matched case-control approach for compiling pre-crash and normal traffic data (first proposed by Abdel-Aty et al., 2004 and then followed by many), use logistic regression, Bayesian approaches or vector machine to model the problem. Also, most studies opted for one detector both upstream and downstream or four in the upstream and two in the downstream as the detector layouts of choice to extract data. For pre-crash traffic conditions, most studies also extracted data for 30 minutes from the time of crash occurrence and sliced it into 6 five-minute segments.

**Table 5. Design pathways of reviewed RTCPMs**

| Study ID | Pathway | Study ID | Pathway |
|---|---|---|---|
| 1 | A1-B2-C4-D1b-E3-F3-G3-H6-I1 | 40 | A1-B4-C2-D1avii:2cii-E1b-F2-G3-H17-I2b |
| 2 | A1-B2-C2-D1b-E3-F3-G3-H1-I1 | 41 | A1-B4-C1d-D1ax:1n:1P-E1c-F2-G3-H6-I1 |
| 3 | A1-B2-C1a-D1e-E1f-F3-G3-H11-I1 | 42 | A2-B4-C2-D1p-E2-F2-G3-H6-I2b |
| 4 | A1-B2-C2-D2aii:1g-E1f-F2-G3-H1-I1 | 43 | A1-B2-C1j-D2bi-E2-F2-G3-H11-I1 |
| 5 | A1-B2-C2-D1ai-E3-F2-G3-H17-I1 | 44 | A5-B2-C1l-D1avi-E1a-F2-G3-H5-I2b |
| 6 | A1-B2-C1J-D1f-E2-F3-G2-H10-I2b | 45 | A1-B2-C1b-D1av:1avi-E1a-F2-G3-H6-I1 |
| 7 | A1-B2-C2-D1d:2aii-E1f-F3-G3-H16-I1 | 46 | A6-B2-C2-D1avii-E1c-F3-G1-H16-I1 |
| 8 | A1-B2-C2-D1ai-E2-F3-G2-H17-I1 | 47 | A1-B6-C1b:g-D2bi:f-E1g-F2-G3-H10-I1 |
| 9 | A1-B2-C2-D1ai-E1a-F2-G3-H17-I1 | 48 | A1-B5-C1b-D1avii-E1c-F2-G3-H17-I1b |
| 10 | A1-B1:2-C1a-D1d-E3-F2-G3-H6-I1 | 49 | A1-B2-C1m-D2bii-E1e-F3-G1-H17-I1 |
| 11 | A1-B1:2-C1a-D1g:2aii-E3-F2-G3-H17-I1 | 50 | A1-B2-C1i-D1axi-E1b-F3-G3-H7-I2a |
| 12 | A1-B2-C2-D1ai-E3-F2-G3-H1:16-I1 | 51 | A8-B4:7-C1f-D3b-E1b-F3-G3-H3-I1 |
| 13 | A1-B2-C2-D1aviii-E1e-F2-G3-H1:15-I2a | 52 | A1-B1-C2-D1j:2aiii-E2-F3-G1-H1:16-I2a |
| 14 | A1-B2-C2-D1ai:2g-E3-F2-G3-H3-I1 | 53 | A4-B7-C1f-D3b-E2-F2-G3-H17-I1 |

| | | | |
|---|---|---|---|
| 15 | A1-B2-C1k-D1h-E3-F3-G3-H12-I2a | 54 | A1-B2-C1h-D2ciii-E1b-F2-G2-H17-I1 |
| 16 | A1-B2-C2-D1k-E2-F3-G3-H14-I1 | 55 | A1-B2-C2-D1av:vii-E1b-F3-G3-H5-I1 |
| 17 | A5-B2-C2-D1a-E3-F2-G3-H1-I2b | 56 | A1-B2-C1n-D1aii-E1e-F2-G3-H17-I1 |
| 18 | A6-B2-C2-D3a-E2-F3-G3-H6:16-I1 | 57 | A5-B2-C2-D1avi-E1a-F2-G2-H7-I2a |
| 19 | A1-B3-C4-D3f-E2-F2-G3-H1-I1 | 58 | A1-B2-C2-D2bi-E1b-F2-G3-H17-I1 |
| 20 | A1-B2-C2-D1l:2av-E2-F2-G3-H1:16-I2b | 59 | A1-B1:2-C2-D2civ-E1a-F2-G2-H17-I1 |
| 21 | A1-B2-C1b-D1l:2av-E1b:1c-F2-G1-H1:16-I1 | 60 | A1-B2-C1b-D2bii-E1b-F2-G3-H17-I1 |
| 22 | A2-B2-C2-D1aiii-E1a-F2-G3-H6-I2b | 61 | A1-B2-C1b-D2cii-E1c-F3-G1-H1-I1 |
| 23 | A5-B2-C2-D1aiv:xiv-E2-F2-G3-H3:16-I2b | 62 | A3-B1:2-C2-D2cii:f-E1a-F2-G2-H17-I1 |
| 24 | A5-B2-C1b-D1aii:2ci-E1a-F2-G3-H6-I2b | 63 | A3-B5-C1e:i-D1a:3e-E3-F3-G3-H11-I1 |
| 25 | A1-B2-C3-D1i:o-E1a-F2-G3-H6-I2b | 64 | A1-B5:6-C1d-D1a:2av:2cii-E2-F2-G3-H17-I1 |
| 26 | A6-B2-C2-D1c-E1a-F2-G3-H13-I2b | 65 | A2-B2:5:6-C1i-D1m-E1g-F3-G3-H3-I1 |
| 27 | A6-B2-C1b-D2ci-E1b-F1-G3-H17-I1 | 66 | A1-B2-C1b-D1ai:2cii-E1a-F3-G3-H6-I1 |
| 28 | A1-B2:6-C2-D1aix-E1d-F3-G3-H13-I2b | 67 | A1-B5-C1e:i-D1ai:3d-E3-F3-G3-H11-I1 |
| 29 | A1-B6-C1b-D1ai-E1d-F3-G1-H6-I1 | 68 | A3-B5:6-C1i:o-D1avii:q-E3-F3-G3-H1-I1:2a |
| 30 | A1-B2-C1c-D2bi-E1c-F1-G3-H17-I1 | 69 | A1-B2-C2-D1r-E3-F2-G3-H1-I2a |
| 31 | A1-B6-C2-D1avii-E2-F2-G3-H6-I1 | 70 | A3-B6-C1d-D2aiii:e-E2-F2-G1-H1:16-I2a |
| 32 | A1-B2-C4-D2cii-E1c-F2-G3-H17-I2b | 71 | A1-B2-C3-D2bi:ii:3c-E1a-F2-G2-H16-I1 |
| 33 | A1-B2:6-C1b-D1ai-E1d-F3-G3-H13:16-I2b | 72 | A6-B4-C2-D1axii:s-E3-F2-G3-H3-I2a |
| 34 | A5-B2-C1b:c-D2ci-E1c-F1-G3-H17-I1 | 73 | A1-B2-C2-D1axiii-E2-F2-G2-H1-I2a |
| 35 | A1-B4:6-C1d-D2e-E1d-F2-G3-H13-I1 | 74 | A1-B3-C2-D1av:vi-E1-F2-G3-H3-I1 |
| 36 | A1-B2-C1c-D2bi-E1b-F1-G2-H17-I1 | 75 | A1-B2-C1p-D1a:2aiv-E1b-F3-G3-H17-I1 |
| 37 | A1-B2:5-C1d-D2ai-E1d-F2-G1-H16-I1 | 76 | A1-B5-C1a-D1r-E2-F3-G3-H6-I2b |
| 38 | A1-B2-C2-D1aiv-E1c-F3-G2-H11-I2b | 77 | A1-B3-C2-D1av-E3-F2-G3-H1:18-I2b |
| 39 | A1-B2-C1b-D1m:2d-E1b-F2-G2-H12-I2a | 78 | A1-B2-C2-D2bi-E1a-F2-G2-H17-I1 |

