Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 3179207, 16 pages
https://doi.org/10.1155/2018/3179207

*Research Article*

# Predicting Work Zone Collision Probabilities via Clustering: Application in Optimal Deployment of Highway Response Teams

**Przemysław Sekuła** [1,2] **Zachary Vander Laan** [1]
**Kaveh Farokhi Sadabadi**,[1] **and Mirosław J. Skibniewski**[1,3,4]

[1] *Department of Civil and Environmental Engineering, University of Maryland, College Park, MD, USA*
[2] *Faculty of Informatics and Communication, University of Economics in Katowice, Katowice, Poland*
[3] *Institute for Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, Poland*
[4] *Chaoyang University of Technology, Taichung, Taiwan*

Correspondence should be addressed to Przemysław Sekuła; psekula@umd.edu

Received 28 February 2018; Revised 25 July 2018; Accepted 30 August 2018; Published 18 September 2018

Academic Editor: Richard S. Tay

This paper proposes a clustering approach to predict the probability of a collision occurring in the proximity of planned road maintenance operations (i.e., work zones). The proposed method is applied to over 54,000 short-term work zones in the state of Maryland and demonstrates an ability to predict work zone collision probabilities. One of the key applications of this work is using the predicted probabilities at the operational level to help allocate highway response teams. To this end, a two-stage stochastic program is used to locate response vehicles on the Maryland highway network in order to minimize expected response times.

## 1. Introduction

Work zone collisions accounted for approximately 1.2 to 1.7 percent of all 2006-2013 crashes in the United States, amounting to 67,523 incidents in 2013 [1]. These collisions resulted in numerous injuries (e.g., over 47,000 in 2013), as well as fatalities that occurred in about 1 percent of the cases. Mohan and Gautam [2] performed a thorough analysis of work zone accidents in the US between 1995 and 1997 and estimated the direct costs resulting from worker and passenger injuries to be approximately $6.2 billion, which composes only part of the overall cost. Although this value may be somewhat lower now given that annual work zone collisions have mirrored the gradual downward trend of overall collisions [1], there are still enormous safety and financial benefits that can be achieved by avoiding or lessening the severity of such accidents. Here we consider two ways in which transportation agencies may mitigate the impact of work zone collisions:

(i) *Optimally allocating response teams.* Transportation agencies could allocate response vehicles in accordance with anticipated risks at work zones and thereby reduce the time it takes to reach the scene of an accident. This would improve their ability to assist injured motorists, clear blocked traffic lanes, and help prevent additional secondary accidents.

(ii) *Adjusting work zone characteristics.* Transportation agencies could identify risky work zone projects in the planning phase and adjust work zone characteristics in order to improve their safety and operational performance.

Both of these applications depend on the ability to *quantify how work zone characteristics affect collision occurrence risk in work zones*, which is precisely the focus of this paper. We briefly describe how this fits into the broader context of work zone research and focus in particular on two types of approaches which may be employed to model this relationship: statistical modeling and machine learning.

Existing work zone research tends to be oriented towards mobility or safety applications. The mobility-oriented research generally focuses on the effect of work zones on traffic capacity, queuing delays, and other performance metrics [3–8], as well as optimizing controllable work zone parameters [9–14]. Research focused on safety commonly describes how work zones affect collision frequency [15–18], severity [19–23], or both [24–28]. In particular, this paper examines how work zone characteristics (e.g., duration, time of day, and number of lanes) impact the likelihood of work zone collisions occurring. Here we review the two prevailing methodologies used to quantify this relationship, considering each approach separately.

(i) *Statistical Modeling*. The majority of literature in this area uses regression techniques to model the relationship between explanatory variables describing work zones and the frequency with which incidents occur, employing traditional maximum likelihood and Bayesian methods to estimate the parameters. However, least squares regression is not ideal for representing the relationship between work zone characteristics and crash frequency, because it assumes normally distributed errors and the dependent variable (i.e., number of crashes) is not continuous [29]. While isolated examples of linear regression models exist in the literature (e.g., [30]), researchers have found that discrete models with Poisson-distributed responses are better suited for modeling crash frequency. Accordingly, the Poisson and related models (e.g., Negative Binomial (NB), Poisson-lognormal, and zero-inflated Poisson and NB) are more prevalent in the literature [29]. The Poisson model's primary deficiency is that it has poor estimation performance when the variance is greater than the mean (i.e., overdispersion), which the NB model addresses by specifying a gamma distribution for the error term [29]. There are numerous applications of the NB model [31–36], focusing on different aspects of crash frequency modeling (e.g., location within a work zone, time of day, duration, fatal/injury/property damage crash classifications). For example, Venugopal and Tarko [32] estimate work zone crash frequency based on traffic levels, work zone lengths, work duration, on/off ramps, and classification of work zones, observing that there is not a statistically significant difference in crash rate for day and nighttime work zones. From a slightly different perspective, Srinivasan et al. [33] use the NB models to predict crash modification factors, while Yang et al. [37] incorporate concepts of measurement error in the NB modeling framework. However, NB models are susceptible to underdispersion for small datasets, which Poisson-lognormal models seek to counter [24, 25, 38]. Finally, zero-inflated models attempt to address the fact that work zone datasets do not contain many incidents (i.e., the response variable primarily takes the value of zero), which can cause overdispersion. The zero-inflated approach generally uses two states: one where accidents can occur and the other where they cannot. Binary choice models (e.g., binary probit or logit) are often used to determine which system the state is in, and Poisson and NB approaches are used to model crash frequency for the state in which accidents are allowed [39–43].

Another class of advanced statistical techniques includes random parameter [44–47], generalized additive [48, 49], Markov switching [50–52], and hierarchical models [53, 54]. We do not spend extensive time discussing these methods here, primarily because existing literature in this area focuses on non-work zone crash frequency modeling applications, which is not the focus of this paper.

The models described thus far are primarily useful for predicting the number of crashes that will occur in a work zone over a period of time. As Yang, Ozbay, Xie, and Bartin [55] point out that sometimes it is useful to frame the problem in terms of crash probability (also referred to as crash risk) instead, where the goal is to predict the probability that a work zone will encounter a collision based on its characteristics. In this case, the response variable is binary (i.e., accident or no accident) rather than a positive integer (i.e., number of collisions). Logistic regression is a popular approach for handling binary responses, examples of which include Al-Ghamdi [56], Harb et al. [57], and Bham et al. [58]. However, when modeling crash risk for short work zone durations, the estimates are often biased because there are rarely crashes during the specified time period [55]. Being able to estimate crash probabilities for short durations is one of the core objectives of this paper and has not been studied extensively in the literature. To the best of our knowledge, the only paper which addresses this problem for short durations is Yang, Ozbay, Xie, and Bartin [55], which implements a rare event logistic regression model to account for the biases. They illustrate this approach by considering 466 work zones along a 25-mile section of the New Jersey Turnpike.

(ii) *Machine Learning*. An alternative approach to model relationships between work zone characteristics and incident risks is through machine learning techniques, including neural and Bayesian neural networks, support vector machines, and clustering. However, existing work zone research employing these techniques tends to focus on characterizing work zone capacity, delays, incident duration, and other metrics rather than quantifying crash risks (e.g., [59–64]). For example, Jiang and Adeli [61] use neural network models to assess work zone capacity, concluding that they can estimate the WZ capacity with less than 10% error with this approach. In terms of work zone safety, the literature deals primarily with crash counts (not crash probability) for generic roadway incidents that do not consider the specific characteristics of work zones. Examples of this are Xie et al. [65] and Abdulhafedh [66], both of which compare neural network approaches to traditional NB models, noting that the neural networks show superior predictive capability in their case studies.

Li et al. [67] perform a similar comparison using support vector machines (SVM) and find that the SVM approach better estimates crash frequency in comparison to an NB model. They also point out that although neural networks and SVMs provide similar prediction capabilities, the SVM approach does not overfit the data. From a clustering approach, Ma and Kockelman [20] group road segments with similar characteristics and use a linear regression model to estimate crash frequency within each cluster. Other clustering applications include roadway crashes [68, 69] and safety projects [70], but to our knowledge current literature does not include papers which cluster work zones based their characteristics in order to infer crash risk.

Despite many different modeling approaches and perspectives, the majority of literature focuses on either work zones or crash frequency independently, with comparatively little research dedicated to modeling collision frequency/risk based on work zone characteristics. Within this niche body of research, most papers conclude that increased vehicle demand, work zone length, and duration all increase the number of collisions that occur at a work zone over a period of time [71]. However, we are more interested in determining the probability of a collision rather than the number of collisions that occur. This poses problems when considering short time durations with classical statistical approaches due to the fact that accidents occur so infrequently. The one paper that addresses this issue (i.e., the rare event logistic regression approach by Yang, Ozbay, Xie, and Bartin [55]) showed good performance in a case study, although the model was tested on a single corridor with less than 500 work zones. While machine learning models appear promising based on closely related studies, none of them have been applied specifically to estimating collision probabilities based on work zone characteristics. Accordingly, we focus on this underresearched area and present a clustering approach to estimate the probabilities of work zone collisions, which is useful for determining where to best locate response teams and identifying work zone characteristics that can be modified to reduce collision probabilities. We make the following contributions:

(i) We propose a scalable, unsupervised learning approach to predict the probability of a collision occurring at a short-term work zone. In contrast to most classical regression approaches which do not explicitly calculate crash risk/probability and are not appropriate for short-term durations, this machine learning approach clusters all work zones based on salient characteristics, calculates a collision probability for each cluster, and assigns new work zones to existing ones with similar features. Furthermore, it yields collision predictions without having access to current traffic volumes or work zone lengths and would easily scale to larger work zone datasets (e.g., see [72] for a clustering application involving 1 billion data points).

