Research Article

Statistical Safety Performance Models considering Pavement and Roadway Characteristics

Joshua Q. Li, Wenyao Liu, Xue Yang, Pan Lu, and Kelvin C. P. Wang

1School of Civil and Environmental Engineering, Oklahoma State University, Stillwater 74078, USA
2Department of Transportation and Logistics, North Dakota State University, Fargo 58102, USA

Correspondence should be addressed to Joshua Q. Li; qiang.li@okstate.edu

Received 23 December 2021; Revised 14 March 2022; Accepted 21 April 2022; Published 9 May 2022

1.Introduction

Transportation systems play crucial roles in providing the foundation for economic and social development around the world. Although transportation agencies strive to improve transportation safety, nearly 6.6 million vehicles encounter road traffic crashes every year. Crashes result in tremendous financial, corporeal, and emotional affliction to the affected families, as well as productivity reduction to the whole society [1]. In the United States, the National Highway Traffic Safety Administration (NHTSA) estimated that 38,680 people perished in traffic crashes in 2020, a 7.2% increase compared to the 36,096 fatalities reported in 2019 [2]. Identifying the significant contributing factors to crashes and recognizing their effects are essential for proactive decision-making to reduce roadway crashes [3].

Researchers have put substantial efforts into investigating the intrinsic correlation between crash frequency and risk factors, including implementing traffic crash prediction models as valuable tools for safety analysis [3]. Recently, there has been an ever-growing investigation on utilizing machine learning (ML) models for predicting crashes [4]. However, statistical models remain the mainstream as the amount of crash data is usually small and inadequate for machine learning [5]. Besides, statistical models have the advantage of clearly defining the functional forms and revealing the significance of the contributing factors [6]. Various statistical models can predict the crash frequency and identify the contributing factors [7]. Several recognized models include the Poisson regression model [8], the negative binomial model [9, 10], and the zero-inflated Poisson/negative binomial model [11]. However, choosing the appropriate statistical model is challenging and confusing on many occasions.

Contributing factors to roadway crashes include human behaviors, road geometry, pavement characteristics, traffic
volume, vehicle performance, and environmental conditions [12]. A desirable model for analyzing the correlation regarding crash probabilities should be constructed with detailed driving data (vehicle and driver behaviors), roadway features, environmental factors, and crash data. However, such comprehensive data are seldom available as crashes happen so rarely and randomly that it is costly, even impossible to monitor all the information [7]. As an alternative, researchers have framed approaches to building up safety performance functions (SPFs) to investigate the factors that affect roadway crashes occurring in specific geological spaces (such as roadway segments or intersections) over a particular period (weekly, monthly, annually, or several years). This way of revealing a correlation between crashes and pavement features—the aspect the transportation agencies can control—is supported as research shows that roadway factors contributed to 34% of vehicle crashes [13, 14]. Several pavement variables were found to be statistically significant on crash severity, while their effects on frequency are not consistent [15]. In recent years, pavement properties including skid resistance [16], roughness [17], rutting, and surface texture in terms of mean profile depth (MPD) have been investigated using traditional and advanced models [18–20].

For decades, the Oklahoma Department of Transportation (ODOT) has been managing several rigorous database systems, including the roadway safety database, skid testing program, and pavement management data sets. This paper investigates the appropriate approaches to estimating Oklahoma crash frequency to (1) develop general forms of various models, (2) investigate and compare the model performance to address issues such as the over-dispersion and preponderance of zeros, and (3) interpret the correlation between crashes with roadway characteristics, including pavement surface friction. The procedure and results of this research are expected to provide a reference for transportation agencies for the selection of statistical models when building SPFs and evaluation of pavement friction significance regarding crashes.

1.1. Data Compilation and Description

1.1.1. Data Source and Compilation. Factors contributing to roadway crashes include three categories: human behavior, vehicle, and roadway environment [21]. In detail, the Highway Safety Manual Data Needs Guide [22] suggests analyses should include data on traffic volume, roadway geometry, site characteristics, and roadway surface conditions for safety analysis and modeling. This study extracted the crash data and the relevant available data sets provided by ODOT. These data sets include safety data from the Statewide Analysis for Engineering & Technology (SAFE-T) database, skid resistance from the Skid Studies Program, pavement surface characteristics, roadway conditions from the Pavement Management System (PMS), and construction records and material properties from the SiteManager® management system.

SAFE-T is a statewide database of crash information gathered from collision reports presented by law enforcement agencies [23]. Since 1998, crash data in Oklahoma have been recorded as point location events with various relevant information. Because crashes are rare and random events naturally fluctuate over time [21], four-year crash data were analyzed to decrease the regression-to-the-mean effect. This effect describes a statistical phenomenon that pavement sections with more crashes in a specific period tend to have fewer during the following period, even without disturbance or measures. In particular, crash data on the Interstate Highway System, US highways, and the state highways from 2012 to 2015 in Oklahoma were authorized and obtained from the SAFE-T database.

The Strategic Asset & Performance Management (SAPM) Division at ODOT historically managed the Skid Studies Program for monitoring the pavement friction in Oklahoma. The friction was measured using a locked-wheel skid tester operating at 80 km/h (50 mph), reported at an interval of approximately 0.8 km (0.5 miles), and saved in separate files by control sections. In recent years, ODOT had downsized the scope of skid studies from the entire statewide highway system to a small portion of the network, including annual testing only on US-69 (a truck corridor), interstate highways, as well as the Strategic Highway Research Program (SHRP) sites. With the changing scopes of testing sites and evolving data reporting formats, numerous data preprocessing efforts were made to combine those friction files into one universal database. Friction data from 2012 to 2015 were gathered and summarized in Table 1. It is notable that only a handful of control sections on US-69 were tested in 2013.