1

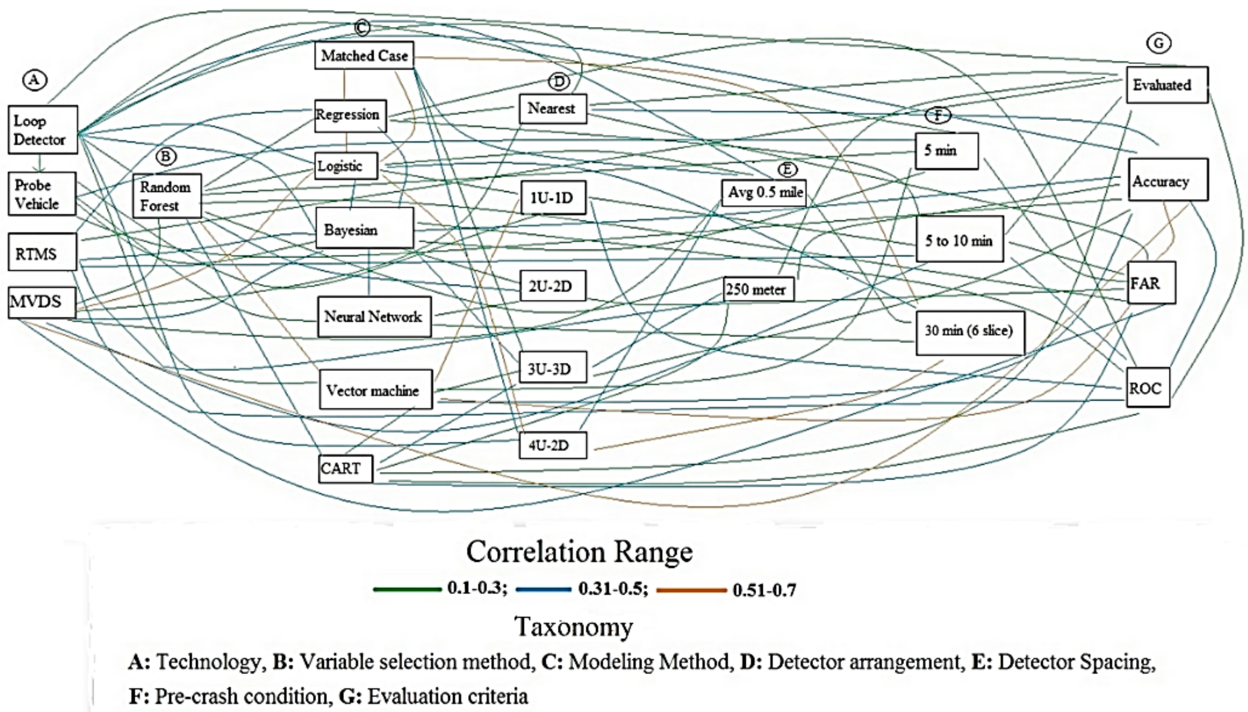



**Figure 5. Correlation analysis of design pathways**

# Ubiquitous Design (UD) Requirements and the State-of-the-Art

## UD requirements

The concept of ubiquitous design (UD) is an evolving paradigm adopted in many fields from art to science and engineering. By UD requirements, this manuscript does not postulate developing a one size-fits all situation model, rather it seeks to encourage the development of RTCPMs that are transferrable, usable and applicable to the widest range of existing (mostly loop detector, infra-red or ultra sound sensors) and future infrastructures (video image processors) to identify hazardous traffic conditions, gain insight into crash mechanism, as well as apply interventions. The review of the problem statements, objectives, limitations and future scopes of the 78 catalogued manuscripts suggest that most of the studies unanimously expected RTCPMs to have high accuracy in hazardous traffic condition detection with low false alarm, be able to explain the underlying determinants of crash using the model predictors, require low sample size to train, be able to predict risk early enough to apply the intervention and be transferrable to other expressways with little effort. Some studies have indicated the importance of using real-time modelling methods, flexibility in including new variables as more data become available, workability during detector failure,. RTCPMs have high resemblance with incident detection systems as both use high resolution sensor data and model the problem for dichotomous outcome and perform in real-time. Incident detection concept is now available as a commercial technology. Abdulhai and Richie (1999) have outlined the UD requirements of an incident detection system with several capabilities and attributes which were re-classified and aggregated by Zhang and Taylor (2006). By combining findings from the literature on RTCPM, theoretical reasoning, knowledge from incident detection systems and experience, this manuscript envisions RTCPMs to possess these capabilities and attributes – practicality, performance, knowledge generation ability, flexibility, transferability, adaptability and timeliness and robustness. As the general meaning of these terms can be overlapping, the following subsections present the contexts in which this manuscript has catalogued them.

*Practicality (PR)*

"Practicality" has several dimensions for RTCPMs. This includes:

i) Detector layout and spacing: a practical RTCPM is expected to bind the crash risk with both time and space. The current modelling paradigm expects RTCPMs to be implemented on existing instrumented highways or future highways that will be equipped with various kinds of traffic sensors. However, as a cost effective solution, it is essential to consider highways that currently do not have sensors but may install those to monitor their hotspots or locations of high interest. Therefore, RTCPMs are expected to come with recommended detector layout and spacing with allowable deviation (minimum-average-maximum-standard deviation) that can be implemented to an existing instrumented highway or highway authorities can install detectors based on the supplied specifications to monitor locations of interest,

ii) Intervention friendliness: RTCPMs have no practical meaning if they do not provide ample time for an intervention to make an impact by improving safety after detecting an evolving unsafe traffic condition considering human cognitive ability to adapt to an intervention in the form of variable message sign (VMS), variable speed limit (VSL) or ramp metering. For example, studies focused on real-time interventions to reduce crash risk recommended to maintain a 5 to 10 minute lead time for the intervention to take effect. Lee et al. (2004) experimented with various variable speed limit (VSL) strategies for both short (2 min) and long (5 to 10 min) durations and concluded that the former situation increased crash potential due to more frequent speed limit changes. However, the later strategy was found to maximize safety benefits for the freeway segment examined in the study. Abdel-Aty et al. (2007) found that sudden reduction in speed limit by 15 mph two miles directly upstream through VSL and subsequent raising of the speed limit by 15 mph two miles directly downstream of the station of interest starting 5 to 15 minutes prior to crash reduces the crash potential most efficiently for moderate to high-speed traffic operations. In a later study, Abdel-Aty et al. (2008) recommended to maintain a buffer of a minimum 5 to 10 min to let VSL make significant impact in reducing crash risk. Therefore, it is recommended that the RTCPMs shall allow a buffer time of 5 to 15 minutes after an intervention has been applied;

iii) Predictors: RTCPMs should be developed based on the variables that are readily available. Most of the existing sensor technologies can yield data relating to flow, speed and occupancy. However, sophisticated surveillance systems such as video based detection systems can yield headway, time mean speed, space mean speed and lateral distance between vehicles. They can provide data based on each lane and also for very short time window as well. However, these technologies are expensive and not seen often on existing instrumented highways. Hence, the model should be based on variables that are easy to be yielded by most common existing sensors.

*Performance (PE)*

RTCPMs are expected to have high detection rate triggering low false alarm rate, which is essential to avoid unnecessary introduction of interventions. From the existing literature, it was found that the 85$^{th}$ percentile value of successful crash detection was 81.4% whereas 15$^{th}$ percentile value of false alarm was 6.02%. This suggests that 15% of the studies reported crash detection rate to be higher than 81.4%. At the same time, around 15% of the existing literature could develop RTCPMs with less than 6.02% false alarm rate. Also, it is expected that RPCMS will report their prediction capability through ROC (Receiver Operating Characteristic) curves. This way, the concerned authorities will have the flexibility to set the threshold to choose between high tolerance for false alarm to prevent severe crashes and for property damage or opt for low false alarm triggering interventions only for high risk traffic conditions.

*Knowledge Generation (KG)*

A RTCPM is normally expected to be constructed with a small sample size as crashes are rare events and it is also challenging to obtain a large dataset with synchronized crash and sensor data. This creates a dilemma as in one hand, rare events leave little opportunity to learn about the phenomena and on the other hand, the prediction model has to train itself to draw inference about the probability of a crash occurring using very few cases. However, once in operation, a RTCPM can be continuously fed with new data and whether the data is associated with crash or no crash situation is also revealed almost instantaneously. Therefore, it is expected that the adopted modeling methods will have the capacity to learn from new data as it is being fed into the system and be able to enrich its insight about the crash mechanism. This will facilitate understanding why crash happens leading to arming the RTCPMs with more appropriate variables which will eventually enable such models to perform better and applying countermeasures through adaptive dynamic operational models more appropriately.

*Flexibility (FX)*

Different studies employed different sets of variables to build the RTCPMs based on the data that were available to them. At times, the newly introduced variables were surrogate in nature to capture a specific attribute of which data were not available. Moreover, it is expected that the available data on all the variables may not come from the same time period. RTCPMs should have flexibility to add new variables with little effort, i.e., without needing to re-build or re-calibrate the whole model. Moreover, it is expected to have the capability to update itself with partially available data.

*Transferability, Adaptability & Timeliness (TAT)*

Building an RTCPM from the scratch is resource demanding and infeasible to perform frequently. Therefore, such models must not be bounded by spatiotemporal constraints. Both the theory and the logic should be accommodative enough to be transferred to a new expressway with limited effort. Moreover, traffic characteristic on urban expressways can be influenced by its surrounding urban development. Hence, the models are expected to have the capability to both learn (from new data) and fed away the older prior beliefs in short time intervals to address the timeless issue. Various space state models have recently developed along with adaptation and fading algorithms to accommodate such requirements. A few studies have demonstrated the issue of transferability (e.g. Abdel-Aty et al. 2005; Abdel-Aty and Pande, 2004; Hellinga and Simimi, 2007). For instance, Abdel-Aty et al. (2008). Later, transferability issues were studied by Shew et al. (2013), Xu et al. (2014c), Sun and Sun. (2015, 2016), and Xu et al. (2015). Katrakzas et al. (2017) used the k-nearest neighbour method which is easily transferrable because they do not require prior knowledge of any datasets. Quite recently, Roy et al. (2018) used CTM to present a framework addressing spatial transferability issue where the existing detector layout can be supplied as an input yielding simulated traffic flow data for a predefined detector layout as output which was eventually used to construct the RTCPMs.

*Robustness (RB)*

Detector failure is a common event resulting in extraction of data for only a subset of model variables. RTCPMs are expected to acknowledge this hindrance and be able to make inferences under such circumstances. Moreover, in the case of a complete detector failure, the model must be able to use data from alternative detector layouts to continue predicting the crash risk without substantially compromising its overall accuracy (e.g., Ahmed and Abdel-Aty, 2013). The first requirement can be addressed by employing modelling methods that can make inferences when data on some variables are missing. ROC curves should be produced evaluating the model performance for distinct situations when one of the detectors fails to yield speed or occupancy or flow data. At the same time, as these models extract data from a specific set of detectors, they should also identify the second and the third best detector layouts and report their performance when data from these detectors are used for prediction. For example, the most prominent (7 studies) choice of detector layout has been to extract data from four detectors - two from the nearest upstream and two from the nearest downstream (7:30,32,34,46,48,49,61) with respect to the crash location. Now, in case one of these detectors, say, the second nearest downstream detector fails, then the data from the third nearest downstream can be extracted replacing the variables of the second nearest detector. During the model building process, results from such alternative detector layouts should also be reported in the form of ROC curves. Now, it is quite natural to expect that when data from a variable will be missing or the second or the third best detector layout will be used to make inferences, performance of both the detection and false alarm rates may be compromised. However, when the corresponding ROC curves are provided, the relevant traffic authorities will have option to decide whether to make an inference under such circumstances.