(ii) We present an integrated approach which combines both work zone collision risk predictions and actionable response recommendations, which is intuitive and applicable for practitioners. This approach takes historic information about work zones and returns the optimal allocation of highway response teams, which could be readily implemented by an agency such as the Coordinated Highways Action Response Team [73], an example of which is illustrated in the case study using the previously described approach and a stochastic optimization model.

The remainder of this paper is organized in the following manner. We begin by discussing the historical work zone and collision dataset that is used for model development in subsequent sections. Next, we describe clustering methods and explain how we determine the number of clusters and quantify model performance. We then apply this clustering framework to the work zone data set, noting how different approaches affect the model performance. Afterwards, we discuss example applications, focusing on optimally locating highway response vehicles by using the work zone incident probabilities as an input to a stochastic optimization model. Finally, we draw conclusions and suggest future steps to extend the research.

## 2. Materials and Methods

*2.1. Data.* Data describing work zones (WZs) and collisions that take place in their proximity (i.e., within 1 mile distance) were collected from the Regional Integrated Transportation Information System [77] and Coordinated Highways Action Response Team [73], whose joint objective is to improve operations of Maryland's highway system. We focus on those WZs that were set up and cleared in less than one day, which account for about 99% of all the WZs observed during 2010-2015. The resulting dataset considered in our analysis includes 54,463 WZs (Figure 1) and 380 WZ collisions. In most cases, a WZ had one or two lanes closed, and those closures took place along either the main or shoulder lanes (Figure 2). Furthermore, maintenance work lasted for 6 hours on average (Figure 3). In order to minimize negative effects on mobility and safety, only about 7% of the maintenance work was conducted during the peak-periods (7-9 AM and 4-6 PM). It is worth noting that about 72% of the work was done in daylight (Figure 3) and that WZs were typically set up on weekdays (Figure 4)

The observed 380 WZ collisions occurred at a very small subset of all the recorded WZs. There were hardly any WZs with more than one collision; however, a few observed as many as three collisions (Figure 5). The highest number of WZ collisions took place on Thursdays, while far fewer occurred on Saturdays and Sundays (Figure 6), which is unsurprising due to lower maintenance activity over the weekends (Figure 4). The seasonal distribution of WZ collisions (Figure 6) implies that most took place in summer and fall, which is expected due to higher maintenance activities during these seasons (Figure 4). Additionally, Figure 7 shows a visual representation of Average Annual

Table 1: Input variables considered in the analysis of WZ collisions.

(1) Day of the week when the WZ was opened (nominal)
(2) Season of the year when the WZ was operational (nominal)
(3) Road class (nominal: MD, I, US)
(4) Duration of the WZ during the peak hours (continuous)
(5) Duration of the WZ during the off-peak hours (continuous)
(6) Duration of the WZ in daylight (continuous)
(7) Duration of the WZ at night (continuous)
(8) Total number of lanes (ordinal)
(9) Number of main lanes closed (ordinal)
(10) Number of shoulder lanes closed (ordinal)
(11) Number of median lanes closed (ordinal)
(12) Average annual daily traffic (AADT) (ordinal)

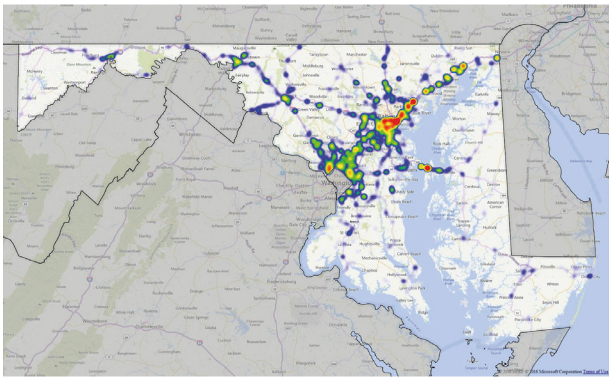


Figure 1: Heat map showing locations of over 54 thousand WZs that took place in Maryland during 2010-2015. The pie chart indicates that majority of WZs were setup along MD and Interstate roads.

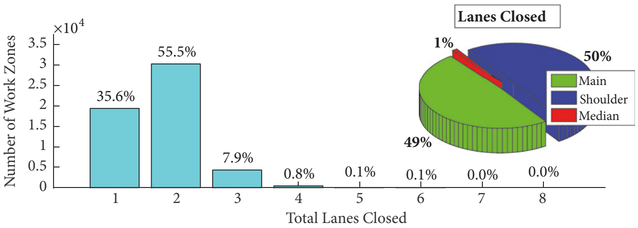


Figure 2: Most WZs would have one or two lanes closed, while closures of more than four lanes were very rare and typically occurred close to toll plazas. For the most part, closures affected main and shoulder lanes.

Daily Traffic (AADT) on the Maryland road network in 2013, indicating that a wide range of AADT values are observed, and highlighting the spatial distribution of high-volume roads. Based on this descriptive analysis of available data, we selected twelve input variables (Table 1) that might affect safety levels in the proximity of a WZ and thereby influence occurrence of a collision. The correlation between a subset of these variables and the number of WZ collisions is provided in Figure 8 and indicates relatively weak correlation between the WZ collisions and other variables.

*2.1.1. Data Processing.* Data preprocessing can significantly improve performance of the clustering algorithm presented in the following section. Categorical variables were modeled using their binary counterparts, and the peak and day duration of a WZ were divided by its overall duration because we are interested in predicting the one-hour probability of a collision. The off-peak and night durations were removed from the model as redundant information (e.g., relative peak and relative off-peak add up to one), and all the variables were normalized in order to enhance performance of the clustering approach presented in the next section. The WZ features used for clustering are presented in Table 2.

Note that WZ lengths were not available for the 54,000 WZs considered in this study, but since WZ lengths have been shown to influence crash frequency [71], we expect that including this variable in the future would improve the model accuracy. Additionally, a variety of other temporal variables would likely enhance model performance (e.g., speed variance and weather), but these characteristics would be unavailable ahead of time, and the proposed approach focuses on predicting the crash probability at a future work zone. Along these lines, it may be worth considering using hourly volume profiles [78] instead of AADT, as it would allow the model to account for the average hourly volumes during the specific work hours (e.g., 7 AM to 9 AM). Of course, like the AADT, the volume profiles would only help give a rough idea about the future volumes, because these volumes would be affected by the presence of the work zone.

*2.2. Methodology.* The objective is to use historical data in order to predict the probability of a collision occurring within

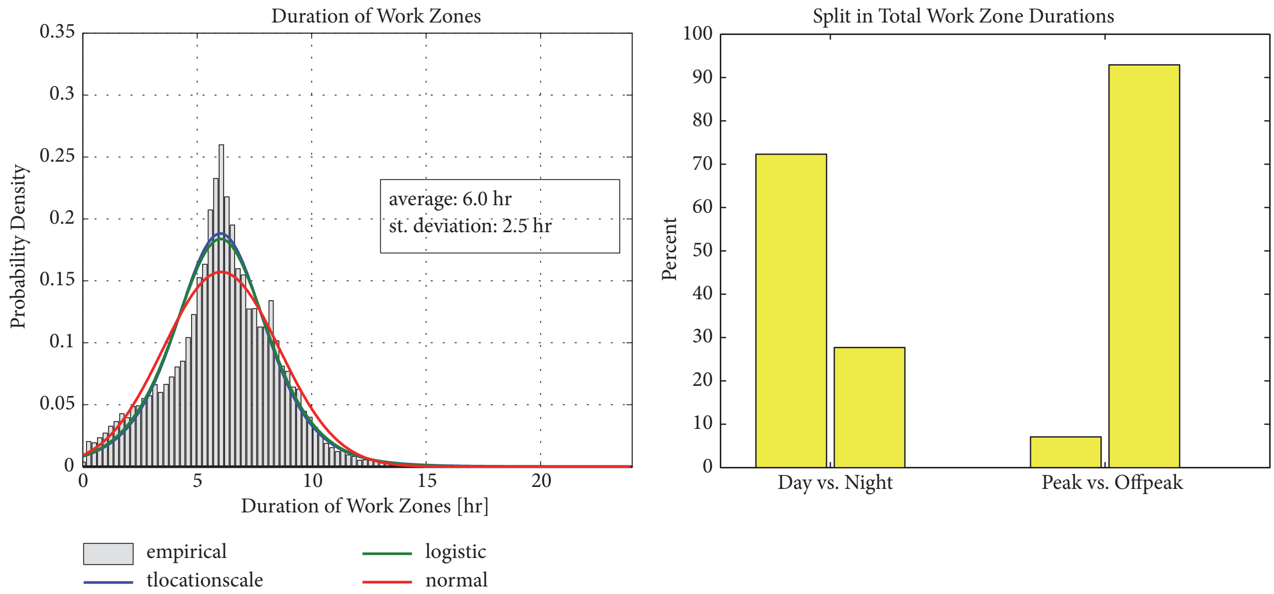


FIGURE 3: Duration of WZs and split in total work zone hours in day/night and peak/off-peak. Seventeen continuous distributions were fitted to WZ duration [74], and the probability density functions of those distributions that best fit in terms of Bayesian information criterion are shown in the figure on the left. Sunrise and sunset times for the dates of maintenance work and geographic locations of WZs are computed using MATLAB toolbox from SCRIPPS Institute of Oceanography [75].

