11-14-2023

Exploring Statistical and Machine Learning-Based Missing Data Imputation Methods to Improve Crash Frequency Prediction Models for Highway-Rail Grade Crossings

Muhammad Umer Farooq

Aemal Khattak

IRF R2T Conference & Exhibition
November 14 – 17, 2023
Phoenix, AZ

PAPER TITLE	Exploring Statistical and Machine Learning-Based Missing Data Imputation Methods to Improve Crash Frequency Prediction Models for Highway-Rail Grade Crossings		
TRACK			
AUTHOR (Capitalize Family Name)	**POSITION**	**ORGANIZATION**	**COUNTRY**
Muhammad Umer FAROOQ	Post Doctoral Research Associate, Mid-America Transportation Center (MATC)	University of Nebraska-Lincoln	USA
CO-AUTHOR(S) (Capitalize Family Name	**POSITION**	**ORGANIZATION**	**COUNTRY**
Aemal KHATTAK	Professor and Director, Mid-America Transportation Center (MATC)	University of Nebraska-Lincoln	USA
E-MAIL (for correspondence)	mfarooq2@unl.edu		

KEYWORDS:

Missing data imputation, Crash frequency prediction, Highway-rail grade crossings, Statistical methods, Machine learning-based methods

ABSTRACT:

Highway-rail grade crossings (HRGCs) are critical spatial locations of transportation safety because crashes at HRGCs are often catastrophic, potentially causing several injuries and fatalities. Every year in the United States, a significant number of crashes occur at these crossings, prompting local and state organizations to engage in safety analysis and estimate crash frequency prediction models for resource allocation. These models provide valuable insights into safety and risk mitigation strategies for HRGCs. Furthermore, the estimation of these models is based on inventory details of HRGCs, and their quality is crucial for reliable crash predictions. However, many of these models exclude crossings with missing inventory details, which can adversely affect the precision of these models. In this study, a random sample of inventory details of 2000 HRGCs was taken from the Federal Railroad Administration's HRGCs inventory database. Data filters were applied to retain only those crossings in the data that were at-grade, public and operational (N=1096). Missing values were imputed using various statistical and machine learning methods, including Mean, Median and Mode (MMM) imputation, Last Observation Carried Forward (LOCF) imputation, K-Nearest Neighbors (KNN) imputation, Expectation-Maximization (EM) imputation, Support Vector Machine (SVM) imputation, and Random Forest (RF) imputation. The results indicated that the crash frequency models based on machine learning imputation methods yielded better-fitted models (lower AIC and BIC values). The findings underscore the importance of obtaining complete inventory data through machine learning imputation methods when developing crash frequency models for HRGCs. This approach can substantially enhance the precision of these models, improving their predictive capabilities, and ultimately saving valuable human lives.

Exploring Statistical and Machine Learning-Based Missing Data Imputation Methods to Improve Crash Frequency Prediction Models for Highway-Rail Grade Crossings

Dr. Muhammad Umer Farooq and Dr. Aemal Khattak [1]

[1]Mid-America Transportation Center, University of Nebraska Lincoln, USA
Email for correspondence, e.g. mfarooq2@unl.edu

1 INTRODUCTION

Missing data imputation becomes essential when it is crucial to utilize all the available information and avoid discarding records containing missing values. Missing values in data can pose significant challenges in predictive modelling, particularly in the context of crash frequency modelling. Crash frequency modelling is a crucial aspect of highway safety analysis, as it helps identify high-risk areas and allocate resources effectively for preventive measures. However, the presence of missing values in crash data can lead to biased estimates and inaccurate predictions, making it essential to explore and evaluate different missing data treatments to improve the performance of predictive models in crash severity modelling.

When there are missing values in the data points, they create gaps in the dataset, which can lead to a smaller sample size and introduce potential bias in the findings. As a consequence, the accuracy and reliability of predictive models can be adversely affected, as they may not be able to fully capture the complete patterns and relationships within the data. Moreover, missing values can introduce noise and distort the statistical properties of the dataset, affecting the model's ability to make accurate predictions. Several statistical and machine learning approaches have been created to tackle this issue. Upon reviewing the existing literature, it becomes apparent that the effectiveness of these methods is heavily influenced by the problem's domain characteristics, such as the number of cases, variables involved, and the patterns of missing data. Consequently, no definitive evidence points to a single method as superior to others (Royston 2004; Ye & Wang, 2018; Abdulhafedh, 2016; Deb & Liew, 2016; Imprialou & Quddus, 2018; Asgharpour et al., 2023; Farooq, 2023; Farooq et al., 2023).

