Accident Analysis and Prevention 179 (2023) 106880

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Predicting intersection crash frequency using connected vehicle data: A framework for geographical random forest

Yangsong Gu, Diyi Liu, Ramin Arvin, Asad J. Khattak *, Lee D. Han

*Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, USA*

ARTICLE INFO

*Keywords:*
Connected vehicle
Crash frequency prediction
Geographical random forest
Variable importance

ABSTRACT

Accurate crash frequency prediction is critical for proactive safety management. The emerging connected vehicles technology provides us with a wealth of vehicular motion data, which enables a better connection between crash frequency and driving behaviors. However, appropriately dealing with the spatial dependence of crash frequency and multitudinous driving features has been a difficult but critical challenge in the prediction process. To this end, this study aims to investigate a new Artificial Intelligence technique called Geographical Random Forest (GRF) that can address spatial heterogeneity and retain all potential predictors. By harnessing more than 2.2 billion high-resolution connected vehicle Basic Safety Message (BSM) observations from the Safety Pilot Model Deployment in Ann Arbor, MI, 30 indicators of driving volatility are extracted, including speed, longitudinal and lateral acceleration, and yaw rate. The developed GRF was implemented to predict rear-end crash frequency at intersections. The results show that: 1) rear-end crashes are more likely to happen at intersections connecting minor roads compared to major roads; 2) a higher number of hard acceleration and deceleration events beyond two standard deviations in the longitudinal direction is a leading indicator of rear-end crashes; 3) the optimal GRF significantly outperforms Global Random Forest, with a 9% lower test error and a substantially better fit; and 4) geographical visualization of variable importance highlights the presence of spatial non-stationarity. The proposed framework can proactively identify at-risk intersections and alert drivers when leading indicators of driving volatility tend to worsen.

## 1. Introduction

Intersections are often considered to be crash hotspots due to vehicular conflicts. i.e., the merge and split of traffic flow from different approaches. According to the statistics from Federal Highway Administration, more than 50 percent of the combined total of fatal and injury crashes occur at or near intersections (FHWA, 2021). The different types of crashes, such as rear-end, cause a major loss of people's lives and property. Therefore, crash frequency is needed not only for evaluating the risk of a crash but also for providing instructions in proactive countermeasures.

Contributing factors and predictive models used to be two foci of crash frequency prediction. Different factors are evaluated for their effects on crash frequency, and the results are used to select appropriate countermeasures. In many previous studies, macro-level data have been extensively exploited in proposed models. Those data, such as roadway traffic and geometry design, mainly describe and explain the influence of driving context on crash frequency. More recently, with the emergence of big data, the focus of attributes selection has been shifted to micro-level driving behaviors. Those driving behaviors in essence quantify the variation of driving motion and manipulation, which are more straightforward surrogate measures of driving status compared against earlier macro attributes. Like several previous studies in crash analysis, this study processed the data from the Safety Pilot Model Deployment (SPMD) project, which collected the driving vehicular data (e.g., speed, lateral acceleration, etc.) from connected vehicles (CV) at a level of 10 Hz (Henclewood et al., 2014). Such high-resolution data enables researchers to innovate and investigate the instantaneous behaviors relating to collision risk.

Regarding models, conventional statistical and machine learning models are two common techniques for crash frequency prediction. Among statistical models, Poisson Regression (Kwon et al., 1856) is the most widely applied baseline model that approaches the linear relation between the logarithm of the average crash frequency and associated

* Corresponding author.
*E-mail addresses:* ygu17@vols.utk.edu (Y. Gu), dliu27@vols.utk.edu (D. Liu), rarvin@vols.utk.edu (R. Arvin), akhattak@utk.edu (A.J. Khattak), lhan@utk.edu (L.D. Han).

https://doi.org/10.1016/j.aap.2022.106880
Received 22 March 2022; Received in revised form 6 October 2022; Accepted 20 October 2022
Available online 4 November 2022

factors. However, this relation could be non-linear. Hence, machine learning models (e.g., Random Forest (hereinafter called RF)) are adopted to seize different relations. In most cases, these models produce higher accuracy than statistical models but are weak in model explanation (Pu et al., 2020; Chang, 2005). However, in recent years, more and more findings suggest that the spatial heterogeneity and spatial dependency of crash frequency should be properly considered. Thus, some studies integrated geographically weighted models with statistical models (Arvin et al., 2019; Tang et al., 2020; Xu and Huang, 2015). Although spatial models handle the spatial non-stationary of crash frequency well, only a set of uncorrelated predictors are allowed in the model, while some other predictors that could potentially lead to crash might be abandoned. In addition, the essence of existing regression methods is building the linear relationship between crash frequency and factors (Lord and Mannering, 2010); the underlying non-linearity cannot be untangled well. In terms of predictive capability, machine learning models are superior to regression models due to their flexible and non-linear nature. On the other hand, they are not able to catch spatial bias, namely, they are aspatial algorithms in essence. Ignoring the spatial heterogeneity might diminish the predictive performance when dealing with the spatially dependent data. The machine learning models used to be a "black-box" and the interpretability is a concern. Fortunately, existing effort has shown some applicable techniques to uncover this "black-box". For instance, a study found that the Variable Importance indicators based on bagged conditional forests appear to reach a sound balance between identification of significant variables and avoiding unnecessary flagging of correlated variables (Auret and Aldrich, 2011). As a result, it is promising to put forward a framework that can perform machine learning models in a geographical fashion, ultimately for a both predictive and explanatory tool.

This study aims to utilize driving volatility extracted from big CV data as the major indicator to predict crash frequency at intersections. As one of the most frequent crash types at intersections (Wang and Abdel-Aty, 2006), rear-end crash frequency is selected as the outcome variable. The main goals of this study are to:

1. Incorporate high-resolution large-scale CV data into transportation safety analysis.
2. Investigate the framework of GRF for rear-end crash frequency prediction using new artificial intelligence and spatial analysis techniques.
3. Evaluate the spatial dependency of rear-end crash frequency.

The remainder of this paper is organized as follows. The next section reviews the past works on crash frequency prediction. In the methodology section, the framework and algorithm of GRF are elaborated. Then, the model is evaluated and discussed, followed by the result summary. Finally, the conclusion is drawn, and future work is proposed.

## 2. Literature review

This section reviews the influencing factors associated with rear-end crashes and two types of predominant approaches including statistical and machine learning models for crash frequency prediction.

### 2.1. Crash contributing factors

Researchers have identified various contributing factors associated with rear-end crashes, which generally can be classified into two categories: macro and micro. The macro-level generally depicts the driving context and traffic exposure, which are fundamental features relating to crashes. Some representative factors are 1) Roadway characteristics or geometry: number of lanes, signal, percent of the freeway, curve radius (Wang and Abdel-Aty, 2006; Dong et al., 2014; Haleem et al., 2010; Zhang et al., 2021); 2) Traffic-related information: Annual Average Daily Traffic (AADT), speed limit, vehicle miles traveled (Arvin et al., 2019; Tang et al., 2020; Wang and Abdel-Aty, 2006; Dong et al., 2014; Wali et al., 2018); 3) Socio-demographics: race, land use, population (Tang et al., 2020; Xu and Huang, 2015). In terms of micro-level variables, surrogate safety measures (SSMs) are extracted from traffic conflicts that are directly linked to crashes. Some representative measures applied in many studies are time to collision, modified time-to-collision, and post encroachment time (Fu and Sayed, 2021; Katrakazas et al., 2021). Notably, these measures may only be able to capture the risk in terms of time proximity, while other risk aspects (e.g., speed and deceleration) are not reflected (Wang et al., 2021). Benefited from the connected vehicle pilot project in the US, massive vehicular motion data (e.g., BSM). Thereafter, the new term "driving volatility" was introduced to characterize the driving behaviors. The fundamental idea is to capture the magnitudes and extent of variation in driving decisions as larger variations in instantaneous decisions by the driver can not only influence their own safety but also the surrounding vehicles. In the literature, this driving volatility was quantified through several functions that are employed to vehicle speed (Arvin et al., 2019; Kamrani et al., 2018), longitudinal acceleration (Arvin et al., 2019; Liu and Khattak, 2016), lateral acceleration (Wali et al., 2020; Dong et al., 2019) and yaw rate (Arvin et al., 2019). Some of these studies have demonstrated that volatility measures such as coefficient of variation in speed, number of speed points lying beyond two standard deviations, and coefficient of variation volatility of positive longitudinal acceleration are highly correlated with rear-end crashes (Arvin et al., 2019; Arvin et al., 2021). Although a series of different driving volatilities has been proposed in previous studies, most of them were eliminated due to the multicollinearity between them. Thus, how other driving volatilities affect the crash frequency is still unknown.

### 2.2. Statistical models

Numerous statistical models have been investigated for crash frequency prediction. Given that the crash frequency is always a non-negative integer, most of the work exploited the Poisson Regression (Kwon et al., 1856) model as a starting point (Lord and Mannering, 2010). However, it is difficult for real-life data to satisfy the assumption that the mean and variance should be equal in the PR model. Thus, using the pure PR model would lead to erroneous prediction if such criterium is not met. To get rid of the overdispersion issue, the PR is generalized by including a gamma noise variable which has a mean of 1 and a scale parameter of $\lambda$ (Hilbe, 2011), which releases the relation between mean and variance. This enhanced model is called Negative Binomial (NB), which is widely implemented in crash frequency estimation and prediction (Abdulhafedh, 2016; Mohammadi et al., 2014; Naznin et al., 2016). In addition, recent studies noticed that some contributing factors like AADT may exist in unobserved heterogeneity. Hence, to address this issue, random parameters estimation (Arvin et al., 2019; Xu and Huang, 2015; Anastasopoulos and Mannering, 2009) and geographically weighted methods are integrated into statistical models (Tang et al., 2020; Xu and Huang, 2015) for more accurate estimation. The above-mentioned models all originated from the fundamental PR model. As a result, the essence of PR that it transforms linear relations between the outcome variable and contributing factors through probability is inherited in all models. Hence, they are not able to capture the non-linear relations. Additionally, although much enhancement has been done, the statistical models discard the highly correlated contributing factors to avoid overfitting, which could lead to the loss of potential critical factors. A recent study integrated the Bayesian Additive Regression Trees into negative binomial regression for accident hotspot identification. To the authors knowledge, the proposed model is the state-of-the-art statistic method in dealing with both spatial dependency, unobserved heterogeneity, and non-linearity (Krueger et al., 2020).

