Accident Analysis and Prevention 165 (2022) 106501

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A Review of Incident Prediction, Resource Allocation, and Dispatch Models for Emergency Management

Ayan Mukhopadhyay [a,*], Geoffrey Pettet [a], Sayyed Mohsen Vazirizade [a], Di Lu [b], Alejandro Jaimes [b], Said El Said [c], Hiba Baroud [d], Yevgeniy Vorobeychik [e], Mykel Kochenderfer [f], Abhishek Dubey [a]

[a] Electrical Engineering and Computer Science, Vanderbilt University, USA
[b] Dataminr, USA
[c] Tennessee Department of Transportation, USA
[d] Civil and Environmental Engineering, Vanderbilt University, USA
[e] Computer Science and Engineering, Washington University in St. Louis, USA
[f] Aeronautics and Astronautics, Stanford University, USA

ARTICLE INFO

Keywords:
Resource allocation for smart cities
Incident prediction
Computer aided dispatch
Decision making under uncertainty
Accident analysis
Emergency response

ABSTRACT

In the last fifty years, researchers have developed statistical, data-driven, analytical, and algorithmic approaches for designing and improving emergency response management (ERM) systems. The problem has been noted as inherently difficult and constitutes spatio-temporal decision making under uncertainty, which has been addressed in the literature with varying assumptions and approaches. This survey provides a detailed review of these approaches, focusing on the key challenges and issues regarding four sub-processes: (a) incident prediction, (b) incident detection, (c) resource allocation, and (c) computer-aided dispatch for emergency response. We highlight the strengths and weaknesses of prior work in this domain and explore the similarities and differences between different modeling paradigms. We conclude by illustrating open challenges and opportunities for future research in this complex domain.

1. Introduction

Emergency response management (ERM) is a challenge faced by communities across the globe. First responders need to respond to a variety of incidents such as fires, traffic accidents, and medical emergencies. They must respond quickly to incidents to minimize the risk to human life (Jaldell, 2017; Jaldell et al., 2014). Consequently, considerable attention in the last several decades has been devoted to studying emergency incidents and response. Data-driven models help reduce both human and financial loss as well as improve design codes, traffic regulations, and safety measures. Such models are increasingly being adopted by government agencies. Nevertheless, emergency incidents still cause thousands of deaths and injuries and result in losses worth billions of dollars directly or indirectly each year (Hattis, 2015). This is in part due to the fact that emergency incidents (like accidents, for example) are on the rise with rapid urbanization and increasing traffic volume.

ERM can be divided into five major components: (1) mitigation, (2) preparedness, (3) detection, (4) response, and (5) recovery. While most prior work has identified mitigation, preparedness, response, and recovery as the primary components of ERM systems (Mukhopadhyay, 2019; U. Department of Homeland Security, 2019), crowd-sourced information and additional sensors have motivated deployment of technology to provide early detection of incidents (before someone calls for help). Mitigation involves sustained and continuous efforts to ensure safety and reduce long-term risks to people and property. It also involves understanding *where* and *when* incidents occur and designing predictive models for both risk and incident occurrence. Preparedness involves creating infrastructure that enables emergency response management. This stage involves selecting stations for housing responders, ambulances, and police vehicles as well as designing plans for response. The third phase seeks to use automated techniques to detect incidents as they happen in order to expedite response. The fourth phase, arguably the most crucial, involves dispatching responders when incidents happen or are about to occur. Finally, the recovery phase ensures that impacted individuals and the broader community can cope with the effects of incidents. While most prior work in ERM has studied these problems

* Corresponding author.

https://doi.org/10.1016/j.aap.2021.106501
Received 23 June 2021; Received in revised form 14 November 2021; Accepted 15 November 2021
Available online 18 December 2021

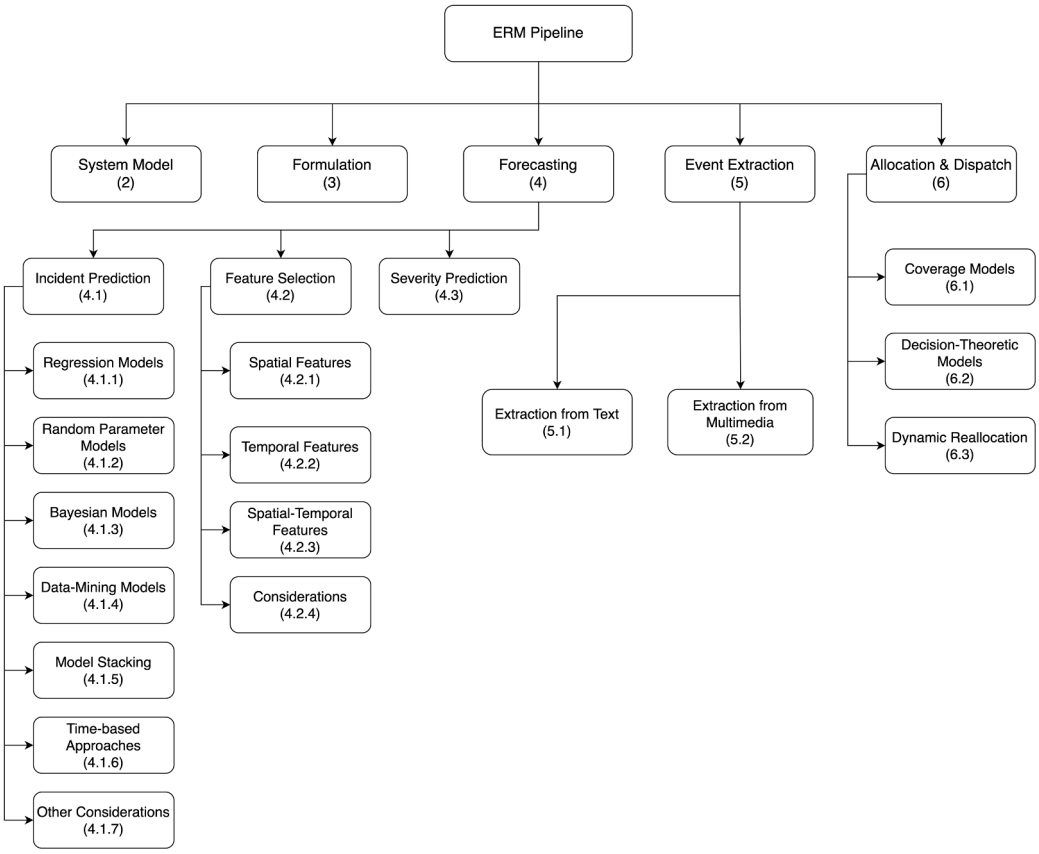

Fig. 1. Outline for the survey: Numbers in each box link the topic to a specific section in this paper.

independently, these stages are actually inter-linked, with the output of one stage serving as input for another. For example, predictive models learned in the *preparedness* stage are used in planning *response* strategies. Therefore, it is crucial that ERM pipelines are designed in a manner that considers such intricate inter-dependencies.

In this survey, we cover prior work on some of the most widely explored approaches that fall into the categories of mitigation, preparedness, detection, and response, and explain how the overall ERM pipeline functions. First, we address the scope of the problem and precisely define the incidents that we consider in the paper. One way to categorize incidents is by the rate at which they occur and how they affect first responders. For example, some incidents happen often, and addressing them is part of the day-to-day operations of first-responders. Examples of such incidents include crimes, accidents, calls for medical services, and urban fires. A second category consists of comparatively less frequent incidents, which include natural calamities like earthquakes, floods, and cyclones. To scope our research we primarily focus on principled approaches to address frequent urban incidents like road accidents. Having a narrow focus regarding the type of the incident enables us to explore the large spectrum work dedicated to designing principled approaches to ERM. Our primary reason to focus on urban emergency incidents is simply the alarming extent of the damage such incidents cause and the sheer frequency of their occurrence. Globally, about 3,200 people die every day from road accidents alone, leading to a total of 1.25 million deaths annually (Control et al., 2019). In fact, it is noted that without appropriate measures, road accidents are set to be the fifth largest cause of death worldwide by 2030 (Association for Safe International Road Travel, 2019). Calls for emergency medical services (EMS) are also a major engagement for first responders, and there are more than 240 million EMS calls made annually in the United States alone (National Emergency Number Association, 2019). Therefore, it is imperative that we design principled approaches to understand the spatial and temporal characteristics of such incidents and investigate algorithmic methods that can mitigate their effects.

In this survey, we explore models and approaches to design principled approaches to ERM from various fields like operations research, transportation engineering, statistics, and machine learning in order to understand commonalities and differences among them. We seek to provide a unified perspective on such systems. There are comprehensive reviews on crash prediction models (Nambuusi et al., 2008; Yannis et al., 2017; Kiattikomol, 2005), emergency facility location approaches (Li et al., 2011), and dispatch strategies (Bohm and Kurland, 2018). In particular, the work by Lord and Mannering (2010) provides a particularly insightful summary of crash prediction models. However, to the best of our knowledge, there is no comprehensive study that links prediction models from different persepectives and investigates covariates of relevance, modeling paradigms, and planning approaches comprehensively. We treat the ERM system in its entirety and provide a comprehensive survey of prior work done in the fields of predictive modeling, event extraction, and planning approaches to aid emergency response. This survey provides a framework for future research on integrated emergency response management pipelines for smart and connected communities.

We show a brief outline of the survey in Fig. 1. We begin by providing an overview of the system model for ERM pipelines in Section 2 and mathematical formulations for each of the sub-problems in Section 3. Then, we discuss approaches to incident prediction (Section 4), event extraction (Section 5), and allocation and dispatch (Section 6). We highlight key takeaways for practitioners at the end of each section. Finally, we identify key challenges, knowledge gaps, and opportunities for future work in Section 7.

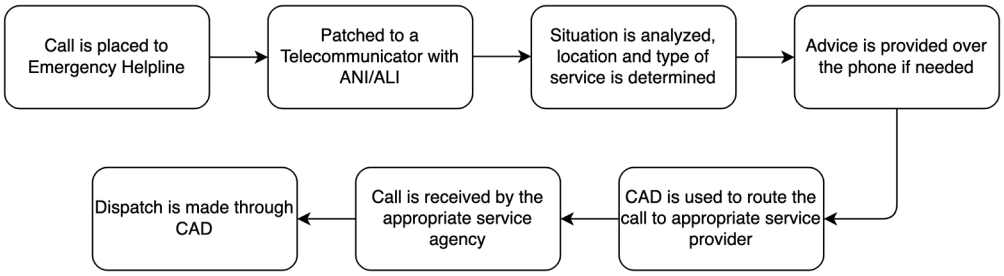

Fig. 2. Typical Emergency Dispatch Helpline Model: first responders analyze the type of call, offer immediate help and use computer-aided dispatch to allocate specific resources to incidents.

2. Understanding emergency response systems

We study the problem of optimally responding to emergency incidents in urban areas. Incidents are reported to central emergency response agencies, which have streamlined mechanisms for processing the request. For example, in the United States, emergency helpline calls are placed by dialing 911. Calls can be made by many *agents*. People in need of assistance or other people who might have observed an incident can report it to the concerned authorities. Such a report is typically referred to as a "call" for emergency services. It is also possible for first responders to automatically extract reports about incidents through social media or video feeds. We show the steps that follow a call for assistance in Fig. 2 (Ciy of Rochester, 2019). The call is appended with automatic name and location information (ANI/ALI) and patched to a trained telecommunicator. The telecommunicator analyzes the situation and the type of response needed (EMS or fire, for example). In some cases, such as those requiring cardiopulmonary resuscitation (CPR), guidance might be provided through the phone before first-responders reach the scene. The call is then transferred to the concerned agency (such as the fire department) by a computerized mechanism. The agency then uses its computer-aided dispatch (CAD) system to dispatch a responder to the scene. This set of events defines an ERM system, and it governs the pipeline of incident response, including detecting and reporting incidents, monitoring and controlling a fleet of response vehicles, and finally dispatching responders when incidents occur. In many cases there are multiple organizations governing this pipeline for an urban area; for example, ambulances and police cars might be dispatched from different departments.

Agents[1] which respond to incidents like accidents and fires include ambulances, police vehicles, and fire trucks (among others) and are referred to as *responders*. Responders are typically equipped with devices that facilitate communication to and from central control stations. In many cases, especially in the United States, responders like ambulances are equipped with computational devices like laptops as well. Once an incident is reported, responders are dispatched by a human agent to the scene of the incident (guided by some algorithmic approach like a CAD system). This process typically takes a few seconds,[2] but can be longer if dispatchers are busy. If no responder is available, the incident typically enters a waiting queue and is attended once a responder becomes free. Each responder is located in a specific *depot* (fire-station, for example), which are situated at various points in the spatial area under consideration. Once a responder has finished servicing an incident, it is directed back to the depot and becomes available to be re-dispatched by the dispatcher while en-route. An aspect that plays a key role in dispatch algorithms is that if there are any free responders available when an incident is reported, one must be dispatched to attend to the incident. This constraint is a direct consequence of the bounds within which emergency responders operate, as well as the critical nature of the incidents.

The components of ERM that we focus on are shown in Fig. 3. ERM pipelines typically use data from historical incidents and the environment, including weather, road geometry, traffic patterns, and socio-economic data. It is also possible to use textual and video data to extract information about the occurrence of incidents. We divide an ERM system into five major components: (1) predictive models for incident occurrence, (2) event extraction models to detect incidents, (3) models for environmental features like traffic and weather, (4) allocation models to optimize the spatial locations of responders and depots, and (5) dispatch models to create algorithmic approaches to respond to incidents when they occur. These components are intricately linked, and the performance of each plays a crucial role in the overall performance of the ERM pipeline.

Incident prediction models form the basis of an ERM system. In order to mitigate the effects of incidents, it is important to understand *where* and *when* such incidents occur. Incident models are typically designed using historical incident data, but such models often use historical environmental data as well; for example, it is common for accident prediction models to use historical traffic data. Allocation models are then used to allocate responders in time and space in anticipation of future incidents. Finally, allocation and prediction models are used to create dispatch models, which can be thought of as a policy that guides real-time response. A rather recent trend in designing ERM systems is to include event extraction models in the pipeline, which can automate the discovery and reporting of incidents.

Significant prior work has focused on understanding and designing algorithmic approaches for each of the modular components. This article studies models for incident prediction, allocation, extraction, and dispatch. While we do not discuss models of relevant environmental factors, they are important to the development of the overall pipeline.