In the context of crash prediction, one common approach to handling missing data is complete-case analysis (CC), which involves discarding crash records with missing information on any of the variables (Ye & Wang, 2018). While this approach is straightforward, it may result in a loss of valuable data and potentially biased estimates. Another approach is inverse probability weighting (IPW), which adjusts for bias by including weights in estimation based on the probability of a crash-record being a complete case (Ye & Wang, 2018). This method can help mitigate the impact of missing values but relies on assumptions about the missing data mechanism. Furthermore, Multiple imputation (MI) is another widely used method for handling missing data, where missing values are imputed based on the conditional distribution of the variable with missing information (Royston, 2004; Ye & Wang, 2018). In comparison to CC and IPW, MI takes a more versatile approach. It preserves the information present in the incomplete records and creates multiple imputed datasets to effectively handle the uncertainty associated with the imputation process. However, existing literature shows that efforts to enhance crash prediction modelling through data imputation have been conducted for highway-crash data in the past. Nevertheless, there has been a noticeable absence of studies exploring missing value data imputation and its association with crash frequency prediction modelling, specifically for Highway-rail grade crossings (HRGCs) inventory data.

In this work, different statistical and artificial intelligence-based data imputation techniques are applied to address the issue of missing values in HRGCs inventory data. Statistical methods, such as mean, mode, and median imputation, along with iterative Expectation-Maximization (EM) imputation and Last Observation Carried Forward (LOCF) imputation, have been employed. Furthermore, artificial intelligence-based techniques, including K-Nearest Neighbor (KNN) imputation, Support Vector Machines (SVM), and Random Forest (RF) imputation, have been utilized to effectively handle the missing values in the HRGCs inventory dataset. Using these imputations, Zero-inflated Negative Binomial (ZINB) models are estimated for each imputed dataset, and their model fitness is compared based on Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values. This analysis seeks to ascertain the most effective imputation method for generating optimally fitted crash prediction models. The objective of this study is to enhance the precision of predictive analyses concerning crashes at HRGCs, thereby facilitating improved policymaking for enhancing the safety of HRGCs. The next section discusses in detail statistical and machine learning-based data imputation techniques, followed by description on inventory dataset and filtration process. The next section covers

methodology and gives details on packages and functions used in R (open-source programming language) for data imputation. The subsequent sections present the results of the ZINB model based on different imputed datasets. Finally, the last section discusses the conclusions and recommendations.

2 DATA IMPTUATION USING STATISTICAL METHODS

Statistical methods for data imputations are essential for dealing with missing values in datasets, enabling researchers to utilize complete data for analysis and predictive modeling. Several techniques are commonly employed for this purpose. Mean/Mode/Median imputation replaces missing values with the mean, mode, or median of the observed data in the same column (Royston, 2009). Regression imputation builds a regression model using complete data as the dependent variable and other variables as predictors to predict missing values. This technique is commonly used in data analysis and is especially valuable when dealing with large datasets that have missing values, as it preserves as much information as possible while handling the missing entries. The process of regression imputation involves several steps. First, the dataset is divided into two groups: the complete cases, which contain all the required variables without any missing values, and the cases with missing values that need to be imputed. The complete cases are used to train the regression model, with the dependent variable being the one with missing values, and the other variables serving as predictors. Once the regression model is trained, it can be applied to the cases with missing values to predict their values based on the relationships learned from the complete data. The predicted values are then used to replace the missing entries, effectively imputing the missing data. However, it's essential to acknowledge that regression imputation has its limitations. For instance, it assumes that the relationships between variables remain constant across the complete and incomplete cases. If this assumption is not met, the imputed values might not accurately reflect the true missing data (Royston, 2009; Templ et al., 2011).