## Universal Design Requirements Evaluation and State-of-the-art

This section evaluates the design pathways of reviewed RTCPMs to identify the extent to which they fulfil the UD requirements based on these criteria: variable space, detector layout and spacing, prediction lag time for intervention, modelling method and model performance evaluation process. Each criterion was associated with certain set of capabilities and attributes as outlined in Table 4. To illustrate, Table 6 suggests that the criteria 'modelling method' will primarily be judged by its knowledge generation ability, flexibility, transferability, adaptability and timeliness, and robustness capabilities.

**Table 6. Evaluation criteria and Universal Design requirements**

| Evaluation Criteria | PR | PE | KG | FX | TT | RB |
|---|---|---|---|---|---|---|
| Variable space | ✓ | | ✓ | ✓ | ✓ | ✓ |
| Detector spacing | ✓ | | | | ✓ | ✓ |
| Prediction lag time for intervention | ✓ | | | | | |
| Modeling method | | | ✓ | ✓ | ✓ | ✓ |
| Model performance | | ✓ | | | | |

The performance of each criterion was arranged as high (H), medium (M) or low (L). Figure 6 presents the performances of the reviewed studies in this manuscript evaluated against the UD requirements. As some of the capabilities and attributes are spanned over multiple criteria, e.g., PR is evaluated for variable space, detector layout and spacing and prediction lag time for intervention, the grades are presented with 3 letters with HLM for example, meaning high for variable space, low for detector layout and spacing and medium for prediction lag time for intervention. The following subsection presents the grading system along with corresponding rationales.

*Variable space*

Performance (PR) - The variable space of a manuscript for PR is rated to be 'H' if it has only utilized speed, flow or vehicle count and occupancy data and their various statistical forms (e.g., standard deviation, coefficient of variation) or mathematical transformations (e.g., logarithmic) as traffic flow variables along with road geometry (static infrastructure) or simple weather (precipitation in Boolean form) and lighting condition (high/low/medium visibility) as variables. The manuscript falls down to 'M' category if they fulfil the requirements of 'H' category but does not include road geometry, weather, lighting related basic variables as outlined in 'H' category. The remaining studies are termed as 'L'.

Knowledge Generation (KG)- Several studies explaining crash phenomena suggest that the differences in traffic conditions, both laterally and longitudinally, are associated with crash (8: 2,16,20,48,50,63,67,72). In addition, the association of ramp with crash at close to a ramp zone is well established. Consequently, to be able to provide insight into crash mechanism the models are expected to incorporate data obtained from different sections of road – both longitudinal and lateral sections and consider ramp as a variable - which may be introduced as a dichotomous variable with two outcomes such as near ramp or basic freeway segment or a continuous variable represented by distance from the crash location. The studies fulfilling these requirements are classified as 'H' for KG category. The manuscripts only considering longitudinal differences are given 'M' and the remaining studies are labeled as 'L'.

Flexibility (FX) and Robustness (RB)- To be highly robust (H) and flexible, we expect the models to take input from more than one detector location for both in the upstream and downstream from a crash site. If they have considered more than one detector location for either an upstream or a downstream, those are categorized as 'M' and the remaining studies are termed as 'L'.

Transferability, adaptability and Timeliness (TAT) - For transferability, adaptability and timeliness, the model needs to adjust to a large set of detector arrangements. For TAT, the manuscripts that obtained 'H' for both PR and FX categories are marked as 'H'. If they have received 'L' in any of those two categories then they are labeled as 'L' and the rest are classified as 'M'.

*Detector layout and spacing*

PR - For PR, the studies that provided detector layout and spacing, i.e., number of detectors required and their average distance along with standard deviation, are awarded 'H'. If the average, maximum and minimum values are provided then they are categorized as 'M' and other specifications are labeled as 'L'.

RB and TT – These two UD requirements expect greater flexibility in the detector arrangements to accommodate to the new infrastructures and to continue its operation in case of detector failures. Hence, the studies qualifying as 'H' and involving at least data from two detector locations are classified as 'H', those obtaining 'M' for PR but make use of more than one detector locations are labeled as 'M' and the remaining categories are graded as 'L'.

*Prediction lag-time for intervention*

Existing studies on real-time interventions and most of the reviewed studies have heavily insisted on providing a buffer time of at least 5 minutes for an intervention to take effect. At this moment, due to lack of ample studies on intervention design, it is difficult to comprehend whether the 5 minute time gap between crashes is an over or under estimate for the intervention to set in. Studies acknowledging a minimum lag time between expected crash time and the detection of such evolving situation are graded as 'H' and otherwise as 'L'.

*Modelling method*

KG - The main choices in modelling for RTCPM have been among various types of logit and probit regressions models, different forms of neural networks, Bayesian networks and classifying methods, such as classification and regression trees (CART), SVM, RVM or simple rule-based classifier. From the perspective of knowledge generation, Bayesian network and classification based methods have advantages over other methods. Both the methods have graphical representations, making the interrelationship among variables easy to comprehend. Bayesian network builds a directed acyclic graph using conditional independence and probabilistic parameter estimates where the variables are presented as nodes and their interrelationships are demonstrated with edges. It has several structural learning algorithms that help in understanding the interrelationship among variables. Classification based methods mainly direct in which way to classify an observation. Keeping 'crash' as a dependent variable, it can identify certain combinations of values that different variables can take which will have high association with crash. Li et al. (2012) verified that SVM model can also be used to evaluate the impacts of explanatory variables on crash injury severity using the sensitivity analysis. Qu et al. (2012b) suggested that SVM classifiers regarding roadway and environmental conditions may produce decent accuracies. Katrakazas et al. (2016) stated that RVMs can successfully be employed in real-time classification of traffic conditions and their classification rates are similar to those of SVMs. On the contrary, logit and probit regression models are statistical methods where they identify high association between crash and its predictors. They can also present the odds of a variable being associated with crash. However, traffic flow variables such as speed, flow and occupancy are highly correlated in nature (Gazis, 2002). Therefore, most of the highly correlated variables are dropped revealing the underlying determinants only partially. Finally, neural networks are efficient in making prediction but lack the ability to reveal the interrelationship among variables due to the unexplainable hidden layers. Hence, Bayesian network, Stochastic Boosting Gradient Algorithm, classification based and methods with similar advantages are graded as 'H' from the knowledge generation perspective whereas logistic regression and neural network-based methods are graded as 'M' and 'L' respectively.

FX – For RTCPMs methods that can accommodate new variables in future and learn from new data in course of time without requiring re-building or re-calibrating the whole model are highly desirable. Also, sensors may fail to yield data on some variables in real-time operation. A robust model should be able to perform under such situations. Both Bayesian network and neural network based methods can be easily transformed into real-time models, and hence, graded as 'H' for flexibility. Abdel-Aty et al. (2008b) and Shew et al. (2013) empirically addressed the issue by calibrating and subsequently evaluating the logistic regression-based models with new data for different expressways. However, the approaches were more in line with re-building or re-calibrating, rather porting an existing model to a new expressway and updating it in real-time as new data becomes available. Hence, logistic regression and probit models are categorized as 'M'. Recently, some studies have applied SVM as a real-time modelling tool through improvisation of the algorithms in the hardware level. Nashat et al. (2011) accomplished that by introducing multi-core processor with advanced multiple-buffering and multithreading algorithms. Kyrkou et al. (2016) attained acceleration by cascading SVM through a customized hybrid processing hardware architecture optimized for the cascade SVM classification. Hence, as the advantage of real-time modelling can be obtained mainly through hardware optimization, SVM and RVM based methods have also been assigned 'M' grade. Classification tree based methods lack in these flexibilities and therefore fall into 'L' category.