We focus this survey primarily on roadway accidents. The reason for this choice is twofold. First, most prior work in incident analysis has focused on accidents and crashes, and this presents a rich body of work to survey and draw inferences about. Secondly, prediction, allocation and response to accidents and EMS calls involve similar characteristics and constraints from an algorithmic perspective. Both require dispatching of ambulances as quickly as possible, and from the scene of the incident, injured victims need to be transported to nearby hospitals. Therefore, most of our discussion on accidents can be broadened to EMS calls in general, but focusing on one particular type allows us to discuss various technical approaches in greater detail.

3. General mathematical formulation of the decision problem

To help provide common context, we start by defining a broad mathematical formulation for incident prediction, extraction, and planning problems that we use throughout this survey. Given a spatial area of interest S, the decision-maker observes a set of samples (possibly noisy) drawn from an incident arrival distribution. These samples are

[1] We use the term "agents" as is common in multi-agent systems community.

[2] This is based on our communication with fire departments in the United States (Private Communication, 2018); time taken to dispatch responders presumably varies across the globe.

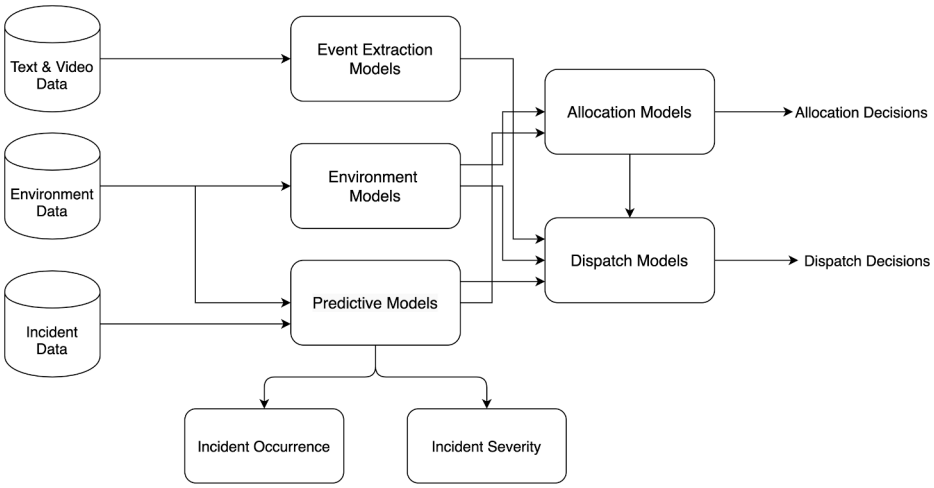

Fig. 3. ERM System Pipeline: historical data from different sources are used to design predictive models for incidents and the environment, which in turn are used to create allocation and response models. Events can be extracted using text and video data to expedite reporting and aid response.

denoted by $\{(s_1,t_1,k_1,w_1),(s_2,t_2,k_2,w_2),\ldots,(s_n,t_n,k_n,w_n)\}$, where s_i,t_i, and k_i denote the location, time of occurrence, and reported severity of the ith incident, respectively, and $w_i \in \mathbb{R}^m$ represents a vector of features associated with the incident. We refer to this tuple of vectors as D, which denotes the input data that the decision-maker has access to. The vector w can contain spatial, temporal, or spatio-temporal features and it captures covariates that potentially affect incident occurrence. For example, w typically includes features such as weather, traffic volume, and time of day. The most general form of incident prediction can then be stated as learning the parameters θ of a function over a random variable X conditioned on w. We denote this function by $f(X|w,\theta)$. The random variable X represents a measure of incident occurrence such as a *count* of incidents (the number of incidents in S during a specific time period) or *time* between successive incidents. The decision-maker seeks to find the *optimal* parameters θ^* that best describe D. This can be formulated as a maximum likelihood estimation (MLE) problem or an equivalent empirical risk minimization (ERM) problem.

We review prior work focused on modeling the function $f(X|w,\theta)$. There have been many different approaches for modeling f. It can be modeled as an explicit probability density or mass (e.g., Poisson distribution), or a function that does not strictly conform to such definitions (e.g., a linear regression approach with X being the dependent variable). Nonetheless, such functions typically have probabilistic interpretations, and we present different approaches of modeling f in Section 4. We first highlight different modeling choices for understanding the spatial–temporal nature of accidents. Then, we focus on the vector w. Arguably, the most crucial part in learning a model over incident occurrence involves choosing w, and we review various covariates in Section 4.2.

The next step in an emergency response pipeline is planning in anticipation of incidents. This involves stationing responders strategically and dispatching them as incidents occur. This process can be broadly represented by the optimization problem $\max_y G(y|f)$, where y represents the decision variable (which typically denotes the location of emergency responders in space), G is a reward function chosen by the decision-maker, and f is the model of incident occurrence. For example, G might measure the total coverage (spatial spread) of the responders, or the expected response time to incidents. Therefore, given f, the decision-maker seeks to maximize the function G.

There are two major approaches for modeling the response problem. First, the planning problem can be represented as a stochastic control process. For example, it can be formulated as a Markov decision process (MDP) (Kochenderfer, 2015). This formulation is particularly relevant for problems seeking to find policies for dispatch. The aim is to find an optimal policy (i.e. control choices for every possible state of the system) that maximizes the expected sum of rewards. The second approach is to directly model the planning problem as an optimization problem according to a specific measure of interest. As an example, a lot of prior work has focused on maximizing the coverage of emergency responders (Toregas et al., 1971; Church and ReVelle, 1974; Gendreau et al., 1997).

It is important to note that emergency incidents can only be responded to after the concerned authorities are notified. However, often times in practice, there is a gap in time between the occurrence of an incident and the time at which a call is made for response. Event extraction models seek to bridge this gap by identifying incidents before they are reported to the responders. Mathematically, this process involves learning a function E which takes relevant data u as input (for example, text data from social media), and outputs information about an event (for example, the location of an accident). This can be represented as $\bar{z} = E(u)$, where $\bar{z}$ denotes an estimate of variable z. In event extraction, z can denote any variable related to the incident such as the location of the incident, the extent of the damage, the number of people involved, and so forth.

4. Incident forecasting

Incident forecasting is necessary to understand the likely demand of the emergency resources in a given region, and forms the basis for approaches to stationing and dispatch. We divide the discussion on incident forecasting into three major parts: (1) approaches to incident prediction, (2) features used in incident prediction and feature selection, and (3) predicting incident severity.

4.1. Approaches to incident prediction

Prior work has involved learning spatial–temporal models of incident occurrence. From our definition of incident prediction models described in Section 3, forecasting models correspond to determining the function f. An important method in incident prediction is known as 'crash frequency analysis', which uses the frequency of incidents in a specific discretized spatial area as a measure of the inherent risk the area possesses (Deacon et al., 1974). Deacon et al. (1974) identified key questions that practitioners should answer while designing predictive models for incident occurrence, and their work is still relevant to decision-makers and policy designers. This approach also forms the basis of *hotspot* analysis (Cheng and Washington, 2005; Eck et al., 2005), which is widely used today as a relatively simple and fast method to visualize incident data. A shortcoming of frequency analysis is that it

neglects fluctuations in incident occurrence and requires a large volume of incident data to infer accurate characteristics of occurrence (Yu et al., 2014; Ryder et al., 2019). The core idea behind frequency analysis continues to be in use today; although it is common to use it in conjunction with other covariates of relevance and frame the overall problem as a regression model.

4.1.1. Regression models

One of the earliest regression models used to model incident occurrence involved multiple linear regression models with Gaussian errors (Frantzeskakis et al., 1994; Jovanis and Chang, 1986). However, modeling accident count by linear regression can be inaccurate, as the response variable is discrete and strictly positive. In addition, it has also been shown that linear regression models fail to model the sporadic nature of emergency incidents (Miaou and Lum, 1993; Joshua and Garber, 1990). Linear regression models with multiplicative effects have also been investigated but have shown to be inaccurate compared to other models (Miaou and Lum, 1993). The inaccuracies of linear regression methods in the context of accident prediction is investigated and summarized by Miaou and Lum (1993). Rakha et al. (2010) revisited this problem recently, and used data aggregation techniques to satisfy assumptions made by linear regression. While such an approach has shown performance on par with other regression models (Poisson regression, for example), it needs further validation before it is widely adopted.

The inaccuracies of linear regression and the suitability of Poisson models for count data led to the widespread use of Poisson regression for modeling incident occurrence (Jovanis and Chang, 1986). Each incident is considered a result of an independent Bernoulli trial. Given that all the trials are generated by the same stochastic process, the series of trials can be modeled by a binomial distribution. As the number of trials becomes large and the probability of success is very small, the probability distribution over the count of incidents takes the form of a Poisson distribution (Lord et al., 2005). To accommodate the feature vector w, Poisson regression assumes that the logarithm of the expected value of the distribution is a linear combination of w. This methodology has been used extensively for emergency incident analysis (Bonneson and McCoy, 1993; Maher and Summersgill, 1996; Sayed and Rodriguez, 1999; Joshua and Garber, 1990; Miaou and Lum, 1993).

An issue with using Poisson regression is that the expected value of the response variable (count of incidents) equals its variance. This is typically not the case with crash data, which is over-dispersed, meaning that the variance of the data is greater than its mean (Lord et al., 2005). There are examples of incident data being under-dispersed as well (Ye et al., 2018). Therefore, the broader argument against the use of Poisson regression is that it might not be able to model real-world crash data, which can be under-dispersed or over-dispersed. An approach to accommodate over-dispersion is to use Poisson–hierarchical models (Deublein et al., 2013). Poisson–hierarchical models (as well as Poisson models) fall under the broader category of generalized-linear models (GLM), which is a family of distributions used widely in statistics and machine learning. From this family, the Poisson-gamma (also called negative binomial) and Poisson-lognormal models are particularly relevant.

The Poisson-gamma model is a Poisson distribution whose mean parameter follows a gamma distribution. It has been shown that the Poisson-gamma model fits crash data better than Poisson models, and it has been extensively used for crash prediction (Quddus, 2008; Akin, 2011; Ladron de Guevara, 2004; Caliendo et al., 2007; Ackaah and Salifu, 2011; Dissanayake and Ratnayake, 2006). While the Poisson-gamma model solves the problem of over-dispersion, it performs poorly on under-dispersed data and is particularly problematic to use with small sample sizes and with data with low sample mean (Lord et al., 2008; Aguero-Valverde and Jovanis, 2008). The Poisson-lognormal model is conceptually the same as Poisson-gamma model, but it uses the lognormal distribution for the mean parameter rather than the gamma distribution (Aguero-Valverde, 2013; Ma et al., 2008; Park, 2019; Shirazi, 2019). The lognormal distribution is a heavy tail distribution and provides more flexibility for over-dispersion. Recently, the Poisson-inverse-gamma model has been used in crash modeling (Khazraee et al., 2018). However, such models do not have closed-form MLE solutions unlike the Poisson-gamma models (Lord and Mannering, 2010).

Despite the success of Poisson and Poisson–hierarchical models, a common shortcoming is that both models fail to adequately handle the prevalence of zero counts in crash data (Lord et al., 2005). A remedy to this problem is to use zero-inflated models, and both zero-inflated Poisson and zero-inflated Poisson-gamma models have been used to model accident data (Qin et al., 2004; Lord et al., 2007; Huang and Chin, 2010). Zero-inflated models can be described as having dual states, one of which is the *normal* state, and the other the *zero* state. The excess zeros that cannot be explained by standard count-based models can then be considered to have arisen due to the presence of a separate state. Zero-inflated models result in improved statistical fit to accident data. However, Lord et al. (2005) note that most prior works justify the use of zero-inflated models by improved likelihood, and therefore automatically assume that crash data is generated by a dual-state process (except work by Miaou and Lum (1993), which uses a zero-inflated model to justify misreporting of incidents). Through empirical data and simulations, they show that excess zeros could arise due to various other factors like low traffic exposure and the choice of spatial and temporal scales by the model designer. As a result, it is not clear if the statistical backing to using dual-state models is accurate or not. In our opinion, the work by Lord et al. (2005) is particularly profound, and the argument that statistical fit should not be the only consideration for fitting models to crash data (and other data in general) is extremely cogent.

4.1.2. Random-parameter models

Accounting for unobserved heterogeneity (i.e., factors affecting incident frequency but not captured in the data) has dominated recent statistical modeling development, with random-parameter (RP) models being among the most widely used approaches (Mannering et al., 2016). Unobserved heterogeneity introduces a variation in the effect of observed variables on the outcome. The outcome is typically the likelihood and severity of a crash. For example, a highway's design speed limit is a commonly used variable in the prediction of the likelihood of crashes. However, this may introduce unobserved heterogeneity if the vehicle's actual speed is not considered which may be different than the design speed limit across different drivers. Environmental conditions are also commonly used to explain crash occurrence and severity such as time of the day and weather variables. However, the same amount of precipitation may lead to different outcomes in the likelihood and severity of accidents depending on the geographic area and the different ways drivers respond to adverse conditions. Additionally, unobserved heterogeneity can result from the spatial or temporal aggregation of accidents. Since these events are rare, they are often aggregated over time (e.g., number of accidents per 4 h) or space (e.g., number of accidents per road segment) before they are modeled. The lack of consideration for unobserved heterogeneity will lead to biased estimates because the effect of an observed variable will be the same across all observations for a particular instance (Mannering et al., 2016). RP models address heterogeneity by allowing the estimated parameters to vary across observation according to a continuous distribution. A significant portion of RP models in the literature are based on the assumption that random parameters follow a distribution with a common mean and no mutual dependence (El-Basyouny and Sayed, 2009; Milton et al., 2008). However, lack of consideration of cross-correlation and mutual dependence can lead to biases in the estimation of parameter variances (Conway and Kniesner, 1991).