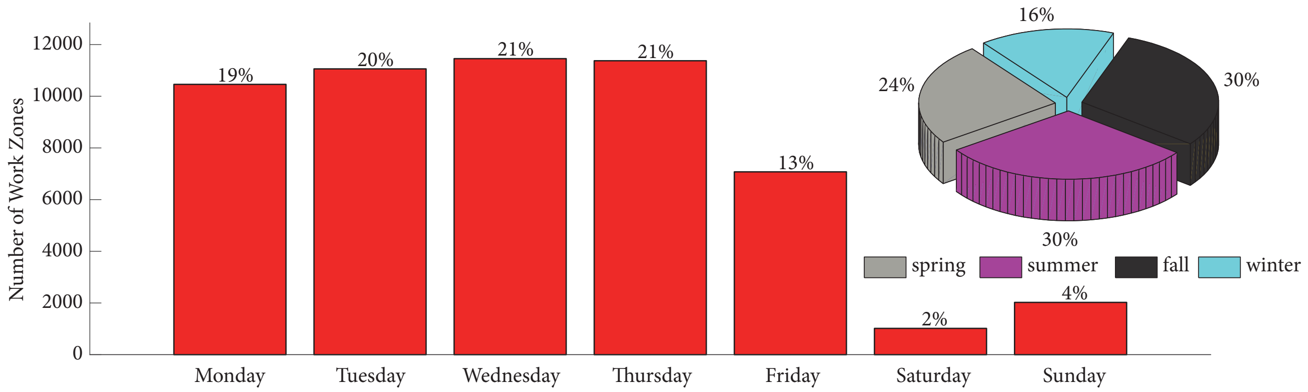


FIGURE 4: Most maintenance work was carried out on weekdays and during summer and fall.

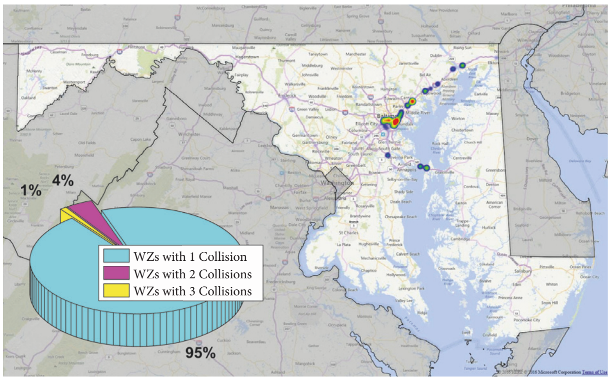


FIGURE 5: Heat map showing locations of 359 WZs within which 380 collisions occurred. The pie chart indicates that very few WZs had more than one collision.

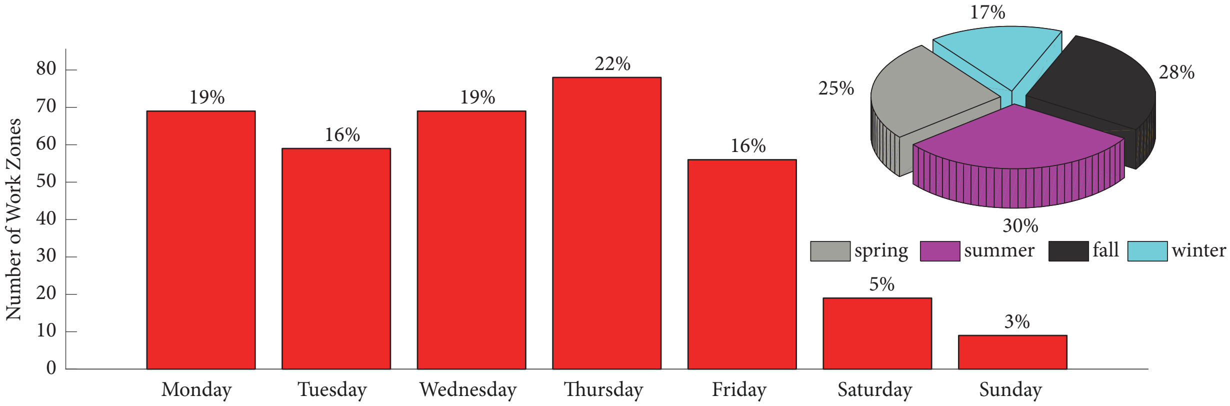


Figure 6: Most collisions took place on weekdays and during summer and fall, when high maintenance activity was observed.

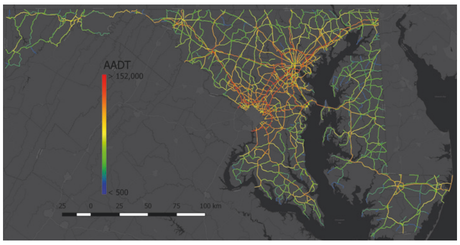
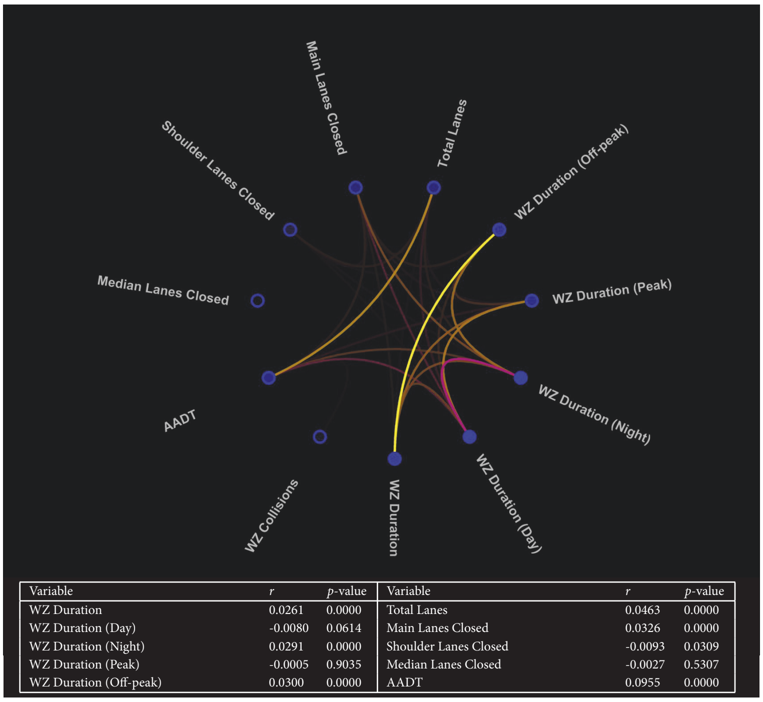


Figure 7: Visual representation of AADT in 2013 for the road links in Maryland.

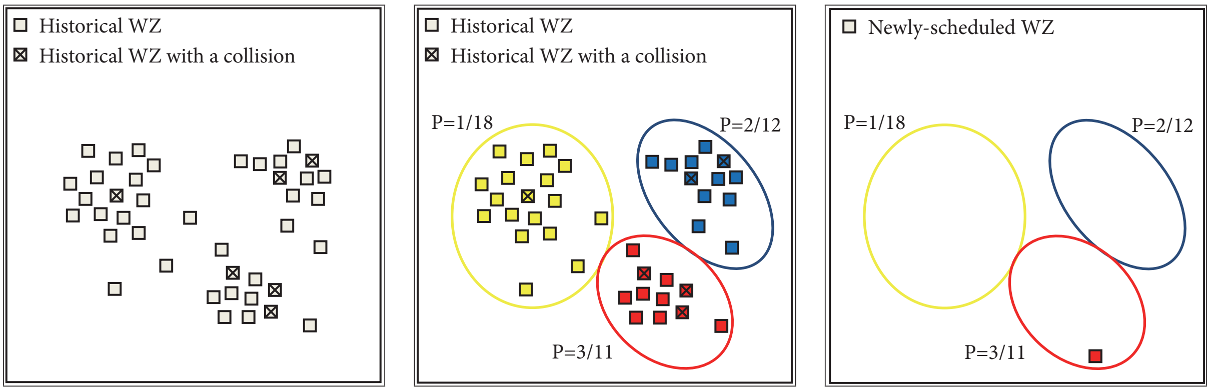


| Variable | *r* | *p*-value | Variable | *r* | *p*-value |
|---|---|---|---|---|---|
| WZ Duration | 0.0261 | 0.0000 | Total Lanes | 0.0463 | 0.0000 |
| WZ Duration (Day) | -0.0080 | 0.0614 | Main Lanes Closed | 0.0326 | 0.0000 |
| WZ Duration (Night) | 0.0291 | 0.0000 | Shoulder Lanes Closed | -0.0093 | 0.0309 |
| WZ Duration (Peak) | -0.0005 | 0.9035 | Median Lanes Closed | -0.0027 | 0.5307 |
| WZ Duration (Off-peak) | 0.0300 | 0.0000 | AADT | 0.0955 | 0.0000 |

Figure 8: Numeric values indicate correlation between (continuous and ordinal) input variables and number of WZ collisions. Yellow and magenta denote positive and negative correlation, respectively. Brighter links indicate greater correlation, while less visible links imply low correlation. Unfilled tokens indicate that a variable has low total absolute correlation with other covariates [76].

Table 2: Input variables used for WZ clustering.

| Variable | Description | Columns |
|---|---|---|
| (1) Season | Season of year | 3 |
| (2) Road type | Type of road (I, US, MD) | 3 |
| (3) Weekend | Equals 1 for Saturday or Sunday, otherwise 0 | 1 |
| (4) Lanes | The number of all, main, shoulder and median lanes | 1 |
| (5) AADT | Average annual daily traffic | 1 |
| (6) Peak | Peak duration / overall duration | 1 |
| (7) Day | Day duration / overall duration | 1 |



(a) Historical (unclustered) work zones

(b) Clustered WZs and corresponding collision probabilities

(c) Newly scheduled WZ is attributed to the closest cluster

Figure 9: Graphical description of the methodology via a trivial example including 2 dimensions, 3 clusters, and 41 data points (i.e., WZs). After clustering historical WZs, for each cluster, we compute the number of WZs with at least one collision over the total number of WZs within the cluster. A newly scheduled WZ is attributed to the cluster with the most similar features. The predicted probability of a collision occurring at this WZ corresponds to the ratio computed for the cluster it was attributed to.

a future WZ (e.g., road maintenance work scheduled for the next week), based on the underlying assumption that WZs with similar features should have comparable safety levels. Accordingly, we take the set of 54,000 WZs observed in the past (Section 2.1) and partition it into clusters of WZs with similar features (Table 2). After grouping WZs with similar characteristics, we calculate the collision probability for each cluster by counting the number of WZs where at least one collision occurred and dividing it by the total number of WZs within that cluster. Thus, the probability of a collision occurring at a future WZ can be predicted by assigning it to the closest cluster (in terms of its features) and looking up that cluster' observed collision probability, a process that is illustrated in Figure 9. The relation between predicted overall collision probability denoted by $P$ and one-hour predicted collision probability delivered by the model (i.e., $P_h$) is computed as

$$P = 1 - \left(1 - P_h\right)^D, \tag{1}$$

where $D$ is duration of the observed WZ.