The PMS is another database maintained by ODOT recording pavement surface conditions (e.g., surface type, roughness, rutting, and macrotexture), geometry features (e.g., grade and curvature), and other pavement characteristics for the National Highway System (NHS) and State Highway System (SHS) routes [24]. The field pavement data were recorded every 0.01 miles with location labels including highway name, mile point, control section number, control subsection, and GPS. Subsections are the subsets of a control section generated by multiple criteria defined in the ODOT Road Inventory Manual [25]. These criteria include the changes in state highway junctions, political jurisdictions, urban area boundaries, surface widths or type change, and traffic volumes. It is reasonably justified that the traffic characteristics and surface conditions for one subsection are relatively uniform. Hence, it is considered the basic analysis unit in this study, and all crashes, roadway, and related supporting data were gathered and compiled by subsections for further safety analysis.

The maintenance, rehabilitation, and reconstruction (MR&R) activities within the four-year analysis period could affect roadway crash frequency and patterns. Therefore, the sites that had received MR&R work from 2012 to 2015 were excluded from the SPF development. These sites were identified based on the data in the SiteManager® construction database, a database that recorded the contracts of MR&R activities in Oklahoma and provided the project and location description, project number, and allocated days.
1.1.2. Data Summary and Correlation. After significant effort compiling various data items, all required data were connected by subsection and travel direction for each pavement segment. Consequently, 1,811 subsections were screened out with complete data sets for SPF modeling. The distribution of the number of crashes—the dependent factor—at these subsections over four years was presented in Figure 1. As observed, there exists an excess number of zeros, and it is a distinct right-skewed distribution.

The statistical descriptions of the compiled data sets are summarized in Table 2. The contributing factors in the compiled data sets include traffic volume in terms of the average annual daily traffic (AADT), pavement surface characteristics (friction, texture, surface type, and roughness), and roadway geometry features (longitudinal grade, horizontal curvature, and presence of shoulder or median). The segment length and AADT were transferred to the natural logarithm values to fit in the generalized linear regression [19, 26]. The classification of IRI (international roughness) is based on the FHWA recommendations, where 95 in/mile (1.5 m/km) is the upper threshold of “good” ride quality, 170 in/mile (2.7 m/km) is the upper threshold for “acceptable” ride quality, and IRI beyond 170 in/mile (2.7 m/km) is classified as “poor” ride quality [27, 28]. Following the AASHTO recommendation, the rut depth threshold is set as 0.5 in (12.7 mm) between high and medium rut severity and 0.25 (6.4 mm) between medium and low rut severity [29, 30]. It is observed that the proportion of segments with “good” pavement conditions is smaller than the national average according to FHWA Conditions and Performance Report to Congress [28]. This results from the filtering process used in this paper, which had excluded the segments with maintenance, rehabilitation, and repair (MR&R) work from 2012 to 2015 in order to eliminate its effect on changing crash patterns. As a result, the proportion of roadway segments in “poor” conditions was higher than expected. In addition, some research found that such data filtering may also help handle the unbalanced feature of crash data [26].

Table 3 summarized the correlation coefficients among the influencing variables and the corresponding significance level. The Pearson’s correlation coefficients shown in the lower triangular cells evaluate the linear relationship between two continuous variables. The Spearman correlation coefficients shown in the upper triangular cells evaluate the monotonic relationship between two variables. The coefficients with an absolute value larger than 0.5 were marked with bold fonts. For example, the positive correlation coefficient (0.56) between the existence of median and the natural log of traffic volume shows that segments with higher traffic volume are often constructed with a median. Meanwhile, the significance level is calculated to determine whether the correlation between variables is significant. In general, the significance level with a p-value lower than 0.05 indicates a correlation exists between the two variables, and a correlation coefficient greater than 0.7 demonstrates the multicollinearity between the variables [31–33]. As observed in Table 3, some variables are correlated with a p-value less than 0.05, while no correlation coefficients have an absolute value higher than 0.7, indicating the chosen predictors have no evidence of manifest multicollinearity and can be used for building generalized linear models.

2. Methodologies
Crash occurrence, statistically considered as count data, can be appropriately modeled using several methods; the most popular applied are Poisson and negative binomial (NB) models. The hurdle negative binomial (HNB) and zero-inflated negative binomial (ZINB) regression models were also developed and implemented for highway safety analysis to address the phenomena of zero inflation.

2.1. Poisson Model. Poisson models often behave well in approximating non-negative integer count data that follows the Poisson distribution [34]. In the crash prediction case, the probability of subsection \( i \) having \( y_i \) (0, 1, 2, \ldots) crashes is given by

\[
P(y_i) = \frac{e^{(-\lambda_i)} \cdot (-\lambda_i)^{y_i}}{y_i!},
\]

where Poisson parameter \( \lambda_i \) is the predicted or the expected number of crashes per year at subsection \( i \). Poisson model identifies \( \lambda_i \) as a function of explanatory variables. The commonly selected functional form (or function link) is in the log-linear form:

\[
\log(\lambda_i) = \beta \mathbf{X}_i,
\]

where \( \beta \) and \( \mathbf{X}_i \) are vectors of estimable parameters and explanatory variables independently. Maximized log-likelihood method is applied to estimate the regression coefficients \( \beta \). The expected number of crashes on a specific subsection is given by \( E(Y_i) = \lambda_i = \exp(\beta \mathbf{X}_i) \).