A few recent studies have considered cross-correlated RP models and compared their performance to fixed-parameters and uncorrelated RP models. The correlated RP negative binomial model resulted in an

improved log-likelihood compared to the fixed-parameters model (Venkataraman et al., 2011) and better statistical performance and predictive power compared to the uncorrelated model (Coruh et al., 2015). In another study, correlated RP Tobit model was shown to outperform both fixed-parameters and uncorrelated RP Tobit models (Yu et al., 2015). However, these results are still not conclusive as other studies have found the relative statistical performance between uncorrelated and correlated RP count models to be comparable (Saeed et al., 2019). Therefore, additional research is needed to determine the advantages of correlated RP models. In addition to cross-correlations and improved statistical performance, another advantage of using correlated RP models is the ability to account for the heterogeneous effects of covariates across roadway segments as they apply to crash frequency analysis on multilane highways (Saeed et al., 2019). While the focus of this section is on RP models as they are the most adopted methods, it is worth noting that other approaches have been developed to address unobserved heterogeneity (see the work by Mannering et al. (2016) for an extensive review). For instance, latent-class (finite mixture) models seek to identify groups of observations having homogeneous variable effects (Cerwick et al., 2014). These models do not require a parametric assumption for the distribution of estimated parameters like RP models; however, they still impose a parametric model structure and can be computationally intensive. To account for the variation at both the group and individual observation levels, RP models within each class have been used with mixture models (Xiong and Mannering, 2013). Other approaches address specific heterogeneity issues such as Markov-switching models which have been used for time-dependent unobserved heterogeneity (Xiong et al., 2014). Such a form of heterogeneity can be caused by time-varying factors such as traffic and weather conditions or when the accidents are aggregated over a certain time period.

4.1.3. Bayesian approaches

Bayesian methods (Gilks et al., 1996; Goldstein, 1995) are often used for parameter estimation. Such models result in a distribution over parameters rather than point estimates, which can result in greater robustness to outliers and small sample sizes (Miaou and Lord, 2003). The empirical Bayes method (also known as maximum marginal likelihood) has been used in traffic engineering (Hauer and Persaud, 1983; Heydecker, 2001; Hauer, 1986; Hauer, 1992) (the method as applied to crash prediction is explained particularly well by Hauer et al. (2002)). Bayesian modelling techniques have also been used to assess potential risk factors of spatial regions (MacNab, 2004; Pettet et al., 2017) and to estimate expected crash frequencies (Aguero-Valverde and Jovanis, 2009).

Hierarchical Bayesian estimation of safety performance models have also been explored over the last two decades (Davis, 2001; Lord and Miranda-Moreno, 2008; Ma et al., 2008; Miaou and Song, 2005; Park, 2019; Schlüter et al., 1997). Recently, the Poisson-gamma and Poisson-lognormal models have also been estimated using Bayesian methods (Ackaah and Salifu, 2011; Akin, 2011; Basu and Saha, 2017; Caliendo et al., 2007; Dissanayake and Ratnayake, 2006; Aguero-Valverde, 2013; Khazraee et al., 2018; Ladron de Guevara et al., 2004; Quddus, 2008; Shirazi, 2019). A caveat regarding Bayesian models is that the crucial choice of priors in the predictive models. The underlying information for designing priors might be available from previous models, engineering judgement, etc., and prior distributions can also be chosen to be non-informative or weakly informative. An important investigation in this context, specifically regarding crash prediction, has been done by Song et al. (2006), who study the performance of various Bayesian multivariate spatial models with different prior distributions. It has also been shown that using non-informative priors may result in a high bias for the dispersion parameter in models, especially with small sample sizes (Park et al., 2010).

4.1.4. Data mining approaches

With improved sensor technology and easier storage, data-mining methods have successfully been used for crash prediction. Random forests (Yu, 2014; Abdel-Aty et al., 2008), support vector machines (Li et al., 2008; Yu, 2013; Zhang and Xie, 2007), and neural networks (Chang, 2005; Pande and Abdel-Aty, 2006; Riviere et al., 2006; Abdel-wahab and Abdel-Aty, 2002) have recently been used to model crashes. Bayesian neural networks have also been explored, which address overfitting of neural-networks in crash modeling (Xie et al., 2007). Deep learning techniques have also been used in various studies (Zhu et al., 2018; Bao et al., 2019). One model that may be of interest to practitioners was developed by Basak et al. (2019), who used a spatio-temporal convolution long short-term memory network (LSTM) to predict short-term crash risks, including propagation of traffic congestion. While the network structure was a combination of various complex networks, the accuracy of hourly predictions was limited, which highlights the inherent difficulty of predicting crash frequency at low temporal and spatial resolutions. It also makes a case against the use of complex models in this domain because they are harder to generalize.

4.1.5. Model stacking

Ensemble methods use multiple trained models to improve prediction compared to what can be obtained from individual models. The most straightforward approach is averaging the prediction of two or more models. However, a better approach is to use a meta-learning algorithm to learn the best combination of the predictions from multiple models, which is known as stacking or stacked generalization (Witten, 2016). Big data and the surge in availability of computational resources have paved for more sophisticated approaches such as model stacking in incident prediction. Various stacking models with different numbers of layers and assorted types of models (Iqbal et al., 2021; Tang et al., 2019; Xiao, 2019; Ma et al., 2021; Behura and Behura, 2020; Singh and Mohan, 2018; Chen et al., 2018) have been used during recent years to predict and detect incidents. The main caveats of using ensemble models are overfitting and the data size required for testing and training.

4.1.6. Modeling time to incidents

A somewhat different approach for predicting emergency incidents is to directly model inter-incident time as a function of relevant covariates. In this case, the variable X corresponds to the time between consecutive incidents. Mukhopadhyay et al. (2017) describe an example of such models by using uncensored (parametric) survival models to estimate time between accidents. It has been since used to model different incident types (Pettet et al., 2017; Mukhopadhyay et al., 2019; Mukhopadhyay et al., 2018). A key advantage of such methods is that planning problems are often modeled as continuous-time processes, and as a result, the incident prediction models can be easily used by planning models. While time-based models are not the most commonly used approaches to model the occurrence of crashes, continuous-time models are often used for other purposes in ERM pipelines. For example, such models are widely used for predicting the duration of crashes and the delay that crashes cause in traffic and congestion (Jiang et al., 2014; Tajtehranifard et al., 2016; Li, 2014; Chung and Recker, 2012; Basak et al., 2019; Zhan et al., 2011). Hazard-based approaches have also been used to evaluate the time it takes to report, respond to, and clear incidents (Nam and Mannering, 2000). While such algorithmic directions of work are crucial to the overall ERM pipeline, they lie outside the scope of this paper.

Another way to directly model time between incidents is to use time-series based forecasting. While approaches like survival models assume that inter-dependencies between successive incidents (or related in the feature space) can be modeled by designing appropriate features, time-series based approaches explicitly consider that consecutive observations are statistically *dependent*. Typically, algorithmic approaches applicable to stationary time-series (defined as a series whose mean, variance, and auto-correlation are constant over time) have been used for forecasting roadway accidents (Al-Ghamdi, 1995; Khasnabis and Lyoo, 1989; Al-Hasani et al., 2019). While non-stationary time-series

data is more common in practice, such data can typically be converted to a stationary series by differencing (Al-Ghamdi, 1995). The combination of time-series and data-mining based approaches have also been explored for forecasting traffic accidents (Shao et al., 2019).

4.1.7. Other considerations

Most approaches to incident prediction assume that estimated model parameters do not change over time – i.e. the parameters are temporally stable. However, several studies have found temporal instability in incident and injury-severity models' parameter estimates (Behnood and Mannering, 2015; Venkataraman et al., 2016; Marcoux et al., 2018; Alnawmasi and Mannering, 2019; Al-Bdairi et al., 2020; Islam and Mannering, 2020). There are many reasons to expect model paramters to shift over time. Driver behavior has been shown to be influenced by factors such as macroeconomic conditions, cognitive biases that affect risk perception, and drivers' attitudes toward safety (Mannering, 2018). All of these factors change over time, suggesting that driver behavior is also temporally unstable. Additionally, the dynamics of urban environments evolve due to factors such as population shifts and roadway construction. It is important that such changes are taken into account by forecasting methodologies. Recently, the development of online models for predicting accidents has been explored that update learned models continuously using incoming streams of data (Mukhopadhyay et al., 2019). See Mannering (2018)'s work for a detailed discussion of the potential causes and implications of temporal instability in accident data (Mannering, 2018).

An important consideration when evaluating various modeling approaches is the ability of each to reveal underlying causal relationships between features and the risk of incident occurrence. Often, there is a tradeoff between a model's causal inference ability, scalability to large datasets, and predictive capability (Mannering et al., 2020). To properly understand causality, statistical models must consider factors such as potential endogeneity in the data (discussed in Section 4.2.4) and unobserved heterogeneity with techniques such as random parameter models (Section 4.1.2). Unfortunately, it is challenging to apply these methods to large datasets due to the complexity of estimating their parameters. Data-mining methods, on the other hand, scale very well to big-data applications, and have shown excellent predictive performance. However, this comes at the cost of causal inference, and their black-box nature makes it difficult to separate correlation from causation. Practitioners should be aware of these tradeoffs, and chose a modeling approach with strengths that align with the goals of their analysis.

As incidents like accidents evolve in space and time, it is particularly important to identify the spatial and temporal resolutions that predictive models can accommodate. Naturally, changes in the degree of discretization affect the distributions of the dependent and the independent variables. On one hand, since forecasting the exact time and location of incidents like crashes is virtually impossible, high-resolution models are very difficult to construct. On the other hand, reducing the resolution may result in aggregation bias and unobserved heterogeneity (Washington et al., 2020). A specific problem with fine-grained spatial and temporal discretization is the prevalence of zero counts in the resulting data, which might pose problems with convergence while statistical learning (Lord et al., 2005). Similarly, it has also been shown that increasing the resolution of spatial and temporal discretization can lower the accuracy of various methods, including deep learning, tree-based, and econometric models (Bao et al., 2019). As a result, it is crucial that designers balance the spatial and temporal discretization of their models according to the specific needs of an area and explore the sensitivity of the model with changes in resolution.

4.2. Feature selection

An important part of developing predictive models is data collection and feature engineering. Since various factors are involved in causing an accident, an important step in accident prediction pipeline is to collect as much as data as possible about relevant determinants. Collecting and cleaning data might be very challenging due to the size and incompatibility of datasets from different resources (Vazirizade et al., 2021). Furthermore, some micro-level features such as age of the drive or type of the car might not be available due to privacy and legal reasons. In general, the performance of models strongly depends on the selected features. Consequently, they should be chosen strategically. Missing relevant explanatory features may result in an inaccurate model. On the other hand, including too many features requires more computational resources and may cause overfitting and erroneous prediction.

Different approaches have been used to select a subset of all available features. Filter-based methods (Durduran, 2010), wrapper methods (Lee and Wei, 2010; Tambouratzis et al., 2014; Ke et al., 2017; Krishnaveni and Hemalatha, 2011), embedded methods (Chen et al., 2019; Fu et al., 2019), and combination of multiple methods (Haruna et al., 2019; Ramani and Selvaraj, 2016; Shanthi, 2012; Wang et al., 2020) are the most common approaches for feature selection. SHAP (SHapley Additive exPlanations), a framework originally used for model interpretability, has also be used for evaluating the importance of the features in the model (Parsa et al., 2020; Wen et al., 2021).

In general, features for accident prediction can be categorized into temporal, spatial, or a combination of both. For example, one can choose to use time of day as a feature in order to understand how it affects accident rates. This is an example of a temporal feature. The geometry of a specific road segment, on the other hand, is a spatial feature, as it is a characteristic property of a particular spatial unit. Spatio-temporal features measure spatial properties that change with time. For example, traffic congestion in a specific part of the city falls under this category since it is characterized by both space and time. Unfortunately, not all of the underlying factors involved in an accident are measurable. The features available for crash analysis are usually restricted to the information on the crash report, weather and environmental conditions, roadway geometry, and traffic information. It is also possible to categorize features into static or dynamic (Qi et al., 2007), but we choose to follow the categorization with respect to spatio-temporal characteristics of the features.

4.2.1. Temporal features

Weather (Songchitruksa and Balke, 1959; Mukhopadhyay et al., 2017; Qi et al., 2007) and visibility range (Abdel-Aty et al., 2012) have been proven to be useful in predicting accident rates, especially features like fog, rain, and snow. Weather data can also include seasonality features, temperature, light, etc. Time of the day and day of the week are also important predictors of accident rates (Mukhopadhyay et al., 2017; Huang et al., 2008; Qi et al., 2007).

4.2.2. Spatial features

Roadway geometry is also known to be an effective predictor of crash frequency (Chin, 2003; Khazraee et al., 2018; Poch and Mannering, 1996; Shankar et al., 1995). The most commonly used features in this regard are the number of lanes, annual average daily traffic (AADT), segment length, width of the lanes, features regarding shoulders, horizontal turns and slopes (Ma et al., 2008; Zeng et al., 2017; Wen et al., 2021), the presence of uncontrolled left-turn lane, the presence of bus stops and surveillance cameras, median widths, speed limit (Khazraee et al., 2018; Huang et al., 2008), and features specific to intersections (Chin, 2003; Huang et al., 2008). Population density (Parsa et al., 2020), road density (Bao et al., 2019), and socio-economic features can also be important predictors of accidents rates, for example, the density of the bars in a region has been used in crash prediction, in particular hit and run accidents (Kuo and Lord, 2020).

4.2.3. Spatio-temporal features

Crashes exhibit strong spatial–temporal incident correlation. Past incidents are an important predictor of future incidents. For example,

areas that have typically experienced a relatively high concentration of incidents in the recent past are more likely to have incidents in the future (Mukhopadhyay et al., 2017; Mukhopadhyay et al., 2018; Mukhopadhyay et al., 2019). The average speed of vehicles naturally serves as a predictor for the likelihood of crashes (Shi, 2015). The role of traffic congestion has been studied in the context of crash analysis. Traffic congestion has been shown to increase the likelihood of rear-end crash (Shi, 2015), while there have been studies showing that congestion has no or negative effect on crash frequency (Wang et al., 2009; Baruya, 1998). Several other features like peak hour (Qi et al., 2007) and traffic volume (Xie et al., 2007), which are indirect measures of congestion, can also be used as covariates to model crash occurrence.

4.2.4. Other considerations

In addition to accounting for different types of features (spatial, temporal, spatio-temporal), there are various considerations that need to be addressed. One of the most important but often ignored challenges is endogeneity. The source of endogeneity can be broadly classified as omitted variables (unobserved heterogeneity), simultaneity, and biased sampling. Unobserved heterogeneity refers to features that are not recorded in the data but that affect the occurrence of accidents or are correlated to other observed features (Mannering and Bhat, 2014). For instance, a recent study found that the positive correlation effect between tangent segments and a wider left shoulder signifies increased crash occurrence on multilane highways. This can be the result of heterogeneous driving behavior where the risk perception of driving on tangent segments with wider left shoulders can lead drivers to speed (Saeed et al., 2019). Simultaneity arises when the explanatory variable causes the response variable and the response variable causes the explanatory variable. The classic example of simultaneity is the influence of ice-warning signs on increasing accident rates. In reality, these signs are installed in locations where the accident rate is high due to icy surfaces (Lord and Mannering, 2010). Finally, the data collected for accident prediction might be biased due to various reasons. Under reporting of crashes with no severe injuries is a prime example of biased sampling (Yasmin et al., 2014). Failure to consider the aforementioned challenges can lead to biased parameter estimates and incorrect inferences on accident rates.