The historical data can be partitioned into clusters of WZs with similar features via classical $k$-means clustering [79, 80]. Given a set of observations $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, where observation $\mathbf{x}_i$ is a $d$-dimensional vector of variables describing $i$th WZ, the $k$-means clustering seeks to partition these observations into $k$ sets $\mathbf{S} = \{S_1, \dots, S_k\}$ so as to minimize the within-cluster sum of squares. More formally, WZs are clustered by solving an optimization problem

$$\underset{\mathbf{S}}{\arg\min} \quad \sum_{l=1}^{k} \sum_{\mathbf{x} \in S_l} \left\| \mathbf{x} - \boldsymbol{\mu}_l \right\|^2, \tag{2}$$

where $\boldsymbol{\mu}_l$ is the mean of points in $S_l$. Since finding an exact solution to the above problem is NP-hard, many efficient heuristic techniques have been devised over the past 50 years. We apply $k$-means++ algorithm [81], which employs a randomized seeding technique to select initial cluster centers in a way that reduces the chance of the algorithm getting stuck in a suboptimal solution (Algorithm 1). However, even the $k$-means++ algorithm is not guaranteed to find an optimal solution in instances that include many dimensions and data points. To alleviate this problem, we perform clustering 100 times using different initial cluster centers selected by the $k$-means++ technique and choose the one with the smallest cost function for further analysis.

To evaluate the model, the dataset is first separated into training and testing datasets. Upon selecting the number of clusters (discussed below), the $k$-means++ algorithm is used to cluster the training data into $k$ discrete groups. Each point in the test dataset is then associated with its most similar cluster from the training dataset in order to compute the

```
**Procedure** *k-means++*
  Select centers for $k$ clusters by using randomized seeding technique
  **repeat**
    Assign each point to the closest cluster center
    Compute new cluster centers using assigned points
  **until** *Convergence (cluster centers did not change during the last step)*
```

Algorithm 1: Outline of the $k$-means++.

forecasted probability of a collision. These points are then ordered from lowest to highest probability (i.e., decreasing in safety) and separated into a specified number of groups, referred to as quantiles. For quantile $i$, the mean of the forecasted probability can be calculated (denoted $F_i$), which can be compared to the actual probability of collision (found by computing the ratio of work zones from quantile $i$ which had at least one collision to the number of work zones in $i$, denoted $A_i$). More specifically, the symmetric mean absolute percentage error ($SMAPE$) is used to quantify the error between the actual and forecasted probabilities across all quantiles, which is computed as

$$SMAPE = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \frac{|F_i - A_i|}{|F_i| + |A_i|}, \quad (3)$$

where $\tilde{n}$ is the number of quantiles.

In order to determine the number of clusters to use, we employ three different methods: Elbow, Silhouette, and Cross-Validation. In the Elbow method, the number of clusters is increased until the improvement in the objective function becomes marginal. The Silhouette method [82] requires computing each point's similarity to both its own cluster and the others. This measure is computed as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad (4)$$

where $a_i$ represents the average dissimilarity of data point $i$ with all other data within the same cluster, and $b_i$ is the lowest average dissimilarity of the data point $i$ to any cluster of which $i$ is not a member. A value of $s_i$ close to 1 implies that the point was appropriately clustered; values close to 0 mean that the points are on the border of two clusters, whereas negative values imply misclustering. The third approach is based on cross-validation, where the training dataset is divided into training and cross-validation sets. The model is then trained using the new, reduced training dataset, and evaluated by using the cross-validation dataset. After selecting the number of clusters, the model is retrained using the entire (undivided) training set.

Additionally, these three methods are helpful in determining the relevant model scenarios. In the base scenario, we assume that all the variables are of equal importance and thus are normalized from 0 to 1. To explore the influence of individual variables, we inflate their normalized values and study how this affects the clustering results. Thus, the Elbow, Silhouette, and Cross-Validation methods help provide insight into the number of clusters and appropriate variables to use.

## 3. Results and Discussion

The proposed clustering approach may overfit the data (i.e., memorize data points rather than detect patterns). In order to check for possible overfitting, it is helpful to train models with various sizes of training data and then check the behavior of the test error. When overfitting occurs, the error should significantly decrease as the training dataset size increases, but if the model is able to generalize well (i.e., does not overfit), further changes in training dataset size should not result in a meaningful reduction of test error. The results for the base scenario are presented in Figure 10 and imply that the model does not overfit the data for training sets greater than 50% of the entire dataset. Thus, the employed clustering approach may be used to predict WZ collision probabilities.

*3.1. Number of Clusters.* Selecting the number of clusters is essential for obtaining satisfactory results. After some preliminary experiments, the lower and upper bound on the number of clusters were set to 8 and 21, respectively. It should be noted that if a model includes few clusters, the WZs in each cluster may be quite diverse. This results in relatively small differences among mean collision probabilities in each cluster. Consequently, even if the accuracy of clustering is very high, the model cannot be employed for predictions. On the other hand, having many clusters should (in theory) improve the prediction accuracy. However, if some clusters do not include enough data points with collisions (due to relatively few collisions in the entire dataset), then the estimated collision probabilities may be inaccurate for these clusters. Therefore the upper bound on the number of clusters should correspond to the size of the training set and the number of collisions in the entire dataset.

As argued before, the authors considered three methods to select the number of clusters: Elbow, Silhouette, and Cross-Validation. The accuracy of each was estimated using a relative difference between the error of the model indicated by the method and error of the best model selected using a test set. The results for each method were computed for different scenarios and for different training dataset sizes (due to the overfitting analysis only the results for training dataset sizes greater or equal of 50% of entire dataset size were taken into consideration), which are summarized in Table 3.

Table 3: Aggregated accuracy metrics of various methods for selecting the best model (smaller numbers indicate better accuracy).

| | Silhouette | Cross-Validation | Elbow |
|---|---|---|---|
| Mean | 0.36365 | 0.40575 | 0.44222 |
| Std. dev. | 0.28405 | 0.42113 | 0.43249 |

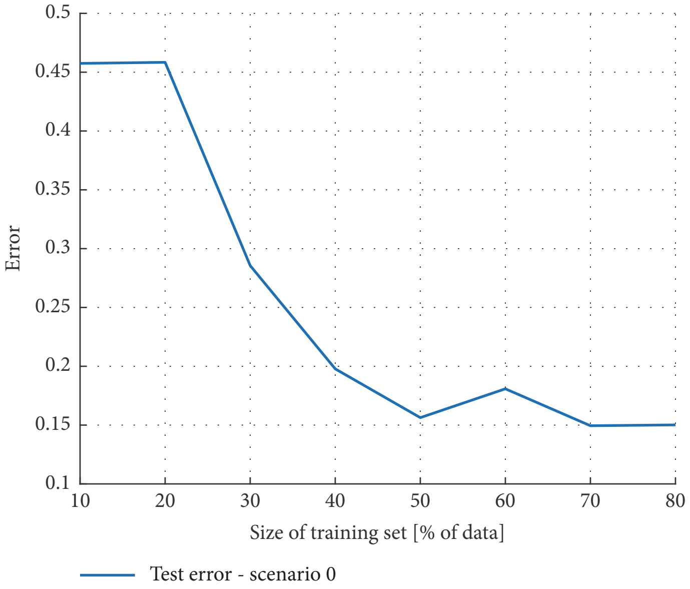


Figure 10: Predicted and actual mean collision probability for the base scenario.

The results indicate that the Silhouette method has the best accuracy and smallest standard deviation amongst the three measures, meaning that it is less prone to changes of scenario or training dataset size and, consequently, more reliable. Accordingly, the Silhouette method was chosen to determine the number of clusters.

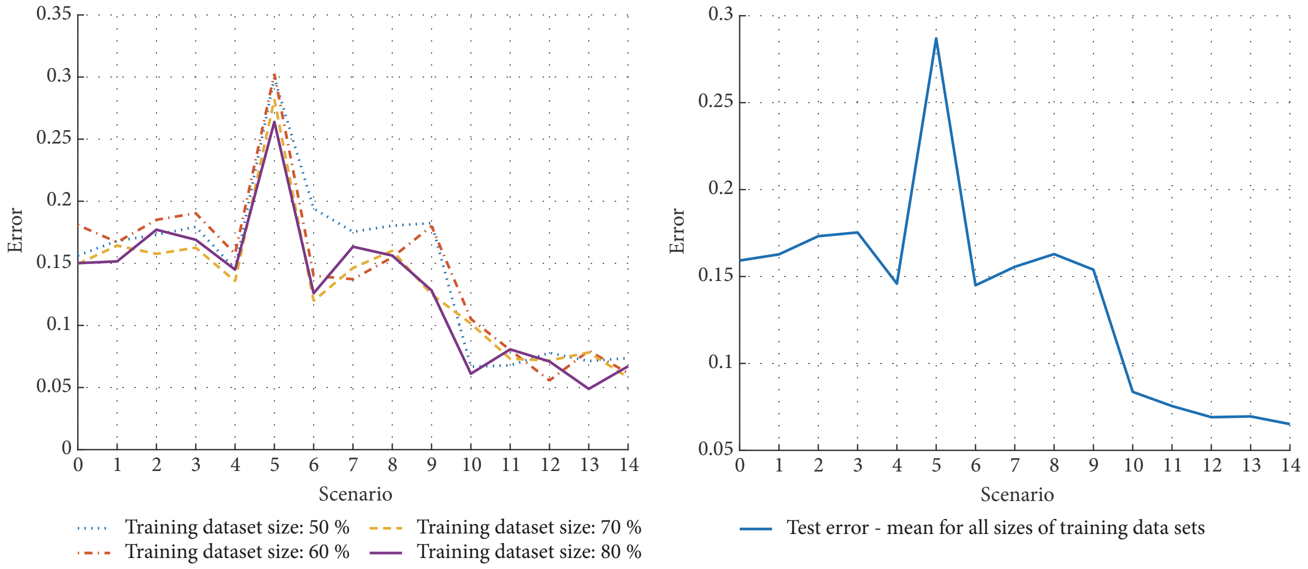
*3.2. Choosing a Scenario.* The base scenario assumes that all the features are equal, but in actuality some may have a larger impact on the predicted values and others. In order to verify this, different scenarios were created and tested. In each scenario some features are more (or less) important than others, and the proximity in dimensions corresponding to these features has a greater impact on the attribution of WZs to certain clusters. A list of tested scenarios is presented in Table 4, while errors associated with the performance of the best model in each scenario are shown in Figure 11.