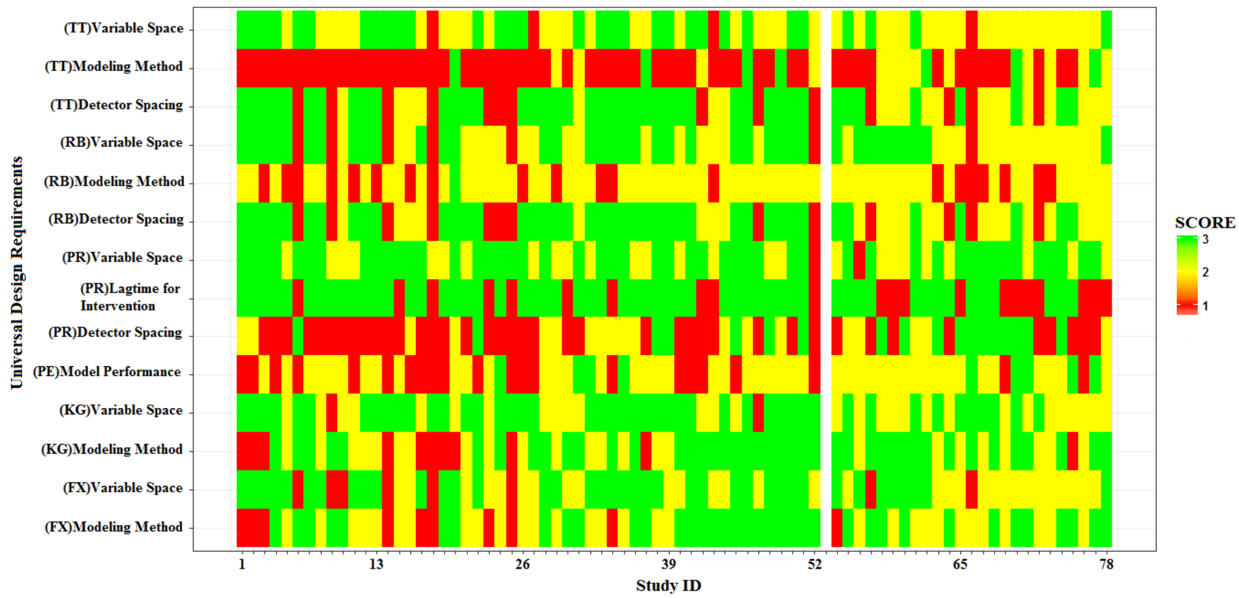
TAT– The requirements to score 'H' in TAT the model has to fulfil the requirements of FX - 'H' and demonstrate its transferability on a different study area. At the same time, it needs to demonstrate or mention how the model is expected to keep itself updated by applying fading and learning algorithms. If only the transferability has been demonstrated, 'M' grade, and otherwise 'L' grade was awarded.

RB – For robustness as well, to obtain 'H', the model is requited to demonstrate its performance during the situation of detector failure. When such demonstration was not provided, if the model's performance was 'H' in FX, it was awarded 'M' and otherwise 'L'.

*Model performance*

Model performance is rated as 'H' if the manuscript has achieved at least 81.4% accuracy in detecting crashes with a false alarm rate less than 6.02%, which are the respective approximate 85$^{th}$ and 15$^{th}$ percentile reported values calculated from the reviewed literature. The values are quite promising as they suggest that researchers have already achieved high accuracy in crash prediction as 15% of the studies considered here could accurately predict at least 81.4% of the crashes and 15% of the studies reported a false alarm less than 6.02%. Also, they are expected to either provide an ROC curve or some form of sensitivity analysis to understand the interaction between false alarm and crash prediction accuracy. If the manuscript has not provided ROC curve or sensitivity analysis, but fulfilled the accuracy aforementioned accuracy requirements they are categorized as 'M' and otherwise as 'L'.

Figure 6 illustrates that the chronological improvements in RTCPM development in the form of a heat map where transition from red to green means progression from a 'L' score to 'H'. At times 'X'=0 grade was assigned for situations when the manuscript did not provide ample information to complete the categorization. A universal RTCPM is expected to score 'H' in all six capabilities and attributes. Figure 6 suggests that RTCPMs over the time have made commendable progress, though a substantial improvement is expected, especially in addressing these UD requirements - flexibility, transferability, adaptability, timeliness and robustness.




(SCORE: H = 3; M = 2; L = 1; X = 0)

[*NB: Study ID# 53 is on Meta-analysis*]

**Figure 6. Heat map of state-of-the-art and UD requirements**

## Proposed Framework for a Universal RTCPM

From the preceding discussion, it can be synthesized that recent RTCMPs have made significant improvements over their predecessors in terms of practicality, performance and knowledge generation though there is still a long way to progress in terms of flexibility, transferability, adaptability & timeliness and robustness. Building on the achievements till date and addressing shortcomings, a framework for constructing a universal RTCPM is proposed as illustrated by Figure 7.

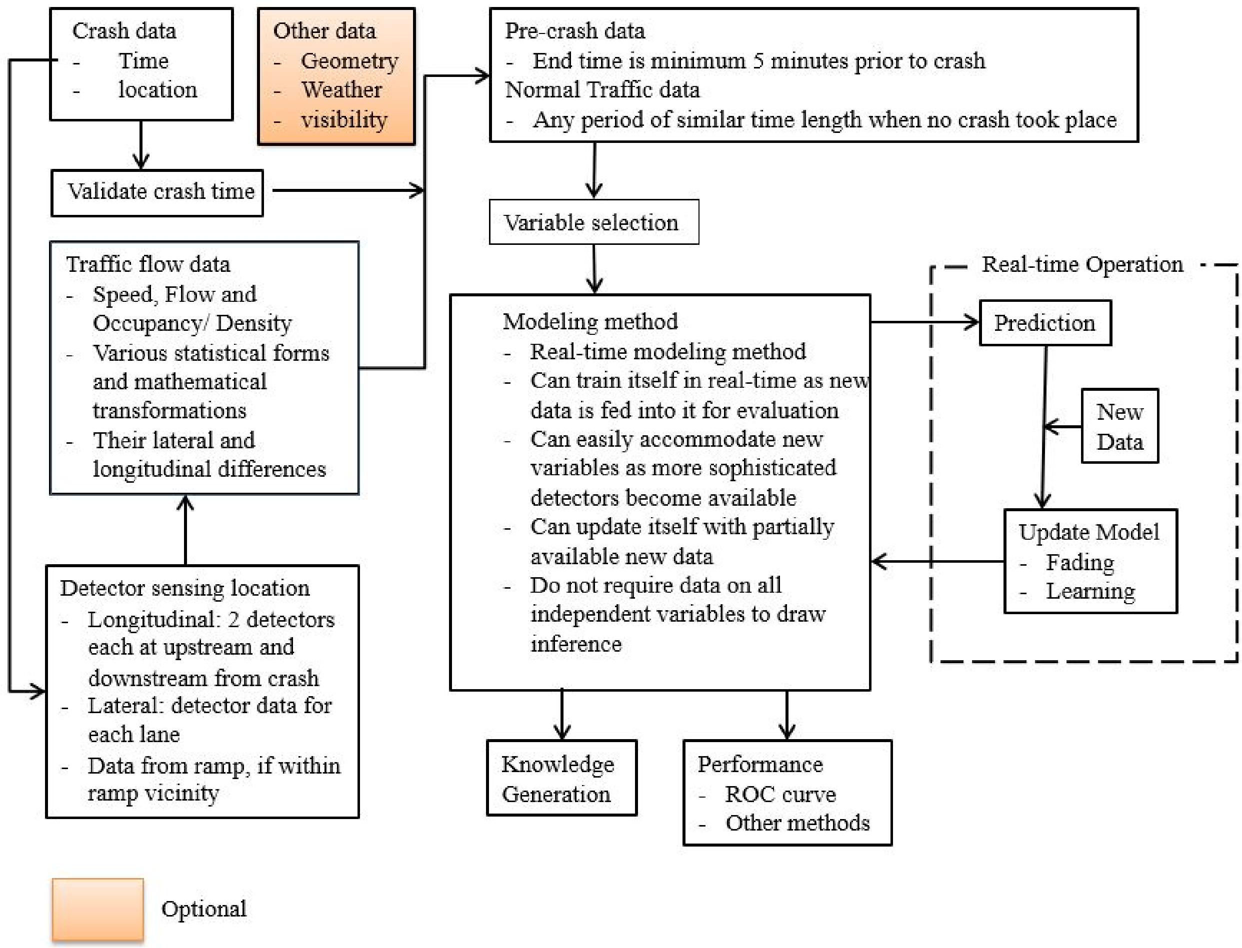



**Figure 7. Proposed Framework for a Universal RTCPM**

The process of developing RTCPMs in the proposed framework commences with collecting crash data containing at least crash time and location information. If there is no camera installed then the accurate crash time can be verified using various methods, such as, detecting sudden drop in speed - often by plotting speed profile, identifying backward-forming shockwave upstream of the crash location, applying shock-wave and rule-based methods, spotting speed and flow variation between adjacent lanes, drawing speed contour plots, etc. For each crash, at least two nearby locations both in the downstream and the upstream will be identified. These locations will either be equipped with detectors or probe vehicles (or connected and autonomous vehicles as discussed in a later section) will be in operation in those areas to supply real-time traffic state data. For crashes within ramp vicinity, the nearest location on the ramp should also be identified and their corresponding data will also be collected.

In the next step, pre-crash and normal traffic conditions will be defined. A substantial number of studies have suggested that data for the first five minutes before the reported crash time should not be mingled with pre-crash data to provide buffer from any crash time errors as well as for the intervention to set it. If the trust associated with the crash time is not high then the actual crash time should be determined empirically. Pre-crash traffic conditions can be further broken down into small chunks of a suitable time, e.g., 5 minutes for up to 30 minutes before the time of crash. Researchers can take the liberty in choosing the normal traffic condition from the detector database for any typical time of day when no crash took place. It is important to ensure that the various congestion levels are well represented in the normal traffic condition data. Next, from the chosen detector location, the pre-crash and normal traffic data should be extracted. The basic model is expected to be constructed with speed, flow and occupancy data, their various statistical forms and mathematical transformations as well as their differences in longitudinal and lateral spaces. However, when available, data on basic road geometry, real time weather (can be a Boolean representation of precipitation data), and visibility (can be categorical – clear, low and very low) may improve model prediction performance as well as unveil underlying relationships among traffic states, geometry and environment. As such, a variable space is expected to be substantially large as compared to the sample size, an appropriate variable selection method such as random forest, random multinomial logit models, and classification and regression trees can be used to identify the most important variables. Afterwards, the problem can be modelled with a method dealing with a dichotomous outcome.