Furthermore, Multiple imputation (MI) generates multiple plausible imputed datasets and combines results to provide more accurate estimates and measures of uncertainty (Royston, 2009; Puri & Gupta, 2017). Expectation-Maximization (EM) imputation iteratively estimates missing values based on available data and updates the imputed values until convergence (Royston, 2009). Bayesian imputation uses probabilistic models, incorporating prior knowledge about the data generating process and uncertainty (Royston, 2009; Buuren, 2007; Jerez et al., 2010). Furthermore, Time Series (TS) imputation is used when dealing with time-series data. It involves using historical data or neighboring time points to impute missing values in the time series (Afrifa-Yamoah et al. 2020). In advance statistical methods, one notable method of data imputation is Last Observation Carried Forward imputation (LOCF), where the last observed value before the missing entry is carried forward to fill the gap (Woolley et al., 2009; Waljee et al., 2013; Hedeker et al., 2007). While numerous statistical imputation methods exist in current practice, their application is contingent upon the type and complexity of the dataset. It is expected that detailed-understanding of the dataset will aid in selecting the most appropriate statistical imputation methods to be employed.

3 DATA IMPUTATION USING MACHINE LEARNING METHODS

Machine learning and artificial intelligence-based missing data methods offer powerful approaches for imputing missing values in datasets (Waljee et al., 2013). For instance, K-Nearest Neighbors (KNN) imputation leverages the similarity between data points to impute missing values (Waljee et al., 2013; Malarvizhi & Thanamani, 2012; Pujianto et al., 2019; García-Laencina et al., 2009). It identifies the k-nearest neighbors of a sample with missing data and computes the average or weighted average of their values to fill in the missing entry (Waljee et al., 2013). Random Forest (RF) imputation utilizes the Random Forest algorithm to predict missing values by considering other features in the dataset as predictors (Jing et al., 2022; Tang & Ishwaran, 2017; Waljee et al., 2013; Pantanowitz & Marwala, 2009). RF constructs multiple decision trees and aggregates their predictions to impute missing values more accurately (Waljee et al., 2013). Other machine learning-based methods include Support Vector Machines (SVM) imputation, which uses SVM to predict missing values, and Deep Learning imputation, where deep neural networks are trained to impute missing data points (Waljee et al., 2013; Pelckmans et al., 2005; Zhang & Liu, 2009). These AI-driven approaches have proven effective in handling missing data and enhancing the reliability of analyses and modeling tasks (Waljee et al. 2013). However, it is crucial to carefully consider the suitability of each method for specific data types and research objectives to achieve accurate imputations (Waljee et al., 2013).

4 DATA DESCRIPTION

Present study utilized FRA-provided inventory and crash data (2018-2022) for HRGCs. Initially, 2000 crossing's data points were randomly selected from the national database, representing all 50 states. Subsequently, filters were applied to include only public, at-grade, operational crossings that intersected a highway. This step yielded a total of 1096 rows of data on physical and dynamic characteristics of HRGCs (N=1096). Subsequently, the 'naniar' package in R (open-source programming language) was utilized to create heat maps of missing values, providing insight into the distribution of missing data and the percentage of missing values within the dataset. Figure 1 shows one of such heatmaps that was generated to explore the percentage of missing values in the inventory dataset. While the inventory dataset contains numerous physical and dynamic factors of HRGCs, the detailed investigation into missing values was limited to variables that had been previously studied or were part of the FRA 2020 Accident Prediction Model (Brod et al., 2020; Farooq, 2023; Khattak & Farooq, 2023).