Scenarios 1-7 imply that AADT can deteriorate model accuracy. Specifically, increasing the importance of AADT inflates errors regardless of the data size. Conversely, scenarios 4, 6, and 7, which correspond to lanes, day/night, and peak/off-peak, are characterized by relatively smaller errors. Scenarios 8 and 9 and to some extent 11 show that AADT has no significant impact on model accuracy. In fact, reducing the significance of AADT or completely removing it does not affect the accuracy. This is unsurprising because AADT represents an average annual measure, which may be a poor indicator of traffic volumes during particular hours of road maintenance work. Moreover, scenarios 10-12 indicate that WZ hours of operations are highly relevant, as models that emphasize day/night and peak/off-peak values provide more accurate predictions. Finally, scenarios 13 and 14 show that the number of lanes, day/night, and peak/off-peak data have a crucial impact on prediction accuracy. However, performance of all these models is comparable, so we use the Silhouette method to select both the number of clusters and the best scenario.

*3.3. The Best Model.* The error for the model whose specifications are determined using the Silhouette method for 3 quantiles is 2.95%, indicating a high level of model accuracy. The accuracy for each quantile is shown in Figure 12(a), while the *SMAPE* errors based on the number of quantiles for the selected model are shown in Figure 12(b). While the number of quantiles (the resolution of the model) is increasing, the *SMAPE* errors also increase, meaning accuracy of the model is reduced. It is worth noticing that for 4, 5, 6 and 7 quantiles the *SMAPE* errors are almost constant (from 9.84% to 10.88% with 10.5% mean), thus implying that if 11% error is acceptable, it is possible to increase the resolution of the model up to 7 quantiles. When the resolution of the model increases above 7 quantiles, the error grows to an unacceptable level as a result of an insufficient amount of training data. Since the predicted probability of collision in each quantile is computed by dividing the number of work zones with a collision occurrence by the total number of WZs in the quantile, when there are not enough WZs in a quantile, even a single collision can significantly change predictions.

Table 4: List of tested scenarios.

| Scenario | Description |
|---|---|
| 0 | Base scenario, all features equally important |
| 1 | Importance of seasons increased |
| 2 | Importance of road type increased |
| 3 | Importance of weekday/weekend increased |
| 4 | Importance of number of lanes increased |
| 5 | Importance of AADT increased |
| 6 | Importance of peak/off-peak increased |
| 7 | Importance of day/night increased |
| 8 | Importance of AADT decreased |
| 9 | AADT data removed |
| 10 | Importance of peak/off-peak and day/night increased |
| 11 | Importance of peak/off-peak and day/night increased, AADT data removed |
| 12 | Uses only peak/off-peak and day/night data |
| 13 | Importance of number of lanes, peak/off-peak and day/night increased |
| 14 | Uses only number of lanes, peak/off-peak and day/night data |



(a) Errors for different sizes of training data

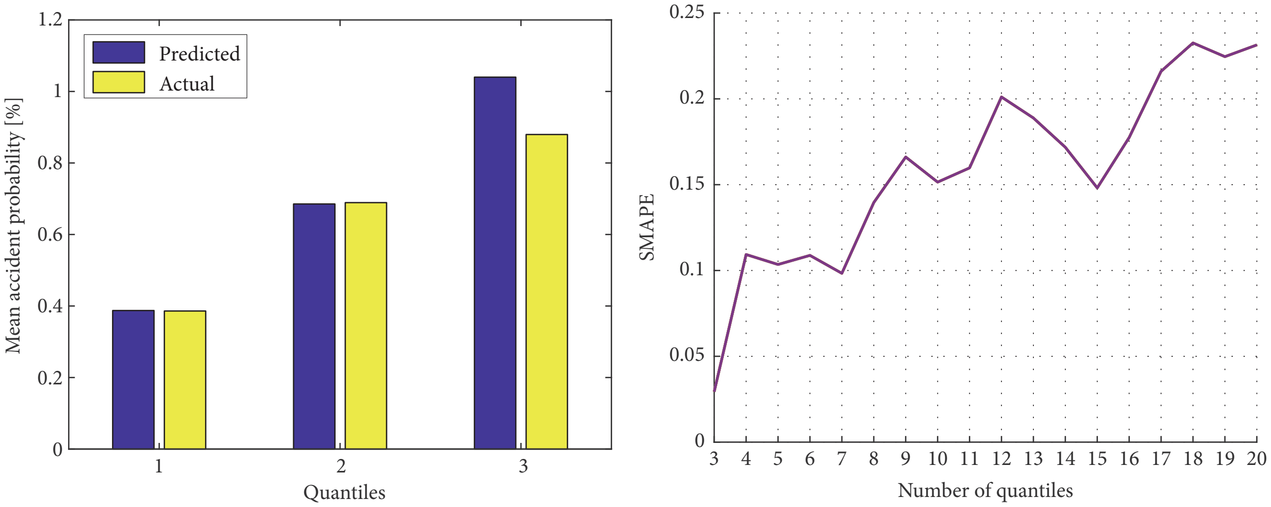
(b) Mean errors

Figure 11: *SMAPE*-based errors associated with the best model in each scenario.

Accordingly, it is possible to increase the accuracy of the model for higher resolutions by increasing the size of the training dataset.

*3.4. Example Application.* Using the previously proposed clustering approach to predict the probability of a collision occurring within a WZ, we now provide an illustrative application of the proposed model. This hypothetical case study pertains to the jurisdiction of the Coordinated Highways Action Response Team, whose objective is to improve operations of Maryland's highway system. Suppose that this agency has a list of planned maintenance work for the following day and is interested in deploying a fixed number of response units to tackle collisions that may happen within these WZs. Clearly, we can assign these WZs to the clusters derived in the previous section and consequently estimate collision probability for each of the WZs scheduled for the following day (Figure 9). Once these probabilities are computed, the allocation of response units becomes a stochastic facility location problem, which has been tackled extensively in the operations research literature [83, 84]. In order to solve this response team allocation problem based on collision probabilities obtained from clustering, we formulate a two-stage stochastic program [85], the details of which are described in the Appendix.

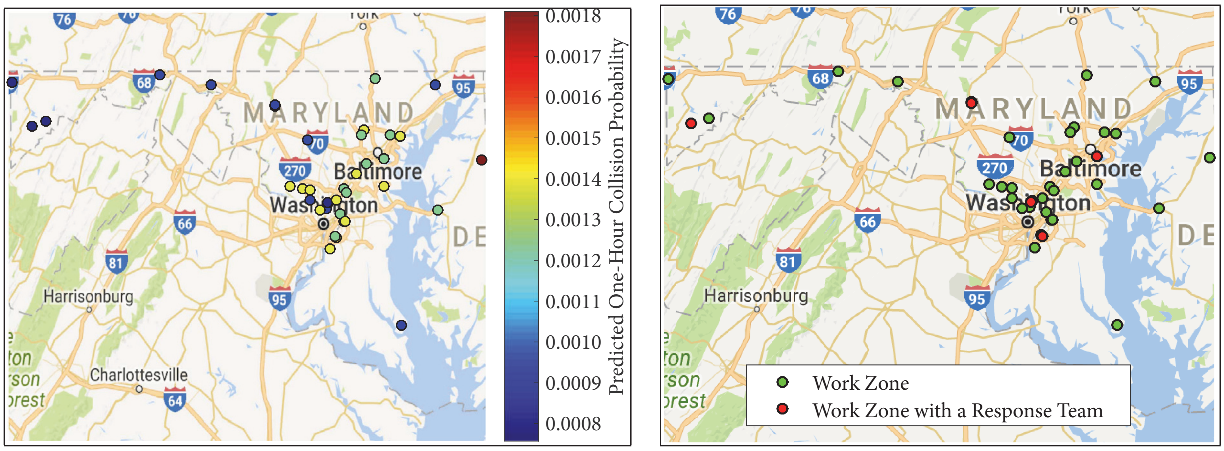
In this illustrative example we randomly sample the 40 WZs that were used to test the clustering methodology and pretend they represent the maintenance work scheduled for



(a) Accuracy for 3 quantiles

(b) *SMAPE* errors regarding the number of quantiles

Figure 12: Accuracy and *SMAPE* errors for the selected model.