As the Poisson model assumes the dependent variable follows the Poisson distribution, the variance of the number
of crashes is assumed to be equal to its mean; thus, the dispersion parameter $\phi = \text{Var}(Y)/E(Y)$ is fixed at 1. If this assumption is violated, on condition of either overdispersed ($\phi > 1$) or underdispersed ($\phi < 1$), the resulting parameter estimate will be biased [35].

2.2. Negative Binomial Model. The negative binomial (ZB) model is derived from the Poisson model to solve the overdispersion phenomenon in count data sets [36]. ZB model assumes that

$$\log(\lambda_i) = \beta X_i + \epsilon_i,$$  

where $\epsilon_i$ is a gamma-distributed error term and $\beta$ and $X_i$ are defined in the Poisson model. The addition of $\epsilon_i$ relaxes the constraint of equal mean and variance as follows:

$$\text{Var}(Y) = E(Y)[1 + \alpha E(Y)],$$

$$E(Y) + \alpha E(Y)^2,$$  \hspace{1cm} (4)

where $\alpha$ is often referred to as the dispersion parameter. Poisson model can be regarded as a negative binomial model with $\alpha$ of 0. The NB model is predominant in traffic crash data analysis, especially when the dependent variable is overdispersed distributed [35]. Nevertheless, the NB model is limited in handling an excess of zeros [37].

2.3. Hurdle Negative Binomial Model. Hurdle models can depict data with overdispersion and excess-zero observation by combining two components: a truncated count component (Poisson or negative binomial) for modeling positive

| Description                      | Mean  | Standard deviation | Minimum | Maximum |
|----------------------------------|-------|--------------------|---------|---------|
| **Response variables**           |       |                    |         |         |
| (1) N_Crash: number of crashes   | 16.6  | 46.2               | 0       | 642     |
| **Continuous variables**         |       |                    |         |         |
| (2) Ln_Lgth: natural log of the segment length in feet | 8.3   | 0.9                | 5.6     | 10.9    |
| (3) Ln_AADT: natural log of average annual daily traffic | 9.8   | 1.0                | 7.4     | 12.0    |
| (4) Lq_MPD: largest quartile of mean profile depth | 0.7   | 0.3                | 0.2     | 2.5     |
| (5) Av_Frict: average friction value in skid number | 40.1  | 6.8                | 16.9    | 62.0    |
| (6) Av_Grad: average longitudinal grade | 1.0   | 0.6                | 0       | 4.5     |
| (7) Lg_Dgcv: largest degree of curve | 1.2   | 3.5                | 0       | 66.6    |
| (8) Ln_Lgcv: natural log of the length of curve in feet | 1.8   | 3.1                | 0       | 9.7     |
| **Indicator variables** (described with proportion for each category) |       |                    |         |         |
| (9) PvType: pavement type        |       |                    |         |         |
| 1. AC:                         | 57.8% |                    |         |         |
| 2. JCP:                       | 30.6% |                    |         |         |
| 3. CRCP:                      | 11.7% |                    |         |         |
| (10) RrloUrB: rural or urban   |       |                    |         |         |
| 1. Rural:                     | 55.5% |                    |         |         |
| 2. Urban:                    | 44.5% |                    |         |         |
| (11) ExsMdn: existence of median |       |                    |         |         |
| 1. No:                       | 10.7% |                    |         |         |
| 2. Yes:                      | 89.3% |                    |         |         |
| (12) ExsShd: existence of shoulder |       |                    |         |         |
| 1. No:                       | 4.1%  |                    |         |         |
| 2. Yes:                      | 95.9% |                    |         |         |
| (13) RnkIRI: rank of international roughness index |       |                    |         |         |
| 1. Good:                     | 4.8%  |                    |         |         |
| 2. Acceptable:               | 19.8% |                    |         |         |
| 3. Poor:                    | 75.4% |                    |         |         |
| (14) RnkRut: rank of maximum rutting depth |       |                    |         |         |
| 1. Good:                     | 6.4%  |                    |         |         |
| 2. Acceptable:               | 23.3% |                    |         |         |
| 3. Poor:                    | 70.3% |                    |         |         |

Note. AC: asphalt pavement; JCP: jointed concrete pavement; and CRCP: continuously reinforced concrete pavement.
Table 3: Correlation coefficients between predictor variables (1,811 segments).