### 2.3. Machine learning models

Compared to statistical count models, machine learning models are more powerful in that they can capture complex relationships between input and output, and are more capable of accommodating correlated features. Abdel-Aty et. Al. applied the multivariate adaptive regression splines (MARS), and results show that MARS generates a lower mean square prediction error than the fitted NB model (Abdel-Aty and Haleem, 2011). Besides, Pu et al. employed RF to capture the non-linear correlation between crash count and road geometry, which obtained a more accurate prediction than the random effect negative binomial (RENB) model (Pu et al., 2020). Another typical model, ANN, was considered in Abdulhafedh (Abdulhafedh, 2016) and Change et al. (Chang, 2005); they compared it with statistical models and found that ANN is a consistent alternative method for analyzing crash frequency. More recently, ensemble learning gets more attention as a state-of-art technology. It is capable of reducing prediction error by aggregating the results from multiple learners (Polikar, 2012). Wu et. Al incorporated ensemble learning into NB, support vector machine (SVM), and back propagation neural network (BPNN) to significantly decrease the crash frequency prediction error (Wu et al., 2019). Although these machine learning models are well-performing in aspatial prediction, they cannot address spatial heterogeneity of parameters.

### 2.4. Research gap

In recent years, the geographical random forest has emerged in the geography field, e.g., remote sensing (Georganos et al., 2021) and landslide susceptibility mapping (Quevedo et al., 2021). Domain-wise, the analysis and prediction of crash frequency in a transportation network are qualitatively different from the analysis of application domains in geography. As noted in the literature, crash frequencies correlate with numerous and complex sets of factors (Arvin et al., 2019; Wang et al., 2021; Arvin et al., 2019; Dong et al., 2019; Khattak and Wali, 2017). Extracting variables from new kinds of data is important to improve predictions. Improving machine learning methods points to the overfitting issue. If this issue is not addressed, then overfitting can potentially diminish prediction performance and can cause misleading results. Additionally, differences are apparent in terms of the shortest path distance used in this study as opposed to the Euclidean distance used in the Georganos et al. study (Georganos et al., 2021). In summary, the gaps in previous studies include: 1) How to use microscopic level trajectory data and driving volatility measures to explain crash frequency; 2) Existing statistical models only account for generalized linear relationships between the explanatory variables and response variable; 3) Solely relying on statistical models would discard potential important indicators; 4) Machine learning models cannot handle spatial dependence of the crash frequency. To fill in these gaps, this study developed a GRF framework for better predicting rear-end crash frequency. By processing rich connected vehicles' motion data, 30 measures of driving volatilities were proposed and calculated. Then, the GRF was trained and tested in a form of local machine learning models in geographical space. In the end, the optimal model was tuned according to the accuracy measures.

## 3. Methodology

### 3.1. Geographical random forest

In this study, the developed GRF is based on the knowledge of RF and spatial disaggregation rule of geographical models. RF is chosen for geographical regression because it works well for high-dimensional predictors even with a small number of samples through the bagging technique on resampling (Oshiro et al., 2012; Gromping, 2009). In addition, RF not only can capture the complex relationships but also assess the feature importance, which can be used to explain the contribution of features to the response variable (Gromping, 2009). What's more, RF is capable of preserving all features which benefits greatly from its two merits: one is RF applied bagging learning mechanism that builds a large collection of de-correlated trees thereby eliminating the overfitting; the other is each tree of RF is built from randomly selected subsets of features of sample data so that a part of features will be employed for making decisions (Hastie and Tibshirani, 2009). However, pure RF may not fix the spatial non-stationarity of the effect of features. Therefore, a more explicit way of distinguishing spatial dependence should be considered. The widespread use of Geographically Weighted Regression (GWR) has proved its capability of capturing the spatial variation of parameters by executing local regression models with a certain number of neighbors in a geographical location (Xu and Huang, 2015; Brunsdon et al., 1998). In this study, this fashion is integrated into the RF model, making the RF able to untangle the spatial variation.

Fig. 1 presents the entire framework of implementing developed GRF for rear-end crash frequency prediction. Generally, it works in a sequence of data preparation, model training, model testing, and tuning.

Training RF in a geographical fashion is the core step of the developed GRF. As opposed to global RF which runs on the whole intersections (i.e., whole dataset), a local RF is performed on each intersection $i$ and its neighbors (i.e., a portion of dataset) that meets spatial distance or density criterion with target intersection $i$. This step includes two sub-steps, the first sub-step is choosing neighbors of a geographic unit where the sub-model operates on. Without losing the generality, we introduced the bandwidth to select the qualified neighbors. Like the GWR model, two types of bandwidths including adaptive and fixed kernel were applied to select observations for model input. Bandwidth is the most important parameter to consider for the geographical model, which can significantly influence the model's performance (Guo et al., 2008). Specifically, the fixed kernel selects neighbors within a certain distance while the adaptive kernel filters the neighbors with a definite number. For instance, in the fixed kernel model, bandwidth = 2000 m means that model will choose the neighbors which have a distance to that intersection that is less than or equal to 2000 m. In contrast, bandwidth = 20 (unit) in the adaptive kernel model represents that the 20 nearest neighbors will be engaged in fitting the local model. Intuitively, when intersections are sparsely distributed, the adaptive kernel can ensure a certain number of observations for fitting the local model. However, if the intersections are densely distributed, the fixed bandwidth will consider more observations. Notably, to better capture the association of intersections, Dijkstra's shortest path distance (Wikipedia, 2021) between two intersections is computed and used for bandwidth increment.

The other sub-step is training RF geographically through the bagging technique. A number of trees are built by randomly resampling training data with replacement. Then, the tree grows with node splitting by the best among a random subset of features. The purpose of the training process is to tune the number of trees and predictors randomly selected at each candidate split (hereinafter called *mtry*). According to the previous study, the prediction performance may not be significantly improved while the computation time could increase exponentially as the number of trees grows (Perner, 2012). Hence, it is prespecified by considering the size of the training set and total features. Regarding the *mtry*, it is usually determined through minimizing the out-of-bag (OOB) error (James et al., 2013). Even though bandwidth restricts the model complexity to some extent, plenty of predictors could lead to potential overfitting issues. Hence, the maximum *mtry* (hereinafter, called *maxmtry*) is further set to control the maximum number of predictors used for splitting trees. When the model does not apply any bandwidth, it turns to a global Random Forest. To implement the basic RF, we used the R package "Caret" (Kuhn et al., 2008). The 10-fold cross-validation is performed to obtain the optimal number of predictors for growing trees. In the process of building trees, two-thirds of training set is randomly assigned to each tree, and rest one-thirds of training set is used for

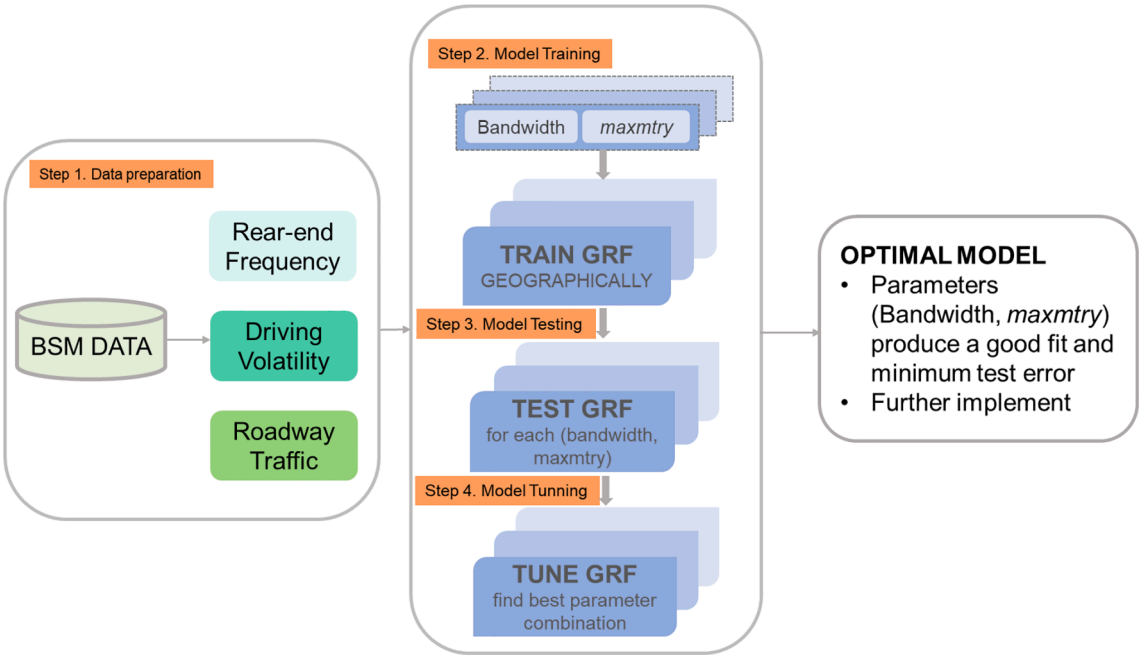


**Fig. 1.** Framework of Geographical Random Forest in Crash Frequency Prediction.

**Algorithm:** Geograpical Random Forest

**Input:**

$\Phi_a$: a set of bandwidth of adaptive kernel, size = $n$
$\Phi_f$: a set of bandwidth of fixed kernel, size = $m$
$\Psi_{mm}$: a set of mtry for both adaptive and fixed kernel, size = $k$
$\Omega_{train}$: a set of id of intersections in training set
$\Omega_{test}$: a set of id of intersections in test set

**Initialize** $AE_{train} \leftarrow [\,]_{n\times m}$, $AE_{test} \leftarrow [\,]_{n\times m}$, $RF_{train} \leftarrow [\,]_{n\times m}$, $ntree = 50$
**for** *each* $\psi_{mm}$ in $\Psi_{mm}$
  **for** *each* $\phi_a$ **in** $\Phi_a$ // $\Phi_a$ *is the number of neighbors in adaptive kernel*
// **for** *each* $\phi_f$ **in** $\Phi_f$ // $\Phi_f$ *is the distance threshold in fixed kernel*
    **for** *each intersection* $i$ **in** $\Omega_{train}$
      $S \leftarrow$ the closest $\phi_a$ intersections to intersection $i$
      // $S \leftarrow$ the intersections within the bandwidth $\phi_f$
      $RF_i \leftarrow$ train random forest, given $\psi_{mm}, S$ *and* $ntree$
      $RF_{train}.append(RF_i)$
      $AE_i \leftarrow OOB\ error$ // *one out-of-bag absulote error*
      $AE_{train}.append(AE_i)$
    **end for**
    **for** *each intersection* $i$ **in** $\Omega_{test}$
      find the closet intersection $j$ in $\Omega_{train}$
      $AE_i \leftarrow$ implement $RF_j$ to make predictions
      $AE_{test}.append(AE_i)$
    **end for**
  **end for**
**end for**

**Output:**

$MAE_{train} \leftarrow$ average $AE_{train}$ // *Mean asboluate error of training set*
$MAE_{test} \leftarrow$ average $AE_{test}$ // *Mean asboluate error of test set*

**Fig. 2.** Pseudo Algorithm of GRF.

testing the model performance under specific parameters.