In order to develop universal RTCPMs, one needs to consider a number of factors outlined above. The wish list includes making inference in real-time, producing a high prediction success with a low false alarm, making inference with missing or surrogate data, forgetting and relearning when transferred into a new environment, adding and dropping variables to suite the requirements of a new environment, recalibrating itself to draw inference giving emphasize on newer datasets, i.e., ability to learn and at the same time fade/unlearn/forget the prior belief earlier than a prescribed time. The literature suggests that logistic regression had been a method of choice in many of the early RTCPMs. However, their use has reduced in recent literature due to their various limitations as compared to AI based methods, such as, inability to model with highly correlated variables, lacking real-time updating ability, drawing inferences in case of missing data, updating model with partial data or easily dropping or incorporating new variables. Researchers have addressed some of these issues by adopting real-time modelling methods that include various forms of Bayesian networks, neural networks or advanced real-time implementations of SVM, etc. The results exhibited high prediction accuracy. In many fields, researchers have started to employ deep learning (Deep Neural Network, DNN) to improve on their prediction performance over ANN and it is expected that application of such methods in RTCPMs may further boost the model performance. Apart from that, a universal RTCPM is expected to generate knowledge by providing insight into the underlying mechanism of crash from real-time traffic states and thereby facilitate the design of relevant interventions. Although this conflation between predictive and explanatory modeling methods are common, often the best performing models do not serve both prediction and explanatory purposes equally well as where the former is focused on measuring the value of ‘y’ accurately, the latter is more concerned about finding a relationship with the set of ‘x’ that best represents ‘y’. The dilemma is, as suggested by Shmueli (2010), that often relatively less structured models can outperform the true explanatory model with respect to model performance. For example, neural network-based models lack the causal theory but are excellent in prediction. A solution to this can be the use of probabilistic graphical models such as Bayesian Network or Dynamic Bayesian Network which can do both prediction and the exploration of underlying mechanisms. Such models are robust against missing data, have the flexibility to both learn and forget when transferred to a new environment and fed with new data, can easily add and drop variables through partial calibration of their condition probability tables. At the same time, they are equipped with sophisticated supervised and unsupervised learning algorithms that can help in producing causal diagrams for knowledge discovery. Another solution is to develop models for prediction by employing, for example, DNN or such methods developed mainly for high prediction accuracy and then developing separate models, for example, classification trees, to unveil the crash mechanism. A similar approach can also be followed for prediction and subsequent intervention design where static optimization models can be employed for prediction and dynamic operational models, which are often adaptive in nature, can be used for introducing interventions to bring the hazardous traffic conditions back to normal. Finally, the models should come with an ROC curve for the decision makers to prioritize their objectives, i.e., lower tolerance for hazardous traffic, low false alarm, etc.

## The Future of RTCPM

The idea of road transportation, as it is known today, is likely to transform radically due to the accomplishments in the fields of information technology, vehicle automation, rapid urban densification, challenges in the energy sector, and of course, due to the growing needs for environment friendly sustainable living. With this future transformation in mind, it is important to explore how RTCPMs can still play a major role in improving traffic safety.

At present, most of the existing models are developed for interstate freeways and expressways as they are highly access controlled and traffic flow on these types of roads are uninterrupted – reducing variability and complexity of model construction. However, these types of roads represent a very small share of the existing road-based transportation network. It is expected that in near future studies such as Yuan et al. (2018) dealing with arterials and Dimitriou et al. (2018) considering urban streets and intersections will grow in number and expand into most of the major road classes (e.g., arterial, collectors, rural roads, etc.) and locations on road (e.g., at intersections, rail crossings, bus stops, etc.). At the same time, a quest for further improving the level of accuracy will continue as new methods, such as, deep learning (Yang et al., 2018b), DBN (Roy et al., 2018a), emerge. Apart from that, Abdel-Aty et al. (2018) postulated that in near future, RTCPMs will also be used in conjunction with congestion pricing as well as in route choice decision and this manuscript agrees that such developments may take place soon.

Another technological factor that is expected to complement the RTCPMs in near future is the introduction of a disruptive technology – Connected and Autonomous Vehicles (CAVs). Both these technologies follow some basic procedures: real world environment perception and model building, path planning and decision making and motion control (Cheng, 2011). It may be argued that in future all the vehicles may become connected and autonomous making RTCPMs obsolete. However, prior to that, it is quite likely that during the transition period CAVs and human driven vehicles will be sharing the same roads for years as road infrastructure, CAV technology and related legislations will need to be standardised. Some studies in these directions have already emerged. Wang et al. (2017) compared traffic state for mixed human and automated traffic flows.

Nilsson et al. (2017) compared performance of lane change maneuvers using automated driving approach and manual driving to improve driving automation.

RTCPMs, AVs and CVs are safety enhancing technologies having similar data requirements and their underlying models heavily depend on situation awareness. It is expected that these concepts in future will be complementing each other – where CAVs can be a great source of high resolution accurate data for RTCPMs and the RTCPM can act as an input to further enhance the decision making and risk assessment of CAVs. Trajectory planning in CAVs involves real-time planning of actual vehicle transition from one feasible state to the following, satisfying the vehicle's kinematics limit (Katrakazas et al., 2015). This planning evaluates the safety state in each time stamp and generates safety warnings and alerts whenever the vehicle transition is found to be risky. In line with this idea, Liu and Khattak (2016) explored the potential of using Basic Safety Messages (BSMs) that is transmitted by CVs. It will be interesting to investigate how the real-time crash probability can be used as an input to further improve trajectory planning of CAVs, especially in the area of risk assessment. In addition, the drawbacks in the current RTCPMs are mainly related to flexibility, transferability, adaptability & timeliness and robustness – all of which directly depend on reliable sources of data and modelling methods to accommodate a large variable space, of which, the former can be addressed as more autonomous vehicles and CVs become available in the network. Khan et al. (2017) combined the data of CV with artificial intelligence and demonstrated that their method could generate density data with minimum 85% accuracy when CV penetration reaches at least 20%. Grumert and Tapani (2018) combined the speed and position data of connected vehicles with sparsely located stationary detector data to estimate speed, density and traffic state, such as, lower speed and flow and higher density and suggested that the outcome of the study can be used to formulate suitable traffic control strategies. In the future, as the RTCPMs reach closer to being practice ready, the field is expected to see substantial effort to be put in RTCPMs based intervention design to capitalize the benefit of being able to predict crash in real time. With the possibility of 75% newly manufactured vehicles to be equipped with some sort of connected vehicle technology by 2020 (Coppola and Morisio, 2016), in future, driving algorithms of CAVs may also influence – even direct and control the driving pattern of partially or fully human driven vehicles to reduce crash probability in real-time and be the part of a formidable intervention strategy alongside ramp metering, VMS and VSL.

# Conclusion

Driving a vehicle or being in it is one of the most dangerous activities that people in the motorized societies perform on daily basis. With the advent of sophisticated ITS based technologies, researchers and road authorities are devoting substantial effort to make travelling safer for road users. A RTCPM is one of such initiatives transitioning from its infancy to an applicable technology. Once this is mature, it can become an integral part of real-time proactive road safety management system where the safety hazards can be identified well in advance and interventions can be applied to return the traffic back to normal. At present predicting crash risk in real-time is still limited within an idea which is not ready for deployment. This study conducted a systematic review on the state-of-the-art of real-time crash prediction models to synthesize and find coherence among the existing ideas and to identify the key components of a RTCPM along with outlining the design pathways followed by various studies. Six capabilities and attributes were defined – practicality, performance, knowledge generation ability, flexibility, transferability, adaptability & timeliness and robustness – as the universal design requirements for such a model based on the limitations and future recommendations outlined by the literature as well as by investigating universal requirements of similar models. Afterwards, it evaluated the existing literature against the newly proposed universal design requirements. It was observed that the chronological development in real-time crash prediction has been encouraging and substantial progress has been made in practicality, performance and knowledge generation perspectives. However, the state-of-the-art lacks in flexibility, transferability, adaptability & timeliness and robustness. The discussion in this manuscript also suggests that a solution to these existing limitations are mainly attributed to reliable real-time data availability and modelling methods used. Researchers can explore the opportunities that integration of AV and CV technologies have to offer by acting as source of real-time data. In fact, RTCMPs and CAVs in future may get interlinked to extract symbiotic benefits to both the technologies as RTCPMs may assist CAVs in improved trajectory planning and CAVs may complement RTCPMs by providing data and assisting in intervention designs. Regarding modelling methods, dilemma between predictive and explanatory modelling were highlighted as the models specialized in prediction are not necessarily the best in knowledge discovery and vice versa. In this regard, the benefits and flexibilities of AI based cutting edge modelling methods, such as, dynamic Bayesian network, deep learning were discussed. At the same time, the possibility to use separate models for prediction and knowledge generation were also discussed. As a guidance towards solution of the remaining challenges, the manuscript also proposes a framework to construct a universal RTCPM. It is to be noted that the framework is a demonstration on how to accommodate the remaining challenges rather than a stringent guideline and the authors acknowledge that future researchers may follow different pathways to fulfill these universal requirements. The authors do not claim the proposed framework to be the best and rather would like it to be considered as a framework that leads to the development of a RTCPM that fulfils the minimum universal requirements.

The study expects to be a one stop knowledge source for future and continuing researchers and hopes that the presented framework for developing a universal RTCPM will reduce their learning curve and ensure a faster transition of RTCPM from idea to technology.

**Acknowledgement**

The research study was supported by JSPS KAKENHI Grant [KIBAN(C)-#25420535].