|       | 2   | 3    | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  |
|-------|-----|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| (2) Ln_Lgth | 1   | −0.09*** | 0.13*** | 0.06** | 0.01 | 0.04 | 0.01 | −0.11*** | −0.37*** | −0.03 | 0.14*** | 0.09*** | 0.18*** |
| (3) Ln_AADT | −0.12*** | 1   | −0.11*** | −0.27*** | 0.08*** | 0.07*** | 0.06** | 0.28*** | 0.41*** | 0.48*** | 0.10*** | −0.02 | 0.05** |
| (4) Lq_MPd  | 0.13*** | −0.01 | 1   | −0.23*** | −0.03 | −0.13*** | −0.08*** | −0.63*** | −0.12*** | −0.13*** | 0.03  | −0.08*** | 0.03  |
| (5) Av_Frict | 0.07*** | −0.27*** | −0.21*** | 1   | −0.10*** | −0.09*** | −0.14*** | 0.20*** | −0.17*** | −0.02 | −0.03 | 0.01  | −0.07*** |
| (6) Av_Grad  | −0.05** | 0.03  | −0.02 | −0.09*** | 1   | 0.04* | 0.02 | 0.02 | 0.07*** | 0.01  | 0.04* | 0.04* | 0.06** |
| (7) Lg_Dgcv  | −0.01 | −0.02 | −0.05** | −0.03 | 0.01 | 1   | **0.62*** | 0.09*** | 0.21*** | −0.16*** | −0.06*** | 0.16*** | 0.15*** |
| (8) Ln_LgcV | 0.01  | 0.06** | −0.09*** | −0.16*** | 0.00 | 0.40*** | 1   | 0.07*** | 0.23*** | −0.19*** | −0.06*** | 0.15*** | 0.13*** |
| (9) PVTyP   | −0.10*** | 0.30*** | **0.52*** | 0.15*** | 0.03 | 0.04* | 0.07*** | 1   | 0.14*** | 0.25*** | −0.02 | 0.04* | 0.01  |
| (10) RlOuRb | −0.38*** | 0.42*** | −0.09*** | −0.19*** | 0.06** | 0.12*** | 0.23*** | 0.13*** | 1   | 0.09*** | −0.17*** | 0.06** | 0.03  |
| (11) ExsMdN | −0.04*  | **0.56*** | −0.02 | −0.02 | 0.00 | −0.16*** | −0.20*** | 0.24*** | 0.09*** | 1   | 0.16*** | −0.04* | 0.00  |
| (12) ExsShd | 0.15*** | 0.08*** | 0.06** | −0.03 | 0.04 | −0.06* | −0.06* | −0.02 | −0.17*** | 0.16*** | 1   | −0.09*** | −0.08*** |
| (13) RnkIRI | 0.11*** | −0.01 | −0.09*** | 0.01 | 0.04* | 0.09*** | 0.14*** | 0.05** | 0.05** | −0.04* | −0.08*** | 1   | **0.56*** |
| (14) RnkRut | 0.19*** | 0.06** | 0.05** | −0.07*** | 0.05* | 0.09*** | 0.13*** | 0.02 | 0.03 | −0.00 | −0.08*** | 0.57*** | 1   |

Note. 1. The abbreviation of variables is consistent with Table 2. 2. Lower triangular cells report Pearson’s correlation coefficient and upper triangular cells report Spearman’s rank correlation. 3. Significance indicated with p-value: ***p < 0.01, **p < 0.05, and *p < 0.1.
Zero-inflated models are also the Hurdle model all come from its zero module. The zero-inflated model can be structured as follows:

\[ P(Y_i = y_i) = \begin{cases} 
    p_i, & y_i = 0, \\
    (1 - p_i) \cdot \frac{p(y_i; \mu_i)}{1 - p(y_i = 0; \mu_i)}, & y_i > 0,
\end{cases} \]  

where \( p_i \) resembles the probability of one pavement segment belonging to the zero modules, \( p(y_i; \mu_i) \) represents a probability mass function for a normal count distribution with parameters vector \( \mu_i \), and \( p(y_i = 0; \mu_i) \) is the count distribution evaluated at zero \([39]\). The zero predictions in the Hurdle model all come from its zero module.

2.4. Zero-Inflated NB Models. Zero-inflated models are also two-component mixture models that deal with overdispersed data with excess zero counts, but the sampling zeros occur by chance from two components \([40]\). The zero-inflated model can be structured as follows:

\[ P(Y_i = y_i) = \begin{cases} 
    p_i + (1 - p_i) p(y_i = 0; \mu_i), & y_i = 0, \\
    (1 - p_i) p(y_i; \mu_i), & y_i > 0,
\end{cases} \]  

where \( p_i \) represents the probability of a structural zero with \( p(y_i; \mu_i) \) and \( p(y_i = 0; \mu_i) \) defined in the Hurdle model. As such, zero numbers come from both "structural" origins and "sampling" component \( p(y_i; \mu_i) \) \([39]\).

3. Results and Discussion

3.1. Model Results. As the first attempt, the Poisson model was fitted to seize the relationship between the number of crashes and available pavement characteristics. A stepwise process was conducted to optimize the modeling results. In each step, a variable was subtracted from the set of explanatory variables by minimizing Akaike’s Information Criteria (AIC) \([41]\). The best-fitted Poisson model has the smallest AIC in the stepwise processing, and the model results on studied 1,811 segments are summarized in the first column of Table 3. All variables in the fitted Poisson model are statistically significant in determining the crash likelihood. However, a Poisson model requires the dispersion parameter to be approximately 1, with equal mean and variance. However, the actual dispersion parameter in this study was calculated to be 31.7, which is significantly larger than 1, indicating that distinctive overdispersion characteristics exist in the data sets, and thus, the Poisson regression assumption is violated.

To address the observed overdispersion of the crash data, attempts were made to analyze crash frequency using the NB model, hurdle negative binomial model (HNB), and ZINB regression techniques. The NB model was chosen for its popularity in tacking overdispersion; ZINB and HNB were selected to deal with both overdispersion and the preponderance of zeros. Similarly, for each type of model, the stepwise variable selection method was applied to optimize the fitting performance by minimizing the AIC of the models. It is noteworthy that ZINB and HNB are composed of two model parts so the stepwise optimization should be conducted on each model part separately. The regression results of all investigated models are presented in Table 4 with the estimated coefficients and their corresponding significance levels \([42]\). As shown in Table 4, the subsection length, traffic volume, average grade, and length of the curve are statistically significant with positive estimated coefficients across all the models. Positive coefficients indicate that crash frequency increases with these parameters. In other others, fewer crashes are expected on segments subsections with lower traffic volumes, smaller grades, and shorter lengths of curves. This observation is consistent with the analysis in former research \([15]\).