4.3. Incident severity

Severity of accidents plays a crucial role in planning approaches for allocation as well as for dispatching resources when incidents occur. Naturally, decision-makers plan to prioritize incidents with higher severity over the ones with relatively lower severity. Since it is difficult to gauge the severity of an incident based on a call for assistance, it is common in practice to dispatch the responder closest to the scene of the incident. However, understanding spatial and temporal patterns in severity and its relationship with incident occurrence models is crucial in optimizing the allocation of responders. Understanding covariates that affect severity, and creating models for predicting severity of crashes have attracted a lot of attention. While there are different definitions of severity, it can usually be categorized into five levels: (1) no-injury or just property damage, (2) possible injury, (3) non-incapacitating injury, (4) incapacitating injury, and (5) fatal (Savolainen et al., 2011). Most of the prior work in severity prediction has focused on using a similar ordinal categorization. Savolainen et al. (2011) present a detailed review regarding modeling severity of accidents, which is self-contained, complete, and comprehensive. Much of this section is informed by their work; we identify crucial insights from it and also focus on models that have been introduced since then.

Let incident severity be represented by the random variable K. From the perspective of the formulation in Section 3, designing models for incident severity can be represented in two ways. First, there is significant work on creating marginal models over severity. These models have the form $h(K|\boldsymbol{w},\theta)$, where h is a distribution over K, $\boldsymbol{w}$ is a set of covariates that impact incident severity, and θ denotes the model parameters. Note that $\boldsymbol{w}$ could include information about the crash itself, such as information from post-crash reports. The other approach is to model a joint distribution that governs incident occurrence and the resulting severity. In this scenario, given incident data, the decision-maker seeks to learn a joint distribution over incident occurrence and severity, which can be represented by $h(X,K|\boldsymbol{w},\theta)$.

The relationship between traffic flow and accident severity is well-explored (Jadaan and Nicholson, 1992; Martin, 2002; Turner and Thomas, 1986; McGuigan, 1987; Hall and Pendleton, 1990). Crash severity has been explored using multinomial logit and probit models (Ye and Lord, 2014; Khattak et al., 1998; Shibata and Fukuda, 1994; Fan et al., 2016; Kockelman and Kweon, 2002), decision trees (Chang and Wang, 2006), random forests (Zheng et al., 2018, 2017, 2018), and neural networks (Moghaddam et al., 2011; Chimba and Sando, 2009). Additionally, studies have used the RP models (correlated and uncorrelated) and Bayesian model to evaluate the impact of roadways design and weather on crash injury-severity (Fountas et al., 2018; Huang et al., 2008).

One natural way to account for correlation between crash frequency and severity is to learn an independent regression model for each category of severity. Multiple regression models (Bijleveld, 2005; Ma and Kockelman, 2006; Song et al., 2006) as well as neural networks (Zeng et al., 2016) have been used to this end. Although such a paradigm captures inherent correlation (to some extent) between incident arrival and severity, it does not model an explicit joint distribution. Mukhopadhyay et al. (2017) present an approach that forms a bridge between marginal and joint models. They assume that the joint distribution can be decomposed into a marginal distribution over incident arrival, followed by a conditional distribution over severity given incident arrival.

In the last two decades, there has also been significant interest in jointly modeling incident arrival (frequency) and severity (Aguero-Valverde and Jovanis, 2009; Ma et al., 2008; Ma and Kockelman, 2006; Park, 2019; Pei et al., 2011). This includes multivariate Poisson regression (Ma and Kockelman, 2006) and multivariate Poisson log-normal regression models (Park, 2019). Pei et al. (2011) model the joint distribution explicitly and use a fully Bayesian approach to learn the model. While such models are promising, a crucial (potential) limitation is identified by Savolainen et al. (2011). Jointly modeling crash arrival and severity limits the use of data related to the specific crash while learning the model. On the other hand, marginal models can use detailed post-crash data to infer insights about severity (Savolainen et al., 2011). It is worth mentioning that factors such as the type of the vehicle, age of the driver, and using seat belt are useful in severity prediction (Hadjidimitriou et al., 2020), which are usually not included in the collected environmental data.

Finally, there are two orthogonal areas of work in severity prediction that can be combined with both marginal models and joint models. The first approach is rather recent and focuses to identify spatial relationships between different levels of severity (Kuo and Lord, 2020). The other approach seeks to tackle inherent heterogeneity in crash data by identifying clusters of incidents (not necessarily spatial) to better understand the relationship between crash data and covariates (Sun et al., 2019; Sasidharan et al., 2015; Sivasankaran and Balasubramanian, 2020).

4.4. Key takeaways

Based on the approaches surveyed, we summarize steps that practitioners and model designers should take in Fig. 4. Specifically, we recommend practitioners, model designers, and planners to:

1. Be aware of advances made in predictive modeling in the context of different types of incidents.

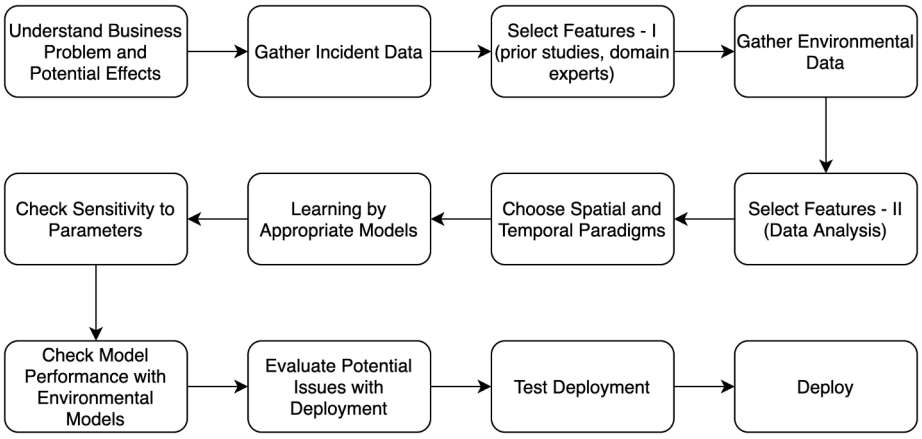

Fig. 4. Incident Prediction Model Design Pipeline.

2. Seek the help of domain experts (researchers, fire-fighters, policy-makers, etc.) to design the feature space w, which is a crucial factor in the performance of predictive models.
3. Start by using well-defined paradigms that have been shown to work on multiple datasets (e.g. hierarchical Poisson models and zero-inflated models).
4. Be aware of flaws and shortcomings of models, and carefully evaluate the possible costs of inaccurate predictive models.
5. Be aware that many different kinds of predictive models have been used to model crash occurrence, and models can be sensitive to a variety of factors like the granularity of spatial discretization, choice of covariates, etc. As a result, it is important that practitioners test multiple models and evaluate their comparative performance to understand the model that best suits a particular scenario.

5. Event extraction

Response to incidents like accidents and medical emergencies must be dispatched as soon as possible. For decades, the pipeline depended on a human reporting the incident, after which responders were dispatched to the scene. However, with the advent of a variety of sensors available in smart cities and the wide corpus of information on social media, it is now possible to detect incidents before they are reported. For example, consider a fire in an urban area. An observer might share the observation on social media, or the incident might be captured by video cameras installed at traffic intersections. The goal of event extraction is to use such data to detect the occurrence of incidents in order to reduce the overall time for response.

Event extraction algorithms seek to identify as much information about an incident as possible, with a focus on specific details such as the location, time, and the agents involved. For different events, the *who, where, what, when*, and *why* information vary a lot, but similar events usually share the same event template. For example, in an *accident* event, there are usually *entities* involved in the accident, the *location* of the event, and its *time* of occurrence. Researchers in Natural Language Processing (NLP) community have developed event ontologies to define the templates for various events (Baker et al., 1998; Schuler, 2005; Doddington et al., 2004). Event extraction can then be defined as a task to convert unstructured data into event-centered structured data based on a specific event ontology. This is greatly beneficial for first responders since it is much easier for users to manage and query structured data. We focus our attention on two major categories of models for event extraction (EE): (1) EE from textual information, and (2) EE from multimedia information. We point out that there has also been recent work in extracting information about events by using crowdsourcing applications such as Waze (Senarath et al., 2021), who focus on optimizing the balance of practitioner-centric parameters (e.g., spatial and temporal localization of alerts) with learning outcomes (e.g., accuracy).

5.1. Event extraction from textual information

The goal of EE using NLP is to identify events from text and classify them into pre-defined categories, as well as identify event participants (for example, the victim in an accident), and event attributes (for example, the location and time of an event).

There are two subtasks in EE: (1) trigger classification, which aims to identify the word/phrase that clearly expresses the occurrence of an event and seeks to classify it based on pre-defined categories. For example, the word 'crash' in Fig. 5 is the trigger word of a Transportation-Accident event; (2) event argument classification, which aims to identify the event participants and event attributes. For example, 'A614 road', and 'a car' are Location argument and Entity argument of the Transportation-Accident event respectively (Fig. 5).

The early work on EE mainly relies on feature engineering and adapts a pipeline framework (Grishman et al., 2005; Ahn, 2006; Ji and Grishman, 2008; Hong et al., 2011; McClosky et al., 2011). The input sentence is first tokenized into a sequence of tokens. For each token, various features are used, which can be divided into three categories: (1) lexical features such as n-grams, lemma, and synonyms of tokens, and brown clusters, (2) syntactic features such as dependent and governor tokens, and (3) entity features such as entity type. Then a classifier is trained based on statistical models (for example, logistic regression) to predict event triggers. Subsequently, another classifier is trained to predict the event arguments. During inference, the argument classifier receives the prediction from the trigger classifier as input; therefore, the errors from the former classifier are easily propagated into the latter one. To mitigate the error propagation issue in pipeline models, a joint model can be learned. For example, Li et al. (2013) proposed using structured perceptron (Collins, 2002) with beam search to learn a joint model by leveraging the dependencies between arguments and triggers.

The limitation of models based on feature engineering is that they rely on handcrafted features and language-specific resources such as part-of-speech (POS) taggers and dependency parsers. As a result, it is hard to adapt such models to new languages or domains. This problem manifests frequently in working with textual data from social media because of language variations and informal grammar used in such platforms. Indeed, POS taggers and dependency parsers perform much worse in the context of social media than in the structured news domain. With the idea of using word embeddings (Mikolov et al., 2013), deep neural networks have become an attractive choice for researchers because no handcrafted features are required. For example, Chen et al. (2015) applied convolutional neural networks (CNN) in a two-step pipeline system, in which the tokens are converted into pre-trained word vectors. Nguyen et al. (2016) proposed a joint framework with

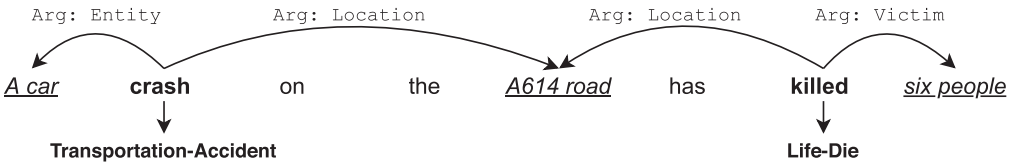

Fig. 5. An example of Transportation-Accident and Life-Die events and their arguments.

bidirectional recurrent neural networks.

However, a potential issue with the use of simple CNN models is that they can hardly capture the syntactic relations between words, which are very important features for argument classification. Nguyen and Grishman (2018) applied Graph Convolutional Networks (GNN) based on dependency trees to generate word representations by leveraging the information from other words with close syntactic relations.

The models mentioned above first identify event trigger and entity and then perform argument role classification. There has also been recent work that bypasses the entity recognition step. For example, Wadden et al. (2019) learn entity span representations instead of explicitly labeling the entity using BIO schema, and Du and Cardie (2020) model event extraction as question answering and extracts the spans of event arguments with certain role types.

5.2. Event extraction from multimedia information

Event extraction is also possible from video and image data (Gan et al., 2015; Chang et al., 2016; Ma et al., 2017; Li et al., 2020). In the context of emergency response, cities are increasingly trying to use traffic cameras to track and monitor congestion and detect incidents that need response. Gan et al. (2015) applied a CNN pre-trained on Imagenet (Deng et al., 2009) to perform keyframe detection. A weighted-sum of the representations learned from different multimedia archives to generate better representations for event videos has also been used (Chang et al., 2016). Lower level video attributes learned from external multimedia datasets can also be used to improve complex event detection in videos (Ma et al., 2017). A potential drawback of such approaches is that they fail to be applicable in complex scenarios that have event arguments. More recently, Li et al. (2020) propose a new task for multimedia EE, in which they extend the EE task from textual setting into the multimedia setting and target complex events. The extension is done by applying weakly supervised training to project the structured semantic representations from textual and visual data into a common space. Event extraction is an emerging field of research and with the rapid growth of smart and connected communities, it promises to play a vital role in the design of principled ERM pipelines.

6. Responder allocation and dispatch

There are two important steps in an ERM system that come into effect *after* the decision-maker gains an understanding of when and where incidents are likely to happen. These involve allocating resources (also referred to as the stationing problem (Pettet et al., 2020)) in anticipation of incidents and dispatching resources when calls for service are received (also referred to as the stationing problem). While prediction problems are primarily formulated as *learning* problems, allocation, and response are commonly modeled as optimization problems. As discussed in Section 3, an allocation or dispatch problem can be represented as $\max_y G(y|f)$, where y represents the decision variable, G is a reward function chosen by the decision-maker, and f is the model of incident occurrence. For allocation problems, y typically refers to the location of emergency responders in space. For dispatch problems, the decision variable is a mapping between responders and specific calls for service.