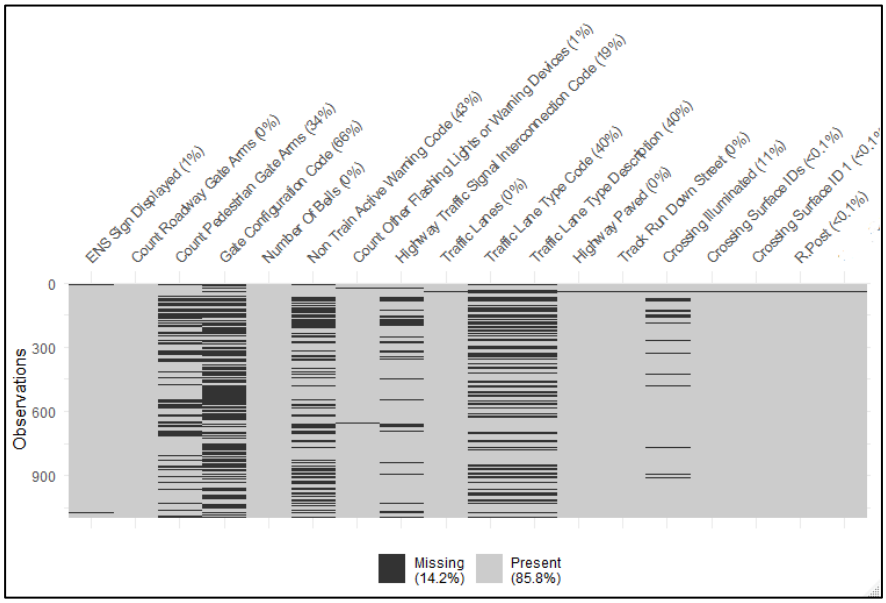

Figure 1. Sample Heatmap of Missing Values in HRGCs Inventory Data

Table 1 also presents the number and percentage of missing values for key variables in the datasets that had been previously observed to be associated with crash frequency at HRGCs.

Table 1. Range of Key Variables and their Percentage of Missing Values (N=1096)

Key Variables	Range ('i' for indicator variables)	Number of Missing Values	Percentage of Missing Values
In/Near City	0-1 (i)	0	0%
Development Type	0-18	0	0%
Number of passenger train per day	0-180	66	6.02%
Latitude	-	0	0.00%
Longitude	-	0	0.00%
Total Daylight Thru Trains	0-120	1	0.09%
Total Nighttime Thru Trains	0-120	1	0.09%
Total Switching Trains	0-40	1	0.09%
Total Transit Trains	0-180	21	1.92%
Movements Per Day Code	0-1 (i)	23	2.10%
Trains Per Week	1-30	889	81.11%
Maximum Timetable Speed	0-75	2	0.18%
Typical Minimum Speed Over Crossing	0-60	2	0.18%
Typical Maximum Speed Over Crossing	0-60	2	0.18%
Number Of Main Tracks	0-4	1	0.09%
Number Crossbuck Assemblies	0-7	0	0.00%
Number Stop Signs	0-4	1	0.09%

Pavement Marking	1-3	642	58.58%
ENS Sign Displayed	0-1 (i)	15	1.37%
Count Roadway Gate Arms	0-6	141	12.86%
Gate Configuration	1-3	141	12.86%
Count Of Flashing Light Pairs	1-18	0	0.00%
Highway Traffic Signals	0-1 (i)	447	40.78%
Storage Distance	0-62	937	85.49%
Traffic Lanes	0-7	2	0.18%
Traffic Lane Type	1-3	440	40.14%
Highway Paved	0-1 (i)	4	0.36%
Count of Bells	1-6	2	0.18%
Crossing Surface IDs	1-20	1	0.09%
Road At Crossing (Rural/Urban)	0-1 (i)	7	0.64%
Road At Crossing Type	1-7	9	0.82%
Highway Speed Limit	0-65	209	19.07%
Annual Average Daily Traffic Count	1-37,900	0	0.00%
Estimated Percent Trucks	0-95	27	2.46%
School Bus Route	0-1 (i)	13	1.19%
Principal Warning Device	1-8	0	0.00%

Note: Some indicator variables are derived from categorical variables in data.

From Table 1, it can be observed that most key variables for crash prediction modeling have a low percentage of missing values in the dataset sample. However, a few variables showed a higher percentage of missing values, such as posted highway speed, storage distance, highway traffic signals, and pavement markings near HRGCs. Additionally, some key variables did not have any missing values, such as Average Annual Daily Traffic (AADT), development type, HRGCs location (latitude and longitude), and the number of crossbuck assemblies. For modeling purposes, several categorical variables were converted into indicator variables, and missing values were imputed for instances with absent categorical data.

5 METHODOLOGY

In this study, a comprehensive comparative analysis is undertaken to address missing data in HRGCs inventory data. Both statistical and machine learning-based methods are employed for imputation. For the statistical methods, mean, median, and mode imputation are performed using the "zoo" package in R, specifically the "na.aggregate" function. Three different samples of imputed data are obtained using these three imputation techniques. Additionally, two more advanced statistical imputation methods, namely Last Observation Carried Forward (LOCF) Imputation and Expectation Maximization (EM) Imputation, were utilized. These methods were implemented using the "imputeTS" and "impute" packages in R (open-source programming language), respectively. Furthermore, three machine learning imputation methods, namely K-Nearest Neighbor (KNN) Imputation, Random Forest (RF) Imputation, and Support Vector Machines (SVM) Imputation, were employed to obtain three additional imputed datasets of crash data. To carry out these imputations, “imputeTS", "impute", and "randomForest" packages were used. Table 2 displays the R functions utilized for the imputation of missing values.