On the contrary, the presence of shoulders has a negative coefficient in all the models, indicating that the presence of the shoulder can decrease the expected crash frequency. Likewise, using the asphalt concrete (AC) surface type as the benchmark for comparison, the negative coefficient for the pavement type parameter indicates that continuously reinforced concrete pavements (CRCP) and jointed concrete pavements (JCP) are associated with a lower level of crash frequencies. This indicates the pavement type does have an impact on the probability of crashes. However, some parameters exhibit different levels of statistical significance or even opposite trends of estimates on the crash frequency among the four types of models. Such differences could be originated from unobserved heterogeneity within the data sets in this study. On the other hand, such differences could provide necessary information for model performance evaluation so that the best model could be recommended.

3.1.1. Comparisons and Interpretation. An intuitive investigation of the performance of fitted models is the plotting of the predicted and observed density distributions, as shown in Figure 2. The observed distribution of the number of crashes is presented in the histogram. Meanwhile, the predictions of the four fitted count models are plotted in lines with different patterns and colors. As observed, the Poisson model prediction results show a distinct difference in Figure 2. The observed distribution of the number of crashes is presented in the histogram. Meanwhile, the predictions of the four fitted count models are plotted in lines with different patterns and colors. As observed, the Poisson model prediction results show a distinct difference in the histogram form the observed distribution of the number of crashes. Primarily when it is used to predict sections with a small number of crashes. Such deviations are primarily caused by the overdispersion of the dependent variable, which could be handled better through the NB, HNB, and ZINB models.

Another approach to evaluating the overdispersion phenomenon is the rootogram, as shown in Figure 3. Rootograms graphically compared the frequencies of fitted and observed probability. This method plotted the predicted frequency with the thick red line and observed counts with bars hanging from the red line of expected counts \([43]\). The horizontal line at 0 frequency distinguishes where the model is over- or underpredicting. If the hanging bar (or rootogram) does not reach the zero line, the model overpredicts the occurrence at that count range. In contrast, if the bar exceeds the zero line, the model underpredicts the crashes.

As observed in Figure 3, the rootograms for the Poisson model dramatically exceed the zero line when the crash counts exceed the zero line.
Table 4: Estimated coefficients of fitted count regression models.

|        | Poisson  | NB   | Truncated NB | HNB    | Zero hurdle | NB   | ZINB   |
|--------|----------|------|--------------|--------|-------------|------|--------|
| (2) Ln_Lgth | 0.566*** | 0.560*** | 0.510*** | 0.359*** | 0.568*** |
|         | (0.008)  | (0.048)  | (0.040)  | (0.059)  | (0.037)  |
| (3) Ln_AADT | 1.401*** | 1.275*** | 1.231*** | 0.452*** | 1.217*** |
|         | (0.012)  | (0.061)  | (0.044)  | (0.070)  | (0.039)  |
| (4) Lq_MPDP | 0.082*** | 0.404*** | −0.701*** |          | 0.744*** |
|         | (0.023)  | (0.139)  | (0.161)  |          | (0.188)  |
| (5) Av_Frict | −0.004*** | −0.010*  |          | −0.017** |          |
|         | (0.001)  | (0.006)  |          | (0.008)  |          |
| (6) Av_Grad | 0.275*** | 0.231*** | 0.172*** | 0.142*   | 0.215*** |
|         | (0.011)  | (0.066)  | (0.062)  | (0.087)  | (0.058)  |
| (8) Ln_Lgcv | 0.040*** | 0.403*** | 0.034*** |          | 0.045*** |
|         | (0.002)  | (0.013)  | (0.011)  |          | (0.011)  |
| (9) PvType2 | −0.176*** | −0.186** | −0.184** | −0.391*** | −0.324*** |
|         | (0.022)  | (0.133)  | (0.118)  | (0.103)  | (0.074)  |
|          |          |          |          |          |          |
| (10) RtolUrb2 | 0.274*** | 0.168*   | 0.270*** |          | 0.247*** |
|         | (0.018)  | (0.098)  | (0.089)  |          | (0.083)  |
| (11) ExsMdn2 | −0.530*** | −0.645*** |          | −1.027*** |          |
|         | (0.041)  | (0.168)  |          | (0.219)  |          |
| (12) ExsShd2 | −1.463*** | −1.388*** | −1.034*** | −2.487*** | −1.190*** |
|         | (0.033)  | (0.199)  | (0.148)  | (0.524)  | (0.145)  |
| (13) RnkIRI2 | 0.264*** | 0.319*   |          |          |          |
|         | (0.045)  | (0.165)  |          |          |          |
| (14) RnkRut2 | −0.150*** |          |          |          |          |
|         | (0.042)  |          |          |          |          |
| RnkRut3 | −0.175*** |          |          |          |          |
|         | (0.043)  |          |          |          |          |
| Intercept | −14.996*** | −13.165*** | −13.876*** | −2.368*** | −13.521*** | −1.723*** |
|         | (0.146)  | (0.761)  | (0.565)  | (1.045)  | (0.492)  | (0.169)  |
| Observations | 1,811   | 1,811   | 1,811     | 1,811    | 1,811    | 1,811    |
| No. of parameters | 16      | 12      | 21        | 12       |          |          |
| No. of predicted zeros | 142     | 585     | 625       | 634      |          |          |
| AIC     | 46,152.6 | 10,792.8 | 10,586.2  | 10,661.7 | 10,661.7 |
| BIC     | 46,240.6 | 10,858.9 | 10,701.7  | 10,727.7 | 10,727.7 |
| Log likelihood | −23,060.3 | −5,385.4 | −5,272.1  | −5,318.8 |          |