Due to the limited prior knowledge on how *maxmtry* and bandwidth would influence the prediction accuracy, this study enumerated the different combinations of bandwidth and *maxmtry* while training the RF. Fig. 2 depicts the pseudocode of the developed GRF in this study. Firstly, a set of candidates of bandwidth were initiated for adaptive $\Phi_a$ and fixed $\Phi_b$ kernels. The specific bandwidth can be a series of incremental values. Meanwhile, a set of *maxmtry* $\Psi_{mm}$ should be given for iterating optimal *maxmtry*. Considering the total number of predictors is 40, we selected each *maxmtry* with increments of 5 (i.e., 5, 10, 15, …, 40) and pre-specified 50 trees for resampling. Given a $\psi_a$ in $\Psi_{mm}$, a $\phi_a$ in $\Phi_a$, or a $\phi_f$ in $\Phi_f$, the GRF trains local RF for each intersection $i$ in the intersection set $\Omega_{train}$. In this process, the adaptive kernel selects the closest $\Phi_a$ intersections to intersection $i$ for growing independent trees. If a fixed kernel is used, then the intersections within the bandwidth $\phi_f$ of intersection $i$ will be considered in the local model. Next, the absolute error (Chen et al., 2015) of one OOB is computed and saved for subsequent model evaluation. As compared to other performance indicators, the MAE is more robust to outliers than the root mean square error (RMSE) (Willmott and Matsuura, 2005), and the MAE is more feat to handle mass zero crash counts than Mean Absolute Percentage Error (MAPE). Thus, MAE has been leading prediction performance measure in previous crash frequency studies (Xu et al., 2021; Zhang et al., 2019; Tang et al., 2021; Xie and Zhang, 2061). The MAE of the training set ($MAE_{train}$) is the average of all AE of training intersections. After all local models in $\Omega_{train}$ are trained, the crash frequencies of intersections in the test set $\Omega_{test}$ are predicted using the closest intersection's model in $\Omega_{train}$, as Fig. 1 step 3 shows. Similarly, the AE of each test intersection is recorded, and the MAE of the test set ($MAE_{test}$) is obtained from the average of all AEs. So far, the procedure of getting MAE of the training set and test set is completed, and the remaining sections will focus on how to evaluate the MAEs for optimal model selection.

### 3.2. Model evaluation

During the training and testing process, MAE (Eq. (1)) is used to assess the model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{1}$$

where $y_i$ and $\hat{y}_i$ are observed and predict crash frequency at intersection $i$, respectively, and n is the sample size.

In addition, variable importance is calculated to understand the importance of each predictor on rear-end crashes. When a tree is grown, the OOB samples (i.e., test samples) are passed down the tree, and the prediction accuracy of this tree is recorded as $MSE_o$. Then, the value of a predictor $j$ in OOB samples is permuted randomly, leading to a new OOB set. The new prediction accuracy is computed using the new OOB set. Finally, the variable importance of predictor $j$, which is expressed as the percent of increased MSE (Equation (2)) averaged over all trees, can be written as Eq. (3). The randomization of permutation tends to make the variable importance more uniformly, compared to Gini index (Hastie and Tibshirani, 2009; James et al., 2013; RcolorBrewer and Liaw, 2018).

$$MSE = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \tag{2}$$

$$VI_j = \frac{MSE_j - MSE_o}{MSE_o} \times \tag{3}$$

where $MSE_j$ is averaged prediction error of all trees after permuting predictor $j$, and $MSE_o$ is the original prediction error. Besides, in Eq. (2), n is the number of observations in a OOB set. The more important the variable is, the larger the increased MSE will be expected to be.

In the GRF model, since the features are randomly selected for the split of trees, the predictors may not be involved in the prediction of all intersections. Thus, the variable importance of a predictor $j$ at specific intersection $i$ ($VI_{i,j}$) is measured by averaging all predictor $j$'s importance from the intersections that considered predictor $j$, which can be written as Eq. (4):

$$VI_{i,j} = \frac{1}{n} \sum_{k=1}^{n} VI_{k,j} \tag{4}$$

where n is the total amount of intersections that consider the predictor $j$.

### 3.3. Model tuning

The iterative fashion creates a lot of results from different combinations of bandwidth and *maxmtry*. Thus, the last step of GRF (i.e., step 4 in Fig. 1) is designed to develop a strategy to identify the optimal model that can avoid under-and overfitting issues and obtain the minimum test error at the same time. To this end, the ratio between training MAE and test MAE is introduced (Eq. (5)) to evaluate the tendency of under- and overfitting. As the ratio approaches 1, the training MAE tends to line up with test MAE, indicating a good fit, and vice versa. However, this case would happen when both training MAE and test MAE are relatively high (Hastie et al., 2009). Hence, another criterion is added that a good model should have both relatively low training and test MAE. With this requirement, a threshold (Eq. (6)) is given to allow some models to be candidates of the final optimal model. For instance, with the threshold of 0.02, the models whose ratio drops between 0.98 and 1.02 will be seen as the candidates. Within this candidate pool, the optimal model is selected based on the smallest test MAE. The criteria of model fitness are summarized in Table 1.

$$R = \frac{MAE_{train}}{MAE_{test}} \tag{5}$$

The Threshold is defined as:

$$Th = |R - 1| \tag{6}$$

## 4. Data

### 4.1. Data source

This study integrates the data from BSM with the road inventory and intersection crash data. The BSM data obtained from the ITS datahub (https://www.its.dot.gov/data/) provides the micro-seconds vehicular information such as coordinates (e.g., longitude and latitude), instantaneous motion (e.g., acceleration and yaw rate), and contextual data (e.g., distance to the front object). Over 2.2 billion BSM were collected through over 2800 connected vehicles deployed along 70 miles roadways on October 2012 and April 2013 using standard protocols by UMTRI at the University of Michigan (6). The complete two-month data were exploited in this study. Notably, a data error-check was conducted and erroneous data records such as "-9.81", "9.81","19.62", or "-19.62" from acceleration measures were removed due to the default placeholder of 'g' or '2g'. After data cleaning, 1496 connected vehicles were selected for computing 30 measures of driving volatilities at 167 CV-involved intersections.

In addition, to account for the effect of the traffic exposure and driving context, road inventory information at 167 intersections was collected from Google Maps and the Michigan Metropolitan Planning

**Table 1**
Criteria of Assessing GRF Model.

| | Underfitting | Overfitting | Good Fit |
|---|---|---|---|
| $R$ | $R < 1 - Th$ | $R$ greater than $1 + Th$ | $1\text{-}Th < R < 1+Th$ |
| $MAE_{train}$ and $MAE_{test}$ | One of them is high | One of them is high | Both are low |

Organization Website (http://semcog.org/). Those road inventory data include AADT on major or minor roads, speed limit, number of lanes, etc. Correspondingly, the historical rear-end crashes that occurred in the data collection period were obtained from the Michigan Metropolitan Planning Organization Website as well. However, those rear-end crashes and connected vehicles data were independently collected thereby matching them together is a challenge in this study. To mitigate this issue, two remedies were claimed. For one thing, 1 year of rear-end crash frequency data from October 2012 to 2013 was collected to obtain a reliable inference and prediction. For the another, we assumed that driving behaviors of CV were representative of the majority of drivers, considering those rear-end crashes could happen among common vehicles.

### 4.2. Driving volatility

In order to calculate the driving volatility indicators at the intersection level, BSM data dropping into a 150-ft circle measured from the center of intersections was extracted and aggregated for four groups of volatilities: 1). Speed volatility, 2) Longitudinal acceleration volatility, 3) Lateral acceleration volatility, and 4) Yaw-rate volatility. Five types of measures were applied and defined via Eqs. (7)-(11).

*Standard Deviation.* This is a common and basic measure used to quantify the variation of data points, it can be written as:

$$S_{dev} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

where $x_i$ is observation $i$, $x_i$ is the mean value of observations, and $n$ is the sample size.

*Coefficient of Variation.* This measure is defined as the ratio of standard deviation to average (Brown, 1998), it can be formulated as:

$$C_v = \frac{S_{dev}}{\bar{x}} * 100 \tag{8}$$

*Mean Absolute Deviation.* This measure quantifies the distance between observations and central tendency (here, mean), which can be formulated as (Huber, 2004):

$$D_{mean} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{9}$$

*Quantile Coefficient of Variation.* This indicator measures the skewness of sampling distribution (e.g., nonnormal distribution) (Bonett, 2006), which is formulated by the difference between 25th and 75th percentile of data points and the sum of 25th and 75th percentile of data points, as written in Eq. (10):

$$Q_{cv} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100 \tag{10}$$

where $Q_1$ and $Q_3$ are 25th and 75th percentiles of sample data, respectively.

*Percent of extreme values.* This measure describes the density of extreme values via counting the number of observations beyond one or two standard deviations, it can be written as (Kamrani et al., 2018; Hoseinzadeh et al., 2020):

$$\%T_z = \frac{x_i > \bar{x} \pm z^* s_{dev}}{n} * 100 \tag{11}$$

where $z^* s_{dev}$ indicates the distance threshold between the mean and a point, $z$ can be 1, 2, 3, etc.

## 5. Results

### 5.1. Descriptive statistics

Table 2 presents the descriptive statistics of rear-end crashes and predictors, among which the driving volatility of speed, longitudinal acceleration, lateral acceleration, and yaw rate were calculated from Eqs. (7)–(11) using complete two-month BSM data at 167 intersections. Four statistics, including the minimum, maximum, standard deviation (S.D.), and mean values were provided. Fig. 3 depicts the spatial heterogeneity of rear-end crashes which are distributed more in peripheral areas than the downtown area of Ann Arbor.