Table 2. Description of R-Programming Language Packages and Functions Used for Missing Value Imputations

Type Of Imputation	Packages Used	Functions Used
Mean Imputation	“zoo”	*na.aggregate()* Example: na.aggregate(Impute1, FUN = mean)
Median Imputation	“zoo”	*na.aggregate()* Example: na.aggregate(Impute1, FUN = median)
Mode Imputation	“zoo”	*na.aggregate()* Example: na.aggregate(Impute1, FUN = mode)
Last Observation Carried Forward (LOCF) Imputation	“zoo”	*na.locf()* Example: na.locf(Impute1, na.rm = FALSE) . “False” is to insure NAs are not removed.
Expectation Maximization (EM) Imputation	"imputeTS"	i*mputeEM()* Example: imputeEM(Impute1)
K-Nearest Neighbor (KNN) Imputation	"impute"	*knn()* Example: Time=1096 Value= C(1,2, NA, NA, .., 4) data_imputed <- knn(Impute1, k = 3), ‘k’ = the number of nearest neighbors considered

Random Forest (RF) Imputation	"randomForest"	*randomForest()*
		Example:
		Data frame with missing values:
		Impute_1_no_missing <- Impute_1[complete.cases(Impute_1),]
		Data frame without missing values:
		Impute_1_missing <- Impute_1[!complete.cases(Impute_1),]
		Train the RF
		# Train the Random Forest model
		rf_model <- randomForest(value ~ time, data = Impute_1_no_missing)
		Predict Missing
		# Predict the missing values
		Impute_1_missing$value <- predict(rf_model, newdata = Impute_1_missing)

Support Vector Machines (SVM) Imputation	"imputeTS"	*imputeSVM()* Example: Impute_1_imputed <- imputeSVM(Impute_1)

Subsequently, Zero-inflated Negative Binomial (ZINB) models were estimated based on all eight imputed datasets. The selection of ZINB models was driven by the over-dispersion observed in the response variable, which represents the number of crashes at HRGCs from 2018-2022. Model performance comparison was based on AIC and BIC values (Khattak et al., 2023; Farooq & Khattak, 2023). Figure 2 presents several attributes, limitations, and assumptions of the ZINB model. Transportation safety analysts commonly prefer the use of Zero-inflated (ZI) models when analyzing crash data due to their demonstrated superiority in achieving a better statistical fit compared to traditional Poisson and Negative Binomial (NB) models. ZI models explicitly address the issue of excess zeros often present in crash data, enabling them to yield more precise and dependable estimates of the actual crash frequency. Additionally, these models play a crucial role in identifying the contributing factors to crashes on the surface transportation network, providing valuable insights for improving transportation safety (Khattak & Farooq, 2023; Sharma & Landge, 2013).

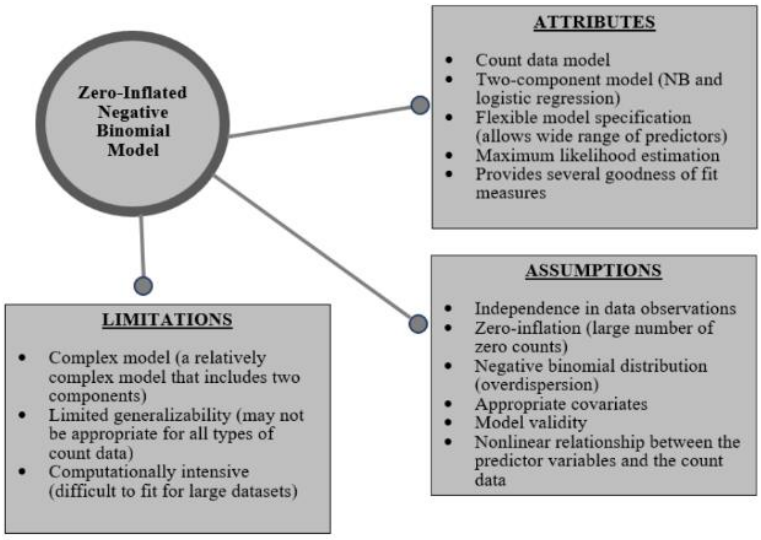

Figure 2. Key Attributes, Assumptions and Limitations of ZINB model (Farooq 2023)