The distinction between allocation and dispatch problems can be hazy since any solution to the allocation problem implicitly creates a policy for response. For example, consider an algorithm that allocates ambulances across an urban area in a manner that minimizes expected response times to incidents according to an incident arrival model f. Now, when an incident occurs in the jurisdiction of a specific station, naturally, a responder (if available) is dispatched from the station, without the need for an explicit dispatch model. While this is generally true for allocation models, there are finer subtleties involved. As noted by Mukhopadyay et.al. (Mukhopadhyay et al., 2017; Mukhopadhyay et al., 2018; Mukhopadhyay et al., 2019), implicit response strategies are not always optimal. For example, consider a situation where an incident occurs close to a station that has no available responders. Should the incident enter a waiting queue? How does the potential severity of the concerned incident affect this decision? If a nearby station has a free responder, should it be dispatched? How do response time guarantees from the allocation model change in such scenarios? Answering such questions is critical for an efficient ERM system, which motivates the design of principled approaches for dispatching responders. This section discusses prior work on the problems of responder allocation and response.

We first introduce the metrics used to allocate emergency response stations and responders. The three most common metrics are coverage (Toregas et al., 1971; Church and ReVelle, 1974; Gendreau et al., 1997), distance between facilities and demand locations (Mukhopadhyay et al., 2019), and patient survival (Erkut et al., 2008; Knight et al., 2012; McCormack and Coates, 2015). *Coverage* measures the proportion of spatial locations that are within some predefined distance of the responders (or depots). It is measured with respect to demand nodes, which are discretized spatial units that can potentially generate calls for service. Of the three metrics, it is the most straightforward to examine as it is generally binary. A demand node is considered to be *covered* by some facility if it is within the predefined distance, and otherwise considered to be uncovered. It also lines up well with the broader objective of many EMS providers, which is to limit the number of calls that are responded to *late*, i.e. that have a response time higher than some threshold (the distance often serves as a proxy for the response time, for example see Mukhopadhyay et al. (2017)). These factors contributed to coverage being a prevalent metric in early EMS allocation research.

The distance between potential demand nodes and their nearest facilities is another metric that can be used for optimization of the spatial distribution of stations and responders. These metrics are more difficult to use since they are not binary, but recent advances in computational capabilities have made them more accessible. Both coverage and distance to potential demand locations actually approximate the true objective of EMS policies, which is increasing patient survival. Erkut et al. (2008) argue that it is more appropriate to use expected patient survival directly by incorporating a survival function that captures the relationship between response times and survival rates.

Most early ERM allocation approaches modeled the allocation problem as an integer or linear optimization problem (Toregas et al., 1971; Church and ReVelle, 1974; Gendreau et al., 1997). These models are relatively straightforward and can be solved by a large body of optimization techniques. Exact methods such as branch-and-bound have been applied to small instances of the problem (Swoveland et al., 1973; Marianov and Revelle, 1994) but do not easily scale to realistic environments. As a result, most prior work relies on heuristic approaches,

such as genetic algorithms (Jia et al., 2007; Rajagopalan et al., 2007) and tabu search (Gendreau et al., 1997; Rajagopalan et al., 2007; Gendreau et al., 2001; Rajagopalan et al., 2008). Recently, decision theoretic models such as Markov decision processes (MDPs) have gained traction as efficient solution methods have evolved (Maxwell et al., 2009; Mukhopadhyay et al., 2019).

6.1. Coverage models

Early allocation approaches also generally tackled *static* allocation. Depots (also referred to as 'facilities' or 'stations') are assumed to be immobile, so the model determines the optimal locations for the depots without allowing for temporal redistribution. In such models, responders are often used synonymously with facilities. The two seminal static facility allocation models are the Location Set Covering Problem (LSCP) (Toregas et al., 1971) and the Maximal Covering Location Problem (MCLP) (Church and ReVelle, 1974). Both models have similar assumptions, including that stations act independently, response is deterministic, that at most one ambulance is at each facility, and that there is one type of ambulance. The primary difference between the two is in the optimization objective. LSCP finds the least number of facilities that *cover* all demand nodes, while MCLP maximizes the demand covered by a given number of facilities. LSCP can be useful for planning a lower bound on the number of facilities needed for a given coverage standard, while MCLP better captures the constraints of real world use cases where the number of facilities is heavily constrained by cost. It is also common to introduce constraints on secondary objectives like waiting times in optimization problems that seek to maximize coverage. For example, Silva and Serra (2008) and Mukhopadhyay et al. (2017) define optimization frameworks for maximizing coverage with upper bounds on waiting times, and can accommodate different levels of incident severity.

There are a number of extensions to LSCP and MCLP, many of which relax some of their strong assumptions. Aly and White (1978) consider a spatially continuous demand model, rather than the discrete demand nodes. Jia et al. (2007) introduce different quality levels for facilities (which can represent each facility's available services or equipment), with demand points having different coverage constraints for each level. Erkut et al. (2008) incorporated a survival function into the optimization function of MCLP, which maps response times to survival rates, to formulate the Maximum Survival Location Problem (MSLP).

LSCP, MCLP, and many of their extensions all have a common shortcoming in that they assume deterministic system behavior in regards to response. Resources at a facility are considered to be always available, and the models assume that a station is able to service all demand nodes that it covers. In the real world, there are finite resources at each station, and calls from a specific demand node might need to be answered by a station other than the closest one. For example, it is common for other stations to respond to a call if the closest one is busy. One way to address this is by increasing the number of stations that cover each demand point, i.e. using a *multiple coverage* metric.

A key example is the Double Standard Model (DSM) (Gendreau et al., 1997), which incorporates two distance standards r_1 and r_2, where $r_1 < r_2$. The model adds the constraint that all demand must be covered within r_2, similarly to LSCP, ensuring that each point has *some* coverage. It also specifies that some proportion α of the demand is covered within r_1. Given those constraints, the objective is to maximize the demand covered *by at least two stations* within r_1. Essentially, this maximizes the demand nodes that have nearby facilities while ensuring that all demand nodes have adequate coverage. While this approach helps mitigate the issue of station unavailability, there can still be situations where both facilities covering some demand points are busy. Accounting for such situations requires modeling facility availability explicitly.

There is a large body of research on probabilistic coverage models, which model the stochastic nature of station availability. Two foundational probabilistic models are the Maximum Expected Covering Location Model (MEXCLP) and Maximum Availability Location Problem (MALP). MEXCLP was introduced by Daskin (1983) and extends MCLP, modifying the optimization function to account for station availability. It assumes that each facility has the same probability of being busy, which simplifies computation but does not accurately represent the real world where facilities near incident hot spots are unavailable for a greater proportion of the time. Also, it inherits many of the assumptions of MCLP, and assumes that facilities act independently. MALP, proposed by Revelle and Hogan (1989) maximizes the demand covered by facilities with some exogenously specified probability. The first version, MALP-I (ReVelle and Hogan, 1989) is similar to MEXCLP in that it assumes equal probabilities for being busy for facilities. MALP-II (ReVelle and Hogan, 1989), however, removes this assumption. The proportion of time that facilities are busy is computed as a ratio between the total demand generated by demand points and the availability of facilities covering them.

There have been several extensions to the above probabilistic models to relax some of their simplifying assumptions and make them better match the real world. TIMEXCLP, developed by Repede and Bernardo (1994), introduces temporal variations in travel times between points to MEXCLP. Adjusted MEXCLP (AMEXCLP) (Batta et al., 1989) relaxes MEXCLP's assumption that facilities are independent by treating them as servers in a hypercube queuing system (Larson, 1974) with equal busy factions. The Queuing Probabilistic Location Set Covering Problem (QPLSCP) (Marianov and Revelle, 1994) makes a similar extension to MALP by computing each individual facility's busy fraction using a queuing model and feeding them into MALP-II. We summarize the different coverage models from prior work, their objective functions, and features in Table 1.

6.2. Decision-theoretic models

Having discussed how the ERM planning problem can be framed as an explicit optimization problem, we now discuss an alternate approach which models the planning problem as a stochastic control problem, and then optimizes over the set of control choices to maximize expected reward. The most commonly used model in this regard is the Markov decision process (MDP). MDP-s can be used as a general framework for sequential decision problems under uncertainty given a model of the concerned system (Kochenderfer, 2015). In such a formulation, an agent chooses an action at a given state of the system and receives a specific reward based on a pre-defined utility function. The system then transitions to a new state probabilistically. The *Markovian* assumption means that the subsequent state depends only on the current state and the action taken. MDP-s have been used extensively to model the EMS dispatch process (Carter et al., 1972; Keneally et al., 2016; Mukhopadhyay et al., 2018; Mukhopadhyay et al., 2019; Mukhopadhyay et al., 2017; Pettet et al., 2020).

Carter et al. (1972) demonstrate one of the earliest examples of using an continuous-time MDP to aid emergency response by using a queuing model for calls. The general framework of such an approach has been used in several studies since then (Mukhopadhyay et al., 2018; Mukhopadhyay et al., 2019; Mukhopadhyay et al., 2017). Keneally et al. (2016) also model the optimal dispatch problem as a continuous-time MDP and consider different levels of priorities for the incidents while dispatching. They assume that the state transition function for the EMS system can be expressed in closed-form and use canonical policy iteration to solve the problem. A shortcoming of such a model is that it assumes that state transitions follow a memoryless distribution. Real world transitions are not necessarily memoryless, and this is addressed by Mukhopadhyay et.al. (Mukhopadhyay et al., 2018), who formulate the problem as a semi-Markovian decision problem (SMDP) instead. In the absence of closed-form expressions for state transitions, a black-box simulator can be used to learn an optimal dispatch policy. However, such an approach does not scale well to realistic scenarios. An approach to alleviate this problem is to focus on finding an action for the current

Table 1
Coverage Models.

Approach	Objective	Description	Reference
LSCP	Minimize number of facilities to completely cover demand.	Finds a lower bound on the number of facilities needed for a given coverage standard. Assumes facilities can service all demand they cover.	Toregas et al. (1971)
LSCP with continuous regions	Minimize number of facilities to completely cover demand.	Extends LSCP to consider spatially continuous regions rather than discrete demand points.	Aly and White (1978)
MCLP	Maximize demand covered by given number of facilities.	Represents problems where the number of facilities is heavily constrained by costs. Assumes facilities can service all demand they cover.	Church and ReVelle (1974)
MCLP with multiple quality levels	Maximize the quality-weighted demand covered by facilities with various quality levels.	Extends MCLP by introducing multiple quality levels representing each facility's available services or distance to demand points.	Jia et al. (2007)
MSLP	Maximize patient survival with given number of facilities.	Extends MCLP by mapping response times to patient survival rates.	Erkut et al. (2008)
DSM	Maximize the demand covered at least twice with available facilities.	Accounts for facilities being busy by ensuring demand is covered by at least two facilities.	Gendreau et al. (1997)
MEXCLP	Maximize demand covered by given number of facilities.	Extends MCLP to account for station availability. Assumes all facilities have same busy probabilities and that facilities act independently.	Daskin (1983)
TIMEXCLP	Maximize demand covered by given number of facilities.	Extends MEXCLP by introducing temporal variations in travel times between locations.	Repede and Bernardo (1994)
AMEXCLP	Maximize demand covered by given number of facilities.	Extends MEXCLP, relaxing the assumption that facilities are independent.	Batta et al. (1989)
MALP-I	Maximize demand that is covered by facilities with a exogenously specified probability.	Assumes all facilities have same busy probabilities and that facilities act independently.	ReVelle and Hogan (1989)
MALP-II	Maximize demand that is covered by facilities with a exogenously specified probability.	Relaxes assumption that all facilities have the same busy probabilities. Busy probabilities are computed as a ratio between demand generated at demand points and the availability of facilities covering them.	ReVelle and Hogan (1989)
QPLSCP	Maximize demand that is covered by facilities with a exogenously specified probability.	Extends MALP-II by relaxing the assumption that facilities are independent.	Marianov and Revelle (1994)

state of the world instead of aiming to find a policy for the entire state-space (Mukhopadhyay et al., 2019). Given a generative model for the EMS system, heuristic search approaches like Monte-Carlo tree search (MCTS) can be used to find promising actions for the current state of the MDP (Mukhopadhyay et al., 2019; Pettet et al., 2020). An advantage of using MCTS is that the Markovian assumption can be relaxed, and high-fidelity simulators can be used to estimate utilities from different actions.

An important aspect of using decision-theoretic models is to carefully design the utility function. As discussed earlier, a threshold on response time can be used to penalize the number of calls that are responded to late. However, an explicit relationship between patient survival and response time can be directly used to design the reward function for MDP-s (Bandara et al., 2012). It has also been highlighted that a consideration of priorities of calls is crucial to take into account while designing the utility function for decision-theoretic approaches to EMS (Keneally et al., 2016; Bandara et al., 2012).

6.3. Dynamic and proactive reallocation

A potential shortcoming of algorithmic dispatch approaches is important to ponder over. Based on conversations with first responders, Pettet et al. (2020) point out that the moral constraints in emergency response dictate that the nearest responder be dispatched to the scene of an incident, particularly when the severity of an incident cannot be gauged from the call for service. As a result, the decision variable that can be optimized exogenously is the location of responders in anticipation of incidents; on the other hand, choosing which responder to dispatch is done in a greedy manner. Pettet et al. (2020) create an approach to optimize over the spatial distribution of responders *between* incidents, while always dispatching the closest available responder to attend to incidents. This process alleviates two major issues. First, it does not waste crucial time *after* an incident has occurred to optimize over which responder to dispatch. Second, the moral constraint of always sending the closest responder to an incident is not violated. Dynamically reallocating emergency responders has actually been investigated earlier, to maximize the overall efficacy of response; such approaches are not necessarily motivated by the use of greedy dispatch policies. An approach to tackle the allocation problem dynamically is to formulate an integer-program and solve it in real-time every time reallocation needs to be made (Maxwell et al., 2010; Kolesar and Walker, 1974). Dynamic reallocation has also been addressed by decision-theoretic formulations, which seek to find optimal policies that govern when and where specific units should be moved. The utility function typically consists of expected time taken to serve requests in the long run. Optimal policies can be found by dynamic programming or approximate dynamic programming techniques (Maxwell et al., 2010; Berman, 1981).

A problem that decision-theoretic methods pose for dynamic reallocation (and dispatch as well, depending on the size of the geographic area) is that the large state and action spaces can render standard approaches to be intractable. In such cases, approximation methods and intelligent heuristics can be used to leverage the structure of the specific problem to ensure scalability. One way is to use the canonical *divide and conquer* approach by segregating the area under construction into sub-regions (Pettet et al., 2020). Then, a smaller decision-theoretic problem can be solved for each sub-region. Another way to ensure scalability is to use decentralized planning (Pettet et al., 2020), in which each agent plans for itself with locally available information. Such an approach can also be particularly useful for situations where communication systems are down (for example, in 2020, an explosion in Nashville, a city in the United States broke communication systems and hampered emergency response for four days (USA Today, 2020)). Both approaches have their own merits that we summarize in the next section (see Fig. 6).