## References


Abdel-Aty, M., Abdalla, M. F., 2004. Linking Roadway Geometrics and Real-Time Traffic Characteristics to Model Daytime Freeway Crashes. *Transportation Research Record: Journal of the Transportation Research Board*. 1897, 106–115.

Abdel-Aty, M., Cunningham, R. J., Gayah, V. V., Hsia, L., 2008a. Dynamic Variable Speed Limit Strategies for Real-time Crash Risk Reduction of Freeways. T*ransportation Research Record: Journal of the Transportation Research Board*. 2078, 108-116.

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006b. Evaluation of Variable Speed Limits for Real-time Freeway Safety Improvement. *Accident Analysis and Prevention*. 38(2), 335-345.

Abdel-Aty, M., Dilmore, J., Hsia, L., 2006a. Applying Variable Speed Limits and the Potential for crash mitigation. *Transportation Research Record: Journal of the Transportation Research Board*. 1953, 21-30.

Abdel-Aty, M., Gayah, V., 2010. Real-Time Crash Risk Reduction on Freeways using Co-ordinated and Uncoordinated Ramp metering approaches. *Transportation Engineering*. 136, 410-423.

Abdel-Aty, M., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. *Transportation Research Part C: Emerging Technologies*. 24, 288–298. http://doi.org/10.1016/j.trc.2012.04.001

Abdel-Aty, M., Pande, A., 2004. Classification of Real-time Traffic Speed Patterns to Predict Crashes on Freeways. *TRB Annual Meeting for Journal of Transportation Research Board*, November.

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. *Journal of Safety Research*. 36, 97–108. http://doi.org/10.1016/j.jsr.2004.11.002

Abdel-Aty, M., Pande, A., 2006. ATMS Implementation System for Identifying Traffic Conditions Leading to Potential Crashes. *IEEE Transactions on Intelligent Transportation Systems*. *7*(1), 78–91.

Abdel-Aty M., Pande A., 2007. Crash Data Analysis: Collective vs. Individual Crash Level Approach. *Safety Research*. 38, 581-587.

Abdel-Aty, M., Pande, A., Das, A., Knibbe, W. J., 2008b. Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems. *Transportation Research Record: Journal of the Transportation Research Board.* 2083*,* 153–161. http://doi.org/10.3141/2083-18

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., Santos, C.D., 2007. Crash Risk Assessment using Intelligent Transportation Systems Data and Real-time Intervention Strategies to Improve Safety on Freeways. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. 11(3), 107-120.

Abdel-Aty, M., Pemmanaboina, R., 2006. Calibrating a Real-Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data. *IEEE Transactions on Intelligent*

*Transportation Systems*. 7(2), 167–174.

Abdel-Aty, M., Pemmanaboina, R., Hsia, L., 2006c. Assessing Crash Occurrence on Urban Freeways by Applying a System of Interrelated Equations. *Transportation Research Record: Journal of Transportation Research Board.* 1953, 1–9.

Abdel-Aty, M., Shi, Q., Pande, A., Yu, R., 2018. Chapter 9: Real time traffic operations and safety. In Lord, D and Washington, S (eds). *Safe Mobility: Challenges, Methodology and Solutions*. Emerald Group Publishing Limited. 11.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., Hsia, L., 2004. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board.* 1897, 88–95.

Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. *Transportation Research Record: Journal of the Transportation Research Board*. 1908, 51–58.

Abdel-Aty, M., Wang, L., 2017. Implementation of variable speed limits to improve safety of congested expressway weaving segments in microsimulation. *Transportation Research Procedia*. 27, 577-584.

Abdulhai, B., Ritchie, S.G., 1999. Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. *Transportation Research Part C: Emerging Technologies*. 7, 261-280.

Ahdi, F., Khandani, M.K., Hamedi, M., Haghani, A., 2012. Research Report on Traffic data collection and anonymous vehicle detection using wireless sensor networks. *State Highway Administration*, University of Maryland, College Park. Project # SP009B4H.

Ahmed, M., Abdel-Aty, M. A., 2012. The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction. *IEEE Transactions on Intelligent Transportation Systems*. 13(2), 459–468.

Ahmed, M., Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. *Transportation Research Part C: Emerging Technologies*. 26, 203–213. http://doi.org/10.1016/j.trc.2012.09.002.

Ahmed,M.A., Abdel-Aty, M., Yu, R., 2012. Assessment of Interaction of Crash Occurrence, Mountainous Freeway Geometry, Real-time Weather, and Traffic Data. *Transportation Research Record: Transportation Research Board*. 2280, 51-59.

Al-Ghamdi, A.S., 2007. Experimental Evaluation of Fog Warning System. *Accident Analysis and Prevention*. 39, 1065-1072.

Blei, D., Ng, A., Jordan, M.J., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3, 993-1022.

Bugdol, M., Segiet, Z., Kręcichwost, M., Kasperek, P., 2014. Vehicle detection system using magnetic sensors. *Journal of Transport Problems*. 9(1), 49-60.

Cheng, H., 2011. Autonomous Intelligent Vehicles: Theory, Algorithms, and Implementation. Springer Publication.

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2011. Identifying crash type propensity using real-time traffic data on freeways. *Journal of Safety Research*. 42, 43-50.

Chu, W., Zhang, H., 2017. Real-time crash prediction estimation of freeway safety: A Review, CICTP, ASCE.

Coppola, R., Morisio, M., 2016. Connected Car: Technologies, Issues, Future Trends. *ACM Computing Surveys*. 49(3), Article 46, 1-36.

Das, S., Sun, X., Dutta, A., 2016. Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings. *Transportation Research Record: Journal of the Transportation Research Board*. 2552, 48-56.

Dias, C., Miska, M., Kuwahara, M., Waitra, H., 2009. Relationship between congestions and traffic accidents on expressways: an investigation with Bayesian Belief Networks. *In Proceeding of 40$^{th}$ Annual Meeting of Infrastructure Planning (JSCE)*, Japan.

Dimitriou, L., Stylianou, K., Abdel-Aty, M., 2018. Assessing Rear-End Crash Potential in Urban Locations Based on Vehicle-by-Vehicle Interactions, Geometric Characteristics and Operational Conditions. *Accident Analysis and Prevention*. In press, corrected proof, Available online 2 March 2018.

Fang, S., Xie, W., Wang, J., Ragland, D. R., 2016. Utilizing the Eigenvectors of Freeway Loop Data Spatiotemporal Schematic for Real-time Crash Prediction. *Accident Analysis and Prevention*. 94, 59–64. http://doi.org/10.1016/j.aap.2016.05.013.

Feinerer, I., Hornik, K., 2015. tm: Text Mining Package. R package version 0.6-2. https://CRAN.R-project.org/package=tm.

Fellows, I., 2014. wordcloud: Word Clouds. R package version 2.5. https://CRAN.R-project.org/package=wordcloud.

Gazis, D.C., 2002. Traffic Theory. Kluwer Academic Publishers, USA.

George, C.P., Doss, H., 2018. Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model. Journal of Machine Learning Research, 18, 1-38.

Golob, T.F., Recker, W.W., 2001. Relationships Among Urban Freeway Accidents, Traffic Flow, Weather and Lighting Conditions. California PATH Working Paper *UCB-ITS-PWP-2001-19. Institute of Transportation Studies*, University of California, Berkeley, 2001.

Golob, T.F., Recker, W.W., Alvarez, V. M., 2004. Tool to Evaluate Safety Effects of Changes in Freeway Traffic Flow. *Journal of Transportation Engineering (ASCE)*. 130(2), 222-230.

Grumert, E.F., Tapani, A., 2018. Traffic state estimation using connected vehicles and stationary detectors. *Journal of Advanced Transportation*. 2018, Article ID. 4106086.

Grun, B., Hornik, K., 2011. topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software.* 40(13), 1-30.

Hansen, K.D., Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., Sarkar, D., 2016. Rgraphviz: Provides plotting capabilities for R graph objects. R package version 2.18.0.

Harb, R., Yan, X., Radwan, Y., Su, X., 2009. Exploring Precrash Maneuvers using Classification Trees and Random Forests. *Accident Analysis and Prevention*. 41(1), 98-107.

Hassan, H. M., Abdel-Aty, M. A., 2013. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. *Journal of Safety Research*. 45, 29–36. http://doi.org/10.1016/j.jsr.2012.12.004.

Hellinga, B., Samimi, A., 2007. Safety Evaluations using a Real-Time Crash Potential Model : Sensitivity to Model Calibration. *Proceedings of the ITE Canadian District Annual Conference*, May 6-10, Toronto, Canada.

Hossain, M., Muromachi, Y., 2011. Understanding Crash Mechanisms and Selecting Interventions to Mitigate Real-Time Hazards on Urban Expressways. *Transportation Research Record: Journal of the Transportation Research Board*. 2213, 53–62. http://doi.org/10.3141/2213-08.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis and Prevention*. 45, 373–381. http://doi.org/10.1016/j.aap.2011.08.004.

Hossain, M., Muromachi, Y., 2013a. Understanding crash mechanism on urban expressways using high-resolution traffic data. *Accident Analysis and Prevention*. 57, 17–29. http://doi.org/10.1016/j.aap.2013.03.024.

Hossain, M., Muromachi, Y., 2013b. A real-time crash prediction model for the ramp vicinities of urban expressway. *IATSS Research*. 37(1), 68–79. http://doi.org/10.1016/j.iatssr.2013.05.001.