Note 1. Significance indicated with p-value: * p < 0.1, ** p < 0.05, and *** p < 0.01. 2. Abbreviations of variables are consistent with those in Table 2. Indicator variables were compared to the first category; for example, PvType2 indicates segments with JCP surface compared to AC surface.
frequency is low, indicating that the Poisson model predicted much fewer crashes than the observed values from the field. The rootograms hanging above the zero lines indicated overfitting at larger crash counts (ranging from approximately 6 to 10). In other words, the variability of the observed crash data is much greater than what a Poisson model could predict. The other three models (NB, HNB, and ZINB) perform better as their rootograms converge around the zero lines, which are expected for a preferred crash prediction model. Minor differences were observed among the NB, HNB, and ZINB models from their rootograms, and thus, additional comparisons are needed to identify the best model type.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are the two most well-known criteria for model selection, and the model with smaller AIC and BIC values was preferred. The AIC and BIC values for the four investigated models are summarized in Table 4. The HNB and ZINB models have the smallest AIC and BIC values, followed by the NB model. The Poisson model has much larger AIC and BIC values since the overdispersion cannot be considered. However, the differences between the HNB and ZINB remain insignificant in terms of the AIC and BIC values.

As excess zeros were observed in the crash data, the predicted zero values from the fitted regression models were
also listed in Table 4. There were 625 roadway sections without any observed crashes. Due to its lack of considering the overdispersion distribution of dependent variables, the Poisson model illustrated the worst performance with only 142 zero predictions, consistent with the analysis results shown in Figure 3. The HNB model predicted the desired number of zeros because of its modeling structure. The ZINB model overpredicted the zeros, and the NB model underpredicted the zeros. Since the zero counts in this study are not dominating (625 divided by 1,811, which is 34.5%), the differences between the HNB and ZINB models are not apparent for this case study. However, the zero-state hypothesis in the zero-inflated count models has been criticized by researchers and is not widely accepted in the analysis of crash frequency [26, 44, 45].

Based on the comparisons of model performance, the HNB model is recommended as the superior model for developing the safety performance function in this study. The regression coefficients of the influencing variables, as shown in Table 4, can be used to establish the HNB model for crash frequency prediction and assist decision-making for improved roadway safety. Although the segmentation length and pavement type generally do not change after construction, transportation agencies can take measures to preserve and improve pavement conditions. The mean profile depth, on the other hand, has one negative coefficient in the zero component and a positive coefficient in the count model component, indicating two-sided effects on the predicted number of crashes. It may be caused by the driver’s abstraction on pavements with excellent conditions but being more cautious under poor pavement conditions [15]. The negative estimate of average friction and positive coefficients in the IRI ranking suggest countermeasures can be taken by improving friction and decreasing roughness. It should be noted that the observed crashes in segments decrease slowly with friction, between 2.5% and 0.9% per unit of friction in skid number. This observation is consistent with former research and provides support in maintaining a proper level of friction that helps improve transportation safety [18].

3.1.2. Limitations and Discussions. Despite the investigated models revealing the statistically significant relationship between the roadway characteristics and crashes, it is worthwhile to note that traffic crash prediction is a complex topic, and a variety of influencing factors are not available during data collection [18, 46]. These related factors include driver behavior, vehicle conditions, and environmental variations, which were generally assumed identical and characterized as assumptions of the generalized linear models. However, these assumptions may diverge from real transportation observations and result in a high likelihood of unobserved heterogeneity [47, 48].

Even for the listed independent factors, the investigated models in this paper were all fixed models, which make an inherent assumption on homogeneous effects of the observed factors. This assumption may not hold firmly when inspecting the data sets closely. Obeying this assumption means that in a generalized linear model, a linear relationship with identical slopes and intercepts across different groups should exist between the linked variable and independent parameters [18, 19, 49]. Figure 4 shows the natural log value of crash numbers (all zero crashes were added by 1 to be mathematically logical) versus the natural log of traffic volume (a) and average segment friction (b). For each pavement type, their correlations to independent variables were shown with regression lines and confidence bands (different colors). It is observed that the slopes and intercepts vary across different pavement types, indicating the
homogeneous assumption is challenged and unobserved heterogeneity presents in the data sets.

Another assumption of the investigated models is that all independent parameters were assumed to have no mutual dependence and no correlation. The correlation coefficients shown in Table 3 reveal no severe multicollinearity, but there are some degrees of correlations between independent parameters. In fact, correlations between pavement characteristics are often observed. For example, a pavement segment with poor conditions often simultaneously shows elevated rutting and roughness. Neglecting cross-correlations among investigated factors may result in the downward bias of the estimated parameter variances [19].

The application of uncorrelated and correlated random parameters models has the potential to partially account for unobserved heterogeneity and cross-correlations. Several studies have revealed the statistical superiority of the correlated random parameters count model [19, 20, 50]. However, these models involve computation and modeling complexity when facing many independent factors. More advanced models including uncorrelated and correlated random parameters models will be investigated and applied in future research efforts.

4. Conclusions

The contribution of this paper is twofold in both the modeling process and results. First, the SPFs for roadway segments in the AASHTO HSM mainly include two categories of crash-related indicators: the annual average daily traffic (AADT) and the segment length, while the roadway infrastructure-related characteristics, which have potential effects on highway safety, have not been thoroughly investigated for predicting crash rates. This study acquired and compiled extensive data sets in Oklahoma and included various roadway infrastructure-related characteristics in the model. Pavement surface characteristics (friction, IRI, rutting, and surface type) and roadway geometry (longitudinal grade, horizontal curvature, length of curve, number of lanes, presence of shoulder, and median) were identified to be statistically significant to roadway crashes. The developed SPF with roadway infrastructure factors assists knowledgeable decisions to select pavement preservation and maintenance practices.

Secondly, this paper implemented and compared the performance of four statistical counting models—Poisson model, negative binomial (NB) model, hurdle negative binomial (HNB) model, and zero-inflated negative binomial (ZINB) model—for the development of SPFs using the Oklahoma data sets. Various model comparison and evaluation methods were applied for the recommendation of the best model type for roadway crash modeling and safety analysis, with the following key findings:

(i) The overdispersion problem can be addressed by the NB, HNB, and ZINB models.
(ii) The HNB and ZINB models can better predict the excessive zeros in the crash data than the NB model.

Based on comparisons between the goodness of fit and the number of independent variables, the HNB is recommended as the preferred model.

(iii) The HNB regression results indicate that improving pavement surface friction and adding outer shoulders could be effective MR&R candidates in reducing the number of crashes and improving roadway safety performance.

It should be acknowledged that the data used in this study are limited in size and scope. Additional data items with a wider range of roadway facilities should be collected and used to develop comprehensive SPFs for the state. Besides, advanced models such as correlated random parameter models should be investigated and compared in future research.

Data Availability

All data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors’ Contributions

Joshua Q. Li and Wenyao Liu conceptualized the study; Wenyao Liu did the formal analysis; Joshua Q. Li contributed to collecting resources; Wenyao Liu contributed to writing the original draft and did the revision; Joshua Q. Li and Pan Lu reviewed and edited the manuscript; Wenyao Liu and Xue Yang contributed to visualization; and Kelvin C. P. Wang supervised the study. All authors have read and agreed to the final version of the manuscript.

References

[1] A. Abdulhafedh, "Road crash prediction models: different statistical modeling approaches," Journal of Transportation Technologies, vol. 7, no. 2, pp. 190–205, 2017.
[2] National Center for Statistics and Analysis, Traffic Safety Facts: Early Estimate of Motor Vehicle Traffic Fatalities in 2020 Summary, National Highway Traffic Safety Administration (NHTSA), Washington DC, USA, 2020.
[3] X. Wen, Y. Xie, L. Wu, and L. Jiang, “Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP,” Accident Analysis & Prevention, vol. 159, Article ID 106261, 2021.
[4] F. Mannering, C. R. Bhat, V. Shankar, and M. Abdel-Aty, “Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis,” Analytic Methods in Accident Research, vol. 25, Article ID 100113, 2020.
[5] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, “Crash injury severity analysis using a two-layer Stacking framework,” Accident Analysis & Prevention, vol. 122, pp. 226–238, 2019.
[6] W. Cheng, G. S. Gill, R. Dasu, M. Xie, X. Jia, and J. Zhou, “Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections
based on crash types,” Accident Analysis & Prevention, vol. 99, pp. 330–341, 2017.

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

[8] S.-P. Miaou and H. Lum, “Modeling vehicle accidents and highway geometric design relationships,” Accident Analysis & Prevention, vol. 25, no. 6, pp. 689–709, 1993.

[9] D. Lord, “Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter,” Accident Analysis & Prevention, vol. 38, no. 4, pp. 751–766, 2006.

[10] S. P. Miaou, “The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions,” Accident Analysis & Prevention, vol. 26, no. 4, pp. 471–482, 1994.

[11] J. Lee and F. Mannering, “Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis,” Accident Analysis & Prevention, vol. 34, no. 2, pp. 149–161, 2002.

[12] G. Yang, Q. J. Li, Y. Zhan, W. Yu, K. C. P. Wang, and Y. Peng, “Field performance evaluation of high friction surface treatments (HfST) in Oklahoma,” Canadian Journal of Civil Engineering, vol. 46, no. 12, pp. 1142–1150, 2019.

[13] J. Treat, N. Tumbas, S. McDonald, and D. Shinar, “Tri-level study of the causes of traffic accidents: final report Executive summary,” Transportation Research Institute (UMTRI), 1979.

[14] Fhwa, “Pavement friction - safety | federal highway administration,” 2021, https://www.fhwa.dot.gov/research/tracts/pavement_friction.

[15] Y. Li, C. Liu, and L. Ding, “Impact of pavement conditions on crash severity,” Accident Analysis & Prevention, vol. 59, pp. 399–406, 2013.

[16] J. N. Ivan, N. Ravishanker, E. Jackson, B. Aronov, and S. Guo, “A statistical analysis of the effect of wet-pavement friction on highway traffic safety,” Journal of Transportation Safety & Security, vol. 4, no. 2, pp. 116–136, 2012.

[17] A. Elghriany, P. Yi, P. Liu, and Q. Yu, “Investigation of the effect of pavement roughness on crash rates for rigid pavement,” Journal of Transportation Safety & Security, vol. 8, no. 2, pp. 164–176, 2016.

[18] S. Chen, T. U. Saeed, and S. Labi, “Impact of road-surface condition on rural highway safety: a multivariate random parameters negative binomial approach,” Analytic methods in accident research, vol. 16, pp. 75–89, 2017.

[19] T. U. Saeed, T. Hall, H. Baroud, and M. J. Volovski, “Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: an empirical assessment of multilane highways,” Analytic methods in accident research, vol. 23, Article ID 100101, 2019.

[20] M. S. Yamany, T. U. Saeed, M. Volovski, and A. Ahmed, “Characterizing the performance of interstate flexible pavements using artificial neural networks and random parameters regression,” Journal of Infrastructure Systems, vol. 26, no. 2, Article ID 04020010, 2020.

[21] Aashto, Highway safety manual, American Association of State Highway and Transportation Officials, Washington, DC, USA, 2010.