**Table 2**
Descriptive Statistics of Outcome and Predictors, N = 167 intersections.

| | Abbreviation | Min | Max | Mean | S.D. |
|---|---|---|---|---|---|
| **Outcome Variables** | | | | | |
| Rear-End Crashes | | 0.00 | 28.00 | 3.52 | 4.79 |
| **Roadway Traffic Characteristics** | | | | | |
| AADT on major road (1000) | AADTMaj | 2.53 | 45.40 | 18.47 | 8.60 |
| AADT on minor road (1000) | AADTMin | 1.10 | 27.40 | 8.85 | 3.87 |
| Speed Limit on major road (mph) | splimmaj | 25 | 45 | 34.52 | 6.52 |
| Speed Limit on minor road (mph) | splimmin | 15 | 45 | 29.28 | 4.37 |
| Signalized intersections (yes = 1) | SigOrnot | 0 | 1 | 0.49 | 0.5 |
| 4-legged intersection (yes = 1) | X4leg | 0 | 1 | 0.47 | 0.5 |
| Total through lanes | TotThrou | 2 | 8 | 4.25 | 1.38 |
| Total left turn lanes | TotLeft | 0 | 6 | 1.38 | 1.37 |
| Total right turn lanes | TotRight | 0 | 4 | 0.84 | 0.80 |
| Total Lanes | TotLanes | 2 | 15 | 6.47 | 2.66 |
| **Volatility Measures** | | | | | |
| **Speed** | | | | | |
| $S_{dev}$(m/s) | SpSD | 4.83 | 16.78 | 10.88 | 2.57 |
| $C_v$(%) | SpCov | 12.34 | 80.81 | 44.48 | 16.00 |
| $Q_{cv}$(%) | SpCovQ | 6.17 | 66.74 | 31.67 | 16.74 |
| $D_{mean}$(m/s) | SpMAD | 3.18 | 12.33 | 7.56 | 2.07 |
| $T_1$(%) | SP1SD | 11.35 | 60.28 | 28.74 | 13.30 |
| $T_2$(%) | SP2SD | 0.00 s | 11.31 | 3.63 | 2.90 |
| **Longitudinal acceleration** | | | | | |
| $S_{dev}$(m/s) | AccAxSD | 0.33 | 1.42 | 0.76 | 0.18 |
| $C_v$ of acceleration (%) | AccAxCov | 43.53 | 74.11 | 58.49 | 5.53 |
| $C_v$ of deceleration (%) | DecAxCov | 52.30 | 120.13 | 65.44 | 8.54 |
| $Q_{cv}$ of acceleration (%) | AccCovAxQ | 21.84 | 50.00 | 38.47 | 5.64 |
| $Q_{cv}$ of deceleration (%) | DecCovAxQ | 22.58 | 59.62 | 43.20 | 7.42 |
| $D_{mean}$(m/s) | AccAxMAD | 0.15 | 0.54 | 0.39 | 0.09 |
| $T_1$(%) | AccAx1SD | 6.50 | 35.87 | 23.44 | 4.71 |
| $T_2$(%) | AccAx2SD | 1.61 | 11.09 | 6.45 | 1.83 |
| **Lateral acceleration** | | | | | |
| $S_{dev}$(m/s) | AccAySD | 0.12 | 2.14 | 1.05 | 0.36 |
| $C_v$ of acceleration (%) | AccAyCov | 28.64 | 225.62 | 87.34 | 38.25 |
| $C_v$ of deceleration (%) | DecAyCov | 57.45 | 221.63 | 138.25 | 34.18 |
| $Q_{cv}$ of acceleration (%) | AccCovAyQ | 21.84 | 93.01 | 45.80 | 14.02 |
| $Q_{cv}$ of deceleration (%) | DecCovAyQ | 15.09 | 92.79 | 57.41 | 20.39 |
| $D_{mean}$(m/s) | AccAyMAD | 0.07 | 4.65 | 0.77 | 0.62 |
| $T_1$(%) | AccAy1SD | 0.00 | 39.57 | 4.80 | 6.10 |
| $T_1$(%) | AccAy2SD | 0.00 | 13.16 | 1.92 | 2.17 |
| **Yaw rate** | | | | | |
| $S_{dev}$(m/s) | SDYaw | 0.42 | 8.88 | 3.64 | 1.45 |
| $C_v$ of positive yaw rate (%) | YawPosCov | 0.22 | 3.13 | 1.64 | 0.54 |
| $C_v$ of negative yaw rate (%) | YawNegCov | 0.33 | 3.21 | 1.64 | 0.51 |
| $Q_{cv}$ of positive yaw rate (%) | YawPosCovQ | 0.08 | 0.92 | 0.57 | 0.20 |
| $Q_{cv}$ of negative yaw rate (%) | YawNegCovQ | 0.10 | 0.92 | 0.55 | 0.20 |
| $D_{mean}$(m/s) | YawMAD | 0.18 | 6.93 | 2.99 | 1.70 |
| $T_1$(%) | Yaw1SD | 0.00 | 0.40 | 0.10 | 0.08 |
| $T_1$(%) | Yaw2SD | 0.00 | 0.13 | 0.04 | 0.03 |

### 5.2. Tuning results

Fig. 4 visualizes the MAE of training and test set with incremental bandwidth and *maxmtry,* which are marked as BW in vertical and MM in horizontal respectively. The *maxmtry* is increased by five each time, and

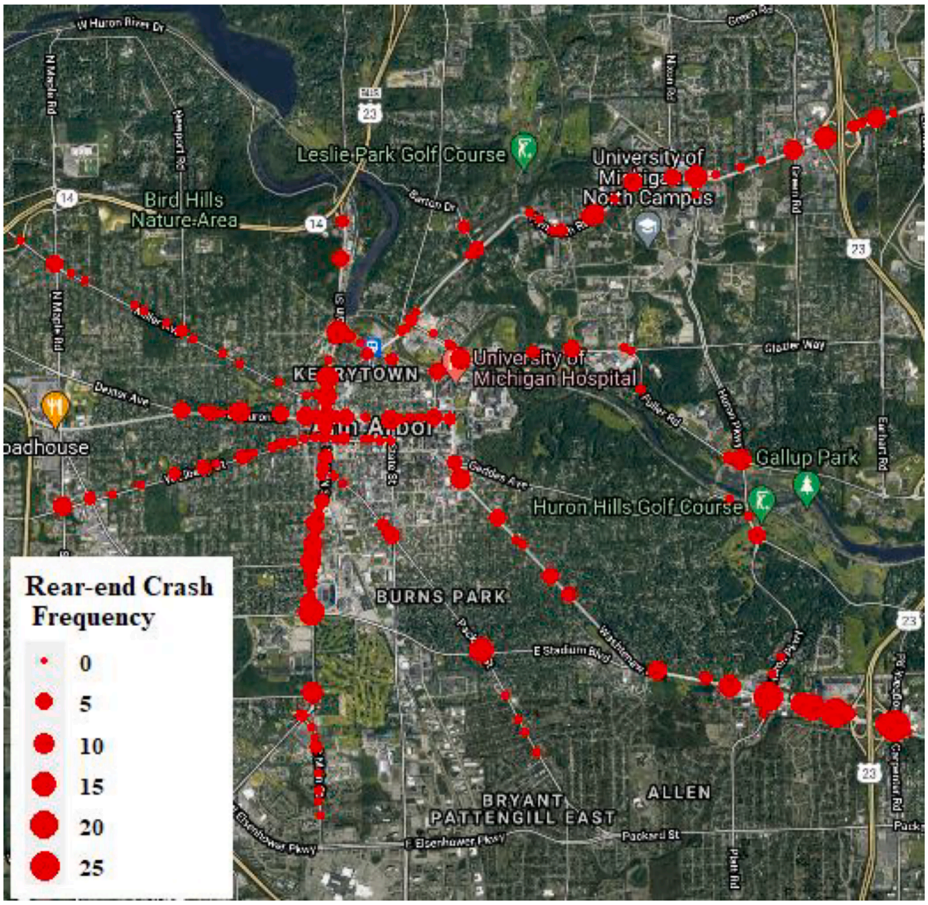


**Fig. 3.** Distribution of Rear-end Crash Frequency at Intersections in Ann Arbor, Michigan.

the bandwidth varies from 10 to 130 with increments of 10 for the adaptive kernel, while the bandwidth of the fixed kernel varies from 2,000 to 10,000, the ends of which are determined by the range of distance. The larger the bandwidth or the larger *maxmtry*, the more neighbors or predictors will be considered in the training model, leading to a more complex model. It should also be noticed that results tuned by this exhaustive search could miss the optimal parameter settings, whereas this issue could be mitigated by setting a higher resolution of bandwidth. The tuned parameters could be viewed as a near-optimal results. The last row of the table records the performance of the global RF, which considers all the neighbors. Notably, the training MAE of the global RF is much smaller than the test MAE for both kernels, suggesting that the global RF is significantly overfitted.

The purpose of drawing the heatmap is to help capture the performance pattern concerning bandwidth and *maxmtry* so that the optimal model can be determined for further prediction. To make the gradient color consistent, both adaptive and fixed kernels utilized the same color scale that is from green to red, corresponding with 2.3 to 3.5. Both color and numbers suggest that the training MAE and test MAE change inversely. For instance, the adaptive kernel has higher training MAE (yellow cells) than test MAE (green cells) when the bandwidth exceeds 90. For the fixed kernel models, low training MAE focuses on the area in which the bandwidth varies from 3500 to 6500, while the corresponding test MAE is sparsely distributed. Hereto, it's hard to find out the optimal model solely dependent on the heatmap. Therefore, the ratio of training MAE to test MAE (Eq. (5)) was computed for balancing the training and test MAE. As shown in the right-most tables in Fig. 4, the ratio of training MAE to test MAE varies around 1, thus dark green is used to depict the ratio close to 1, and red is used to illustrate the ratio far from 1. Comparing two kinds of models, the adaptive kernel model, with more light red cells, has more underestimated or overestimated cases than the fixed kernel model.

Furthermore, Fig. 5 demonstrates all the acceptable models within the threshold 0.01, 0.02, … and 0.1 for adaptive (left) and fixed kernel (right). Among selected models with the adaptive kernel, the test MAE decreased drastically as the ratio increased, and the best model achieves the lowest MAE (2.362) when bandwidth is 110 and *maxmtry* is 30, given a threshold of 0.1. Regarding fixed kernel models, the overall test MAE is smaller than the adaptive kernel, and the MAE varies more mildly. Correspondingly, the lowest test MAE (2.325) was obtained when bandwidth is 5500 and *maxmtry* is 35. Finally, comparing two models, the fixed model with bandwidth 5500, and *maxmtry* 35 outperformed other models with the threshold of 0.1. Compared to the global RF model with the same *maxmtry* condition, the performance is significantly improved by 9 % from 2.541 to 2.325. Besides, it should be noticed that once the threshold is changed, the corresponding optimal model might be changed. The performance of the proposed framework is only compared with the global random forest, which does not account for the spatial heterogeneity of predictors. Note that the comparison is not made between conventional spatial statistical models, e.g., GWR or GWPR. This is because feature pruning is needed for preventing multicollinearity in conventional statistical models, especially it is important to have the 40 predictors used in this study. Otherwise, they will generate misleading parameter results and even overfit the model. On the other hand, they cannot untangle non-linear relationships between outcome and predictors. However, the proposed GRF model attempts to retain all predictors by incorporating a parameter that controls the predictors for training local random forest and a model selection criterion to prevent overfitting or underfitting. Besides, the machine learning technique, embedded in a spatial context can facilitate uncovering the

**Adaptive Kernel**

Training MAE

| BW \ MM | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| 10 | 2.70 | 2.80 | 2.65 | 2.72 | 2.72 | 2.75 | 2.77 | 2.68 |
| 20 | 2.56 | 2.60 | 2.57 | 2.59 | 2.62 | 2.59 | 2.63 | 2.60 |
| 30 | 2.53 | 2.57 | 2.57 | 2.56 | 2.55 | 2.57 | 2.53 | 2.56 |
| 40 | 2.50 | 2.55 | 2.50 | 2.45 | 2.57 | 2.54 | 2.59 | 2.52 |
| 50 | 2.44 | 2.51 | 2.53 | 2.47 | 2.53 | 2.45 | 2.53 | 2.54 |
| 60 | 2.40 | 2.45 | 2.45 | 2.54 | 2.51 | 2.56 | 2.47 | 2.48 |
| 70 | 2.44 | 2.40 | 2.44 | 2.40 | 2.40 | 2.54 | 2.46 | 2.44 |
| 80 | 2.53 | 2.50 | 2.45 | 2.52 | 2.48 | 2.57 | 2.52 | 2.51 |
| 90 | 2.51 | 2.48 | 2.52 | 2.49 | 2.55 | 2.55 | 2.62 | 2.55 |
| 100 | 2.59 | 2.53 | 2.57 | 2.63 | 2.64 | 2.55 | 2.57 | 2.67 |
| 110 | 2.63 | 2.55 | 2.59 | 2.58 | 2.53 | 2.56 | 2.54 | 2.55 |
| 120 | 2.54 | 2.59 | 2.63 | 2.51 | 2.59 | 2.54 | 2.66 | 2.59 |
| 130 | 2.62 | 2.59 | 2.61 | 2.59 | 2.58 | 2.61 | 2.60 | 2.59 |
| Global RF | 1.11 | 1.13 | 1.14 | 1.15 | 1.12 | 1.06 | 1.04 | 1.12 |

Test MAE

| BW \ MM | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| 10 | 3.29 | 3.35 | 3.30 | 3.37 | 3.30 | 3.30 | 3.21 | 3.16 |
| 20 | 2.94 | 2.98 | 2.94 | 2.74 | 2.86 | 2.95 | 2.83 | 2.95 |
| 30 | 2.85 | 3.02 | 3.08 | 3.08 | 3.03 | 3.03 | 2.98 | 2.84 |
| 40 | 2.72 | 2.90 | 2.80 | 2.70 | 2.85 | 2.79 | 2.68 | 2.82 |
| 50 | 2.68 | 2.76 | 2.85 | 2.72 | 2.80 | 2.58 | 2.79 | 2.71 |
| 60 | 2.78 | 2.81 | 2.82 | 2.43 | 2.59 | 2.69 | 2.77 | 2.71 |
| 70 | 2.91 | 2.84 | 2.80 | 2.66 | 2.67 | 2.64 | 2.87 | 2.69 |
| 80 | 2.74 | 2.61 | 2.87 | 2.50 | 2.65 | 2.62 | 2.65 | 2.71 |
| 90 | 2.65 | 2.52 | 2.53 | 2.51 | 2.53 | 2.59 | 2.71 | 2.48 |
| 100 | 2.51 | 2.61 | 2.49 | 2.57 | 2.31 | 2.41 | 2.48 | 2.56 |
| 110 | 2.50 | 2.44 | 2.56 | 2.58 | 2.38 | 2.36 | 2.43 | 2.48 |
| 120 | 2.41 | 2.53 | 2.37 | 2.41 | 2.60 | 2.54 | 2.43 | 2.44 |
| 130 | 2.54 | 2.75 | 2.42 | 2.68 | 2.55 | 2.79 | 2.46 | 2.49 |
| Global RF | 2.77 | 2.64 | 2.43 | 2.50 | 2.49 | 2.56 | 2.54 | 2.42 |

Ratio

| BW \ MM | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| 10 | 0.82 | 0.84 | 0.80 | 0.81 | 0.82 | 0.83 | 0.86 | 0.85 |
| 20 | 0.87 | 0.87 | 0.88 | 0.95 | 0.91 | 0.88 | 0.93 | 0.88 |
| 30 | 0.89 | 0.85 | 0.84 | 0.83 | 0.84 | 0.85 | 0.85 | 0.90 |
| 40 | 0.92 | 0.88 | 0.89 | 0.91 | 0.90 | 0.91 | 0.97 | 0.89 |
| 50 | 0.91 | 0.91 | 0.89 | 0.91 | 0.90 | 0.95 | 0.91 | 0.94 |
| 60 | 0.86 | 0.87 | 0.87 | 1.04 | 0.97 | 0.95 | 0.89 | 0.91 |
| 70 | 0.84 | 0.85 | 0.87 | 0.90 | 0.90 | 0.97 | 0.86 | 0.90 |
| 80 | 0.92 | 0.96 | 0.85 | 1.01 | 0.94 | 0.98 | 0.95 | 0.93 |
| 90 | 0.95 | 0.98 | 1.00 | 0.99 | 1.01 | 0.98 | 0.97 | 1.03 |
| 100 | 1.03 | 0.97 | 1.03 | 1.02 | 1.14 | 1.06 | 1.03 | 1.04 |
| 110 | 1.05 | 1.04 | 1.01 | 1.00 | 1.06 | 1.08 | 1.04 | 1.03 |
| 120 | 1.05 | 1.02 | 1.11 | 1.04 | 1.00 | 1.00 | 1.09 | 1.06 |
| 130 | 1.03 | 0.94 | 1.08 | 0.97 | 1.01 | 0.93 | 1.06 | 1.04 |
| Global RF | 0.40 | 0.43 | 0.47 | 0.46 | 0.45 | 0.41 | 0.41 | 0.46 |

**Fixed Kernel**

Training MAE

| BW \ MM | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| 2,000 | 2.64 | 2.72 | 2.61 | 2.67 | 2.73 | 2.68 | 2.66 | 2.68 |
| 2,500 | 2.63 | 2.69 | 2.64 | 2.63 | 2.70 | 2.62 | 2.64 | 2.66 |
| 3,000 | 2.64 | 2.63 | 2.58 | 2.69 | 2.65 | 2.64 | 2.66 | 2.56 |
| 3,500 | 2.41 | 2.50 | 2.54 | 2.48 | 2.49 | 2.52 | 2.53 | 2.48 |
| 4,000 | 2.41 | 2.42 | 2.50 | 2.48 | 2.45 | 2.47 | 2.51 | 2.43 |
| 4,500 | 2.36 | 2.42 | 2.43 | 2.50 | 2.36 | 2.41 | 2.39 | 2.49 |
| 5,000 | 2.42 | 2.39 | 2.47 | 2.36 | 2.42 | 2.44 | 2.41 | 2.33 |
| 5,500 | 2.45 | 2.49 | 2.45 | 2.39 | 2.45 | 2.41 | 2.48 | 2.53 |
| 6,000 | 2.40 | 2.44 | 2.42 | 2.42 | 2.43 | 2.40 | 2.50 | 2.42 |
| 6,500 | 2.51 | 2.46 | 2.48 | 2.44 | 2.47 | 2.50 | 2.49 | 2.51 |
| 7,000 | 2.53 | 2.54 | 2.55 | 2.49 | 2.46 | 2.47 | 2.47 | 2.48 |
| 7,500 | 2.59 | 2.52 | 2.56 | 2.55 | 2.56 | 2.46 | 2.56 | 2.51 |
| 8,000 | 2.65 | 2.66 | 2.60 | 2.63 | 2.62 | 2.55 | 2.65 | 2.63 |
| 8,500 | 2.62 | 2.61 | 2.63 | 2.62 | 2.64 | 2.56 | 2.68 | 2.68 |
| 9,000 | 2.64 | 2.76 | 2.66 | 2.62 | 2.62 | 2.57 | 2.67 | 2.65 |
| 9,500 | 2.62 | 2.70 | 2.66 | 2.58 | 2.61 | 2.56 | 2.65 | 2.62 |
| 10,000 | 2.67 | 2.70 | 2.67 | 2.61 | 2.61 | 2.55 | 2.63 | 2.62 |
| Global RF | 1.11 | 1.13 | 1.14 | 1.15 | 1.12 | 1.06 | 1.04 | 1.12 |

Test MAE

| BW \ MM | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| 2,000 | 2.67 | 2.71 | 2.60 | 2.63 | 2.73 | 2.72 | 2.67 | 2.74 |
| 2,500 | 2.80 | 2.51 | 2.69 | 2.54 | 2.67 | 2.67 | 2.52 | 2.55 |
| 3,000 | 2.61 | 2.63 | 2.50 | 2.46 | 2.48 | 2.57 | 2.65 | 2.66 |
| 3,500 | 2.68 | 2.58 | 2.54 | 2.46 | 2.49 | 2.55 | 2.55 | 2.58 |
| 4,000 | 2.51 | 2.60 | 2.62 | 2.54 | 2.36 | 2.40 | 2.53 | 2.62 |
| 4,500 | 2.50 | 2.44 | 2.56 | 2.58 | 2.38 | 2.36 | 2.43 | 2.48 |
| 5,000 | 2.56 | 2.46 | 2.48 | 2.46 | 2.36 | 2.36 | 2.38 | 2.39 |
| 5,500 | 2.37 | 2.39 | 2.44 | 2.54 | 2.49 | 2.51 | 2.32 | 2.38 |
| 6,000 | 2.68 | 2.47 | 2.43 | 2.69 | 2.52 | 2.61 | 2.68 | 2.39 |
| 6,500 | 2.54 | 2.38 | 2.37 | 2.57 | 2.63 | 2.50 | 2.71 | 2.50 |
| 7,000 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| 7,500 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| 8,000 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| 8,500 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| 9,000 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| 9,500 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| 10,000 | 2.54 | 2.48 | 2.43 | 2.46 | 2.59 | 2.52 | 2.60 | 2.63 |
| Global RF | 2.77 | 2.64 | 2.43 | 2.50 | 2.49 | 2.56 | 2.54 | 2.42 |

Ratio

| BW \ MM | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| 2,000 | 0.99 | 1.00 | 1.00 | 1.02 | 1.00 | 0.98 | 1.00 | 0.98 |
| 2,500 | 0.94 | 1.07 | 0.98 | 1.03 | 1.01 | 0.98 | 1.05 | 1.04 |
| 3,000 | 1.01 | 1.00 | 1.03 | 1.09 | 1.07 | 1.03 | 1.00 | 0.96 |
| 3,500 | 0.90 | 0.97 | 1.00 | 1.01 | 1.00 | 0.99 | 0.99 | 0.96 |
| 4,000 | 0.96 | 0.93 | 0.95 | 0.98 | 1.04 | 1.03 | 0.99 | 0.93 |
| 4,500 | 0.94 | 0.99 | 0.95 | 0.97 | 0.99 | 1.02 | 0.98 | 1.00 |
| 5,000 | 0.95 | 0.97 | 1.00 | 0.96 | 1.02 | 1.03 | 1.01 | 0.98 |
| 5,500 | 1.03 | 1.04 | 1.00 | 0.94 | 0.98 | 0.96 | 1.07 | 1.07 |
| 6,000 | 0.89 | 0.99 | 1.00 | 0.90 | 0.96 | 0.92 | 0.94 | 1.01 |
| 6,500 | 0.99 | 1.03 | 1.05 | 0.95 | 0.94 | 1.00 | 0.92 | 1.01 |
| 7,000 | 0.99 | 1.02 | 1.05 | 1.01 | 0.95 | 0.98 | 0.95 | 0.94 |
| 7,500 | 1.02 | 1.02 | 1.05 | 1.04 | 0.99 | 0.98 | 0.98 | 0.95 |
| 8,000 | 1.04 | 1.07 | 1.07 | 1.07 | 1.02 | 1.01 | 1.02 | 1.00 |
| 8,500 | 1.03 | 1.05 | 1.08 | 1.07 | 1.02 | 1.02 | 1.03 | 1.02 |
| 9,000 | 1.04 | 1.11 | 1.09 | 1.06 | 1.01 | 1.02 | 1.03 | 1.01 |
| 9,500 | 1.03 | 1.09 | 1.09 | 1.05 | 1.01 | 1.02 | 1.02 | 0.99 |
| 10,000 | 1.05 | 1.09 | 1.10 | 1.06 | 1.01 | 1.01 | 1.01 | 0.99 |
| Global RF | 0.40 | 0.43 | 0.47 | 0.46 | 0.45 | 0.41 | 0.41 | 0.46 |

**Fig. 4.** Model Results with respect to Bandwidth (BW) and *maxmtry* (MM).

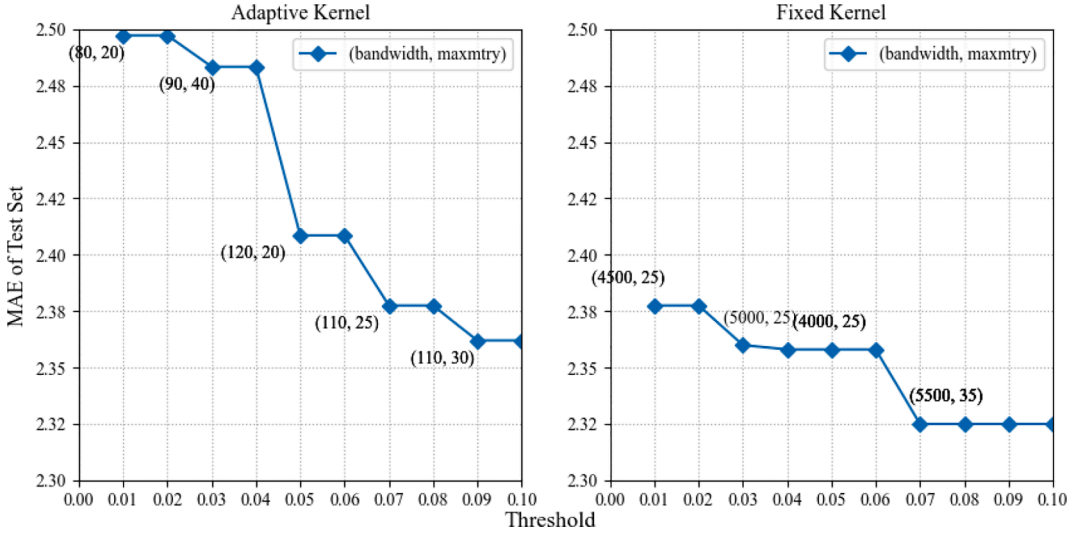


**Fig. 5.** Optimal Model under Different Thresholds.

complex relationship between predictors and outcomes. Notably, the machine learning technique is difficult to compare with conventional spatial models partly because such techniques retain all explanatory variables, whereas conventional GWR models prune variables to avoid collinearity. Overall, this study demonstrated a framework for estimating machine learning geographically. This framework can be extended to other machine learning models.

### 5.3. Discussion

#### 5.3.1. Variable importance

Average local variable importance for each predictor was computed from the optimal training model at each trained intersection, using tuned parameter bandwidth equal to 5500 m, and *maxmtry* equal to 35. Thus, Fig. 6 gives us overall insights on the most important factors that correlate to the rear-end crash frequency for the city of Ann Arbor. As the figure shows, AADT on a minor road (*AADTMin*) from the Roadway Traffic Characteristics group makes the highest contribution to the rear-end crash, followed by the AADT on the major road (*AADTMaj*),

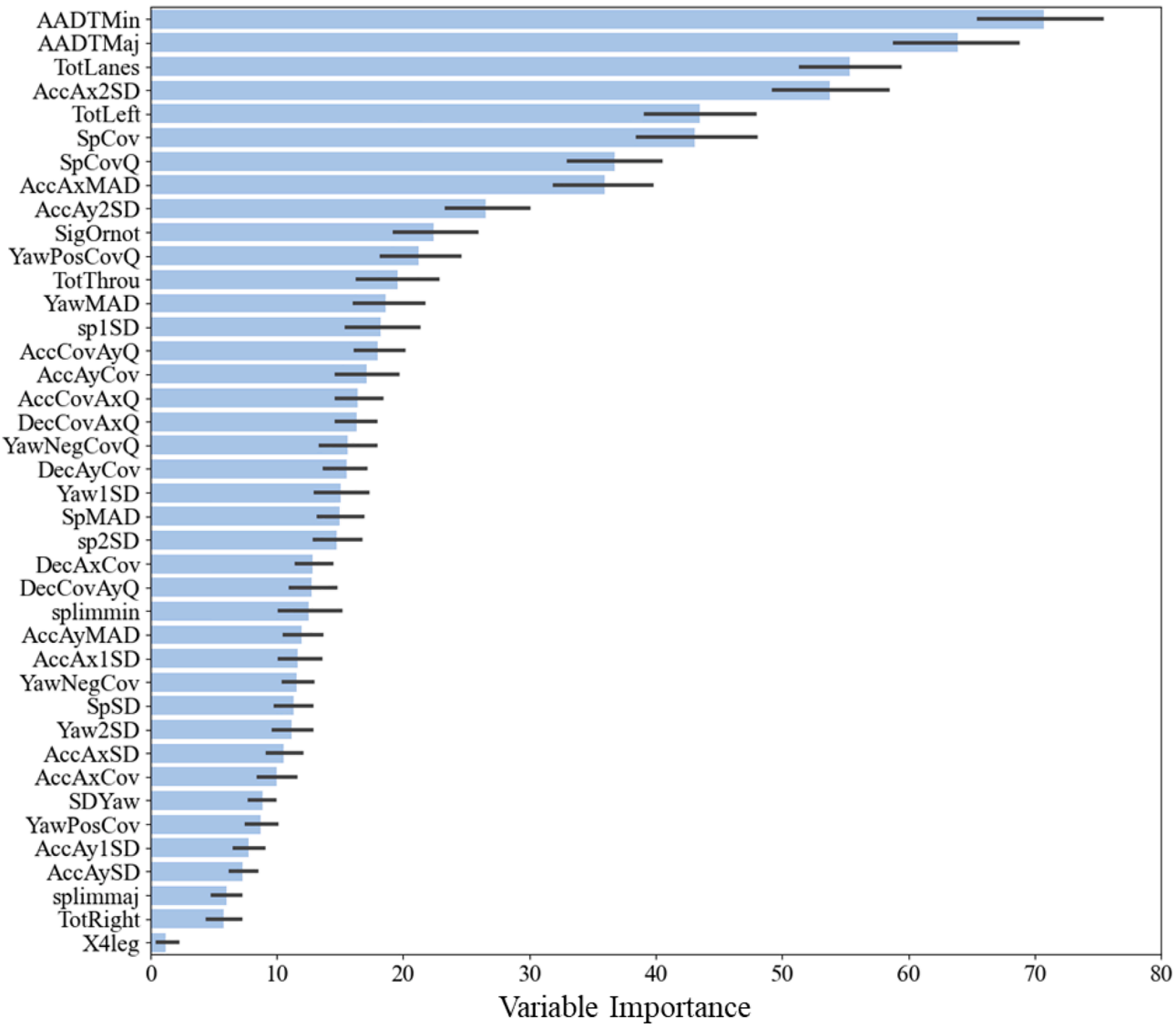


**Fig. 6.** Variable Importance of Fixed Kernel (bandwidth = 5500 m, and *maxmtry* = 35).

indicating that the rear-end crash is more likely to occur on intersections connecting minor roads than major roads. This result is in agreement with the findings of Arvin et al. (6). In addition, we noticed that the driving volatilities such as the percentage of extreme points beyond 2 standard deviations of acceleration in the longitudinal direction (*AccAx2SD*) contribute to over 50 % of rear-end crash frequency. It is logical that the rear-end crash is more likely to happen when the longitudinal acceleration varies too much. As a result, the variable importance can be further used to provide alerts for drivers when such a condition occurs. Notably, these key factors are also in agreement with significant predictors found in previous studies (Arvin et al., 2019; Mohammadnazar et al., 2022).

However, different from the statistical models, the variable importance could only give a sense of the relative contribution of factors to rear-end crashes and cannot explain whether the factors are positively or negatively associated with rear-end crashes. The driving volatility shows the variation of instantaneous driving behavior and status, most of them are positively correlated to the crash frequency from the previous studies (Arvin et al., 2019; Kamrani et al., 2018; Wali et al., 2018). Therefore, rear-end crash risk can be reduced by controlling the variation of driving behaviors in practice.

### 5.3.2. *Spatial variation of local effect*

Fig. 7 illustrates the spatial distribution of the variable importance for the most important predictor in five categories. Redder color indicates higher variable importance while greener color suggests lower importance. Notably, the uneven distribution of point color highlights the evident spatial dependency of rear-end crash frequency. For example, Fig. 7(a) shows that the AADT on minor roads (*AADTMin*) of the downtown area has a larger effect on rear-end crash frequency, compared to the peripheral arterials where the intersections are less dense. Especially, the small points on the northeast arterial interacting with minor roads have a relatively small impact on rear-end crash frequency. This suggests that traffic exposure in city core areas have a larger marginal effect on rear-end crashes than in city peripheral areas. On the other hand, *AADTmin* is not the leading factor in some peripheral areas. Besides, as depicted in Fig. 7(b), redder points are concentrated on the corridors that connect the outer highways except for two southeast direction corridors, while in the downtown area, the impact of coefficient of variation of speed (*SpCov*) is smaller. It is worth noting that the scale local effect on the same corridors is relatively stationary while the spatial dependency of *SpCov* is likely to occur between different corridors. In contrast, Fig. 7(c) shows the evident non-stationary effect of % of extreme points beyond two standard deviations of acceleration

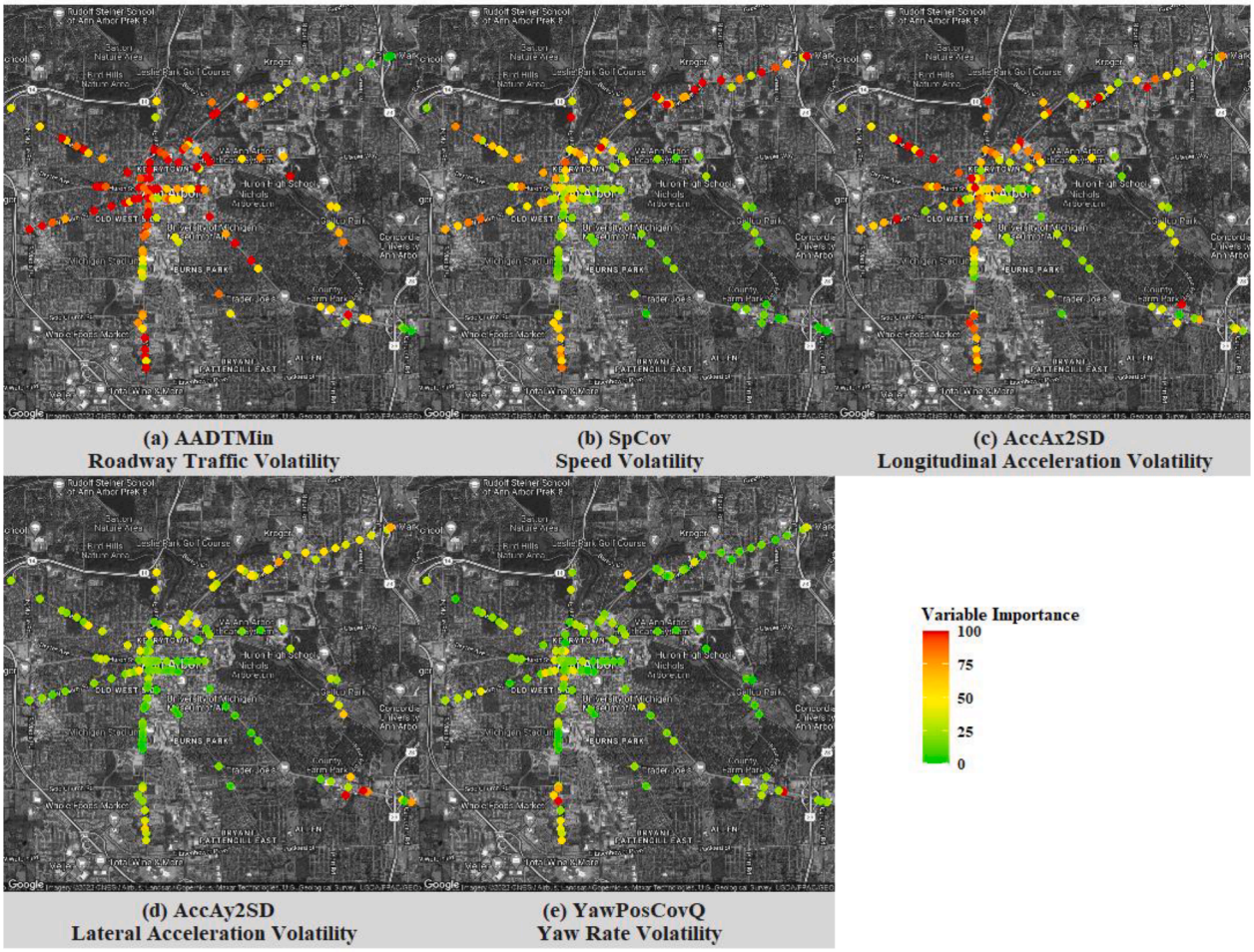


**Fig. 7.** Geographical Distribution of Variable Importance of the Most Important Predictor in Each Group.

(*AccAx2SD*) along the different sections of corridors, as the point size varies significantly. In general, *AccAx2SD* produced a larger effect on rear-end crash frequency at the northwest of the downtown area than other areas, indicating that longitudinal extreme acceleration in driving could lead to rear-end crash with a higher probability in the northwest downtown area. For the same indicator from the lateral side (*AccAy2SD*) shown in Fig. 7(d), the difference of local effect among different corridors is evident. Lateral extreme acceleration generates a higher impact on northeast corridors, followed by the corridor south to the downtown area. Regarding the yaw rate volatility, the association of quantile coefficient of variation of a positive yaw rate (*YawPosCovQ*) with rear-end crash frequency behaves more strongly in the downtown area than peripheral areas, as Fig. 7(e) shows. This might be because the denser street network in downtown areas makes the driver frequently manipulate their steering wheel, leading to higher risks of rear-end crashes. Hence, the yaw rate variation in downtown areas tend to have larger impact on rear-end crashes than other areas.

## 6. Conclusions

This study developed a geographical random forest model for rear-end crash frequency prediction by harnessing large-scale CV data. However, both data and model have some limitations. Regarding data, the CV drivers in the SPMD project are not necessarily representative of all drivers. A more representative sample is desirable. Also, only two months of available CV data were exploited to explore the association with rear-end crash frequency; future work could seek to test a larger dataset. Concerning the model, the GRF only captures the relative contribution of each predictor to rear-end crash frequency, and it cannot explain the positive or negative correlation. For instance, AADT on minor roads is the most important factor to rear-end crash frequency, but we don't know whether AADT increases the rear-end crash frequency or decreases. Finally, iterative fashion over bandwidth and *maxmtry* adds to the complexity of models, leading to a great computation time. This is a trade-off between improved performance and extra computation time, which should be balanced in the future.

This study contributes by harnessing a unique database that combines high-resolution vehicle motion data integrated with crash frequency, traffic flow, and road inventory data. Several driving volatility measures were created and applied to quantify variations in driving behavior at intersections. New AI techniques are applied because using conventional statistical methods cannot retain all predictors due to multicollinearity issues. While most machine learning methods outcompete statistical methods in handling high correlation within predictors and in terms of predictive accuracy, they are not able to capture the spatial heterogeneity of predictors. To this end, this study develops a geographical machine learning technique for predicting the frequency of rear-end crashes at intersections. The proposed model GRF originates from the RF model and inherits the merit of geographical modeling, accounting for spatial heterogeneity.

Regarding methodology, GRF is a disaggregation of RF in geographical space. The use of Dijkstra's shortest path to measure the distance between intersections is more reasonable for a radial road network. To estimate the local RF model, the bandwidth generated from adaptive and fixed kernels is applied to select the neighbors for training. Furthermore, the maximum number of features used to split a tree node

(i.e., *maxmtry*) is applied, which retains all parameters as well as avoids potential overfitting. To accelerate the selection of a combination of bandwidth and *maxmtry*, the ratio of training set error to test set error is introduced as the criterion to assess model goodness of fit. The optimal model is obtained by taking the minimum test MAE, given a threshold of ratio. Thus, the robustness and accuracy of prediction results can be controlled in this framework.

Focusing on the results, the average variable importance computed by the percent of increased MSE quantifies the contribution of each predictor of rear-end crashes, which can be further used for the identification of crucial factors. Moreover, the unbalanced local variable importance suggests that the rear-end crashes have a strong spatial dependency. For example, the number of acceleration points lying beyond two standard deviations in the lateral direction in the central area of Ann Arbor is not as important as the peripheral area. This indicates that excessive lateral acceleration in the peripheral area is more likely to lead to rear-end crashes than in the city central area. This finding could help transportation agencies undertake differential improvements. For predictive models, robustness and accuracy are very important. Compared to the global RF models, the optimal GRF not only reduces the overfitting issue compared to the global RF model but also improves the prediction accuracy by 9 % in terms of MAE.

Ultimately, this study can provide several benefits:

1) This research can be applied to improve the prediction of rear-end crash frequency, and it can be expanded to cover other types of crashes.
2) Upon expansion of this research, transportation agencies can select relevant countermeasures based on variable importance for at-risk intersections.
3) When connected vehicles are widely deployed, the models developed in this study will become more important as they can proactively quantify potential crash risks resulting from improper driving behaviors.

## 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.

## Data availability

The data are publicly available at the website referenced in the paper.

## Acknowledgements

This study is supported by the US Department of Transportation (grant number: 69A3551747113) through the national university transportation center named Collaborative Sciences Center for Road Safety, University of North Carolina-Chapel Hill, and University of Tennessee, Knoxville.

## References

Abdel-Aty, M., Haleem, K., 2011. Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accid. Anal. Prev. 43 (1), 461–470.
Abdulhafedh, A., 2016. Crash frequency analysis. J. Transp. Technol. 6 (04), 169.
Anastasopoulos, P., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.
Arvin, R., M. Kamrani, and A. J. Khattak. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. Acc. Anal. Prevent., Vol. 132, 2019, pp. 105226 %@ 100001-104575.
Arvin, R., A. J. Khattak, and H. Qi. Safety critical event prediction through unified analysis of driver and vehicle volatilities: Application of deep learning methods. Acc. Anal. Prevent., Vol. 151, 2021, pp. 105949 %@ 100001-104575.
Arvin, R., Kamrani, M., Khattak, A.J., 2019. How instantaneous driving behavior contributes to crashes at intersections: Extracting useful information from connected vehicle message data. Accid. Anal. Prev. 127, 118–133.
Auret, L., and C. Aldrich. Empirical comparison of tree ensemble variable importance measures. Chemometr. Intell. Laborat. Syst., Vol. 105, No. 2, 2011, pp. 157-170 %@ 0169-7439.
Brown, C. E. Coefficient of variation. In: Applied multivariate statistics in geohydrology and related sciences, Springer, 1998. pp. 155-157.
Brunsdon, C., S. Fotheringham, and M. Charlton. Geographically weighted regression. J. R. Stat. Soc.: Ser. D (Statistician), vol. 47, No. 3, 1998, pp. 431-443 %@ 0039-0526.
Chang, L.-Y. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci., vol. 43, No. 8, 2005, pp. 541-557 %@ 0925-7535.
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., 2015. Xgboost: extreme gradient boosting. R package version (4-2), 1–4.
Bonett, D.G., 2006. Confidence interval for a coefficient of quartile variation. Comput. Stat. Data Anal. 50 (11), 2953–2957.
Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accid. Anal. Prev. 70, 320–329.
Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2019. Exploring microscopic driving volatility in naturalistic driving environment prior to involvement in safety critical event Concept of event-based driving volatility. Acc. Anal. Prevent. 132.
Fu, C., and T. Sayed. Random parameters Bayesian hierarchical modeling of traffic conflict extremes for crash estimation. Acc. Anal. Prevent., Vol. 157, 2021, pp. 106159 %@ 100001-104575.
Georganos, S., T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. o. Wolff, and S. Kalogirou. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int., vol. 36, No. 2, 2021, pp. 121-136 %@ 1010-6049.
Gromping, U. Variable importance assessment in regression: linear regression versus random forest. Am. Statist., vol. 63, No. 4, 2009, pp. 308-319 %@ 0003-1305.
Haleem, K., Abdel-Aty, M., Santos, J., 2010. Multiple applications of multivariate adaptive regression splines technique to predict rear-end crashes at unsignalized intersections. Transp. Res. Rec. 2165 (1), 33–41.
Hastie, T., R. Tibshirani, and J. Friedman. Random forests. In *The elements of statistical learning*, Springer, 2009. pp. 587-604.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
Henclewood, D., M. Abramovich, and B. Yelchuru. Safety pilot model deployment–one day sample data environment data handbook. *Research and Technology Innovation Administration. Research and Technology Innovation Administration, US Department of Transportation, McLean, VA*, 2014.
Hilbe, J.M., 2011. Negative binomial regression. Cambridge University Press.
Hoseinzadeh, N., R. Arvin, A. J. Khattak, and L. D. Han. Integrating safety and mobility for pathfinding using big data generated by connected vehicles. J. Intell. Transp. Syst., vol. 24, No. 4, 2020, pp. 404-420 %@ 1547-2450.
Huber, P.J., 2004. Robust statistics. John Wiley & Sons.
FHWA. *Intersection Safety*. Federal Highway Administration. https://highways.dot.gov/research/research-programs/safety/intersection-safety. Accessed July 31, 2021.
James, G., D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.
Kamrani, M., Arvin, R., Khattak, A.J., 2018. Extracting Useful Information from Basic Safety Message Data: An Empirical Study of Driving Volatility Measures and Crash Frequency at Intersections. Transp. Res. Rec.: J. Transp. Res. Board 2672 (38), 290–301.
Katrakazas, C., A. Theofilatos, M. A. Islam, E. Papadimitriou, L. Dimitriou, and C. Antoniou. Prediction of rear-end conflict frequency using multiple-location traffic parameters. Acc. Anal. Prevent., vol. 152, 2021, pp. 106007 %@ 100001-104575.
Khattak, A.J., Wali, B., 2017. Analysis of volatility in driving regimes extracted from basic safety messages transmitted between connected vehicles. Transp. Res. Part C: Emerg. Technol. 84, 48–73.
Krueger, R., Bansal, P., Buddhavarapu, P., 2020. A new spatial count data model with Bayesian additive regression trees for accident hot spot identification. Accid. Anal. Prev. 144, 105623.
Kwon, J., P. Varaiya, and A. Skabardonis. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation. Transp. Res. Rec., Vol. 1856, No. 1, 2003, pp. 106-117 %@ 0361-1981.
Guo, L., Z. Ma, and L. Zhang. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Can. J. Forest Res., vol. 38, No. 9, 2008, pp. 2526-2534 %@ 0045-5067.
Liu, J., Khattak, A.J., 2016. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. Transp. Res. Part C: Emerg. Technol. 68, 83–100.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transp. Res. Part A: Policy Practice 44 (5), 291–305.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Software 28 (1), 1548–7660.
Mohammadi, M. A., V. A. Samaranayake, and G. H. Bham. Crash frequency modeling using negative binomial models: An application of generalized estimating equation to longitudinal data. Anal. Methods Acc. Res., Vol. 2, 2014, pp. 52-69 %@ 2213-6657.

Naznin, F., G. Currie, D. Logan, and M. Sarvi. Application of a random effects negative binomial model to examine tram-involved crash frequency on route sections in Melbourne, Australia. Acc. Anal. Prevent., Vol. 92, 2016, pp. 15-21 %@ 0001-4575.

Mohammadnazar, Amin, Patwary, Latif, Moradloo, Nastaran, Arvin, Ramin, Khattak, Asad, 2022. Incorporating driving volatility measures in safety performance functions: Improving safety at signalized intersections. Accident Analysis and Prevention 178. https://doi.org/10.1016/j.aap.2022.106872.

Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest. Springer, pp. 154–168.

Perner, P. Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012, Proceedings. Springer, 2012.

Polikar, R. Ensemble learning. In *Ensemble machine learning*, Springer, 2012. pp. 1-34.

Pu, Z., Z. Li, R. Ke, X. Hua, and Y. Wang. Evaluating the Nonlinear Correlation between Vertical Curve Features and Crash Frequency on Highways Using Random Forests. J. Transp. Eng., Part A: Syst., vol. 146, No. 10, 2020.

Quevedo, R. P., D. A. Maciel, T. D. T. Uehara, M. Vojtek, C. D. Renno, B. Pradhan, J. Vojtekova, and Q. B. Pham. Consideration of spatial heterogeneity in landslide susceptibility mapping using geographical random forest model. Geocarto Int., 2021, pp. 1010-6049.

RcolorBrewer, S., Liaw, M.A., 2018. Package "randomForest". University of California, Berkeley, Berkeley, CA, USA.

Tang, J., W. Yin, C. Han, X. Liu, and H. Huang. A random parameters regional quantile analysis for the varying effect of road-level risk factors on crash rates. Anal. Methods Acc. Res., Vol. 29, 2021, pp. 102213-106657.

Tang, J., Gao, F., Liu, F., Han, C., Lee, J., 2020. Spatial heterogeneity analysis of macro-level crashes using geographically weighted Poisson quantile regression. Accid. Anal. Prev. 148, 105833.

Wali, B., Khattak, A.J., Bozdogan, H., Kamrani, M., 2018. How is driving volatility related to intersection safety? A Bayesian heterogeneity-based analysis of instrumented vehicles data. Transp. Res. Part C: Emerg. Technol. 92, 504–524.

Wali, B., Khattak, A.J., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Analytic Methods Acc. Res.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accid. Anal. Prev. 38 (6), 1137–1150.

Wang, C., Xie, Y., Huang, H., Liu, P., 2021. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. Accid. Anal. Prev. 157, 106157.

Wikipedia. *Dijkstra's algorithm*. https://en.wikipedia.org/wiki/Dijkstra%27s_algorithm. Accessed July 31, 2021.

Willmott, C. J., and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res., vol. 30, No. 1, 2005, pp. 79-82 %@ 0936-0577X.

Wu, P., Meng, X., Song, L., 2019. A novel ensemble learning method for crash prediction using road geometric alignments and traffic data. J. Transp. Saf. Secur. 12 (9), 1128–1146.

Xie, Y., and Y. Zhang. Crash frequency analysis with generalized additive models. Transp. Res. Rec., Vol. 2061, No. 1, 2008, 0361–1981.

Xu, P., H. Zhou, and S. C. Wong. On random-parameter count models for out-of-sample crash prediction: Accounting for the variances of random-parameter distributions. Acc. Anal. Prevent., Vol. 159, 2021, pp. 106237 %@ 100001-104575.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25.

Zhang, Z., X. Li, J. Liu, X. Fu, C. Yang, and S. L. Jones. Localized Safety Performance Functions for Rural 3-Leg Stop-Controlled Intersections in Alabama. In, 2021.

Zhang, X., Waller, S.T., Jiang, P., 2019. An ensemble machine learning-based modeling framework for analysis of traffic crash frequency. Comput.-Aided Civ. Infrastruct. Eng. 35 (3), 258–276.