Hourdakis, J., Garg, V., Michalopulos, P., Davis, G. A., 2006. Real-time detection of crash prone conditions in freeway high crash locations. *Transportation Research Record: Journal of the Transportation Research Board*. 1968, 83-91. DOI: http://dx.doi.org/10.3141/1968-10.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis and Prevention*. 42, 213-224.

Katrakazas, C., Quddus, M., Chen, W.-H., Deka, L., 2015. Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C: Emerging Technologies*. 60, 416–442. http://doi.org/10.1016/j.trc.2015.09.011.

Katrakazas, C., Quddus, M.A., Chen, W.-H., 2016. Real-time classification of aggregated traffic conditions using relevance vector machines. *Presented at the Transportation Research Board 95th Annual Meeting*, January 10–14, 2016, Washington D.C, USA

Katrakazas, C., Quddus, M.A., Chen, W.-H., 2017. A simulation study of predicting conflict-prone traffic conditions in real-time. *Presented at the Transportation Research Board 96th Annual Meeting*, January 8–12, 2017, Washington D.C., USA.

Khan, S. M., Dey, K. C., Chowdhury, M., 2017. Real-Time Traffic State Estimation With Connected Vehicles. *IEEE Transactions on Intelligent Transportation System*. 18(7) 1–13. DOI: 10.1109/TITS.2017.2658664.

Kyrkou, C., Bouganis, C.S., Theocharides, T., Plycarpou, M.M., 2016. Embedded Hardware-Efficient Real-time Classification with Cascade Support Vector Machines. *IEEE Transaction on Neural Network and Learning System*. 27(1), 99-112.

Lee, C., Hellinga, B., Saccomanno, F., 2003b. Real-Time Crash Prediction Model for the Application to Crash Prevention in Freeway Traffic. *Transportation Research Record: Journal of the Transportation Research Board.* 1840*,* 67–77.

Lee, C., Hellinga, B., Saccomanno, F., 2003c. Proactive freeway crash prevention using real-time traffic control. *Canadian Journal of Civil Engineering*. 30(6), 1034-1041.

Lee, C., Hellinga, B., Saccomanno, F., 2004. Assessing Safety Benefits of Variable Speed Limits. *Transportation Research Record: Journal of the Transportation Research Board.* 1897, 183 – 190.

Lee, C., Saccomanno, F., Hellinga, B., 2003a. Analysis of Crash Precursors on Instrumented Freeways. *Transportation Research Record: Journal of the Transportation Research Board.* 1784*,*1–8.

Lee, C., Abdel-Aty, M., Hsia, L., 2006a. Potential Real-Time Indicators of Sideswipe Crashes on Freeways. *Transportation Research Record: Journal of the Transportation Research Board*. 1953, 41–49.

Lee, C., Hellinga, B., Ozbay, K., 2006b. Quantifying Effects of Ramp Metering on Freeway Safety. *Accident Analysis and Prevention*. 38(2), 279-288.

Lee, C., Lee, B. G., Kim, K., Lee, H. S., 2007. A VDS Based Traffic Accident Prediction Analysis and Future Application. *N.T. Nguyen et al. (Eds.): KES-AMSTA 2007*, LNAI 4496, 901–909.

Lee, C., Abdel-Aty, M., 2008. Testing Effects of Warning Messages and Variable Speed Limits on Driver Behavior using Driving Simulator. *Transportation Research Record: Journal of the Transportation Research Board*. 2069, 55-64.

Li, X., Liu, P., Wang, W., Xu, C., 2012. Using Support Vector Machine Models for Crash Injury Severity Analysis. *Accident Analysis and Prevention*. 45, 478-486.

Lin, L., Wang, Q., Sadek, A. W., 2015. A Novel Variable Selection Method Based on Frequent Pattern Tree for Real-time Traffic Accident Risk Prediction. *Transportation Research Part C: Emerging Technologies*. 55, 444–459. http://doi.org/10.1016/j.trc.2015.03.015.

Liu, M., Chen, Y., 2017. Predicting real-time crash risk for urban expressways in China. *Mathematical Problems in Engineering*. Article ID: 6263726.

Liu, J., Khattak, A.J., 2016. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. *Transportation Research Part C: Emerging Technologies.* 68, 83-100.

Luo, L., Garber, N. J., 2006. Freeway Crash Prediction Based on Real-Time Pattern Changes in Traffic Flow Characteristics. *A research project report for the Intelligent Transportation Systems Implementation Center,* UVA Center for Transportation Studies, Research Report No. UVACTS - 15-0-101. January, 2006.

Mimno, D., 2013. mallet: A wrapper around the Java machine learning tool MALLET. R packacge ver 1.0. http://CRAN.R-project.org/package=mallet

Nashat, S., Abdullah, A., Aramvith, S., Abdullah, M.Z., 2011. Support vector machine approach to real-time inspection of biscuits on moving conveyor belt. *Computer and Electronics in Agriculture*. 75(1), 147-158.

National Highway Traffic Safety Administration (NHSTA, 2013); https://www.transportation.gov/briefing-room/us-department-transportation-releases-policy-automated-vehicle-development. Accesed on 21st March, 2017.

Nilsson, P., Laine, L., Jacobson, B., 2017. A Simulator Study Comparing Characteristics of Manual and Automated Driving During Lane Changes of Long Combination Vehicles. *IEEE Transactions on Intelligent Transportation System*. 18(9), 1–11. DOI: 10.1109/TITS.2017.2664890.

Oh, C., Oh, J.S., Ritchie, S.G., 2000. Real-Time Estimation of Freeway Accident Likelihood. Institute of Transportation Studies. University of California, Irvine, CA. Working Paper ID: UCI-ITS-TS-WP-00-8.

Oh, J., Oh, C., Ritchie, S.G., Chang, M., 2005a. Real-Time Estimation of Accident Likelihood for Safety Enhancement. *Journal of Transportation Engineering*. 131(5), 358–363.

Oh, C., Oh, J., Ritchie, S.G., 2005b. Real-Time Hazardous Traffic Condition Warning System : Framework and Evaluation. *IEEE Transactions on Intelligent Transportation Systems*. *6*(3), 265–272.

Paikari, E., Moshirpour, M., Alhajj, R., Far, B.H., 2014. Data Integration and Clustering for Real Time Crash Prediction. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration*. August 13-15.

Pande, A., Abdel-Aty, M., 2005. A Freeway Safety Strategy for Advanced Proactive Traffic Management. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*. 9(3), 145-158. http://doi.org/10.1080/15472450500183789.

Pande, A., Abdel-Aty, M., 2006a. Comprehensive Analysis of the Relationship between Real-time Traffic Surveillance Data and Rear-end Crashes on Freeway*s. Transportation Research Record*: *Journal of the Transportation Research Board*. 1953, 31-40.

Pande, A., Abdel-Aty, M., 2006b. Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions. *Accident Analysis and Prevention*. 38(5), 936–948.

Pande, A., Abdel-aty, M., 2007. Multiple-Model Framework for Assessment of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board*. 2019, 99–107. http://doi.org/10.3141/2019-13.

Pande, A., Abdel-Aty, M., Hsia, L., 2005. Spatiotemporal Variation of Risk Preceding Crashes on Freeways. *Transportation Research Record: Journal of the Transportation Research Board.* 1908, 26–36.

Park, H., Haghani, A., 2015. Real-time Prediction of Secondary Incident Occurrences using Vehicle Probe Data. *Transportation Research Part C: Emerging Technologies*. 70, 69-85. http://doi.org/10.1016/j.trc.2015.03.018.

Park, H., Haghani, A., Samuel, A., Knodler, M.A., 2018. Real-time prediction and avoidance of secondary crashes under unexpected traffic congestion. *Accident Analysis and Prevention*. 112, 39-49.

Pereira, F.C., Rodrigues, F. Ben-Akiva, M., 2013. Text Analysis in Incident duration Prediction. *Transportation Research Part C: Emerging Technologies*, 37, 177-192.

Pham, M., Bhasker, A., Chung, E., Dumont, A., 2010. Methodology for Developing Real-time Motorway Traffic Risk Identification Models Using Individual Vehicle Data. *90 th Annual Meeting of the Transportation Research Board.*

Pirdavani, A., Pauw, E. De., Brijs, T., Daniels, S., Magis, M., Wets, G., 2015. Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data. *Traffic Injury Prevention*. 16:8, 786-791, http://doi.org/10.1080/15389588.2015.1017572.

Qu, X., Wang, W., Wang, W., Liu, P., 2012a. Real-time prediction of freeway rear-end crash potential by support vector machine. *Transportation Research Board 91st Annual Meeting,* Washington DC, USA, January.

Qu, X., Wang, W., Wang, W., Liu, P., 2012b. Real-time Freeway Sideswipe Crash Prediction by Support Vector Machine. *IET Intelligent Transport Systems*. 7(4), 445–453. http://doi.org/10.1049/iet-its.2011.0230.

Roess, R. P., Prassas, E. S., McShane, W. R., 2011. *Traffic Engineering*. Pearson Higher Education, Upper Saddle River, N.J: Prentice Hall.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of Real-time Traffic Characteristics on Freeway Crash Occurrence : Systematic Review and Meta-analysis. *Accident Analysis and Prevention*. 79, 198–211. http://doi.org/10.1016/j.aap.2015.03.013

Roy, A., Kobayshi, R., Hossain, M., Muromachi, Y., 2016. Real time crash prediction model for urban expressway using Dynamic Bayesian Network. *Journal of Japan Society of Civil Engineers, Ser. D3* (Infrastructure Planning and Management). 72(5), 1331-1338.