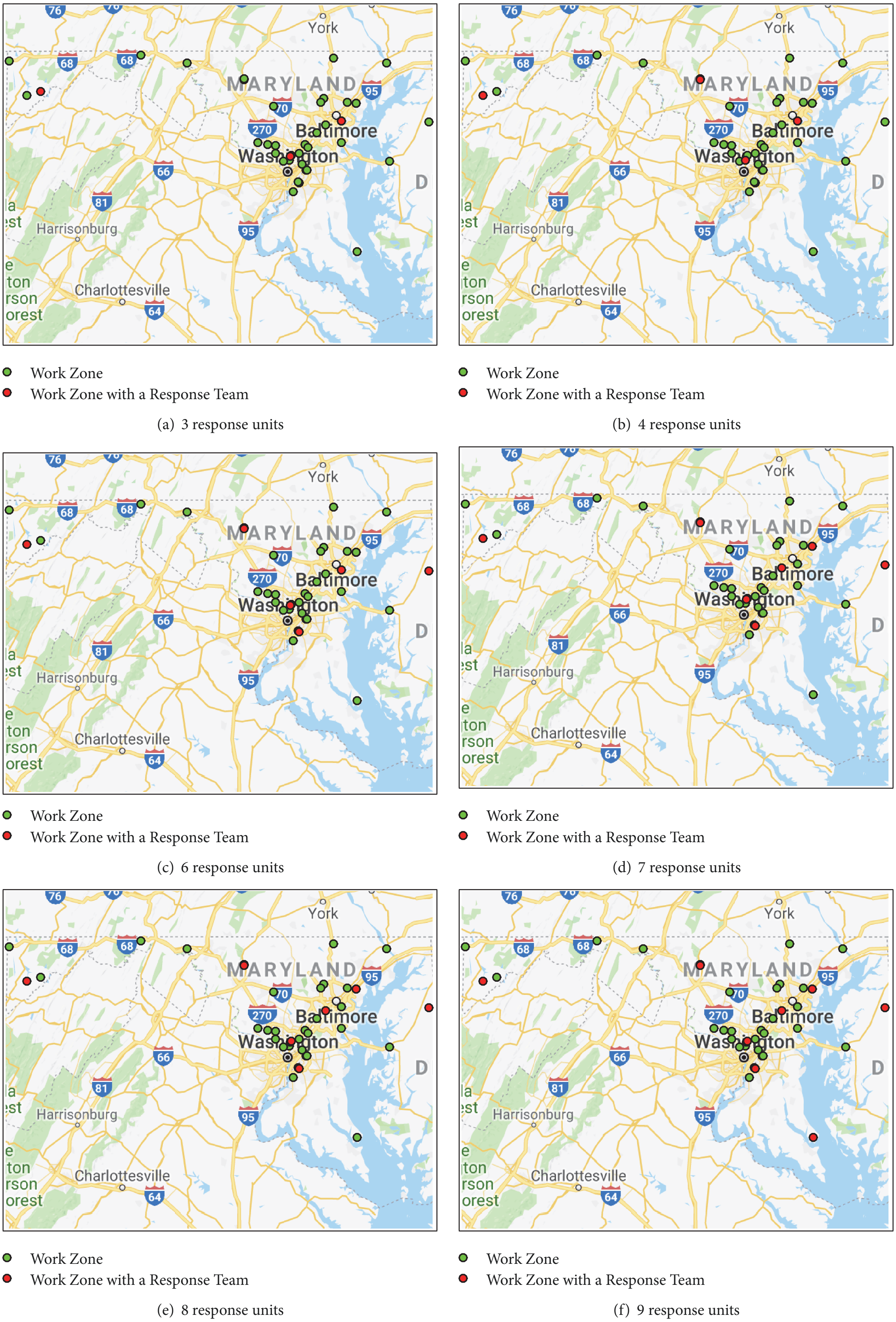
(a) Collision probabilities predicted via clustering

(b) Optimal response team allocation

Figure 13: Collision probabilities at 40 future WZs are estimated via the proposed clustering procedure. These probabilities are used as an input for the two-stage stochastic model to optimally allocate 5 highway response teams.

the following day. Consequently, we assign the 40 WZs to clusters built based on historical WZs (2010-2015), in order to compute the collision probabilities associated with each of the 40 newly-scheduled WZs (Figure 13(a)). Finally, using the optimization model provided in the Appendix, we determine the optimal location of the response units on the Maryland highway network for the observed time period. Figure 13(b) visualizes the optimal allocation when 5 response units are available, and additional allocations are provided in the Appendix (Figure 14) as the number of available response vehicles is perturbed from this value. The optimal allocation can be updated as the situation in the field evolves (e.g., one WZ is cleared and another on is set up), by simply reapplying the model given a new set of inputs (i.e., new configuration of WZs and corresponding collision probabilities). Alternatively, the proposed optimization model can be extended into a multiperiod stochastic program [85], which would better address dynamic relocation of response vehicles. This extended model would also require the collision probabilities estimated in this paper as inputs.

In addition to optimal allocation of highway response units, the proposed clustering method can be used to determine or adjust WZ parameters. Specifically, the presented model could help modify WZ parameters (e.g., lanes closed, day/night, and peak/off-peak duration) in order to meet certain safety levels (e.g., keep the collision probability below a specified threshold). For example, the easternmost WZ in Figure 13(b) has higher collision probability than others, so



● Work Zone
🟥 Work Zone with a Response Team

(a) 3 response units

● Work Zone
🟥 Work Zone with a Response Team

(b) 4 response units

● Work Zone
🟥 Work Zone with a Response Team

(c) 6 response units

● Work Zone
🟥 Work Zone with a Response Team

(d) 7 response units

● Work Zone
🟥 Work Zone with a Response Team

(e) 8 response units

● Work Zone
🟥 Work Zone with a Response Team

(f) 9 response units

Figure 14: Sensitivity analysis showing the optimal allocation of response teams when the number of available units is perturbed from $m = 5$.

its features may be modified to improve its safety level. A government agency might also introduce regulations to keep accident probabilities below certain level or provide financial incentives for those contractors who perform roadwork while maintaining high safety levels.

## 4. Conclusions

This paper proposed a clustering approach to predict the probability of a collision occurring in the proximity of planned maintenance work, which is important for allocation of highway response units. We presented the first application of clustering in the analysis of WZ collisions, which involved a large dataset of over 54,000 WZs in Maryland. The model showed good prediction accuracy, and its potential application was illustrated by optimally allocating response units in the Maryland highway network. Namely, collision probabilities determined via clustering were used as inputs to a two-stage stochastic program to optimally deploy highway response teams. Additionally, the proposed clustering approach can be used to adjust features of WZs to meet specified safety levels.

The proposed clustering method has certain limitations corresponding to the number of quantiles, so in some cases it may be used for classification of WZs rather than to predict the exact collision probabilities. Including more data would allow for additional quantiles to be used (i.e., preparation of higher-resolution models) and including additional WZ features may be useful as well. It would also be interesting to test clustering algorithms other than the $k$-means++, such as $k$-medoids or $c$-means and compare their performance. For the allocation of highway response teams, one could consider a multiperiod extension of the stochastic model used in this paper.

## Appendix

The problem of allocating highway response units is modeled as a two-stage stochastic integer program, with the goal to locate $m$ units in a way that would minimize the expected response time to WZs where collisions occur. In the first stage, we need to decide the location of response units, which may be placed at any of the WZ locations (or in their proximity). In the second stage, we assign these response units to WZs depending on the collision occurrences in each scenario and observe response times. More formally, let $I$ denote a set of WZs indexed by $i$ and $j$. Let $\xi_i$ be a random parameter indicating the probability of a collision occurring at WZ $i \in I$, which was estimated using the proposed clustering approach. A particular realization of the aforementioned random parameters is denoted by $\omega \in \Omega$. Define $x_i$ as an integer variable indicating the number of response units located in the proximity of WZ $i \in I$. We define $J(\omega) \subseteq I$ as a set of WZ locations where collisions occurred in a particular realization of random parameters $\omega \in \Omega$. Let $y_{ij}(\omega)$ be a binary variable which equals 1 if a response unit located at WZ $i \in I$ is dispatched to WZ $j \in J(\omega)$, given a particular realization of random parameters $\omega \in \Omega$. Let $d_{ij}(\omega)$ be the distance (or travel time) from WZ $i \in I$ to WZ $j \in J(\omega)$. We also define $\mathbf{x} = \{x_i \mid i \in I\}$ and $\boldsymbol{\xi} = \{\xi_i \mid i \in I\}$ as vectors of $|I|$ elements.

We formulate the first-stage problem as

$$\min_{x_i \in \mathbb{Z}^+} \quad \mathbb{E}\left[Q\left(\boldsymbol{x}, \boldsymbol{\xi}\right)\right] \tag{A.1}$$

$$\text{s.t.} \quad \sum_{i \in I} x_i \leq m, \tag{A.2}$$

where $Q(\boldsymbol{x}, \boldsymbol{\xi}(\omega))$ is the solution to the following second-stage problem:

$$Q\left(\boldsymbol{x}, \boldsymbol{\xi}\left(\omega\right)\right) = \min_{y_{ij}(\omega) \in \{0,1\}} \quad \sum_{i \in I} \sum_{j \in J(\omega)} y_{ij}\left(\omega\right) d_{ij}\left(\omega\right) \tag{A.3}$$

$$\text{s.t.} \quad \sum_{j \in J(\omega)} y_{ij}\left(\omega\right) \leq x_i \quad \forall i \in I \tag{A.4}$$

$$\sum_{i \in I} y_{ij}\left(\omega\right) = 1 \quad \forall j \in J\left(\omega\right) \tag{A.5}$$

As argued above, the first-stage problem seeks to allocate teams in order to minimize the expected response time subject to vehicle availability constraint. The response time for a particular realization of random parameters is computed in the second stage. Constraints (A.4) ensure that vehicles are dispatched only from locations where they are available, while (A.5) guarantee that a response team is directed to each WZ collision. To ensure feasibility of the second state, we let $m \geq |J(\omega)|$, $\forall \omega \in \Omega$, which implies that the number of response teams is greater or equal than the total number of locations at which collisions occur simultaneously. This condition, however, can be relaxed by extending the model to consider priorities in serving collision locations. Finally, the proposed two-stage stochastic model is implemented in GAMS and solved in the extensive form using CPLEX solver. We observe computation times of several seconds for the case study involving 40 WZs, 5 response teams and 1,000 scenarios (i.e., 1,000 realizations of $\boldsymbol{\xi}$). The corresponding optimal allocation of response teams is provided in Figure 13(b), with additional allocations shown in Figure 14 for different numbers of available response teams.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Disclosure

This research did not receive specific funding, but was performed as part of employment at the Center for Advanced Transportation Technology, University of Maryland.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] FHWA, "Facts and statistics - work zone safety," 2016, http://www.ops.fhwa.dot.gov/wz/resources/facts_stats/safety.htm.