6.4. Key takeaways

Allocation and response models are a crucial component of ERM pipelines, and a variety of algorithmic approaches have been used for allocating responders in anticipation of accidents. Recently, decision-theoretic approaches have been widely used for allocating and dynamically rebalancing the distribution of responders. We summarize the crucial takeaways from different decision-theoretic approaches that

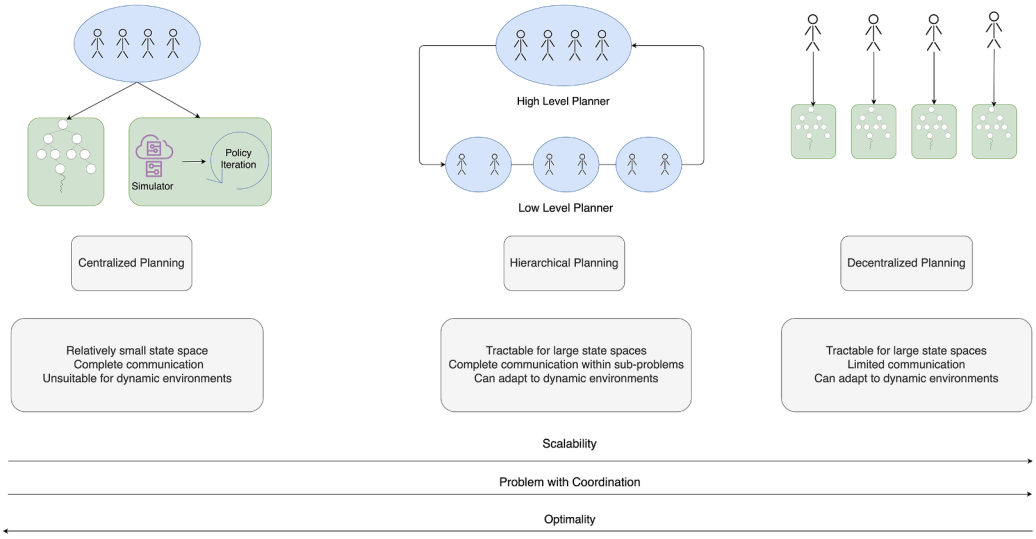

Fig. 6. The spectrum of decision theoretic models that can be used to tackle allocation and response in ERM pipelines. From left to right, scalability increases but coordination among agents and utility (in terms of optimality) decreases.

have been taken for emergency response in Fig. 6. There are three major paradigms that have been used in this context — centralized planning, hierarchical planning, and decentralized planning. Characteristics and features of each method are shown in Fig. 6.

An extremely important area of focus in designing approaches for allocation and dispatch is the choice of the variable or metric that is optimized. Specifically, patient survival is a vital consideration that ambulances need to take into account while designing response models, since ambulances need to transport patients to medical facilities, which in turn increases the overall service time. This effect is naturally manifested in the choice of objectives and variables for allocation models. Despite this difference, there are high-level similarities in response modeling that apply to all emergency incidents (especially in reactive response). Models focusing on increasing coverage and reducing wait-times are common objectives that have been widely used in practice. Finally, the dynamic nature of urban areas must be factored into the models as changes in the traffic and constructions can affect response times by ambulances.

Finally, we recommend model designers and practitioners to:

1. Be well-versed with the different objectives that have been used in response and allocation models, and carefully choose the one that suits the specific needs of the concerned area.
2. Seek the help of domain experts (researchers, fire-fighters, etc.) to understand problems that responders face in the field. For example, the nearest ambulance might be heading in the opposite direction from the demand node on a highway, without the scope of making a turn. This makes it important to consider features that might not be intuitive to researchers.
3. Seek to bridge the gap between theoretical models and realistic environmental constraints. For example, there is a rich body of work that makes the assumption that the environment in which ERM systems operate is static. While such assumptions simplify computational challenges, they might not truly capture the dynamics of actual ERM pipelines.
4. Carefully evaluate the performance of predictive models and simulators before they are used for response and allocation.
5. Consider prior work that uses well-crafted heuristics for scalability.
6. Smart and connected communities must plan in advance to account for events that might destroy communication infrastructure. Decentralized planning with locally available information can be promising in such situations.

7. Summary discussion and conclusion

The field of designing emergency response pipelines has seen tremendous growth in the last few decades. Several factors have contributed to this growth including wider availability of data, the development of data-driven methodologies, increased cognizance, dependence and trust over algorithmic approaches by governments, and increase in computational power. However, there are still challenges in this field that need to be addressed. As we have pointed out, an EMS pipeline consists of an intricate combination of several components for its smooth functioning. There is a need for more research groups to: (i) study EMS pipelines in their entirety, and consider the broader impact of their modular work on ERM systems, (ii) consider and acknowledge the challenges and constraints that first responders face in the field, and (iii) iteratively develop ERM tools by having first responder organizations in the loop. There are nuances that describe such needs throughout this paper. For example, an improved statistical fit for the prediction models does not necessarily mean an overall improvement for the ERM pipeline if the underlying model does not capture the true dynamics of incident occurrence. There is also a need for researchers to make their data and tools available to both the research community and ERM organizations.

In a comprehensive review of statistical methods of crash prediction, Lord and Mannering (2010) pointed out that the wider availability of data is extremely promising for the field of crash prediction. This is particularly true now. Vast volumes of real-time data are now available from electric scooters, automobiles, and ambulances. There is also wider coverage of sensors like video-cameras throughout urban areas. This promise of increased availability of richer data holds true not only for incident data but also for data regarding covariates that potentially affect incident occurrence, like traffic congestion. The net result of an increased stream of data promises a finer understanding of the effect of covariates on incident occurrence. This benefit can be utilized by sharing data and algorithmic approaches between research groups and first responders.

Urban dynamics of accidents and crashes are continuously changing, and hold several opportunities for research. The increase in the number of automobiles and the arrival of autonomous vehicles in the markets across the globe presents the scope of re-evaluating existing models of crash occurrence and designing newer models that accommodate the changing landscape. Litman (2017) lists the various additional planning constraints that need to be taken into account while developing transit systems that can accommodate autonomous vehicles, as well as additional causes for crashes, like software failure and increased overall

travel volume. The potential risk factors caused by the interaction between autonomous and non-autonomous vehicles also pose challenges Jafary et al. (2018) and the need to design newer models of incident prediction.

Incident response also poses fresh challenges and opportunities. First, there is a need to combine the different metrics used in designing dispatch and allocation models. There are several interesting threads of research (cooperative coverage, survival metrics, gradual coverage decay, incorporating multiple resource types with different functionalities, etc.) that, to the best of our knowledge, have not been combined and evaluated together. Also, there has not been much focus on explicitly incorporating measures of patient survival directly in response models. We think that it is crucial that patient survival be studied in more detail and included as a part of objective functions for optimization approaches used in designing allocation and dispatch systems.

A recent development in emergency response systems has been the computational ability of agents. Most modern ambulances and police vehicles are now equipped with laptops, which presents the scope of fast and decentralized decision-making, a particularly exciting area for multi-agent systems. Decentralized decision-making has been explored in the context of urban ERM systems Pettet et al. (2017), but such approaches are probably more relevant for disaster scenarios like floods and hurricanes, where agents might lose connectivity to the central decision-making authority. Algorithmic approaches to aid the strategic redistribution of responders between incidents is extremely promising. While post-incident planning presents many technical challenges, such approaches rarely get implemented in the field. Inter-incident planning, on the other hand, respects the inherent challenges that emergency response faces.

Finally, as smart and connected communities grow, it is crucial to ensure that resources are distributed and allocated in a manner that is equitable. As a result, equity of emergency response is also a concern as accessibility to emergency response can depend on the availability of financial resources and socio-economic backgrounds. For example, prior work has suggested bias in emergency services (specifically in drug administration) against minority communities (NPR, 2019). There is a need to first quantify fairness in the context of emergency response and then explicitly model such a notion in decision-making pipelines. As urban areas grow and witness a rise in population density, the need to design principled approaches to aid emergency response grows as well. This survey identifies how the field has evolved over the last few decades, with the view to aid researchers, policy-makers, and first-responders in designing better ERM pipelines.

CRediT authorship contribution statement

Ayan Mukhopadhyay: Conceptualization, Writing - original draft, Writing - review & editing, Supervision. **Geoffrey Pettet:** Writing - original draft, Writing - review & editing. **Sayyed Mohsen Vazirizade:** Writing - original draft, Writing - review & editing. **Di Lu:** Writing - original draft. **Alejandro Jaimes:** Writing - review & editing, Supervision. **Said El Said:** Writing - review & editing, Funding acquisition. **Hiba Baroud:** Writing - original draft, Writing - review & editing, Supervision. **Yevgeniy Vorobeychik:** Writing - review & editing. **Mykel Kochenderfer:** Conceptualization, Writing - review & editing, Funding acquisition. **Abhishek Dubey:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgement

The authors will like to thank the Center of Automotive Research at Stanford (CARS), the National Science Foundation (Grants: CNS-1640624, IIS-1814958, CNS-1818901) and the Tennessee Department of Transportation (TDOT) for funding the research. We would also like to express our gratitude towards the Nashville Fire Department (NFD), the Metropolitan Nashville Police Department (MNPD), the Nashville Metropolitan Information Technology Services (ITS) and the Tennessee Department of Transportation (TDOT) for invaluable feedback and domain expertise that made us understand the subtleties and intricate challenges of emergency response. We would also like to thank Saideep Nannapaneni and Hemant Purohit for giving feedback about the paper. Finally, we would like to thank the anonymous reviewers who helped us improve the paper.

References

M. Abdel-Aty, A. Pande, A. Das, and W.J. Knibbe. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transportation Research Record, 2083 (1): 153–161, 2008. ISSN 0361–1981.

M.A. Abdel-Aty, H.M. Hassan, M. Ahmed, and A.S. Al-Ghamdi. Real-time prediction of visibility related crashes. Transportation Research Part C: Emerging Technologies, 24: 288–298, 2012. ISSN 0968–090X.

H.T. Abdelwahab and M.A. Abdel-Aty. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record, 1784 (1): 115–125, 2002. ISSN 0361–1981.

Ackaah, W., Salifu, M., 2011. Crash prediction model for two-lane rural highways in the ashanti region of ghana. IATSS Res. 35 (1), 34–40.

Aguero-Valverde, J., 2013. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. Accid. Anal. Prevention 50, 289–297.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transp. Res. Rec. 2061 (1), 55–63.

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate poisson lognormal models for crash severity modeling and site ranking. Transp. Res. Rec. 2136 (1), 82–91.

D. Ahn. The stages of event extraction. In Workshop on Annotating and Reasoning About Time and Events, pages 1–8, 2006.

Akin, D., 2011. Analysis of highway crash data by negative binomial and poisson regression models. International Symposium on Computing in Science and Engineering (ISCSE).

Al-Bdairi, N.S.S., Behnood, A., Hernandez, S., 2020. Temporal stability of driver injury severities in animal-vehicle collisions: A random parameters with heterogeneity in means (and variances) approach. Analytic Methods Accident Res. 26, 100120.

Al-Ghamdi, A.S., 1995. Time series forecasts for traffic accidents, injuries, and fatalities in saudi arabia. J. King Saud University-Eng. Sci. 7 (2), 199–217.

Al-Hasani, G., Khan, A.M., Al-Reesi, H., Al-Maniri, A., 2019. Diagnostic time series models for road traffic accidents data. Int. J. Appl. Stat. Econ. 2, 19–26.

Alnawmasi, N., Mannering, F., 2019. A statistical assessment of temporal instability in the factors determining motorcyclist injury severities. Anal. Methods Accid. Res. 22, 100090.

Aly, A.A., White, J.A., 1978. Probabilistic formulation of the emergency service location problem. J. Operational Res. Soc. 29 (12), 1167–1179.

Association for Safe International Road Travel. Road Safety Facts. url:https://www.asirt.org/safe-travel/road-safety-facts/, 2019.

Baker, C.F., Fillmore, C.J., Lowe, J.B., 1998. The berkeley framenet project. In: International Conference on Computational Linguistics, pp. 86–90.

Bandara, D., Mayorga, M.E., McLay, L.A., 2012. Optimal dispatching strategies for emergency vehicles to increase patient survivability. Int. J. Operational Res. 15 (2), 195–214.

Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prevention 122, 239–254.

Baruya, A., 1998. Speed-accident relationships on European roads. In: International Conference on Road Safety in Europe, pp. 1–19.

Basak, S., Ayman, A., Laszka, A., Dubey, A., Leao, B., 2019. Data-driven detection of anomalies and cascading failures in traffic networks. In: Annual Conference of the Prognostics and Health Management Society.

Basak, S., Dubey, A., Bruno, L., 2019. Analyzing the cascading effect of traffic congestion using lstm networks. In: International Conference on Big Data, pp. 2144–2153.

Basu, S., Saha, P., 2017. Regression models of highway traffic crashes: a review of recent research and future research needs. Procedia Eng. 187, 59–66.

Batta, R., Dolan, J.M., Krishnamurthy, N.N., 1989. The maximal expected covering location problem: Revisited. Transp. Sci. 23 (4), 277–287.

Behnood, A., Mannering, F., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: Some empirical evidence. Anal. Methods Accident Res. 8, 7–32.

Behura, A., Behura, A., 2020. Road accident prediction and feature analysis by using deep learning. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, pp. 1–7.