Roy, A., Muromachi, Y., 2016. The Development of Robust Real-time Crash Prediction Models with Bayesian Network. *Proceedings of Infrastructure Planning*. 53, CD, Japan Society of Civil Engineers.

Roy, A., Hossain, M., Muromachi, Y., 2018a. Enhancing the predcition performance of real-time crash predcition models: A Cell Transmission-Dynamic Bayesian Network approach. *Transportation Research Record: Journal of the Transportation Research Board*, https://doi.org/10.1177/0361198118797802.

Roy, A., Hossain, M., Muromachi, Y., 2018b. Development of Robust Real-time Crash Prediction Models using Bayesian Network. *Asian Transport Studies*. 5(2), 349-361.

Shew, C., Pande, A., Nuworsoo, C., 2013. Transferability and Robustness of Real-time Freeway Crash Risk Assessment. *Journal of Safety Research*. 46, 83–90. http://doi.org/10.1016/j.jsr.2013.04.005.

Shi, Q., Abdel-Aty, M., 2015. Big Data Applications in Real-time Traffic Operation and Safety Monitoring and Improvement on Urban Expressways. *Transportation Research Part C: Emerging Technologies*. 58, 380–394. http://doi.org/10.1016/j.trc.2015.02.022.

Shmueli, G., 2010. To Explain or to Predict. *Statistical Science*. 25(3), 289-310. DOI: 10.1214/10-STS330.

Son, H.D., Kweon, Y., Brian, B.B., 2011. Development of crash prediction models with individual vehicular data. *Transportation Research Part C: Emerging Technologies.* 19, 1353–1363. http://doi.org/10.1016/j.trc.2011.03.002.

Son, H.D., Kweon, Y.J., Park, B.B., 2008. Development of Crash Prediction Models Using Real Time Safety Surrogate Measures. A Research Project Report For the ITS Implementation Center, U. S. DOT Universality Transportation Center, Virginia, USA. Research Report No. UVACTS-15-0-70.

Sun, J., Sun, J., 2015. A Dynamic Bayesian Network Model for Real-time Crash Prediction using Traffic Speed Conditions Data. *Transportation Research Part C: Emerging Technologies.* 54, 176–186. http://doi.org/10.1016/j.trc.2015.03.006.

Sun, J., Sun, J., 2016. Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model. *IET Intelligent Transport Systems.* 10(5), 331-337. doi: 10.1049/iet-its.2014.0288.

Sun, L., and Yin, Y., 2017. Discovering themes and trends in transportation research using topic modeling.*Transportation Research Part C:Emerging Technologies*. 77, 49-66.

Uhlemann, E., 2015. Introducing Connected Vehicle. IEEE Vehicular Technology Magazine. 10(1), 23-31. DOI: 10.1109/MVT.2015.2390920.

Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015. Real-time Crash Prediction for Expressway Weaving Segments. *Transportation Research Part C: Emerging Technologies*. 61, 1–10. http://doi.org/10.1016/j.trc.2015.10.008.

Wang L., Abdel-Aty M., 2017. Implementation of Variable Speed Limits to Improve Safety of Congested Expressway Weaving Segments in Microsimulation. *Transportation Research Procedia*. 27, 577-584.

Wang, L., Abdel-Aty, M., Lee, J.Y., 2017a. Safety analytics for integrating crash frequency and real-time rick modeling for expressways. *Accident Analysis and Prevention*. 104, 58-64.

Wang, L., Abdel-Aty, M., Lee, J.Y., 2017b. Implementation of Active Traffic Management Strategies for Safety of a Congested Expressway Weaving Segment. *Transportation Research Record: Journal of the Transportation Research Board*. 2635, 28-35.

Wang, R., Li, Y., Work, D.B., 2017. Comparing traffic state estimators for mixed human and automated traffic flows. *Transportation Research Part C: Emerging Technologies*. 78, 95–110. http://doi.org/10.1016/j.trc.2017.02.011.

Wu, Y., Abdel-Aty, M., Lee, J.Y., 2017. Crash Risk Analysis during Fog Conditions using Real-Time Traffic Data. Special Issue RS5C. *Accident Analysis and Prevention*. 114, 4-11.

Wu, Y., Abdel-Aty, M., Cai, Q., Lee, J., Park, J., 2018. Developing an Algorithm to Assess the Rear-end Collision Risk under Fog Conditions using Real-time Data. *Transportation Research Part C: Emerging Technologies.,* 87, 11-25.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis and Prevention*. 47, 162– 171.

Xu, C., Liu, P., Wang, W., Li, Z., 2014b. Identification of Freeway Crash-prone Traffic Conditions for Traffic Flow at Different Levels of Service. *Transportation Research Part A: Policy and Practice*. 69, 58–70. http://doi.org/10.1016/j.tra.2014.08.011.

Xu, C., Liu, P., Wang, W., 2016b. Real-time Estimation of Secondary Crash likelihood on Freeways using High-Resolution Loop Detector Data. *Transportation Research Part C: Emerging Technologies*. 71, 406–418. http://doi.org/10.1016/j.trc.2016.08.015.

Xu, C., Liu, P., Yang, B., Wang, W., 2016a. Evaluation of the Predictability of Real-time Crash Risk Models. *Accident Analysis and Prevention*. 94, 207–215. http://doi.org/10.1016/j.aap.2016.06.004.

Xu, C., Tarko, A. P., Wang, W., Liu, P., 2013a. Predicting Crash Likelihood and Severity on Freeways with Real-time Loop Detector Data. *Accident Analysis and Prevention*. 57, 30–39. http://doi.org/10.1016/j.aap.2013.03.035

Xu, C., Wang, W., Liu, P., 2013b. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transanction on Intelligent Transportation Systems*. 14(2), 574–586.

Xu, C., Wang, W., Liu, P., 2013c. Identifying crash-prone traffic coditions under different weather on freeways. *Journal of Safety Research*. 46, 135-144.

Xu, C., Wang, W., Liu, P., Li, Z., 2015. Calibration of Crash Risk Models on Freeways with Limited Real-time Traffic Data using Bayesian Meta-analysis and Bayesian Inference approach. *Accident Analysis and Prevention*. 85, 207–218. http://doi.org/10.1016/j.aap.2015.09.016

Xu, C., Wang, W., Liu, P., Zhang, F., 2014a. Development of a Real-Time Crash Risk Prediction

Model Incorporating the Various Crash Mechanisms Across Different Traffic States. *Traffic Injury Prevention*. 16(1), 28 - 35. http://doi.org/10.1080/15389588.2014.909036

Xu, C., Wang, W., Liu, P., Guo, R., Li, Z., 2014c. Using the Bayesian Updating Approach to Improve the Spatial and Temporal Transferability of Real-time Crash Risk Prediction Models. *Transportation Research Part C: Emerging Technologies*. 38, 167–176. http://doi.org/10.1016/j.trc.2013.11.020

Yang, K., Yu, R., Wang, X., Quddus, M., Xue, L., 2018a. How to determine an optimal threshold to classify rea-time crash prone traffic conditions?. *Accident Analysis and Prevention*. 117, 250-261.

Yang, K., Wang, X., Quddus, M., Yu, R., 2018b. Deep learning for real-time crash predictin on urban expressways. 97th Annual Meeting of Transporation Research Board, Washington D.C.

Yasmin, S., Eluru, N., Wang, L., Abdel-Aty, M., 2018. A joint framework for static and real-time crash risk analysis. *Analytic Methods in Accident Research*. 18, 45-56.

Yeo, H., Jang, K., Skabardonis, A., Kang, S., 2013. Impact of traffic states on freeway crash involvement rates. *Accident Analysis and Prevention.* 50, 713-723.

You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. *Journal of Modern Transportation*. DOI: 10.1007/s40534-017-0129-7

Yu, J., Abdel-Aty, M., 2005. A Combined Approach to Determine the Location and Time of Freeway Crashes using Loop Detectors. *Infrastructure Planning and Management*, Japan Society of Civil Engineers. 786, 157-166.

Yu, R., Abdel-Aty, M., 2013a. Utilizing Support Vector Machine in Real-time Crash Risk Evaluation. *Accident Analysis and Prevention*, 51, 252–259. http://doi.org/10.1016/j.aap.2012.11.027.

Yu, R., Abdel-aty, M., 2013b. Multi-level Bayesian Analyses for Single and Multi-vehicle Freeway Crashes. *Accident Analysis and Prevention*. 58, 97–105. http://doi.org/10.1016/j.aap.2013.04.025.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian Random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis and Prevention*. 50, 371-376.

Yuan, J., Abdel-Aty, M., 2018. Approach-level real-time crash risk analysis for signalyzed intersections. *Accident Analysis and Prevention*. 119, 274-289.

Yuan, J.H., Abdel-Aty, M., Wang, L., Lee, J.Y., Wang, X.S., Yu, R.J., 2018. Real-Time Crash Risk Analysis of Urban Arterials Incorporating Bluetooth, Weather, and Adaptive Signal Control Data. Accepted for presentation at 97th Annual Meeting of the *Transportation Research Board*, TRB No. 18-00590, Washington DC, January 2018.

Zhang, K., Taylor, M.A.P., 2006. Towards universal freeway incident detection algorithms. *Transportation Research Part C: Emerging Technologies*. 14, 68-80. doi:10.1016/j.trc.2006.05.004.

Zheng, Z., Ahn, S., Monsere, C. M., 2010. Impact of Traffic Oscillations on Freeway Crash Occurrences. *Accident Analysis and Prevention*. 42, 626–636. http://doi.org/10.1016/j.aap.2009.10.009