[22] M. National Academies of Sciences, Highway safety manual data Needs Guide, The National Academies Press, Washington, DC, USA, 2008.

[23] A. Adams and M. Warren, ODOT 2017 SAFE-T User Manual, ODOT Traffic Engineering Collision Analysis and Safety Branch, Oklahoma, OK, USA, 2017.

[24] Oklahoma Department of Transportation, 2018-2027 Transportation Asset Management Plan, Oklahoma City, OK, USA, 2019, https://www.odot.org/SAPM/branch/planning/TAMP_and_Approval_Letter.pdf.

[25] Oklahoma Department of Transportation, Road Inventory Manual, Oklahoma City, OK, USA, 2010.

[26] D. Lord, X. Qin, and S. R. Geedipally, Highway Safety Analytics and Modeling, Elsevier, Amsterdam, The Netherlands, 2021.

[27] Fhwa (Federal Highway Administration), Highways Performance Monitoring System Field Manual, US Department of Transportation Washington, Washington, DC, USA, 2010.

[28] Fhwa (Federal Highway Administration) and Federal Transit Administration, “Status of the Nation’s Highways, Bridges, and Transit Conditions and Performance Report to congress,” 2015, https://www.fhwa.dot.gov/policy/23cpr/pdfs/23cpr.pdf.

[29] Aashto (American Association of State Highway and Transportation Officials), AASHTO Report of the AASHTO Joint Task Force on RuttingAASHTO, Washington, DC, USA, 1989.

[30] J. B. Sousa, J. Craus, and C. L. Monismith, “SHRP report A/IR-91-104: summary report on permanent deformation in asphalt concrete,” Report A/IR-91-104, National Research Council, Washington, DC, USA, 1991.

[31] R. C. Rockwell, “Assessment of multicollinearity,” Sociological Methods & Research, vol. 3, no. 3, pp. 308–320, 1975.

[32] C. F. Dormann, J. Elith, S. Bacher et al., “Collinearity: a review of methods to deal with it and a simulation study evaluating their performance,” Ecography, vol. 36, no. 1, pp. 27–46, 2013.

[33] R. Salmerón Gómez, C. B. G. García, and J. G. Pérez, “The multiColl package versus other existing packages in R to detect multicollinearity,” Computational Economics, 2021.

[34] C. N. Kumar, M. Parida, and S. S. Jain, “Poisson family regression techniques for prediction of crash counts using bayesian inference,” Procedia - Social and Behavioral Sciences, vol. 104, pp. 982–991, 2013.

[35] P. Lu and D. Tolliver, “Accident prediction model for public highway-rail grade crossings,” Accident Analysis & Prevention, vol. 90, pp. 73–81, 2016.

[36] J. Q. Li and W. Yu, “Enhanced safety performance function for highway segments in Oklahoma,” Journal of Infrastructure Systems, vol. 27, no. 3, Article ID 04021018, 2021.

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

[38] R. Yu, Y. Wang, M. Quddus, and J. Li, “A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data,” Analytic Methods in Accident Research, vol. 22, Article ID 100092, 2019.

[39] C. X. Feng, “A comparison of zero-inflated and hurdle models for modeling zero-inflated count data,” Journal of Statistical Distributions and Applications, vol. 8, no. 1, pp. 1–19, 2021.

[40] X. Gu, X. Yan, L. Ma, and X. Liu, “Modeling the service-route-based crash frequency by a spatiotemporal-random-effect zero-inflated negative binomial model: an empirical analysis for bus-involved crashes,” Accident Analysis & Prevention, vol. 144, Article ID 105674, 2020.
[41] J. E. Cavanaugh, “Unifying the derivations for the Akaike and corrected Akaike information criteria,” Statistics & Probability Letters, vol. 33, no. 2, pp. 201–208, 1997.

[42] M. Hlavac, “Stargazer: Well-Formatted Regression and Summary Statistics Tables,” 2022, https://cran.r-project.org/web/packages/stargazer/stargazer.pdf.

[43] C. Kleiber and A. Zeileis, “Visualizing count data regressions using rootograms,” The American Statistician, vol. 70, no. 3, pp. 296–303, 2016.

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

[45] D. Lord, S. P. Washington, and J. N. Ivan, “Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory,” Accident Analysis & Prevention, vol. 37, no. 1, pp. 35–46, 2005.

[46] S. Labi, “Efficacies of roadway safety improvements across functional subclasses of rural two-lane highways,” Journal of Safety Research, vol. 42, no. 4, pp. 231–239, 2011.

[47] S. Chen, T. U. Saeed, M. Alinizzi, S. Lavrenz, and S. Labi, “Safety sensitivity to roadway characteristics: a comparison across highway classes,” Accident Analysis & Prevention, vol. 123, pp. 39–50, 2019.

[48] M. Volovski, J. Murillo-Hoyos, T. U. Saeed, and S. Labi, “Estimation of routine maintenance expenditures for highway pavement segments: accounting for heterogeneity using random-effects models,” Journal of Transportation Engineering, Part A: Systems, vol. 143, no. 5, Article ID 04017006, 2017.

[49] M. Waseem, A. Ahmed, and T. U. Saeed, “Factors affecting motorcyclists’ injury severities: an empirical assessment using random parameters logit model with heterogeneity in means and variances,” Accident Analysis & Prevention, vol. 123, pp. 12–19, 2019.

[50] Q. Hou, A. P. Tarko, and X. Meng, “Analyzing crash frequency in freeway tunnels: a correlated random parameters approach,” Accident Analysis & Prevention, vol. 111, pp. 94–100, 2018.