A Zero-inflated Negative Binomial (ZINB) model assumes that zero outcomes arise from two distinct underlying processes. For instance, in the context of crashes at HRGCs, these processes are characterized as follows: (1) the occurrence of a crash at HRGCs and (2) the absence of a crash at HRGCs. In the absence of crashes at HRGCs, the outcome is confined to zero, while the occurrence of a crash is treated as a count process. The zero-inflated model comprises two components: a binary model, often a logit model, responsible for determining the process associated with the zero outcome, and a count model, specifically a negative binomial model, employed to characterize the count process for crashes at HRGCs. The anticipated count is represented as a fusion of the two processes. It is essential to highlight that the ZI-Poisson model bears resemblance to the ZINB model; however, the former assumes that non-zero counts adhere to a Poisson distribution, while the latter assumes a Negative Binomial distribution for non-zero counts (Brod et al., 2020).

The general formula for ZINB model according to Miaou (1994) is presented as:

$$y_i = 0,1,2 \ldots \text{ with probability } \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha*\lambda_i}{1+\alpha*\lambda_i}\right)^{y_i} \quad (1)$$

$$\lambda_i = e^{\beta_i x_i} \quad (2)$$

where x_i is i^{th} independent variable, and β_i is the coefficient of regression

For Zero-inflated Negative Binomial regression

$$y_i = 0, \text{ with probability } p_0 + \left(\frac{1}{1+\alpha*\lambda_i}\right)^{\frac{1}{\alpha}} \quad (3)$$

where p_0 illustrates the probability model that includes the effects of independent variables, such as logit model.

$$p_0 = \frac{e^{r'w_i}}{1+e^{r'w_i}} \quad (4)$$

r is the matrix's coefficient, and w_i is the i^{th} independent variable. Furthermore, $\Gamma(.)$ is Gamma function; and α represents the rate of over dispersion.

Maximum likelihood estimation (MLE) is a widely used method for estimating parameters in Poisson, Negative Binomial, and Zero-inflated regression models (Farooq, 2023). This method involves finding the parameter values that maximize the likelihood function, which measures the probability of the observed data given the model. The MLE method is favored because it has been demonstrated to be effective for a variety of statistical models and offers precise and effective estimates of the model parameters (Farooq, 2023; Sharma and Landge, 2013). Furthermore, to evaluate the performance of the ZINB models, the Akaike Information Criterion (AIC) and BIC (Bayesian Information Criterion) are commonly used. The AIC is a measure of the quality of a model, considering both the goodness of fit and the complexity of the model. A lower AIC value indicates a better model fit, as it penalizes models with a larger number of parameters. However, the idea behind BIC is that the best model is the one that maximizes the likelihood of the data while penalizing for the number of parameters in the model. Therefore, AIC and BIC are often used to compare varied models and select the one that best fits the data (Sharma & Landge, 2013; Anderson et al., 2020; Farooq et al., 2021; Farooq, 2023).

Description of Imputed Data

Table 3 and Table 4 provide descriptive statistics of imputed candidate variables that had missing values in the selected dataset. It was noted that employing mean, median, and mode imputation methods for statistical imputation did not result in a significant change; for some candidate variables, the values of mean and standard deviation remained the same. However, imputation based on EM and LOCF resulted in significantly different means and standard deviations for these variables.

Table 3. Descriptive Statistics of Imputed Candidate Variables (Statistical Imputation)