[2] S. B. Mohan and P. Gautam, "Cost of highway work zone injuries," *Practice Periodical on Structural Design and Construction*, vol. 7, no. 2, pp. 68–73, 2002.

[3] A. G. Beacher, M. D. Fontaine, and N. J. Garber, "Evaluation of the late merge work zone traffic control strategy," Technical Report FHWA/VTRC 05-R6, 2004.

[4] M. Papageorgiou, I. Papamichail, A. D. Spiliopoulou, and A. F. Lentzakis, "Real-time merging traffic control with applications to toll plaza and work zone management," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 5, pp. 535–553, 2008.

[5] N. J. Fudala and M. D. Fontaine, "Work zone variable speed limit systems: Effectiveness and system design issues," Tech. Rep., 27 Nicholas J Fudala and Michael D Fontaine. Work zone variable speed limit systems, Effectiveness and system design issues. Technical Report FHWA/VTRC 10-R20, 2010.

[6] J. Weng and Q. Meng, "Modeling speed-flow relationship and merging behavior in work zone merging areas," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 985–996, 2011.

[7] J. Weng and Q. Meng, "Estimating capacity and traffic delay in work zones: an overview," *Transportation Research Part C: Emerging Technologies*, vol. 35, pp. 34–45, 2013.

[8] B. Du, I. Steven, and J. Y. Chien, "Feasibility of shoulder use for highway work zone optimization," *Journal of Traffic and Transportation Engineering*, vol. 1, no. 4, pp. 235–246, 2014.

[9] A. Tarko, S. Kanipakapatnam, and J. Wasson, "Modeling and Optimization of the Indiana Lane Merge Control System on Approaches to Freeway Work Zones, Part I," Purdue University FHWA/IN/JTRP-97/12-1, 1998.

[10] S. Chien and P. Schonfeld, "Optimal work zone lengths for four-lane highways," *Journal of Transportation Engineering*, vol. 127, no. 2, pp. 124–131, 2001.

[11] S. Chien, Y. Tang, and P. Schonfeld, "Optimizing work zones for two-lane highway maintenance projects," *Journal of Transportation Engineering*, vol. 128, no. 2, pp. 145–155, 2002.

[12] Q. Meng and J. Weng, "Optimal subwork zone length and project start time for short-term daytime work zones from the contractor's perspective," *Transportation Research Part C: Emerging Technologies*, vol. 29, pp. 72–83, 2013.

[13] A. Tympakianaki, A. Spiliopoulou, A. Kouvelas, I. Papamichail, M. Papageorgiou, and Y. Wang, "Real-time merging traffic control for throughput maximization at motorway work zones," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 242–252, 2014.

[14] G. Casal, D. Santamarina, and M. E. Vázquez-Méndez, "Optimization of horizontal alignment geometry in road design and reconstruction," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 261–274, 2017.

[15] J. Wang, W. E. Hughes, F. M. Council, and J. F. Paniati, "Investigation of highway work zone crashes: What we know and what we don't know," *Transportation Research Record*, no. 1529, pp. 54–62, 1996.

[16] J. Daniel, K. Dixon, and D. Jared, "Analysis of fatal crashes in Georgia work zone," *Transportation Research Record*, no. 1715, pp. 18–23, 2000.

[17] N. J. Garber and M. Zhao, "Distribution and characteristics of crashes at different work zone locations in Virginia," *Transportation Research Record*, no. 1794, pp. 19–28, 2002.

[18] S. D. Schrock, G. L. Ullman, A. S. Cothron, E. Kraus, and A. P. Voigt, "An analysis of fatal work zone crashes in Texas," Technical Report FHWA/TX-05/0-4028-1, 2004.

[19] H. T. Abdelwahab and M. A. Abdel-Aty, "Predicting injury severity levels in traffic crashes: a modeling comparison," *Journal of Transportation Engineering*, vol. 130, no. 2, pp. 204–210, 2004.

[20] J. Ma and K. Kockelman, "Crash frequency and severity modeling using clustered data from Washington state," in *Proceedings of the ITSC 2006: 2006 IEEE Intelligent Transportation Systems Conference*, pp. 1621–1626, Canada, September 2006.

[21] Y. Li and Y. Bai, "Highway work zone risk factors and their impact on crash severity," *Journal of Transportation Engineering*, vol. 135, no. 10, pp. 694–701, 2009.

[22] G. Khan, A. R. Bill, and D. A. Noyce, "Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity," *Transportation Research Part C: Emerging Technologies*, vol. 50, pp. 86–96, 2015.

[23] A. J. Khattak and F. Targa, "Injury severity and total harm in truck-involved work zone crashes," *Transportation Research Record*, no. 1877, pp. 106–116, 2004.

[24] E. S. Park and D. Lord, "Multivariate poisson-lognormal models for jointly modeling crash frequency by severity," *Transportation Research Record*, no. 2019, pp. 1–6, 2007.

[25] J. Ma, K. M. Kockelman, and P. Damien, "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accident Analysis & Prevention*, vol. 40, no. 3, pp. 964–975, 2008.

[26] Q. Meng, J. Weng, and X. Qu, "A probabilistic quantitative risk assessment model for the long-term work zone crashes," *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1866–1877, 2010.

[27] C. Xu, A. P. Tarko, W. Wang, and P. Liu, "Predicting crash likelihood and severity on freeways with real-time loop detector data," *Accident Analysis & Prevention*, vol. 57, pp. 30–39, 2013.

[28] Y.-J. Kweon, I.-K. Lim, and M. D. Fontaine, "Work zone safety performance measures for Virginia," Technical Report FHWA/VTRC 16-R10, 2016.

[29] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, 2010.

[30] R. Pal and K. C. Sinha, "Analysis of crash rates at Interstate work zones in Indiana," *Transportation Research Record*, no. 1529, pp. 43–53, 1996.

[31] A. J. Khattak, A. J. Khattak, and F. M. Council, "Effects of work zone presence on injury and non-injury crashes," *Accident Analysis & Prevention*, vol. 34, no. 1, pp. 19–29, 2002.

[32] S. Venugopal and A. Tarko, "Safety models for rural freeway work zones," *Transportation Research Record*, no. 1715, pp. 1–9, 2000.

[33] R. Srinivasan, G. Ullman, M. Finley, and F. Council, "Use of empirical bayesian methods to estimate crash modification factors for daytime versus nighttime work zones," *Transportation Research Record*, no. 2241, pp. 29–38, 2011.

[34] E. Chen and A. Tarko, "Analysis of crash frequency in work zones with focus on police enforcement," *Transportation Research Record*, no. 2280, pp. 127–134, 2012.

[35] O. Ozturk, K. Ozbay, H. Yang, and B. Bartin, "Crash frequency modeling for highway construction zones," in *Transportation Research Board 92nd Annual Meeting*, pp. 13–4555, 2013.

[36] D. Eustace, A. Aylo, and W. Y. Mergia, "Crash frequency analysis of left-side merging and diverging areas on urban freeway segments - A case study of I-75 through downtown Dayton,

Ohio," *Transportation Research Part C: Emerging Technologies*, vol. 50, pp. 78–85, 2015.

[37] H. Yang, K. Ozbay, O. Ozturk, and M. Yildirimoglu, "Modeling work zone crash frequency by quantifying measurement errors in work zone length," *Accident Analysis & Prevention*, vol. 55, pp. 192–201, 2013.

[38] K. El-Basyouny and T. Sayed, "Collision prediction models using multivariate Poisson-lognormal regression," *Accident Analysis & Prevention*, vol. 41, no. 4, pp. 820–828, 2009.

[39] D. Lord, S. P. Washington, and J. N. Ivan, "Further notes on the application of zero-inflated models in highway safety," *Accident Analysis & Prevention*, vol. 39, no. 1, pp. 53–57, 2007.

[40] A. K. Sharma and V. S. Landge, "Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highway," *Internet Journal of Advances in Engineering & Technology*, vol. 5, no. 2, pp. 292–301, 2013.

[41] C. Dong, D. B. Clarke, X. Yan, A. Khattak, and B. Huang, "Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections," *Accident Analysis & Prevention*, vol. 70, pp. 320–329, 2014.

[42] C. Dong, S. H. Richards, D. B. Clarke, X. Zhou, and Z. Ma, "Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression," *Safety Science*, vol. 70, pp. 63–69, 2014.

[43] F. Chen, S. Chen, and X. Ma, "Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models," *International Journal of Environmental Research and Public Health*, vol. 13, no. 6, article no. 609, 2016.

[44] J. C. Milton, V. N. Shankar, and F. L. Mannering, "Highway accident severities and the mixed logit model: an exploratory empirical analysis," *Accident Analysis & Prevention*, vol. 40, no. 1, pp. 260–266, 2008.

[45] P. C. Anastasopoulos and F. L. Mannering, "A note on modeling vehicle accident frequencies with random-parameters count models," *Accident Analysis & Prevention*, vol. 41, no. 1, pp. 153–159, 2009.

[46] E. Chen and A. P. Tarko, "Modeling safety of highway work zones with random parameters and random effects models," *Analytic Methods in Accident Research*, vol. 1, pp. 86–95, 2014.

[47] Z. Zeng, W. Zhu, R. Ke et al., "A generalized nonlinear model-based mixed multinomial logit approach for crash data analysis," *Accident Analysis & Prevention*, vol. 99, pp. 51–65, 2017.