Berman, O., 1981. Dynamic repositioning of indistinguishable service units on transportation networks. Transp. Sci. 15 (2), 115–136.
Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accid. Anal. Prevention 37 (4), 591–600.
Bohm, K., Kurland, L., 2018. The accuracy of medical dispatch - a systematic review. Scandinavian J. Trauma, Resuscitation Emergency Med. 26 (1), 94.
Bonneson, J.A., McCoy, P.T., 1993. Estimation of safety at two-way stop-controlled intersections on rural highways. Transp. Res. Rec. 1401, 83–89.
Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accid. Anal. Prevention 39 (4), 657–670.
Carter, G.M., Chaiken, J.M., Ignall, E., 1972. Response areas for two emergency units. Operations Res. 20 (3), 571–594.
Center of Disease Control and Prevention. Road traffic injuries and deaths – a global problem. url:https://www.cdc.gov/injury/features/global-road-safety/index.html, 2019.
Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Anal. Methods Accident Res. 3, 11–27.
Chang, L.Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Sci. 43 (8), 541–557.
Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accid. Anal. Prevention 38 (5), 1019–1027.
Chang, X., Ma, Z., Yang, Y., Zeng, Z., Hauptmann, A.G., 2016. Bi-level semantic representation analysis for multimedia event detection. IEEE Trans. Cybern. 47 (5), 1180–1197.
Chen, C., Fan, X., Zheng, C., Xiao, L., Cheng, M., Wang, C., 2018. Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In: 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD). IEEE, pp. 328–333.
Chen, T., Shi, X., Wong, Y.D., 2019. Key feature selection and risk prediction for lane-changing behaviors based on vehicles– trajectory data. Accid. Anal. Prevention 129, 156–169.
Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. In International Joint Conference on Natural Language Processing, pages 167–176, 2015.
Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prevention 37 (5), 870–881.
Chimba, D., Sando, T., 2009. The prediction of highway traffic accident injury severity with neuromorphic techniques. Adv. Transp. Studies 19, 17–26.
Chin, H.C., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accid. Anal. Prevention 352, 253–259.
Chung, Y., Recker, W.W., 2012. A methodological approach for estimating temporal and spatial extent of delays caused by freeway accidents. IEEE Trans. Intell. Transp. Syst. 13 (3), 1454–1461.
R. Church and C. ReVelle. The maximal covering location problem. volume 32, pages 101–118, 1974.
Ciy of Rochester. How 911 works - what happens when you dial 911. url:https://www.cityofrochester.gov/article.aspx?id=8589935579, 2019.
Collins, M., 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Conference on Empirical Methods in Natural Language Processing, pp. 1–8.
Conway, K.S., Kniesner, T.J., 1991. The important econometric features of a linear regression model with cross-correlated random coefficients. Econ. Letters 35 (2), 143–147.
Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: The random parameters negative binomial panel count data model. Anal. Methods Accid. Res. 7, 37–49.
Daskin, M.S., 1983. A maximum expected covering location model: Formulation, properties and heuristic solution. Transp. Sci. 17 (1), 48–70.
Davis, G.A., 2001. Bayesian identification of high-risk intersections for older drivers via Gibbs sampling. Transport. Res. Rec. 1746 (1), 84–89.
Deacon, J.A., Zegeer, C.V., Deen, R.C., 1974. Identification of hazardous rural highway locations. Transp. Res. Rec. 543.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition, pp. 248–255.
U. Department of Homeland Security. Phases of Emergency Management. https://www.hsdl.org/?view&did=488295, 2019.
Deublein, M., Schubert, M., Adey, B.T., Köhler, J., 2013. Prediction of road accidents: A Bayesian hierarchical approach. Accid. Anal. Prevention 51, 274–291.
Dissanayake, S., Ratnayake, I., 2006. Statistical modelling of crash frequency on rural freeways and two-lane highways using negative binomial distribution. Adv. Transp. Stud. 9, 81–96.
Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S.M., Weischedel, R.M., 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In: International Conference on Language Resources and Evaluation, pp. 837–840.
X. Du and C. Cardie. Event extraction by answering (almost) natural questions. arXiv preprint arXiv:2004.13625, 2020.
Durduran, S.S., 2010. A decision making system to automatic recognize of traffic accidents on the basis of a gis platform. Expert Syst. Appl. 37 (12), 7729–7736.
J. Eck, S. Chainey, J. Cameron, M. Leitner, and R. Wilson. Mapping crime: Understanding hot spots. Technical report, National Institute of Justice, 2005.
El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accid. Anal. Prevention 41 (5), 1118–1123.
Erkut, E., Ingolfsson, A., Erdoğan, G., 2008. Ambulance location for maximum survival. Naval Research Logistics 55 (1), 42–58.
W.D. Fan, L. Gong, E.M. Washing, M. Yu, and E. Haile. Identifying and quantifying factors affecting vehicle crash severity at highway-rail grade crossings: models and their comparison. Technical report, 2016.
Fountas, G., Anastasopoulos, P.C., Abdel-Aty, M., 2018. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. Anal. Methods Accid. Res. 18, 57–68.
Frantzeskakis, J., Assimakopoulos, V., Kindinis, G., 1994. Interurban accident prediction by administrative area application in greece. Inst. Transp. Eng. J. 64 (1), 35–42.
Fu, K., Ji, T., Zhao, L., Titan, C.-T.Lu., 2019. A spatiotemporal feature learning framework for traffic incident duration prediction. In: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 329–338.
Gan, C., Wang, N., Yang, Y., Yeung, D.-Y., Hauptmann, A.G., 2015. Devnet: A deep event network for multimedia event detection and evidence recounting. In: Conference on Computer Vision and Pattern Recognition, pp. 2568–2577.
Gendreau, M., Laporte, G., Semet, F., 1997. Solving an ambulance location model by tabu search. Location Sci. 5 (2), 75–88.
Gendreau, M., Laporte, G., Semet, F., 2001. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. Parallel Comput. 27 (12), 1641–1653.
Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Introducing markov chain monte carlo. Markov Chain Monte Carlo in Practice 1, 19.
Goldstein, H., 1995. Multilevel Statistical Models. Institute of Education, Multilevel Models Project.
Grishman, R., Westbrook, D., Meyers, A., 2005. Nyu's english ace 2005 system description. In: ACE 2005 Evaluation Workshop.
Hadjidimitriou, N.S., Lippi, M., Dell-Amico, M., Skiera, A., 2020. Machine learning for severity classification of accidents involving powered two wheelers. IEEE Trans. Intell. Transp. Syst. 21 (10), 4308–4317.
Hall, J., Pendleton, O., 1990. Rural accident rate variations with traffic volume. Transp. Res. Rec. 1281, 62–70.
Haruna, L., Sallehuddin, R., Radzi, H.M., 2019. Discrete particle swarm optimization based filter feature selection technique for the severity of road traffic accident prediction. In: International Conference of Reliable Information and Communication Technology. Springer, pp. 298–310.
Hattis, S.H., 2015. Crime in the United States. Lanham Bernan Press.
E. Hauer. On the estimation of the expected number of accidents. Accid. Anal. Prevention, 18 (1): 1–12, 1986. ISSN 0001–4575.
E. Hauer. Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. Accid. Anal. Prevention, 24 (5): 457–477, 1992. ISSN 0001–4575.
Hauer, E., Persaud, B., 1983. Common bias in before-and-after accident comparisons and its elimination. Transp. Res. Rec. 905, 164–174.
E. Hauer, D.W. Harwood, F.M. Council, and M.S. Griffith. Estimating safety by the empirical Bayes method: a tutorial. Transp. Res. Record, 1784 (1): 126–131, 2002. ISSN 0361–1981.
Heydecker, B.G., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. Adv. Eng. Softw. 3210, 859–869.
Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., Zhu, Q., 2011. Using cross-entity inference to improve event extraction. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1127–1136.
Huang, H., Chin, H.C., 2010. Modeling road traffic crashes with zero-inflation and site-specific random effects. Stat. Methods Appl. 19 (3), 445–462.
Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis. Accid. Anal. Prevention 40 (1), 45–54.
Iqbal, Z., Khan, M.I., Hussain, S., Habib, A., 2021. An efficient traffic incident detection and classification framework by leveraging the efficacy of model stacking. Complexity 2021.
Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prevention 108, 27–36.
Islam, M., Mannering, F., 2020. A temporal analysis of driver-injury severities in crashes involving aggressive and non-aggressive driving. Anal. Methods Accid. Res. 27, 100128.
Jadaan, K.S., Nicholson, A.J., 1992. Relationships between road accidents and traffic flows in an urban network. Traffic Eng. Control 33 (9), 507–511.
Jafary, B., Rabiei, E., Diaconeasa, M., Masoomi, H., Fiondella, L., Mosleh, A., 2018. A survey on autonomous vehicles interactions with human and other vehicles. In: International Conference on Probabilistic Safety Assessment and Management.
Jaldell, H., 2017. How important is the time factor? saving lives using fire and rescue services. Fire Technol. 53 (2), 695–708.
Jaldell, H., Lebnak, P., Amornpetchsathaporn, A., 2014. Time is money, but how much? the monetary value of response time for thai ambulance emergency services. Value in Health 17 (5), 555–560.
H. Ji and R. Grishman. Refining event extraction through cross-document inference. In Annual Meeting of the Association for Computational Linguistics, pages 254–262, 2008.
Jia, H., Ordóñez, F., Dessouky, M., 2007. A modeling framework for facility location of medical services for large-scale emergencies. IIE Trans. 39 (1), 41–55.
Jiang, R., Qu, M., Chung, E., et al., 2014. Traffic incident clearance time and arrival time prediction based on hazard models. Math. Problems Eng. 2014.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and poisson regression models. Transp. Planning Technol. 15 (1), 41–58.
Jovanis, P.P., Chang, H.-L., 1986. Modeling the relationship of accidents to miles traveled. Transp. Res. Rec. 1068, 42–51.
Ke, J., Zhang, S., Chen, X.M., 2017. Missing information imputation for traffic incident likelihood prediction for urban expressways. Technical report.
Keneally, S.K., Robbins, M.J., Lunday, B.J., 2016. A markov decision process model for the optimal dispatch of military medical evacuation assets. Health Care Manage. Sci. 19 (2), 111–129.
Khasnabis, S., Lyoo, S.H., 1989. Use of time series analysis to forecast truck accidents. Transp. Res. Rec. 1249, 30–36.
Khattak, A.J., Kantor, P., Council, F.M., 1998. Role of adverse weather in key crash types on limited-access: roadways implications for advanced weather systems. Transp. Res. Rec. 1621 (1), 10–19.
Khazraee, S.H., Johnson, V., Lord, D., 2018. Bayesian Poisson hierarchical models for crash data analysis: Investigating the impact of model choice on site-specific predictions. Accid. Anal. Prevention 117, 181–195.
V. Kiattikomol. Freeway crash prediction models for long-range urban transportation planning. PhD thesis, University of Tennessee, Knoxville, 2005.
Knight, V.A., Harper, P.R., Smith, L., 2012. Ambulance allocation for maximal survival with heterogeneous outcome measures. Omega 40 (6), 918–926.
Kochenderfer, M.J., 2015. Decision Making Under Uncertainty: Theory and Application. MIT press.
Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prevention 34 (3), 313–321.
Kolesar, P., Walker, W.E., 1974. An algorithm for the dynamic relocation of fire companies. Operations Res. 22 (2), 249–274.
Krishnaveni, S., Hemalatha, M., 2011. A perspective analysis of traffic accident using data mining techniques. Int. J. Computer Appl. 23 (7), 40–48.
Kuo, P.F., Lord, D., 2020. Applying the colocation quotient index to crash severity analyses. Accid. Anal. Prevention 135, 105368.
Ladron de Guevara, F., Washington, S.P., Oh, J., 2004. Forecasting crashes at the planning level: Simultaneous negative binomial crash model applied in tucson, arizona. Transp. Res. Rec. 1897 (1), 191–199.
Larson, R.C., 1974. A hypercube queuing model for facility location and redistricting in urban emergency services. Computers Operations Res. 1 (1), 67–95.
Lee, Y., Wei, C.-H., 2010. A computerized feature selection method using genetic algorithms to forecast freeway accident duration times. Computer-Aided Civil Infrastructure Eng. 25 (2), 132–148.
M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang. Cross-media structured common space for multimedia event extraction. arXiv preprint arXiv: 2005.02472, 2020.
Li, Q., Ji, H., Huang, L., 2013. Joint event extraction via structured prediction with global features. In: Annual Meeting of the Association for Computational Linguistics, pp. 73–82.
Li, R., 2014. Traffic incident duration analysis and prediction models based on the survival analysis approach. IET Intel. Transport Syst. 9 (4), 351–358.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prevention 404, 1611–1618.
Li, X., Zhao, Z., Zhu, X., Wyatt, T., 2011. Covering models and optimization techniques for emergency response facility location and planning: a review. Math. Methods Operations Res. 74 (3), 281–310.
Litman, T., 2017. Autonomous vehicle implementation predictions. Technical Report.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transport. Res. Part A: Policy and Practice 445, 291–305.
Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: A bayesian perspective. Saf. Sci. 46 (5), 751–770.
Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accid. Anal. Prevention 37 (1), 35–46.
Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accid. Anal. Prevention 39 (1), 53–57.
Ma, J., Kockelman, K.M., 2006. Bayesian multivariate poisson regression for models of injury count, by severity. Transp. Res. Rec. 1950 (1), 24–34.
Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accid. Anal. Prevention 403, 964–975.
Ma, Z., Chang, X., Xu, Z., Sebe, N., Hauptmann, A.G., 2017. Joint attributes and event analysis for multimedia event detection. IEEE Trans. Neural Networks Learn. Syst. 29 (7), 2921–2930.
Ma, Z., Mei, G., Cuomo, S., 2021. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. Accid. Anal. Prevention 160, 106322.
MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Accid. Anal. Prevention 36 (6), 1019–1028.
Mafi, S., Abdelrazig, Y., Doczy, R., 2018. Machine learning methods to analyze injury severity of drivers from different age and gender groups. Transp. Res. Rec. 2672 (38), 171–183.
Maher, M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accid. Anal. Prevention 28 (3), 281–296.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.
Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Anal. Methods Accid. Res. 25, 100113.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods Accident Res. 11, 1–16.
Marcoux, R., Yasmin, S., Eluru, N., Rahman, M., 2018. Evaluating temporal variability of exogenous variable impacts over 25 years: An application of scaled generalized ordered logit model for driver injury severity. Analytic Methods Accident Res. 20, 15–29.
Marianov, V., Revelle, C., 1994. The queuing probabilistic location set covering problem and some extensions. Socio-Economic Planning Sci. 28 (3), 167–178.
Martin, J.-L., 2002. Relationship between crash rate and hourly traffic flow on interurban motorways. Accid. Anal. Prevention 34 (5), 619–629.
Maxwell, M.S., Henderson, S.G., Topaloglu, H., 2009. Ambulance redeployment: An approximate dynamic programming approach. In: Winter Simulation Conference, pp. 1850–1860.
Maxwell, M.S., Restrepo, M., Henderson, S.G., Topaloglu, H., 2010. Approximate dynamic programming for ambulance redeployment. INFORMS J. Computing 22 (2), 266–281.
Maxwell, M.S., Restrepo, M., Henderson, S.G., Topaloglu, H., 2010. Approximate dynamic programming for ambulance redeployment. INFORMS J. Computing 22 (2), 266–281.
McClosky, D., Surdeanu, M., Manning, C.D., 2011. Event extraction as dependency parsing. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1626–1635.
McCormack, R., Coates, G., 2015. A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. Eur. J. Oper. Res. 247 (1), 294–309.
D.R.D. McGuigan. An examination of relationships between road accidents and traffic flow. PhD thesis, Newcastle University, 1987.
Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transp. Res. Rec. 1840 (1), 31–40.
Miaou, S.-P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prevention 25 (6), 689–709.
Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accid. Anal. Prevention 37 (4), 699–720.
T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119, 2013.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prevention 40 (1), 260–266.
Moghaddam, F.R., Afandizadeh, S., Ziyadi, M., 2011. Prediction of accident severity using artificial neural networks. Int. J. Civil Eng. 9 (1), 41–48.
A. Mukhopadhyay. Robust Incident Prediction, Resource Allocation and Dynamic Dispatch. PhD thesis, Vanderbilt University, 2019.
Mukhopadhyay, A., Vorobeychik, Y., Dubey, A., Biswas, G., 2017. Prioritized allocation of emergency responders based on a continuous-time incident prediction model. In: International Conference on Autonomous Agents and MultiAgent Systems, pp. 168–177.
A. Mukhopadhyay, Z. Wang, and Y. Vorobeychik. A decision theoretic framework for emergency responder dispatch. In International Conference on Autonomous Agents and MultiAgent Systems, pages 588–596, 2018.
Mukhopadhyay, A., Pettet, G., Samal, C., Dubey, A., Vorobeychik, Y., 2019. An online decision-theoretic pipeline for responder dispatch. In: International Conference on Cyber-Physical Systems. ACM, pp. 185–196.
Nam, D., Mannering, F., 2000. An exploratory hazard-based analysis of highway incident duration. Transp. Res. Part A: Policy Practice 34 (2), 85–102.
B. Nambuusi, T. Brijs, and E. Hermans. A review of accident prediction models for road intersections. Technical report, Policy Research Centre for Traffic Safety, 2008.
National Emergency Number Association. 9–1-1 statistics. url:https://www.nena.org/page/911Statistics, 2019.
Nguyen, T.H., Grishman, R., 2018. Graph convolutional networks with argument-aware pooling for event detection. In: AAAI Conference on Artificial Intelligence.
Nguyen, T.H., Cho, K., Grishman, R., 2016. Joint event extraction via recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 300–309.
NPR. Emt vs paramedic. https://www.npr.org/sections/health-shots/2019/01/03/676039371/emergency-medical-responders-confront-racial-bias, 2019.
Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prevention 385, 936–948.
Park, E.S., 2019. Multivariate Poisson-Lognormal models for jointly modeling crash frequency by severity. Transport. Res. Rec.: J. Transport. Res. Board.
Park, B.-J., Lord, D., Hart, J.D., 2010. Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. Accid. Anal. Prevention 422, 741–749.
Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. Accid. Anal. Prevention 136, 105405.