Variable Name	Variable Description	Miss Val.	Mean Imputation		Median Imputation		Mode Imputation		EM Imputation		LOCF Imputation	
			Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D
NOCr_18_22	No. of HRGC Crashes in 5-year Period (2018-2022)	No	-	-	-	-	-	-	-	-	-	-
Nearcity	HRGC situated near city indicator (1 if yes, 0 otherwise)	No	-	-	-	-	-	-	-	-	-	-
ComrclDvlp	HRGC in a Commercial Development (1 if yes, 0 otherwise)	No	-	-	-	-	-	-	-	-	-	-
TotalTrains	Total Number of Day and Night Trains	Yes	6.922	11.276	6.922	11.276	6.922	11.276	6.827	11.911	6.827	11.911
MaxTtSpeed	Maximum Timeable Speed	Yes	31.66	7.524	30.12	6.954	32.11	7.114	31.77	7.81	32.11	7.887
SgnOSgnl	Presence of Signals or Signs (1 if yes, 0 otherwise)	Yes	0.9571	0.202	0.871	0.204	0.912	0.274	0.966	0.194	0.9334	0.214
NofCrsAsmb	Number of Crossbuck Assemblies	Yes	1.416	1.060	1.411	1.041	1.614	1.321	1.514	1.364	1.301	0.960
PvmtMrkg	Pavement Markings (1 if yes, 0 otherwise)	Yes	0.4142	0.492	0.522	0.510	0.400	0.412	0.431	0.471	0.471	0.568
Gates	Presence of Gates (1 if yes, 0 otherwise)	Yes	0.378	0.485	0.410	0.515	0.378	0.485	0.314	0.397	0.412	0.588
FlsLight	Presence of Flashing lights (1 if yes, 0 otherwise)	No	-	-	-	-	-	-	-	-	-	-
XwBells	Crossings with Bells (1 if yes, 0 otherwise)	Yes	0.543	0.464	0.543	0.464	0.543	0.464	0.514	0.421	0.510	0.401
HwyTrSngl	Highway with traffic signals (1 if yes, 0 otherwise)	Yes	0.078	0.269	0.061	0.197	0.064	0.231	0.054	0.112	0.071	0.255
Traffic Lanes	Number of Traffic Lanes	Yes	2.071	1.091	2.071	1.091	2.031	1.042	2.031	1.094	2.071	1.091
TwoWyTrfc	Two-way traffic (Derived from Traffic Lane Type)	Yes	0.511	0.432	0.511	0.432	0.511	0.432	0.577	0.464	0.590	0.512
Rural	HRGC in a rural area – derived from Road at Crossing (1 if yes, 0 otherwise)	Yes	0.622	0.485	0.611	0.477	0.628	0.471	0.635	0.417	0.674	0.427
SpG50	Posted speed limit greater than 50 mph (1 if yes, 0 otherwise)	Yes	0.585	0.492	0.521	0.401	0.497	0.400	0.287	0.487	0.511	0.408
AADT	AADT	No	-	-	-	-	-	-	-	-	-	-

Table 4 provides descriptions for the same variables but imputed using machine-learning methods. It is observed that the imputations based on machine learning techniques resulted in noticeably different means and standard deviations for the candidate variables used in the crash frequency models.

Table 4. Descriptive Statistics of Imputed Candidate Variables (Machine Learning Imputation)

Variable Name	Variable Description	Presence of Missing Values	K-Nearest Neighbor (KNN) Imputation		Random Forest (RF) Imputation		Support Vector Machines (SVM) Imputation	
			Mean	S. D	Mean	S. D	Mean	S. D
NOCr_18_22	No. of HRGC Crashes in 5-year Period (2018-2022)	No	-	-	-	-	-	-
Nearcity	HRGC situated near city indicator (1 if yes, 0 otherwise)	No	-	-	-	-	-	-
ComrclDvlp	HRGC in a Commercial Development (1 if yes, 0 otherwise)	No	-	-	-	-	-	-
TotalTrains	Total Number of Day and Night Trains	Yes	6.900	11.012	6.978	11.281	6.501	11.264
MaxTtSpeed	Maximum Timeable Speed	Yes	31.67	7.728	29.21	6.287	32.79	7.678
SgnOSgnl	Presence of Signals or Signs (1 if yes, 0 otherwise)	Yes	0.9501	0.207	0.9214	0.200	0.921	0.187
NofCrsAsmb	Number of Crossbuck Assemblies	Yes	1.401	1.070	1.487	1.211	1.463	1.862
PvmtMrkg	Pavement Markings (1 if yes, 0 otherwise)	Yes	0.471	0.482	0.401	0.421	0.428	0.480
Gates	Presence of Gates (1 if yes, 0 otherwise)	Yes	0.301	0.447	0.312	0.444	0.357	0.437
FlsLight	Presence of Flashing Lights (1 if yes, 0 otherwise)	No	-	-	-	-	-	-
XwBells	Crossings with No Bells (1 if yes, 0 otherwise)	Yes	0.543	0.464	0.521	0.487	0.574	0.411
HwyTrSngl	Highway with traffic signals (1 if yes, 0 otherwise)	Yes	0.078	0.269	0.066	0.201	0.077	0.274
Traffic Lanes	Number of Traffic Lanes	Yes	2.071	1.091	2.101	1.141	2.187	1.321
TwoWyTrfc	Two-way traffic (Derived from Traffic Lane Type)	Yes	0.487	0.417	0.514	0.471	0.557	0.474
Rural	HRGCs in a rural area (1 if yes, 0 otherwise)	Yes	0.622	0.485	0.422	0.400	0.412	0.471
SpG50	Posted speed limit greater than 50 mph (1 if yes, 0 otherwise)	Yes	0.585	0.492	0.515	0.724	0.571	0.401
AADT	AADT	No	-	-	-	-	-	-