[48] Y. Xie and Y. Zhang, "Crash frequency analysis with generalized additive models," *Transportation Research Record*, no. 2061, pp. 39–45, 2008.

[49] Y. Zhang, Y. Xie, and L. Li, "Crash frequency analysis of different types of urban roadway segments using generalized additive model," *Journal of Safety Research*, vol. 43, no. 2, pp. 107–114, 2012.

[50] N. V. Malyshkina, F. L. Mannering, and A. P. Tarko, "Markov switching negative binomial models: an application to vehicle accident frequencies," *Accident Analysis & Prevention*, vol. 41, no. 2, pp. 217–226, 2009.

[51] N. V. Malyshkina and F. L. Mannering, "Zero-state Markov switching count-data models: An empirical assessment," *Accident Analysis & Prevention*, vol. 42, no. 1, pp. 122–130, 2010.

[52] Y. Xiong, J. L. Tobias, and F. L. Mannering, "The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity," *Transportation Research Part B: Methodological*, vol. 67, pp. 109–128, 2014.

[53] X. Qin, J. N. Ivan, N. Ravishanker, and J. Liu, "Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov chain Monte Carlo modeling," *Journal of Transportation Engineering*, vol. 131, no. 5, pp. 345–351, 2005.

[54] D.-G. Kim, Y. Lee, S. Washington, and K. Choi, "Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models," *Accident Analysis & Prevention*, vol. 39, no. 1, pp. 125–134, 2007.

[55] H. Yang, K. Ozbay, K. Xie, and B. Bartin, "Modeling crash risk of highway work zones with relatively short durations," in *In Transportation Research Board 94th Annual Meeting*, Washington, DC, USA, 2015.

[56] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis & Prevention*, vol. 34, no. 6, pp. 729–741, 2002.

[57] R. Harb, E. Radwan, X. Yan, A. Pande, and M. Abdel-Aty, "Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression," *Journal of Transportation Engineering*, vol. 134, no. 5, pp. 203–214, 2008.

[58] G. H. Bham, B. S. Javvadi, and U. R. R. Manepalli, "Multinomial Logistic Regression Model for Single-Vehicle and Multivehicle Collisions on Urban U.S. Highways in Arkansas," *Journal of Transportation Engineering*, vol. 138, no. 6, pp. 786–797, 2012.

[59] H. Adeli and X. Jiang, "Neuro-fuzzy logic model for freeway work zone capacity estimation," *Journal of Transportation Engineering*, vol. 129, no. 5, pp. 484–493, 2003.

[60] A. Karim and H. Adeli, "Radial basis function neural network for work zone capacity and queue estimation," *Journal of Transportation Engineering*, vol. 129, no. 5, pp. 494–503, 2003.

[61] X. Jiang and H. Adeli, "Clustering-neural network models for freeway work zone capacity estimation.," *International Journal of Neural Systems*, vol. 14, no. 3, pp. 147–163, 2004.

[62] S. Ghosh-Dastidar and H. Adeli, "Neural network-wavelet microsimulation model for delay and queue length estimation at freeway work zones," *Journal of Transportation Engineering*, vol. 132, no. 4, pp. 331–341, 2006.

[63] S. Demiroluk and K. Ozbay, "Adaptive learning in bayesian networks for incident duration prediction," *Transportation Research Record*, vol. 2460, no. 1, pp. 77–85, 2014.

[64] B. Du, S. Chien, J. Lee, L. Spasovic, and K. Mouskos, "Artificial neural network model for estimating temporal and spatial freeway work zone delay using probe-vehicle data," *Transportation Research Record*, vol. 2573, pp. 164–171, 2016.

[65] Y. Xie, D. Lord, and Y. Zhang, "Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 922–933, 2007.

[66] A. Abdulhafedh, "Crash Frequency Analysis," *Journal of Transportation Technologies*, vol. 06, no. 04, pp. 169–180, 2016.

[67] X. Li, D. Lord, Y. Zhang, and Y. Xie, "Predicting motor vehicle crashes using Support Vector Machine models," *Accident Analysis & Prevention*, vol. 40, no. 4, pp. 1611–1618, 2008.

[68] S. Y. Sohn, "Quality function deployment applied to local traffic accident reduction," *Accident Analysis & Prevention*, vol. 31, no. 6, pp. 751–761, 1999.

[69] T. F. Golob and W. W. Recker, "A method for relating type of crash to traffic flow characteristics on urban freeways," *Transportation Research Part A: Policy and Practice*, vol. 38, no. 1, pp. 53–80, 2004.

[70] S. C. Wong, B. S. Y. Leung, B. P. Y. Loo, W. T. Hung, and H. K. Lo, "A qualitative assessment methodology for road safety policy strategies," *Accident Analysis & Prevention*, vol. 36, no. 2, pp. 281–293, 2004.

[71] H. Yang, K. Ozbay, O. Ozturk, and K. Xie, "Work Zone Safety Analysis and Modeling: A State-of-the-Art Review," *Traffic Injury Prevention*, vol. 16, no. 4, pp. 387–396, 2015.

[72] R. Wu, B. Zhang, and M. Hsu, "Clustering billions of data points using GPUs," in *Proceedings of the Combined Workshops on UnConventional High Performance Computing Workshop Plus Memory Access Workshop, UCHPC-MAW '09, Co-located with the 2009 ACM International Conference on Computing Frontiers, CF'09*, pp. 1–5, Italy, May 2009.

[73] CHART, "Coordinated highways action response team," 2016, http://www.chart.state.md.us/.

[74] M. Sheppard, "Allfitdist," MATLAB Central File Exchange, 2012.

[75] R. Pawlowicz, "Function adapted by the SCRIPPS Institute for Oceanography," 2009.

[76] O. Komarov, "MATLAB Central File Exchange," *Retrieved*, pp. 01-02, 2013.

[77] L. P. Michael, R. B. Jeffrey, and A. Steffes, "Overview and status of regional integrated transportation information system in the national capital region," in *Transportation Research Board 87th Annual Meeting*, pp. 08–1299, 2008.

[78] D. Schrank, B. Eisele, T. Lomax, and J. Bak, "2015 urban mobility scorecard," 2015, http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/mobility-scorecard-2015.pdf.

[79] H. Steinhaus, "Sur la division des corps matériels en parties," *Bulletin de l'Academie Polonaise des Sciences*, vol. 4, no. 12, pp. 801–804, 1957 (French).

[80] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," *Biometrics*, vol. 21, Article ID 768769, pp. 768-769, 1965.

[81] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[82] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[83] L. V. Snyder, "Facility location under uncertainty: a review," *IIE Transactions*, vol. 38, no. 7, pp. 547–564, 2006.

[84] M. T. Melo, S. Nickel, and F. Saldanha-da-Gama, "Facility location and supply chain management—a review," *European Journal of Operational Research*, vol. 196, no. 2, pp. 401–412, 2009.

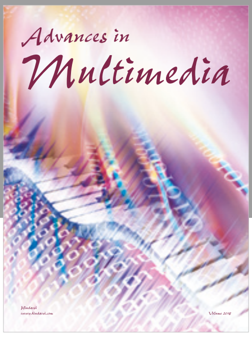
[85] J. R. Birge and F. V. Louveaux, *Introduction to Stochastic Programming*, Springer, Berlin, Germany, 2011.

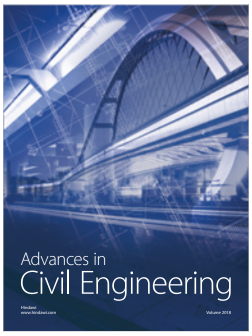
Journal of
Engineering

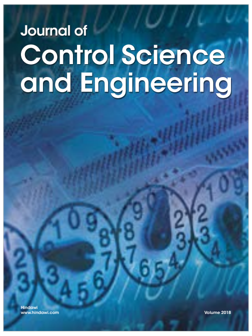
The Scientific
World Journal

International Journal of
Rotating
Machinery

Journal of
Sensors

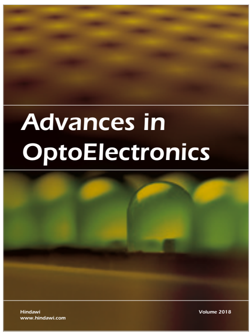

Advances in
Multimedia


Advances in
Civil Engineering


Journal of
Control Science
and Engineering

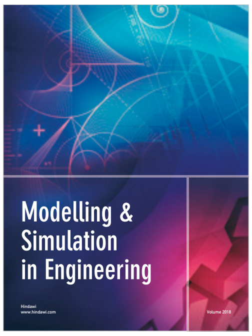
Journal of
Robotics

Journal of
Electrical and Computer
Engineering

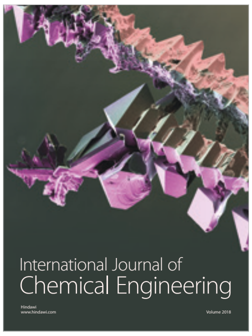

Advances in
OptoElectronics

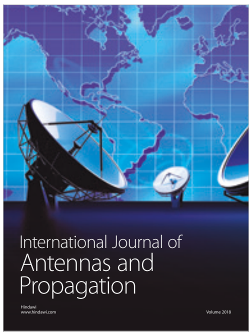
VLSI Design

International Journal of
Navigation and
Observation

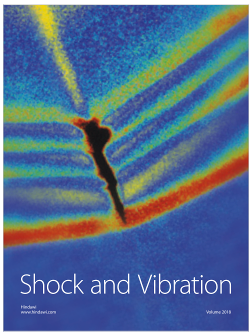

Modelling &
Simulation
in Engineering

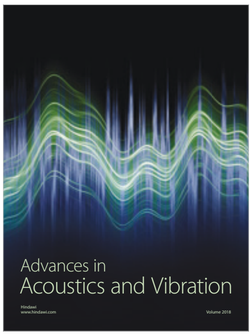
International Journal of
Aerospace
Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and
Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration