Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Safety analytics at a granular level using a Gaussian process modulated renewal model: A case study of the COVID-19 pandemic

Yiyuan Lei, Kaan Ozbay, Kun Xie

C2SMART Center, Department of Civil and Urban Engineering, Tandon School of Engineering, New York University, 6 MetroTech Center, 4th Floor, Brooklyn, NY 11201, USA
Department of Civil and Environmental Engineering, Old Dominion University, 129C Kaufman Hall, Norfolk, VA 23529, USA

ABSTRACT

With the advance of intelligent transportation system technologies, contributing factors to crashes can be obtained in real time. Analyzing these factors can be critical in improving traffic safety. Despite many crash models having been successfully developed for safety analytics, most models associate crash observations and contributing factors at the aggregate level, resulting in potential information loss. This study proposes an efficient Gaussian process modulated renewal process model for safety analytics that does not suffer from information loss due to data aggregations. The proposed model can infer crash intensities in the continuous-time dimension so that they can be better associated with contributing factors that change over time. Moreover, the model can infer non-homogeneous intensities by relaxing the independent and identically distributed (i.i.d.) exponential assumption of the crash intervals. To demonstrate the validity and advantages of this proposed model, an empirical study examining the impacts of the COVID-19 pandemic on traffic safety at six interstate highway sections is performed. The accuracy of our proposed renewal model is verified by comparing the areas under the curve (AUC) of the inferred crash intensity function with the actual crash counts. Residual box plot shows that our proposed models have lower biases and variances compared with Poisson and Negative binomial models. Counterfactual crash intensities are then predicted conditioned on exogenous variables at the crash time. Time-varying safety impacts such as bimodal, unimodal, and parabolic patterns are observed at the selected highways. The case study shows the proposed model enables safety analytics at a granular level and provides a more detailed insight into the time-varying safety risk in a changing environment.

1. Introduction

For traffic safety analytics, contributing factors to crashes such as weather conditions, traffic volumes, and traffic speed are usually collected at the aggregate level. Crash events, however, are discrete in the time dimension. The incompatibility between aggregated covariates and discrete crash data can result in problematic analysis when they are associated without care. Such incompatibility is known as the fundamental dilemma in the traffic safety analysis (Chang and Jovanis, 1990). A common approach to avoid incompatibility is to model crash occurrences by aggregating crash events per month, quarter, or year. Continuous contributing factors are then usually represented by their averages over the time intervals. A limitation of this approach is that aggregations can lead to information loss and reduced sample size (Guo, 2010; Usman et al., 2011). Depending on the level of aggregations, datasets that are further classified into refined subgroups are considered disaggregate data. For example, aggregations by each year can be considered as disaggregate data with respect to data aggregated per three years. Studies show that using disaggregate data can provide a more accurate parameter estimation (Sengupta et al., 2021) and a smaller standard error (Bae et al., 2021) than aggregate one. The granular level is reached when disaggregate crash datasets are broken into the time of the crash. As intelligent transportation systems (ITS) are being more widely deployed, some safety-related factors can now be obtained continuously through advanced data management technologies. Few existing crash models, however, associate crash observations and contributing factors at the granular level, to reduce information loss due to aggregation (homogeneous intensities at aggregated intervals fail to reveal fluctuations).

At the aggregate level, traditional Poisson regression models
which estimate the average crash occurrences in unit time or space are known to be prone to over-dispersion. To address over-dispersion caused by inequalities between mean and variance or excessive zero counts, models such as mixtures of negative binomial regression (or Poisson-Gamma mixture) models (Lord et al., 2005; Zou et al., 2013; 2014), and zero-inflated regression models (Yan et al., 2012; Hall and Tarko, 2019) are employed for crash analysis. These models study the crash intensities at the aggregated time intervals. Instead of studying the crashes at the aggregate level, survival analysis models (or duration models) study the crash occurrences as a sequential process (Jovanis and Chang, 1989; Chang and Jovanis, 1990; Shankar et al., 2008; Chen and Guo, 2016; Xie et al., 2019) at the disaggregate level. Jovanis and Chang (1989) discussed the advantages of using survival theory principles to combine discrete crash data and aggregate exposure data. The widely used survival models such as Cox proportional hazard (semi-parametric) models, however, rely on the assumptions of proportional hazard functions over time, which can be hard to satisfy in practice. Xie et al. (2019) proposed a Bayesian survival analysis model which accounted for the crash hazard functions (Hensher and Manering, 1994; Kalbfleisch and Prentice, 2011) during the consecutive crash times across sites. While this study modeled individual crashes, covariates such as traffic volumes, however, were aggregated during the crash intervals. Since crash events can be rare by nature, the duration between two crashes tends to be long. Aggregate covariates measuring the variations coarsely in a long interval can only capture the aggregated crash occurrences which are spatiotemporally correlated (Lord and Shankar et al., 2008; Chen and Guo, 2016; Xie et al., 2019) at the spatiotemporal scale. Many researchers applied hierarchical models (Park and Lord, 2007; El-Basyouny and Sayed, 2009; Usman et al., 2012; Demirogluk and Ozbay, 2014; Xie et al., 2014; Cui and Xie, 2021) to address the spatiotemporal correlations. For example, Demirogluk and Ozbay (2014) proposed a hierarchical Bayesian model to estimate the spatial correlations using traffic exposure data aggregated at the county level. Usman et al. (2012) applied a multi-level model for crash analysis during storm events using hourly aggregated contributing factors, which addressed correlations within storm events. While these hierarchical models can perform crash analysis well without relying on the i.i.d. assumption by accounting for the spatiotemporal correlations, time-varying crash patterns can be hard to capture. As an alternative approach to relax the i.i.d. assumption and capture the non-homogeneous arrival rates at continuous time, Lasko (2014) proposed a Gaussian-process modulated renewal process (GPMPR) model that drops the i.i.d. assumption (Lasko, 2014). Let the intensity function represent the number of crashes per unit time as \( \lambda(t) \). Denote \( \lambda(t_i) \) as the number of events from 0 to \( t_i \), which can be obtained by the integration of intensity function over time:

\[
\lambda(t_i) = \int_0^{t_i} \lambda(t) dt
\]

In order to obtain a smooth function of \( \lambda(t) \), given two crash occurrence time \( t_i \) and \( t_j \), a Gaussian process with zero mean and squared exponential covariance function \( \Sigma \) is put on the logarithmic intensity function \( f(t) \):

\[
\log(\lambda(t)) = f(t) \sim \mathcal{N}(0, \Sigma(t_i, t))
\]

where a squared exponential matrix \( \Sigma \) with signal variance \( \sigma^2 \) and length scale \( l \) is calculated as:

\[
\Sigma(t_i, t_j) = \sigma^2 e^{-\frac{(t_i-t_j)^2}{2l^2}}
\]

To this end, the general form of crash events’ probability density function \( f_i \) for Bayesian inference can be described as:

\[
f_i(T; \alpha, \lambda(t)) = \frac{1}{\Gamma(\alpha)} \prod_{t_i} \lambda(t_i)^{a_i-1} e^{-\lambda(t_i)}
\]

where \( \Gamma \) stands for the gamma function, \( \alpha \) stands for the shape parameter. Depending on the value of \( \alpha \) with respect to 1, bursty events, memoryless Poisson events and regular events can be modelled (Lasko, 2014). The warped interval \( \lambda(t_i) = \Lambda(t_i) - \Lambda(t_{i-1}) \) is modeled by homogeneous gamma process with shape \( \alpha \), inverse scale parameter \( \beta \), and probability density function as:

\[
f_i(\omega_i; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \omega_i^{\alpha-1} e^{-\beta \omega_i}
\]

where the posterior distributions of intensity function \( \lambda(t) \), shape parameter \( \alpha \), signal variance and length scales are inferred by the Markov Chain Monte Carlo (MCMC) simulation. The objective is to maximize the likelihood equations \( f_i(T; \alpha, \lambda(t_i)) \). To improve computational efficiency, Lasko (2014) applied equally spaced grid intervals \( T \in [0, t_0 + 1] \) and numerical integrations to obtain the intensity functions rather than directly computes the likelihood functions at all observed events \( n \). The intensities at the time of the crash are then obtained from interpolations.
which are guaranteed by the smooth covariance matrix given by equation (3). To begin with, the prior distributions of shape parameter $\alpha$, signal variance $\sigma^2$, and length scale $l$ are assumed to be uniform. Next, the log intensities $f(t)$ given by equation (2) is computed at the given grid intervals using slice sampling. Then, the log intensity of $f(t)$ is calculated by smooth interpolation of $f(\hat{t})$. After that, the warped interval $\omega_i$ is calculated using trapezoid integration method. Finally, the likelihood function given by $f_1(T; \alpha, \lambda(t))$ is calculated so that the shape parameter $\alpha$ and likelihood function will be updated by Metropolis-Hasting algorithms. Notice that the inverse scale parameter $\beta$ is set as 1 to avoid identifiability problem and $\omega_1, \omega_n$ are modeled by $f_2(\omega; 1, 1)$ for computational efficiency (Lasko, 2014).

3. Empirical study

3.1. Framework

As shown in Fig. 1, the empirical study will demonstrate how GPMRP models can be applied for disaggregate crash analysis and to infer the time-varying safety impact of an evolving event. We modeled the interstate highway sections as unit points as an example. The objective is to perform crash analysis to evaluate the time-dependent impact of the pandemic on crash intensities. Let the intervention $Z$ be the dates after March 13, 2020, when the United States declared a national emergency due to the outbreak of COVID-19. The impacts are measured as the difference between the intervention group (the observed crash intensity $\lambda_i$) and the missing counterfactual group that should have been without the intervention (crash intensity $\hat{\lambda}_i$). At step I, crash data sources including both crash data, contributing factors and COVID-19 confirmed cases are collected and cleaned. To avoid information loss due to aggregated covariates during the crash intervals, continuous crash intensities are first inferred using the GPMRP model at step II. Direct comparison of the crash intensity functions before and after the pandemic can have hidden biases, as it ignores exogenous variables, which might affect crashes. Hence, at step III, individual crash intensity ($\lambda_i$) and the exogenous variables at the time of the crash ($X_i$) will be used to restore the counterfactual group. Let the historical crash data from

![Fig. 1. The framework of the empirical study.](image)

(a) Locations of interstate sections used in our study

(b) Interstate sections with the highest crash events since 2017 January in six States (or District)

![Fig. 2. Interstate sections with the highest crash events since 2017 January in six States (or District).](image)
January 1, 2017 to December 31, 2019 be the training dataset, crash data from January 1, 2020 to March 13, 2020 be the validating dataset, crash events after March 13, 2020 be predicted. The function mapping between the discrete counterfactual crash intensities conditional on $X_i$ will be learned by a supervised machine learning algorithm using the training dataset. Selected based on the minimum validation errors, the best machine learning model will be used to predict the expected counterfactual crash intensities conditional on $X_i$ of the intervention group. Finally, the crash intensity evaluations during the pandemic will be evaluated based on the difference between $\lambda_i$ and $\lambda_i^\ast$.

### 3.2. Data sources

The countrywide crash dataset is an open-source crash data augmented with multiple application programming interfaces (API) including Wunderground weather data, TimeAndDate light conditions, crash occurrence time, and locations (Moosavi et al., 2019). The crash data contains over 2 million crashes in the United States with a time span of over 4 years, which meets the threshold to achieve statistically meaningful results (Chang et al., 2017). We modeled the crash occurrences at interstates as a point stochastic process. This case study focuses on the interstate sections with the highest crash counts from January 2017 to December 2020 in 6 districts: New York (NY), California (CA), Washington (WA), Massachusetts (MA), Illinois (IL), and District of Columbia (D.C.). Fig. 2(a) shows the locations of the studied interstate sections. Listed in the order of lengths are Interstate-395 North in District of Columbia (D.C.) with 11.59 miles (I-395N D.C.), Interstate-94 North in Illinois with 61.53 miles (I-94W IL), Interstate-90 East in New York with 386 miles (I-90E NY), Interstate-5 North in California with 799 miles (I-5N CA), Interstate-93 North in Massachusetts with 46.25 miles (I-93N MA), and Interstate-395 North in District of Columbia with 3.48 miles (I-395N D.C.). Fig. 2(b) shows the quarterly aggregated crash times for each of the 6 Interstates. To visualize the aggregated crash occurrences in Fig. 2 (b), the aggregated crash counts are normalized by the estimated annual vehicle-miles traveled extracted from the Office of Highway Policy information of Federal Highway Administration. Reductions can be observed in 2020 Quarter 3 (Q3). For example, intersections such as I-5N CA, has experienced an increasing number of crashes starting from 2019 Q2 and a rapid decrease during 2020 the first half-year, before another drastic increase starting at Q3. The exact time when these changes occur, however, is lost at the process of aggregation. To prevent such information loss, crash intensities (the number of crashes per unit time) will be inferred by the GPMRP model at step II.

Due to the limitation of the data source lacking detailed exposure data such as traffic volumes, vehicle miles traveled, or vehicle hours traveled at granular level, the contributing factors are selected as exogenous variables at the time of the crash, such as temperature, humidity, visibility, wind speed, precipitation and light conditions, which were unaffected by the interventions. Usman et al. (2012) found that temperature and visibility can be significantly negatively correlated with crashes, and wind speed can be significantly positively correlated with crashes. Hourly precipitation, however, showed no evidence of significance. Durduran (2010) and Schlögl (2020) showed that air temperature, air pressure, and precipitation are very important weather covariates for crash analysis. The exogenous dataset has first been cleaned and preprocessed by removing the outliers or miscoded values and replacing the missing values. For example, a data record with the air pressure of 196 kPa in I-5N CA is removed because it is an unreasonable extreme value. The missing values for precipitation (<5%) have been filled with the mode value because the skewed frequency of crash events is the highest at zero precipitation. Table 1 summarizes the summary statistics of the numerical traffic weather data. From the table, we know that humidity has the largest standard deviations (about 22.11) while precipitation has the smallest deviations (about 2.0). The exact time when these changes occur, however, is lost at the process of aggregation. To prevent such information loss, crash intensities (the number of crashes per unit time) will be inferred by the GPMRP model at step II.

### 3.3. Results

#### 3.3.1. Bayesian inference for crash data

Four years of crash events from 2017 to 2020 ($t_{\text{max}} = 1461$) in 6 interstate sections are modeled by using the GPMRP model at Step II. Considering the trade-off between accuracy and computational efficiency, the number of burn-in and iteration times are set as 1000 respectively. Shape parameter $\alpha$ is assumed to be a uniform distribution.
between \([0.8, 25]\), signal variance \(\sigma^2\) a uniform distribution between \([0.01, 7]\), and length scale \(l\) a uniform distribution between \([2, \text{max}]\). The inferred crash intensities per unit mile with mean and 95\% confidence intervals for the 6 interstate sections are shown in Fig. 3. All interstate sections have experienced sharp intensities in the year of 2020. I-5N CA, I-90E NY, and I-5N WA, I-395N DC have increasing crash intensities at the end of 2020. I-5N CA and I-395N DC have especially the sharp slopes in the latter half year of 2020. I-93N MA, I-94W IL appear to be stable in the latter year of 2020 from the figure. However, this does not necessarily mean that those interstates did not experience significant impacts during the pandemic, due to the annual vehicle-miles travelled are estimations based on the proportions of the studied intersections over the total interstate lengths within their states. The proportions for I-94W IL and I-93N MA are about 3\% and 8\%, while other interstates comprised up about over 20\% of the total interstate lengths in the states, the actual vehicle-miles travelled for I-94W IL and I-93N MA may be underestimated. Hence, the evaluations of the change in crash intensities during the pandemic need further analysis. Compared with Fig. 2(b), the model is able to capture the non-homogeneous crash intensity functions. Furthermore, the Gaussian Process covariance functions are able to obtain smoother curvatures which show the value of crash intensities at a continuous time–space so that crash data can be studied at the individual level. To sum up, step II results provide direct inference of crash intensities at a continuous time.

To validate the accuracy of the GPMRP model results, a trapezoid numerical integration method is applied to obtain areas under curve (AUC) of crash times using equation (1). Table 2, shows the inferred crash times compared with actual crash events per year for the six interstates. The results show that the inferred average crash intensities are generally close to the actual crash times. The relative errors between inferred crash cases and actual cases have relatively low inference errors before 2020 (<5\%), however, the inference error for I-5N WA, I-94W IL, and I-395N DC in 2020 have an estimated error rate up to 6.7\%. The possible reason for this is that crash events during the pandemic are non-homogeneous. To reduce the computational cost, our prosed GPMRP model uses a grid time \(T\) to compute the intensity functions and then performs a smooth interpolation at the time of the crash \(T\) to infer the crash intensities. The equally spaced grid time can increase the inference error when the crash events present heterogeneous crash intensities. Considering the actual crash times fall in the 95\% confidence intervals obtained by the GPMRP model, the accuracy of model inference results for crash intensities can be considered as acceptable.

### Table 2

Area under the curve of the mean intensities check.

| Interstate sections | 2017 Inference | 2017 Actual | Error (%) | 2018 Inference | 2018 Actual | Error (%) | 2019 Inference | 2019 Actual | Error (%) | 2020 Inference | 2020 Actual | Error (%) |
|---------------------|---------------|-------------|-----------|---------------|-------------|-----------|---------------|-------------|-----------|---------------|-------------|-----------|
| I-5N CA             | 3434.4        | 3437        | -0.08\%   | 3103.0        | 3131        | -0.89\%   | 4318.2        | 4318        | 0.00\%    | 7353.7        | 7596        | -3.19\%    |
| I-90E NY            | 614.2         | 618         | -0.61\%   | 812.7         | 828         | -1.85\%   | 797.2         | 792         | 0.66\%    | 856.0         | 865         | -1.04\%    |
| I-5N WA             | 1647.5        | 1643        | 0.27\%    | 1638.4        | 1647        | -0.52\%   | 1328.8        | 1329        | -0.02\%   | 1213.9        | 1284        | -5.46\%    |
| I-93N MA            | 358.8         | 350         | 2.51\%    | 358.5         | 359         | -0.14\%   | 284.3         | 274         | 3.76\%    | 478.9         | 464         | 3.21\%     |
| I-94W IL            | 835.1         | 816         | 2.34\%    | 734.0         | 741         | -0.94\%   | 574.9         | 581         | -1.05\%   | 464.4         | 495         | -6.18\%    |
| I-395N DC           | 41.7          | 39          | 6.92\%    | 44.6          | 43          | 3.72\%    | 38.4          | 38          | 1.05\%    | 144.1         | 135         | 6.74\%     |

**Fig. 3.** Bayesian inference of GPMRP crash intensities of six States (Districts).
regression models (PR) and negative binomial models (NB). We can observe that PR/NB models assume that the crash intensities are homogeneous within the quarterly intervals. Thus, detailed time-varying information at granular level cannot be revealed. In contrast, GP results, however, can capture the inhomogeneous properties of the inferred crash intensities within the quarterly interval. For example, from I to 5N CA, the actual crash counts in 2019 November have reduced compared with those in October and December of the same year. Such fluctuation can be successfully captured by GP as shown in the dotted blue lines, however, PR and NB models can only capture a constant intensity in 2019 Q3. Similarly, the increment of crash counts in 2020 May and June for I-5N WA, I-93N MA and I-94W IL can be reflected by our renewal model. These increasing trends are lost by the PR/NB models because the inferreded intensities models by PR and NB models in 2020 Q2 are constant. To further compare our proposed models with traditional count models, we present the residual error box plot of GP, PR, and NB for the 6 interstate sections in Fig. 5. Overall, both biases and variances of GP are smaller than PR and NB, suggesting that the inferred granular crash intensities exhibit less information loss in case of GP models. Outliers for most of the sections shown in Fig. 5 are found to
occur during the second half year of 2020, after the stay-at-home policies in the six districts were imposed. The specific residual error records can be found in Tables A.3.1-A.3.6 in Appendix.

3.3.3. Counterfactual crash intensity restoration

In order to reduce the biases from exogenous variables such as weather conditions, lighting conditions, the inferred crash intensities at the time of the crash are combined with the exogenous variables and categorical variables such as day of the week. To evaluate the impacts on crash intensity ($\lambda$) during the pandemic, the common approach is to compare the difference between the observed intervention group $E[\lambda|Z=1]$ and the missing counterfactual group $E[\lambda|Z=0]$ (Rubin, 1974), where $Z$ is the binary intervention. We make an assumption that the historical training group and the counterfactual group $E[\lambda|Z=0]$ follows a same mapping function between the historical crash intensities and real-time exogenous variables. In other words, if the coronavirus did not exist and no national emergency was ever announced, the crash intensities in 2020 can be predicted by the previous years’ crash intensities given the same traffic conditions such as weather conditions. This mapping assumption is supported by the observations from Fig. 2(b) and Fig. 3: except I-5N CA, all other interstate sections did not experience a sharp increase in their crash intensity functions until March 2020. The fairness of the assumption will be justified by the validation errors from the validation dataset. Under this assumption, statistical models are applied to restore counterfactual crash intensity as a comparison to the intervention group. The restored counterfactual group will be predicted conditioned on exogenous variables $X_{t-1}$. Five machine learning models are selected as follows: historical mean model (Li et al., 2015), support vector regression model (Wu et al., 2004; Ahn et al., 2016), linear regression Model, XGBoost model (Dong et al., 2018; Mei et al., 2018; Shi et al., 2018), random forest model (Wang et al., 2018).

Historical Mean model (HM): this model takes the average of historical inferred crash intensities before March 13, 2020, and the predicted intensities will be the interpolated values at the corresponding time.

Support Vector Regression model (SVM): A radial-based kernel function is used to find the best fitted nonlinear relationships between training attributes (hyperplane) and training crash intensities (boundary lines).

Linear Regression model (LR): Multiple linear regression model predicts the expected traffic intensities conditioned on the exogenous variables by minimizing the least square residuals.

XGBoost model (XGB): A gradient boosting tree model with higher accuracy and lower computational costs which contains tree-pruning and regularizations to prevent overfitting.

Random Forest model (RF): A tree-based prediction model with a subset of randomly selected attributes from the exogenous variables to prevent overfitting.

To justify our assumptions made earlier and evaluate the accuracy of statistical models’ restorations, evaluation metrics needs to be picked. Considering crash events in I-395 N are sparse, thus the inferred crash intensity $\lambda_i$ at the denominator can be smaller than 1, resulting in a large
numerical error if mean absolute percentage error (MAPE) is used. Hence, root mean square error (RMSE) is selected as the evaluation metrics, which is written as:

$$
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{\lambda}_i - \lambda_i)^2}{n}}
$$

(6)

where $n$ is the maximum number of crash events in the validation period, $\hat{\lambda}_i$ is the inferred crash intensity during the validation period, $\lambda_i$ is the predicted crash intensity with training dataset from crash events before 2020. Table 3 shows validation results for the six interstate sections. As highlighted in black, models with the smallest validation error will be selected as the best model for counterfactual restoration.

Counterfactual crash intensities after March 13, 2020, are restored as shown in Fig. 6. The black solid line is the inferred crash intensities. The best prediction models are shown in the other line. The best prediction model for I-5N CA is the XGB model. Notice that I-5N CA crash patterns have higher variations starting from July 1st, 2019, which violates the assumption that crash intensity distributions are identical before March 2020. The training dataset has two distinct probability distributions, leading to dense spikes in its prediction results. The best statistical models for I-5N WA and I-385N DC are also XGB, different from I to 5N CA, no significant violation of our proposed assumptions found in their crash intensity distributions. The best prediction model for I-90E NY is the historical mean model with a RMSE around 0.0652, the restored counterfactual data is shown in the red solid line. RF model is the best prediction model for counterfactual construction with a RMSE around 0.4210, which is shown in the green solid line.

3.3.4. Pandemic impact evaluation

The GPMRP model allows crash data and covariates to be combined at the individual level. The usage of real-time covariates helps to reduce the estimation biases from weather conditions and light conditions. One of the benefits is that the time-varying impacts of crash intensity changes during the pandemic period can be obtained. Fig. 7(a) shows the impacts based on the difference between crash intensities at the intervention group $\lambda_i$ and restored counterfactual group $\hat{\lambda}_i$. The vertical lines are the marked important timestamps as a reference. The normalized Covid-19 case numbers at nationwide is included in Fig. 7 for visual comparisons. While I-5N CA has increased crash intensities up to 50 in the later periods of 2020, other 5 interstate sections have relatively smaller changes in crash intensities (no more than 10/unit time). In order to compare the impacts of six interstate sections, the impacts are normalized to the same scale from $-1$ to 1. Three distinct time-varying patterns can be observed: I-5N CA has a parabolic pattern; I-395N DC and I-5N WA have bimodal patterns, and I-93N MA and I-94W IL have unimodal patterns.

3.4. Discussion

The empirical study section of the paper showed safety analytics performed at a granular level using GPMRP and insights of COVID-19 impacts on safety. The spread of COVID-19 together with stay-at-home policies raises challenges to crash frequency modeling and data aggregation, as using aggregation per quarter may fail to reveal important patterns of crash intensities. Existing aggregated crash modeling methods rely on the i.i.d. assumptions of crash events (see Manning and Bhat, 2014; Abdulhafedh, 2016; Zheng et al., 2020 for a systematic review), crash occurrences, however, did not follow the identical patterns before the pandemic (see Fig. 2b). In addition, models fitted by aggregated data can be insufficient to capture the non-homogeneous phenomena. Following the work of Demirosk and Ozbay (2015; Li et al. (2017); Li et al. (2020) to model non-homogeneous crash intensity, we proposed to apply the GPMRP model, which is a modulated renewal stochastic process that relaxes the i.i.d. assumptions. It uses the Bayesian inference approach to estimate the distribution of crash intensity functions by maximizing the Gamma likelihood functions, which provided confidence intervals for the inferred crash intensity (Fig. 3).

The crash intensities inferred by the GPMRP model present detailed inhomogeneous properties during the pandemic. For instance, from Fig. 2(b) in Section 3.2, one can observe that the number of crashes per unit mile for I-93N MA experienced its peak near Q2 in 2020, which was the time when one of the major cities, Boston, announced the emergency order due to COVID-19 (from 03/24/2020 to 04/07/2020). The detailed crash intensity variation information at the granular level was hidden due to aggregation per quarter. In contrast, based on the GPMRP inference results seen on Fig. 3 and Fig. 4, the crash intensity of I-93N MA first reduced slightly near mid-April 2020 before its intensity continued to climb up. Moreover, different from the coarsely aggregated crash occurrences, the application of the Gaussian Process with a squared exponential covariance matrix is found to be able to infer the crash intensities smoothly considering temporal correlations, presenting direct comparisons of historical intensity during different periods. As an example, the inferred average intensity of I-94W IL in Fig. 3 showed that there were four peaks on average each year from 2017 to 2019 before the spread of COVID-19. When it came to the year 2020, the crash intensities experienced one major peak from 03/01/2020 to 08/01/2020, which was higher than the ones in the previous three years. Similar phenomena can be observed in other studied interstate sections, suggesting that highway traffic safety hazards in 2020 have changed with respect to the previous years. These findings are consistent with the previous stay-at-home related traffic safety studies (NHTSA, 2020; Doucette et al., 2021). Table 2 and Fig. 5 of the case study showed that our proposed model can accurately infer the four-year crash events based on AUC using equation (1). The results in this paper add to the existing literature on how the inhomogeneous crash intensities of the major interstate sections in continuous time.

Furthermore, the time-varying changes during the pandemic in six interstate sections were quantified. Five machine learning algorithms were employed to combine the crash intensities and exogenous variables under the assumption of each crash is a discrete event. The dashed black line in Fig. 7b was the date when the United States declared a national emergency and the dashed color lines represented the corresponding States that started to reopen. It is noted that an increased crash intensity reached its peak for I-5N WA, I-93N MA, and I-94W IL during the stay-at-home policy periods compared with the restored counterfactual crash intensities. 1-90 E NY, I-5N CA, and I-395 DC generally experienced a reduced crash intensity during Q2 and Q3 before their intensity differences started to become positive in Q4. The presence of bimodal patterns of I-395N DC and I-5N WA are shown in Fig. 7b, suggesting the presence of perturbation interrupted from the external field (Moslonka and Sekimoto, 2020). The parabolic pattern of I-5N CA suggests that the crash intensities were increasing rapidly since the third quarter of 2020. This can be expected since different States were proposing “stay-at-home” policies as a response to stop spreading the virus and the total traffic volumes across the country during this period were reported to decrease since March of 2020 (Li et al., 2021; USDOT, 2021). Previous studies provided possible explanations of these COVID-19 impact patterns. Research shows that while the traffic counts in many regions such as California (Vanlaar et al., 2021), Washington, D.C. (Lazo, 2021), and Alabama (Adanu et al., 2021) were found to be reduced significantly over the stay-at-home policy compared with the same periods in previous years, speeding-related crash rates, however, increased significantly by nearly 4 times in New York City, approximately 8 times in Seattle (Liao and Lowery, 2021). Moreover, more than 80% increase in speeding tickets over 100 mph were issued in California during the stay at home policy (Mcgrevey, 2020).

Since the temperature variations in I-90E NY, I-93N MA, I-94W IL, and I-395N DC were found to be about twice larger than those in I-5N CA and I-5NWA (see Table 1). The weather conditions were considered when estimating the crash intensity differences given the observed
fluctuations. The temporal effects such as day of the week were also included as categorical data in the model, however, seasonal trends and local trends were not included due to the irregularity of crash events. Moreover, important variables related to traffic exposure such as traffic volumes, traffic speed could not be included in this case study due to the limitation of datasets. The fitted machine learning models will be expected to have fewer variances (spikes) if a more important contributing factor dataset becomes available (see Fig. 6). It was demonstrated that the application of GPMRP can quantify time-varying impacts and can be used in future studies to analyze the emerging disproportionate impacts. When the inferred crash intensities are combined with covariates at the granular level, a wider range of information collected by ITS in real-time can be used for analysis. For example, vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) technologies provide nearby vehicles’ numbers and speeds at the time of the crash. These emerging data sources can be used to study important contributing factors leading to crash events or near-crash events to reduce the hidden biases (Manning and Bhat, 2014).

4. Conclusions

This paper proposed a Gaussian Process modulated renewal process model for the crash intensities estimations for disproportionate crash analysis. One of the benefits of our proposed model is that it can capture the non-homogeneous properties of crash events and allow crash data to be associated with contributing factors at the granular level to avoid information loss. In our case study, we used the GPMRP model to perform disproportionate analysis for interstate sections with the highest crash rates in 6 regions. Based on the normalized crash intensities from January 2017 to December 2020, we can observe that except I-5N CA, other sections (I-90E NY, I-5N WA, I-93N MA, I-94W IL, and I-395N DC) did not experience a visually drastic increase until 2020 March. Based on this observation, an assumption that the relationship between crash intensities and contributing factors can be modeled by the same mapping functions if COVID-19 has not had happened. We used five machine learning algorithms to model the mapping functions. The best-fitted models were selected based on RMSE. Except for I-5N CA, the counterfactual crash intensities can be properly predicted based on the training dataset, which is as expected because of the fact that I-5N CA violated our mapping assumption (see Section 3.3.3). It experienced a drastic increase from 2019 October. The inferred continuous intensity functions showed their usefulness in estimating the time-varying impacts of COVID-19. Three distinct patterns were found among the six interstate sections that were included in our study: bimodal patterns (I-95N DC, I-90E NY), parabolic pattern (I-5N CA), and unimodal patterns (I-93N MA, I-94W IL). We conclude that the GPMRP model can be applied for disaggregate analysis as it can help identify inhomogeneous properties at the continuous-time dimension when combined with real-time contributing factors.

One of the limitations, however, is that the time-lagging effects between crash events become hard to analyze as the intervals between crash events are unevenly spaced at the individual level. As a future research direction, the temporal correlations, seasonality, and time lagging effects for irregularly sparse crash events shall be explored. Another limitation is that the important real-time traffic exposure datasets and driver information such as age were not included in the contributing factors due to unavailability. With the proliferation of ITS technologies, emerging data sources can be created by advanced sensors and cameras continuously. In addition, with the help of V2V and V2I communication technologies, crash-related information can be stored and accessed efficiently. Better use of continuous-time datasets will help improve traffic safety analytics. One of the practical implications our proposed model is its potential ability to analyze the naturalistic trajectory data at the intersections considering the surrounding vehicles at the time of crash or near crash events. The proposed model is expected to further reduce the estimation biases should more real-time traffic exposure data become available before long.

CRediT authorship contribution statement

Yiyuan Lei: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. Kaan Ozbay: Conceptualization, Methodology, Supervision, Writing – review & editing. Kun Xie: Conceptualization, Visualization, Writing – review & editing.

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.

Acknowledgements

This work was partially supported by the Connected Cities for Smart Mobility towards Accessible and Resilient Transportation (C2SMART), a Tier 1 U.S. Department of Transportation funded University Transportation Center (UTC) led by New York University. The contents of this paper only reflect the views of the authors who are responsible for the facts and do not represent any official views of any sponsoring organizations or agencies.

Appendix

The summary statistics of crash events for the six interstate segments are represented in Table A.1. It contains the minimum (min.) maximum (max.), mean and standard deviation (sd.) of the monthly crash counts as shown in the black lines in Fig. 4.

Based on Deviance and Chi-square tests, we do not reject the null hypothesis H0 (H0: the negative binomial models are good fits) at the significance levels (α) of 0.01. To show the importance of the fitted covariates, the coefficients (Coef), standard error of the coefficients (SE), the 95% confidence intervals (95% CI), Z statistics (Z-score) as well as P-value are listed in Tables A.2.1-A.2.6. Results show that the important variables (statistically significant) are day of week (DayofWeek), light conditions (DayNight), temperature (Temperature), visibility (Visibility) and wind speed.

Table A1
Summary statistics of the crash counts.

|                  | I-5N CA | I-90E NY | I-5N WA | I-93N MA | I-94W IL | I-395N DC |
|------------------|---------|----------|---------|----------|----------|-----------|
| min.             | 40      | 10       | 25      | 12       | 4        | 0         |
| max.             | 1274    | 160      | 207     | 86       | 102      | 29        |
| mean             | 273     | 64       | 121     | 30       | 54       | 5         |
| sd.              | 263     | 30       | 35      | 14       | 25       | 6         |
level of 0.05; Temperature are negatively correlated with crash intensities for I-90E NY, I-5N MA, and I-395N DC. While visibility is negatively correlated with crash intensities for I-5N WA section, the significance of such correlation, however, tends not to be statistically significant for most other intersections. For I-93N MA, I-94W IL, and I-395N DC, the wind speed is found to be positively correlated with crash intensities at the statistical significance level of 0.05. The residuals of Gaussian process modulated renewal model (GP), Poisson regression model (PR) and negative binomial model (NB) for the six interstate sections are listed in Tables A.3.1-A.3.6.

Table A.2.1
The covariates of the fitted Negative Binomial model of I-5N CA.

| Coef     | SE     | 95% CI       | Z-score | P-value |
|----------|--------|--------------|---------|---------|
| Intercept| 10.543 | 2.899        | (4.861, 16.225) | 3.637 | <0.001 |
| (DayofWeek.Tue) | -0.2287 | 0.156        | (-0.534, 0.077) | -1.468 | 0.142 |
| (DayofWeek.Wed) | -0.3770 | 0.152        | (-0.675, -0.09) | -2.48 | 0.013 |
| (DayofWeek.Thu) | 0.3449 | 0.205        | (-0.057, 0.747) | 1.682 | 0.093 |
| (DayofWeek.Fri) | 0.2877 | 0.114        | (0.064, 0.511) | 2.524 | 0.012 |
| (DayofWeek.Sat) | -1.6766 | 0.258        | (-2.183, -1.179) | -6.486 | <0.001 |
| Temperature | -0.006 | 0.042        | (-0.087, 0.075) | -0.144 | 0.885 |
| WindChill | 0.0129 | 0.026        | (0.038, 0.064) | 0.498 | 0.619 |
| DayNight | -0.2094 | 0.05        | (-0.195, 0) | -1.966 | 0.049 |
| WindSpeed | 1.4600 | 0.864        | (-0.228, 3.160) | 1.696 | 0.09 |

Table A.2.2
The covariates of the fitted Negative Binomial model of I-90E NY.

| Coef     | SE     | 95% CI       | Z-score | P-value |
|----------|--------|--------------|---------|---------|
| Intercept| 2.7435 | 1.193        | (0.405, 5.082) | 2.3 | 0.021 |
| (DayofWeek.Tue) | -0.6027 | 0.228        | (-1.050, -0.156) | -2.644 | 0.008 |
| (DayofWeek.Wed) | -0.1649 | 0.207        | (-0.571, 0.242) | -0.795 | 0.436 |
| (DayofWeek.Thu) | -0.5786 | 0.265        | (-1.099, -0.102) | -4.091 | 0.002 |
| (DayofWeek.Fri) | -0.2011 | 0.23        | (-0.652, 0.25) | -0.875 | 0.382 |
| (DayofWeek.Sat) | 0.6698 | 0.363        | (-1.381, 0.042) | -2.763 | 0.006 |
| Temperature | 0.0525 | 0.027        | (-0.104, 0.001) | -1.98 | 0.048 |
| WindChill | 0.0013 | 0.012        | (-0.006, 0.010) | -0.115 | 0.905 |
| DayNight | 0.2642 | 0.094        | (0.079, 0.449) | 2.799 | 0.005 |
| WindSpeed | -0.0303 | 0.045        | (-0.121, 0.055) | -0.738 | 0.461 |
| DayNight | 0.8109 | 0.924        | (-1.000, 2.622) | 0.878 | 0.380 |

Table A.2.3
The covariates of the fitted Negative Binomial model of I-5N WA.

| Coef     | SE     | 95% CI       | Z-score | P-value |
|----------|--------|--------------|---------|---------|
| Intercept| 8.2420 | 0.294        | (6.431, 10.053) | 8.919 | <0.001 |
| (DayofWeek.Tue) | -0.2451 | 0.119        | (-0.478, -0.012) | -2.063 | 0.039 |
| (DayofWeek.Wed) | 0.9915 | 0.088        | (-0.081, 0.264) | 1.029 | 0.299 |
| (DayofWeek.Thu) | 0.2204 | 0.108        | (0.009, 0.432) | 2.029 | 0.041 |
| (DayofWeek.Fri) | 0.0010 | 0.092        | (-0.179, 0.181) | 0.011 | 0.991 |
| (DayofWeek.Sat) | -1.0848 | 0.153        | (-1.385, -0.785) | -7.082 | <0.001 |
| Temperature | 0.0402 | 0.014        | (0.013, 0.067) | 2.947 | 0.003 |
| WindChill | 0.0021 | 0.006        | (-0.009, 0.013) | 0.367 | 0.713 |
| DayNight | 0.2968 | 0.087        | (-0.468, -0.126) | -3.404 | 0.001 |
| WindSpeed | 0.1190 | 0.030        | (0.060, 0.178) | 3.956 | <0.001 |
| DayNight | 0.1301 | 0.271        | (-0.401, 0.661) | 0.481 | 0.631 |

Table A.2.4
The covariates of the fitted Negative Binomial model of I-93N MA.

| Coef     | SE     | 95% CI       | Z-score | P-value |
|----------|--------|--------------|---------|---------|
| Intercept| 3.0571 | 1.013        | (1.072, 5.041) | 3.019 | 0.003 |
| (DayofWeek.Tue) | 0.2409 | 0.190        | (-0.131, 0.612) | 1.267 | 0.205 |
| (DayofWeek.Wed) | 0.3175 | 0.184        | (-0.043, 0.677) | 1.727 | 0.084 |
| (DayofWeek.Thu) | 0.5305 | 0.167        | (0.203, 0.856) | 3.174 | 0.002 |
| (DayofWeek.Fri) | -0.2686 | 0.189        | (-0.638, 0.101) | -1.422 | 0.155 |
| Temperature | -0.0050 | 0.012        | (-0.029, 0.019) | -0.399 | 0.690 |
| WindChill | 0.0115 | 0.011        | (-0.007, 0.037) | 1.341 | 0.180 |
| DayNight | -0.0295 | 0.056        | (-0.136, 0.085) | -0.444 | 0.657 |
| WindSpeed | 0.1061 | 0.031        | (0.046, 0.166) | 3.443 | 0.001 |
| DayNight | -1.5328 | 0.688        | (-2.087, -1.019) | -2.234 | 0.025 |
### Table A.2.5
The covariates of the fitted Negative Binomial model of I-94W IL.

| Coef     | SE of Coef | 95% CI          | Z-score | P-value |
|----------|------------|-----------------|---------|---------|
| Intercept| 2.0943     | 1.061           | (0.015, 4.174) | 1.974   | 0.048   |
| C(DayofWeek.Mon) | 1.5481 | 0.253           | (1.053, 2.044) | 6.123   | <0.001  |
| C(DayofWeek.Tue)  | 1.4011 | 0.253           | (0.906, 1.897) | 5.542   | <0.001  |
| C(DayofWeek.Wed)  | 1.7823 | 0.236           | (1.32, 2.245)  | 7.554   | <0.001  |
| C(DayofWeek.Thu)  | 1.4636 | 0.253           | (0.995, 1.933) | 6.116   | <0.001  |
| C(DayofWeek.Fri)  | 1.0922 | 0.241           | (0.587, 1.531) | 4.396   | <0.001  |
| Temperature      | −0.0067 | 0.013           | (−0.033, 0.019) | −0.506  | 0.613   |
| WindChill        | 0.01022 | 0.006           | (−0.002, 0.022) | 1.657   | 0.097   |
| Visibility       | −0.041 | 0.067           | (−0.171, 0.089) | −0.616  | 0.538   |
| WindSpeed        | 0.1448 | 0.039           | (0.069, 0.22)  | 3.756   | <0.001  |
| DayNight         | −0.2521 | 0.439           | (−1.113, 0.609) | −0.574  | 0.566   |

### Table A.2.6
The covariates of the fitted Negative Binomial model of I-395N DC.

| Coef     | SE of Coef | 95% CI          | Z-score | P-value |
|----------|------------|-----------------|---------|---------|
| Intercept| 0.0695     | 2.769           | (−5.358, 5.497) | 0.025   | 0.98    |
| C(DayofWeek.Mon) | 1.7199 | 0.6           | (0.543, 2.897) | 2.864   | 0.004   |
| C(DayofWeek.Tue)  | 1.5877 | 0.599         | (0.414, 2.761)  | 2.652   | 0.008   |
| C(DayofWeek.Wed)  | 0.7427 | 1.367         | (−1.936, 3.421) | 0.544   | 0.587   |
| C(DayofWeek.Thu)  | 1.7029 | 0.676         | (0.379, 3.027)  | 2.52    | 0.128   |
| C(DayofWeek.Fri)  | 1.098 | 0.721         | (−0.314, 2.51)  | 1.524   | 0.128   |
| Temperature      | −0.0522 | 0.026         | (−0.103, 0.002) | −2.02   | 0.043   |
| WindChill        | 0.058 | 0.027         | (0.005, 0.111)  | 2.154   | 0.031   |
| Visibility       | −0.0121 | 0.176         | (−0.358, 0.333) | −0.069  | 0.945   |
| WindSpeed        | 0.0651 | 0.033         | (0.001, 0.129)  | 1.991   | 0.046   |
| DayNight         | 4.309 | 2.16          | (0.075, 8.543)  | 1.995   | 0.046   |

### Table A.3.1
Information loss comparisons of I-5N CA.

|          | Actual Count | Counts | GP | PR | NB |
|----------|--------------|--------|----|----|----|
| 2017     |              |        |    |    |    |
| Q1       | 1004         | 975    | 1119| 1128|
| Error (%)| −2.89%       | 11.45% | 12.35%| 853|
| Q2       | 826          | 820    | 860 | 853 |
| Error (%)| −0.73%       | 4.12%  | 3.27%| 876 |
| Q3       | 780          | 772    | 743 | 758 |
| Error (%)| −1.03%       | −4.74% | −2.82%| 739 |
| Q4       | 827          | 847    | 706 | 739 |
| Error (%)| 2.42%        | −14.63%| −10.64%| 876 |
| 2018     |              |        |    |    |    |
| Q1       | 893          | 819    | 880 | 876 |
| Error (%)| −8.29%       | −1.46% | −1.90%| 777 |
| Q2       | 765          | 795    | 799 | 777 |
| Error (%)| 3.92%        | 4.44%  | 1.57% | 946 |
| Q3       | 798          | 793    | 960 | 946 |
| Error (%)| −0.63%       | 20.30% | 18.55% | 674 |
| Q4       | 675          | 665    | 674 | 674 |
| Error (%)| −1.48%       | −0.15% | −0.15%| 578 |
| 2019     |              |        |    |    |    |
| Q1       | 549          | 550    | 590 | 578 |
| Error (%)| 0.18%        | 7.47%  | 5.28% | 541 |
| Q2       | 503          | 475    | 582 | 541 |
| Error (%)| −5.57%       | 15.71% | 7.55% | 736 |
| Q3       | 1127         | 1115   | 1187| 1210|
| Error (%)| −1.06%       | 5.33%  | 7.36% | 2293|
| Q4       | 2139         | 2160   | 2249| 2293|
| Error (%)| 0.98%        | 5.14%  | 7.20% | 2306|
| 2020     |              |        |    |    |    |
| Q1       | 2153         | 2169   | 2042| 2016|
| Error (%)| 0.74%        | −5.16% | −6.36%| 382 |
| Q2       | 1560         | 1329   | 1152| 1127|
| Error (%)| −14.81%      | −26.15%| −27.76%| 382 |
| Q3       | 383          | 717    | 382 | 382 |
| Error (%)| 87.21%       | −0.26% | −0.26%| 3046|
| Q4       | 2913         | 3130   | 2962| 3046|
| Error (%)| 7.45%        | 1.68%  | 4.57% |
### Table A.3.2
Information loss comparisons of I-90E NY.

|        | Actual Count | Counts | GP      | PR      | NB      |
|--------|--------------|--------|---------|---------|---------|
| 2017 Q1 | 118          | 107    | 125     | 121     |
| Error (%) | –9.32%     | 5.93%  | 2.54%   |
| Q2     | 89           | 87     | 153     | 144     |
| Error (%) | –2.25%     | 71.91% | 61.80%  |
| Q3     | 170          | 183    | 171     | 175     |
| Error (%) | 7.65%       | 0.59%  | 2.94%   |
| Q4     | 241          | 234    | 233     | 236     |
| Error (%) | –2.90%     | –3.32% | –2.07%  |
| 2018 Q1 | 166          | 156    | 175     | 183     |
| Error (%) | –6.02%     | 5.42%  | 10.28%  |
| Q2     | 191          | 183    | 204     | 206     |
| Error (%) | –4.19%     | 6.81%  | 7.85%   |
| Q3     | 220          | 216    | 159     | 161     |
| Error (%) | –1.82%     | –27.73%| –26.82% |
| Q4     | 251          | 248    | 200     | 320     |
| Error (%) | –1.20%     | 19.52% | 27.49%  |
| 2019 Q1 | 223          | 217    | 165     | 164     |
| Error (%) | –2.69%     | –26.01%| –26.46% |
| Q2     | 229          | 224    | 201     | 200     |
| Error (%) | –2.18%     | –12.23%| –12.66% |
| Q3     | 159          | 166    | 162     | 160     |
| Error (%) | 4.40%       | 1.89%  | 0.63%   |
| Q4     | 181          | 182    | 173     | 175     |
| Error (%) | 0.55%       | –4.42% | –3.31%  |
| 2020 Q1 | 144          | 131    | 171     | 168     |
| Error (%) | –9.02%     | 18.75% | 16.67%  |
| Q2     | 177          | 175    | 184     | 184     |
| Error (%) | –1.13%     | 3.95%  | 3.95%   |
| Q3     | 89           | 92     | 89      | 88      |
| Error (%) | 3.37%       | 0%     | –1.12%  |
| Q4     | 428          | 458    | 403     | 383     |
| Error (%) | 7.91%       | –5.84% | –10.51% |

### Table A.3.3
Information loss comparisons of I-5N WA.

|        | Actual Count | Counts | GP      | PR      | NB      |
|--------|--------------|--------|---------|---------|---------|
| 2017 Q1 | 381          | 375    | 394     | 394     |
| Error (%) | –1.31%     | 3.41%  | 3.41%   |
| Q2     | 350          | 344    | 379     | 377     |
| Error (%) | –1.71%     | 8.29%  | 7.71%   |
| Q3     | 387          | 395    | 361     | 362     |
| Error (%) | 2.07%       | –6.72% | –6.46%  |
| Q4     | 525          | 521    | 490     | 485     |
| Error (%) | –0.76%     | –6.67% | –7.62%  |
| 2018 Q1 | 395          | 383    | 407     | 407     |
| Error (%) | –3.04%     | 3.04%  | 3.04%   |
| Q2     | 406          | 397    | 362     | 363     |
| Error (%) | –2.22%     | –10.84%| –10.59% |
| Q3     | 410          | 415    | 421     | 422     |
| Error (%) | 1.22%       | 2.68%  | 2.93%   |
| Q4     | 436          | 429    | 458     | 460     |
| Error (%) | –1.61%     | 5.05%  | 5.50%   |
| 2019 Q1 | 355          | 351    | 335     | 335     |
| Error (%) | –1.13%     | –6.48% | –5.63%  |
| Q2     | 296          | 290    | 360     | 360     |
| Error (%) | –2.03%     | 21.62% | 21.62%  |
| Q3     | 347          | 337    | 372     | 372     |
| Error (%) | –2.88%     | 7.20%  | 7.20%   |
| Q4     | 331          | 338    | 335     | 334     |
| Error (%) | 2.11%       | 1.21%  | 0.91%   |
| 2020 Q1 | 296          | 286    | 296     | 295     |
| Error (%) | –3.38%     | 0%     | –0.34%  |
| Q2     | 440          | 426    | 376     | 375     |
| Error (%) | –3.18%     | –14.55%| –14.77% |
| Q3     | 130          | 134    | 130     | 129     |
| Error (%) | 3.08%       | 0%     | –0.77%  |
| Q4     | 344          | 360    | 349     | 349     |
| Error (%) | 4.65%       | 1.45%  | 1.45%   |
Table A.3.4
Information loss comparisons of I-93N MA.

| Year | Quarter | Actual Count | Counts GP | PR | NB |
|------|---------|--------------|-----------|----|----|
| 2017 | Q1      | 104          | 103       | 112| 113|
|      | Error (%) | -0.96%       | 7.69%     | 8.65%|    |
|      | Q2      | 49           | 59        | 53 | 53 |
|      | Error (%) | 20.41%       | 8.16%     | 8.16%|    |
|      | Q3      | 95           | 96        | 111| 111|
|      | Error (%) | 1.05%        | 16.84%    | 16.84%|   |
|      | Q4      | 102          | 97        | 98 | 96 |
|      | Error (%) | -4.90%       | -3.92%    | -5.88%|   |
| 2018 | Q1      | 85           | 84        | 96 | 97 |
|      | Error (%) | -1.18%       | 12.94%    | 14.12%|   |
|      | Q2      | 107          | 102       | 102| 102|
|      | Error (%) | -4.67%       | -4.67%    | -4.67%|   |
|      | Q3      | 85           | 89        | 75 | 77 |
|      | Error (%) | 4.71%        | -11.76%   | -9.41%|   |
|      | Q4      | 82           | 78        | 67 | 68 |
|      | Error (%) | -4.88%       | -18.29%   | -17.07%|   |
| 2019 | Q1      | 103          | 105       | 100| 100|
|      | Error (%) | 1.94%        | -2.91%    | -2.91%|   |
|      | Q2      | 67           | 63        | 69 | 70 |
|      | Error (%) | -5.97%       | 2.99%     | 4.48%|   |
|      | Q3      | 49           | 52        | 44 | 45 |
|      | Error (%) | 6.12%        | -10.20%   | -8.16%|   |
|      | Q4      | 55           | 52        | 60 | 89 |
|      | Error (%) | 9.09%        | 61.82%    | 61.82%|   |
| 2020 | Q1      | 96           | 94        | 87 | 87 |
|      | Error (%) | -2.08%       | -9.38%    | -9.38%|   |
|      | Q2      | 224          | 221       | 202| 188|
|      | Error (%) | -1.34%       | -9.82%    | -16.07%|   |
|      | Q3      | 62           | 74        | 74 | 73 |
|      | Error (%) | 19.35%       | 19.35%    | 17.74%|   |
|      | Q4      | 77           | 84        | 57 | 61 |
|      | Error (%) | 9.09%        | -25.97%   | -22.78%|   |

Table A.3.5
Information loss comparisons of I-94W IL.

| Year | Quarter | Actual Count | Counts GP | PR | NB |
|------|---------|--------------|-----------|----|----|
| 2017 | Q1      | 268          | 264       | 250| 248|
|      | Error (%) | -1.49%       | -6.72%    | -7.46%|   |
|      | Q2      | 171          | 173       | 203| 203|
|      | Error (%) | 1.17%        | 18.71%    | 18.71%|   |
|      | Q3      | 166          | 171       | 173| 173|
|      | Error (%) | 3.01%        | 4.22%     | 4.22%|   |
|      | Q4      | 211          | 221       | 198| 204|
|      | Error (%) | 4.74%        | -6.16%    | -3.32%|   |
| 2018 | Q1      | 215          | 204       | 181| 180|
|      | Error (%) | -5.12%       | -15.81%   | -16.28%|   |
|      | Q2      | 170          | 168       | 176| 178|
|      | Error (%) | -1.18%       | 3.53%     | 4.71%|   |
|      | Q3      | 169          | 168       | 141| 141|
|      | Error (%) | -0.59%       | -16.57%   | -16.57%|   |
|      | Q4      | 187          | 185       | 186| 186|
|      | Error (%) | -1.07%       | -0.53%    | -0.53%|   |
| 2019 | Q1      | 172          | 175       | 210| 212|
|      | Error (%) | 1.74%        | 22.09%    | 23.26%|   |
|      | Q2      | 186          | 176       | 181| 178|
|      | Error (%) | -5.38%       | -2.69%    | -4.30%|   |
|      | Q3      | 104          | 99        | 109| 108|
|      | Error (%) | -4.81%       | 4.81%     | 3.85%|   |
|      | Q4      | 119          | 117       | 137| 136|
|      | Error (%) | -1.68%       | 15.13%    | 14.29%|   |
| 2020 | Q1      | 125          | 123       | 119| 120|
|      | Error (%) | -1.60%       | -4.80%    | -4.00%|   |
|      | Q2      | 283          | 272       | 275| 273|
|      | Error (%) | -3.89%       | -2.83%    | -3.53%|   |
|      | Q3      | 17           | 29        | 21 | 20 |
|      | Error (%) | 70.59%       | 23.53%    | 17.65%|   |
|      | Q4      | 28           | 34        | 23 | 24 |
|      | Error (%) | 21.43%       | -17.86%   | -14.29%|   |
Table A3.6
Information loss comparisons of I-395SN DC.

|        | Actual Count | Counts | GP  | PR  | NB  |
|--------|--------------|--------|-----|-----|-----|
| 2017   | Q1           | 11     | 12  | 12  |     |
|        | Error (%)    | 11     | 12  | 12  |     |
| Q2     | 4            | 0%     | 9.09%| 9.09%|     |
| Q3     | 10           | 50%    | 75% | 75% |     |
| Q4     | 14           | –10%   | –20%| –20%| –20%|
| 2018   | Q1           | 23     | 21  | 20  |     |
|        | Error (%)    | –8.70% | –8.70%| –13.04%|     |
| Q2     | 8            | 8      | 5   | 5   |     |
| Q3     | 4            | 0%     | –37.5%| –37.50%|     |
| Q4     | 8            | 25%    | 100%| 100%|     |
| 2019   | Q1           | 5      | 4   | 5   |     |
|        | Error (%)    | 0%     | –20%| 0%  |     |
| Q2     | 15           | 12     | 14  | 13  |     |
| Q3     | 11           | 12     | 11  | 11  |     |
| Q4     | 7            | 9      | 9   |     |     |
| 2020   | Q1           | 10     | 8   | 8   |     |
|        | Error (%)    | 10     | 8   | 8   |     |
| Q2     | 29           | 28     | 28  | 28  |     |
| Q3     | 25           | –3.45%| –3.45%| –3.45%|     |
| Q4     | 69           | 24     | 22  | 22  |     |
|        | Error (%)    | 15.94%| –1.45%| 0%  |     |

References

Abdulhafedh, A., 2016. Crash frequency analysis. J. Transp. Technol. 6 (4), 169–180.
Adanu, E.K., Brown, D., Jones, S., Parrish, A., 2021. How did the covid-19 pandemic affect crashes and crash outcomes in Alabama? Accid. Anal. Prev. 163, 106428.
Ahn, J., Ko, E., Kim, E.Y., (Year). Highway traffic flow prediction using support vector regression and bayesian classifier. In: Proceedings of the 2016 International conference on big data and smart computing (BigComp), pp. 239-244.
Bae, B., Lee, C., Pak, T.-Y., Lee, S., 2021. Identifying temporal aggregation effect on crash-frequency modeling. Sustainability 13 (11), 6214.
Bae, B., Lee, C., Pak, T.-Y., Lee, S., (Year). Identifying temporal aggregation effect on crash-frequency modeling. Sustainability 13 (11), 6214.
Brillinger, D.R., 1994. Time series, point processes, and hybrids. Canadian J. Stat. 22 (2), 177–206.
Chang, A., Sanier, N., Laureyns, A., 2017. Proactive methods for road safety analysis. SAE Technical Paper.
Chang, H.-L., Jovianis, P.P., 1990. Formulating accident occurrence as a survival process. Accid. Anal. Prev. 22 (5), 407–419.
Chen, C., Guo, F., 2016. Evaluating the influence of crashes on driving risk using recurrent event models and naturalistic driving study data. J. Appl. Stat. 43 (12), 2225–2238.
Colonna, P., Intini, P., 2020. Compensation effect between deaths from covid-19 and crashes: The italian case. Transp. Res. Interdiscip. Perspect. 6, 100170.
Demiroluk, S., Ozbay, K., 2015. A doubly stochastic point process model for modeling crashes along a corridor. United States, Washington DC.
Dong, X., Lei, T., Jin, S., Hou, Y., Year. High-way traffic flow prediction based on short-term traffic flow prediction using support vector regression and bayesian classifier. In: Proceedings of the 2016 International conference on big data and smart computing (BigComp), pp. 239-244.
Hensher, D.A., Manering, F.L., 1994. Hazard-based duration models and their application to transport analysis. Transport Reviews 14 (1), 63–82.
Hensher, D.A., Manering, F.L., 1994. Hazard-based duration models and their application to transport analysis. Transport Reviews 14 (1), 63–82.
Inada, H., Ashraf, L., Campbell, S., 2021. Covid-19 lockdown and fatal motor vehicle collisions due to speed-related traffic violations in japan: a time-series study. Injury Prev. 27 (1), 98–100.
Jovianis, P.P., Chang, H.-L., 1989. Disaggregate model of highway accident occurrence using survival theory. Accid. Anal. Prev. 21 (5), 445–458.
Kalbfleisch, J.D., Prentice, R.L., 2011. The Statistical Analysis of Failure Time Data. John Wiley & Sons.
Lasko, T.A., 2014. Efficient inference of gaussian-process-modulated renewal processes with application to medical event data. Uncertain Artif. Intell. 469–476.
Lazo, L., 2021. Traffic Counts Fell During the Coronavirus Pandemic, But Road Fatalities Still Increased. The Washington Post.
Liao, H., Lowry, M., 2021. Speeding and traffic-related injuries and fatalities during the coronavirus pandemic: The cases of seattle and new york city. medRxiv.
Liang, Y., Zheng, Y., Zhang, H., Chen, L., Year. Traffic prediction in a bike-sharing system. In: Proceedings of the 2018 International Conference on big data and smart computing (BigComp), pp. 239-244.
Luo, S., Kao, Y., Year. Methodological frontier and future directions. Analyt. Methods Accid. Res. 1, 1–10.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the analytic methods in accident research: Methodological frontier and future directions. Analyt. Methods Accid. Res. 1, 1–10.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analyt. Methods Accid. Res. 1, 1–10.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analyt. Methods Accid. Res. 1, 1–10.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analyt. Methods Accid. Res. 1, 1–10.
McGreavy, P., 2020. Tickets for speeding in excess of 100 mph surge 67% amid coronavirus shutdown, csp says. URL. www.latimes.com/california/story/2020-04-22/tickets-for-drivers-speeding-more-than-100-mph-surges-87-amid-california-sutdown-csp-says (accessed 11.1. 2021).
Mei, Z., Xiang, F., Zhou-Hui, L., Year. Short-term traffic flow prediction based on combination model of xgboost-lightgbm. In: Proceedings of the 2018 International Conference on Sensor Networks and Signal Processing (SNSP), pp. 322-327.
Moosavi, S., Samavatian, M.H., Parthasarathy, S., Ramnath, R., 2019. A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409.
Mochida, M., Tsur, T., 1995. A new point process model for modeling the occurrences of traffic accidents. Transp. Res. Part B: Methodolog. 29 (1), 31–42.
Moosavi, S., Samavatian, M.H., Parthasarathy, S., Ramnath, R., 2019. A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409.
Moslonka, C., Sekimoto, K., 2020. Memory through a hidden martingale process in progressive quenching. Phys. Rev. E 101 (6), 062139.

Nhtsa, 2020. Early estimates of motor vehicle traffic fatalities and fatality rate by subcategories through June 2020. National Highway Traffic Safety Administration.

Park, E.S., Lord, D., 2007. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. Transp. Res. Rec. 2019 (1), 1–6.

Rapoport, M.J., Chee, J.N., Aljenabi, N., Byrne, P.A., Naglie, G., Ilari, F., Elzohairy, Y., Virgilis, E., Mulsant, B.H., 2021. Impact of covid-19 on motor vehicle injuries and fatalities in older adults in Ontario, Canada. Accid. Anal. Prev. 157, 106195.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 (5), 688–701.

Schlogl, M., 2020. A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach. Accid. Anal. Prev. 136, 105398.

Sengupta, A., Gayah, V.V., Donnell, E.T., 2021. Examining the impacts of crash data aggregation on SPF estimation. Accid. Anal. Prev. 160, 106313.

Shankar, V., Jovanis, P.P., Aguero-Valverde, J., Gross, F., 2008. Analysis of naturalistic driving data: Prospective view on methodological paradigms. Transp. Res. Rec. 2061 (1), 1–8.

Shi, X., Wong, Y.D., Li, M.-Z.-F., Chai, C., 2018. Accident risk prediction based on driving behavior feature learning using cart and xgboost. Transporation Research Board 97th Annual Meeting.

Usdot, 2021. Travel monitoring traffic volume trends. URL. https://www.fhwa.dot.gov/policyinformation/travelmonitoring/tdt.cfm (accessed 11.01.2021).

Usman, T., Fu, L., Miranda-Moreno, L.F., 2011. Accident prediction models for winter road safety: does temporal aggregation of data matter? Transp. Res. Rec. 2237 (1), 144–151.

Usman, T., Fu, L., Miranda-Moreno, L.F., 2012. A disaggregate model for quantifying the safety effects of winter road maintenance activities at an operational level. Accid. Anal. Prev. 48, 368–378.

Vandoros, S., 2021. Covid-19, lockdowns and motor vehicle collisions: Empirical evidence from Greece. Injury Prev.

Vanlaar, W.G.M., Woods-Fry, H., Barrett, H., Lyon, C., Brown, S., Wicklund, C., Robertson, R.D., 2021. The impact of covid-19 on road safety in Canada and the United States. Accid. Anal. Prev. 160, 106324.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R.S., Ahrentzen, S., 2018. Random forest based hourly building energy prediction. Energy Build. 171, 11–25.

Wu, C.-H., Ho, J.-M., Lee, D.-T., 2004. Travel-time prediction with support vector regression. IEEE Trans. Intell. Transp. Syst. 5 (4), 276–281.

Xie, K., Ozbay, K., Yang, H., Yang, D., 2019. A new methodology for before-after safety assessment using survival analysis and longitudinal data. Risk Anal. 39 (6), 1343–1357.

Xie, K., Wang, X., Ozbay, K., Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Analyst. Methods Accid. Res. 2, 39–51.

Yan, X., Wang, B., An, M., Zhang, C., 2012. Distinguishing between rural and urban road segment traffic safety based on zero-inflated negative binomial regression models. Discr. Dyn. Nat. Soc. 2012, 1–11.

Zheng, L., Sayed, T., Manering, F., 2020. Modeling traffic conflicts for use in road safety analysis: a review of analytic methods and future directions. Analyst. Methods Accid. Res. 29, 100142.

Zou, Y., Zhang, Y., Lord, D., 2013. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. Accid. Anal. Prev. 50, 1042–1051.

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Analyst. Methods Accid. Res. 1, 39–52.