Pei, X., Wong, S., Sze, N.-N., 2011. A joint-probability approach to crash prediction models. Accid. Anal. Prevention 43 (3), 1160–1166.
G. Pettet, S. Nannapaneni, B. Stadnick, A. Dubey, and G. Biswas. Incident analysis and prediction using clustering and bayesian network. In Conference on Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation, pages 1–8, 2017.
G. Pettet, A. Mukhopadhyay, M. Kochenderfer, and A. Dubey. Hierarchical planning for resource allocation in emergency response systems. arXiv preprint arXiv: 2012.13300, 2020.
Pettet, G., Mukhopadhyay, A., Kochenderfer, M., Vorobeychik, Y., Dubey, A., 2020. On algorithmic decision procedures in emergency response systems in smart and connected communities. In: International Conference on Autonomous Agents and Multiagent Systems, pp. 1046–1054.
Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. J. Transp. Eng. 122 (2), 105–113.
Private Communication. Nashville Fire Department, 2018.
Qi, Y., Smith, B.L., Guo, J., 2007. Freeway accident likelihood prediction using a panel data analysis approach. J. Transport. Eng. 1333, 149–156.
Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accid. Anal. Prevention 36 (2), 183–191.
Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accid. Anal. Prevention 40, 1486–1497.
Rajagopalan, H.K., Vergara, F.E., Saydam, C., Xiao, J., 2007. Developing effective meta-heuristics for a probabilistic location model via experimental design. Eur. J. Oper. Res. 177 (1), 83–101.
Rajagopalan, H.K., Saydam, C., Xiao, J., 2008. A multiperiod set covering location model for dynamic redeployment of ambulances. Computers Operations Res. 35 (3), 814–826.
H. Rakha, M. Arafeh, A. Abdel-Salam, F. Guo, and A. Flintsch. Linear regression crash prediction models: Issues and proposed solutions. Efficient Transportation and Pavement Systems: Characterization, Mechanisms, Simulation and Modeling, pages 241–256, 2010.
Ramani, G., Selvaraj, S., 2016. Learning through misclassified instances using pipelined voting algorithm for aggregated feature selection (p-vaafs) in the prediction of road accident severity. Adv. Natural Appl. Sci. 10 (16), 1–12.
Repede, J.F., Bernardo, J.J., 1994. Developing and validating a decision support system for locating emergency medical vehicles in Louisville. Kentucky. Eur. J. Operational Res. 75 (3), 567–581.
Revelle, C., Hogan, K., 1989. The maximum reliability location problem and α-reliable p-center problem: Derivatives of the probabilistic location set covering problem. Ann. Oper. Res. 18 (1), 155–173.
ReVelle, C., Hogan, K., 1989. The maximum availability location problem. Transp. Sci. 23 (3), 192–200.
Riviere, C., Lauret, P., Ramsamy Manicom, J.-F., 2006. A Bayesian neural network approach to estimating the energy equivalent speed. Accid. Anal. Prevention 382, 248–259.
Ryder, B., Dahlinger, A., Gahr, B., Zundritsch, P., Wortmann, F., Fleisch, E., 2019. Spatial prediction of traffic accidents with critical driving events – insights from a nationwide field study. Transp. Res. Part A: Policy Practice 124, 611–626.
Saeed, T.U., Hall, T., Baroud, H., Volovski, M.J., 2019. Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: An empirical assessment of multilane highways. Anal. Methods Accid. Res. 23, 100101.
Sasidharan, L., Wu, K.-F., Menendez, M., 2015. Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in switzerland. Accid. Anal. Prevention 85, 219–228.
Savolainen, P.T., Mannering, F.L., Lord, D., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accid. Anal. Prevention 435, 1666–1676.
Sayed, T., Rodriguez, F., 1999. Accident prediction models for urban unsignalized intersections in british columbia. Transp. Res. Record: J. Transp. Res. Board 1665, 93–99.
Schlüter, P.J., Deely, J.J., Nicholson, A.J., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. J. Royal Statistical Soc: Series D (The Statistician) 46 (3), 293–316.
K.K. Schuler. VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania, 2005.
Senarath, Y., Mukhopadhyay, A., Vazirizade, S., Purohit, H., Nannapaneni, S., Dubey, A., 2021. Practitioner-centric approach for early incident detection using crowdsourced data for emergency services. In: International Conference on Data Mining (to appear). IEEE.
Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prevention 27 (3), 371–389.
X. Shao, L.L. Boey, and Y. Luo. Traffic accident time series prediction model based on combination of arima and bp and svm. Journal of Traffic and Logistics Engineering Vol. 7 (2), 2019.
Shanthi, S., Ramani, R.G., 2012. Feature relevance analysis and classification of road traffic accident data through data mining techniques 1, 24–26.
Shi, Q., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C: Emerging Technol. 58, 380–394.
Shibata, A., Fukuda, K., 1994. Risk factors of fatality in motor vehicle traffic accidents. Accid. Anal. Prevention 26 (3), 391–397.
Shirazi, M., 2019. Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. Transportmetrica A: Transport Sci. 152, 1791–1803.
Silva, F., Serra, D., 2008. Locating emergency services with different priorities: the priority queuing covering location problem. J. Operational Res. Soc. 59 (9), 1229–1238.
Singh, D., Mohan, C.K., 2018. Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. IEEE Trans. Intell. Transp. Syst. 20 (3), 879–887.
Sivasankaran, S.K., Balasubramanian, V., 2020. Exploring the severity of bicycle–vehicle crashes using latent class clustering approach in india. J. Safety Res. 72, 127–138.
Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. J. Multivariate Anal. 97 (1), 246–273.
Songchitruksa, P., Balke, K., 1959. Assessing weather, environment, and loop data for real-time freeway incident prediction. Transp. Res. Record: J. Transp. Res. Board 105–113, 2006.
Sun, M., Sun, X., Shan, D., 2019. Pedestrian crash analysis with latent class clustering method. Accid. Anal. Prevention 124, 50–57.
Swoveland, C., Uyeno, D.H., Vertinsky, I., Vickson, R.G., 1973. A simulation-based methodology for optimization of ambulance service policies. Socio-Economic Planning Sciences 7 (6), 697–703.
Tajtehranifard, H., Bhaskar, A., Haque, M.M., Chung, E., 2016. Motorway crash duration and its determinants: do durations vary across motorways? J. Adv. Transp. 50 (5), 717–735.
T. Tambouratzis, D. Souliou, M. Chalikias, and A. Gregoriades. Maximising accuracy and efficiency of traffic accident prediction combining information mining with computational intelligence approaches and decision trees. Journal of Artificial Intelligence and Soft Computing Research, 4, 2014.
Tang, J., Liang, J., Han, C., Li, Z., Huang, H., 2019. Crash injury severity analysis using a two-layer stacking framework. Accid. Anal. Prevention 122, 226–238.
Toregas, C., Swain, R., ReVelle, C., Bergman, L., 1971. The location of emergency service facilities. Operations Res. 19 (6), 1363–1373.
Turner, D., Thomas, R., 1986. Motorway accidents: an examination of accident totals, rates and severity and their relationship with traffic flow. Traffic Eng. Control 27 (7–8), 377–383.
USA Today. Nashville bombing froze wireless communications, exposed 'achilles heel' in regional network. https://www.usatoday.com/story/news/nation/2020/12/29/nashville-bombing-area-communications-network-exposed-achilles-heel/4070797001/, 2020.
S.M. Vazirizade, A. Mukhopadhyay, G. Pettet, S.E. Said, H. Baroud, and A. Dubey. Learning incident prediction models over large geographical areas for emergency response systems. arXiv preprint arXiv:2106.08307, 2021.
Venkataraman, N., Shankar, V., Blum, J., Hariharan, B., Hong, J., 2016. Transferability analysis of heterogeneous overdispersion parameter negative binomial crash models. Transp. Res. Rec. 2583 (1), 99–109.
Venkataraman, N.S., Ulfarsson, G.F., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: exploratory insights from random parameter negative binomial approach. Transp. Res. Rec. 2236 (1), 41–48.
D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi. Entity, relation, and event extraction with contextualized span representations. arXiv preprint arXiv: 1909.03546, 2019.
Wang, C., Quddus, M.A., 2009. Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. Accid. Anal. Prevention 414, 798–808.
Wang, S., Li, Z., Zhang, J., Yuan, Y., Liu, Z., 2020. The crash injury severity prediction of traffic accident using an improved wrappers feature selection algorithm. Int. J. Crashworthiness 1–12.
Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and econometric methods for transportation data analysis. CRC Press, ISBN 9780429534225.
Wen, X., Xie, Y., Wu, L., Jiang, L., 2021. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with lightgbm and shap. Accid. Anal. Prevention 159, 106261.
Witten, I.H.I.H., 2016. In: Witten, Ian, Frank, Eibe, Hall, Mark A., Pal, Christopher J. (Eds.), Data mining [electronic resource]: practical machine learning tools and techniques/, 4th ed. edition,. Elsevier, Amsterdam.
Xiao, J., 2019. Svm and knn ensemble learning for traffic incident detection. Physica A 517, 29–35.
Xie, Y., Lord, D., 2007. Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. Accid. Anal. Prevention 395, 922–933.
Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. Transp. Res. Part B: Methodological 49, 39–54.
Xiong, Y., Tobias, J.L., Mannering, F.L., 2014. The analysis of vehicle crash injury-severity data: A markov switching approach with road-segment heterogeneity. Transp. Res. Part B: Methodological 67, 109–128.
Yannis, G., Dragomanovits, A., Laiou, A., La Torre, F., Domenichini, L., Richter, T., Ruhl, S., Graham, D., Karathodorou, N., 2017. Road traffic accident prediction modelling: A literature review. In: Proceedings of the Institution of Civil Engineers-Transport, volume 170. Thomas Telford Ltd, pp. 245–254.
Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Methods Accident Res. 1, 23–38.
Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Analytic Methods Accident Res. 1, 72–85.

Ye, Z., Xu, Y., Lord, D., 2018. Crash data modeling with a generalized estimator. Accid. Anal. Prevention 117, 340–345.
Yu, R., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prevention 51, 252–259.
Yu, R., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Sci. 63, 50–56.
Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prevention 66, 80–88.
Yu, R., Xiong, Y., Abdel-Aty, M., 2015. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. Transp. Res. Part C: Emerging Technol. 50, 68–77.
Zeng, Q., Huang, H., Pei, X., Wong, S., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accident Res. 10, 12–25.
Zeng, Q., Sun, J., Wen, H., 2017. Bayesian hierarchical modeling monthly crash counts on freeway segments with temporal correlation. J. Adv. Transp. 5391054.
Zhan, C., Gan, A., Hadi, M., 2011. Prediction of lane clearance time of freeway incidents using the m5p tree algorithm. IEEE Trans. Intell. Transp. Syst. 12 (4), 1549–1557.
Zhang, Y., Xie, Y., 2007. Forecasting of short-term freeway volume with v-support vector machines. Transp. Res. Rec. 2024 (1), 92–99.
Zheng, Z., Lu, P., Lantz, B., 2018. Commercial truck crash injury severity analysis using gradient boosting data mining model. J. Safety Res. 65, 115–124.
Zhu, L., Guo, F., Krishnan, R., Polak, J.W., 2018. The use of convolutional neural networks for traffic incident detection at a network level. Transp. Res. Board Annual Meeting.