6 RESULTS

The ZINB model was estimated in R (open-source programming language), using the “zeroinfl ()” function from the “pscl” package. The output in Table 5 and 6 provides information on model fitness, coefficient estimates, and p-values. The first part of the output, "Count model coefficients (negative binomial with log link)", shows the coefficients and their p-values for the count part of the model. The Log(theta) is the logarithm of the dispersion parameter. Furthermore, "Zero-inflation model coefficients (binomial with logit link)", shows the coefficients and their p-values for the zero-inflation part of the estimated ZINB model based on the imputed datasets. Furthermore, the dispersion parameter is denoted by Theta and its estimated value varies for different imputed datasets. In addition, the optimization algorithm used to estimate the coefficients required 100 iterations. The estimated model reveals a positive coefficient

for maximum timetable speed, absence of highway traffic signs, two-way traffic, rural HRGCs and posted speed limit greater than 50 mph. However, negative coefficient estimates are observed for flashing lights, and crossing with bells.

Table 5. ZINB Results for Predicted Crashes Based on Statistical Imputation

Variables	Mean Imputation		Median imputation		Model Imputation		EM Imputation		LOCF Imputation	
	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)
Count model coefficients (negbin with log link)										
(Intercept)	-2.714	0.000	-2.974	0.000	-2.014	0.000	-2.540	0.000	-2.836	0.000
MaxTtSpeed	0.017	0.010	0.021	0.044	0.010	0.021	0.078	0.000	0.015	0.014
FlsLight	-0.057	0.000	-0.071	0.010	-	-	-0.080	0.061	-0.060	0.021
XwBells	-	-	-0.074	0.041	-0.005	0.000	-0.044	0.019	-0.005	0.048
NHwyTrSngl	0.935	0.002	0.935	0.002	0.935	0.002	0.935	0.002	0.935	0.002
TwoWyTrfc	0.114	0.000	0.127	0.004	0.117	0.000	0.144	0.012	0.125	0.000
Rural	-	-	-	-	-	-	-	-	-0.434	0.028
SpG50	-	-	0.514	0.000	0.514	0.000	0.554	0.000	0.506	0.018
Log(theta)	0.758	0.397	0.974	0.300	0.912	0.310	0.987	0.025	0.731	0.384
Zero-inflation model coefficients (binomial with logit link)										
(Intercept)	1.285	0.000	1.201	0.033	1.547	0.011	1.284	0.012	1.285	0.025
TotalTrains	-0.218	0.004	-0.287	0.010	-0.288	0.017	-0.298	0.001	-0.218	0.015
Theta	2.384		2.745		2.199		2.124		2.079	
Log-likelihood	-178.85		-187.25		-198.74		-250.2		-259.9	
AIC	571.87		591.14		578.17		547.74		541.74	
BIC	621.27		608.74		619.77		612.01		596.72	

From Tables 5 and 6, it can be observed that intuitive parameters are obtained. Past research has also revealed that warning devices decrease the likelihood of crashes at HRGCs, while factors such as higher posted speed limits, rural areas, higher train speeds, and two-way traffic increase the likelihood of crashes at HRGCs (Khattak & Farooq 2023; Bord et al. 2020).

Table 6. ZINB Results for Predicted Crashes Based on Machine Learning Imputation

Variables	K-Nearest Neighbor (KNN) Imputation		Random Forest (RF) Imputation		Support Vector Machines (SVM) Imputation	
	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)
Count model coefficients (negbin with log link)						
(Intercept)	-2.104	0.000	-1.710	0.000	-1.640	0.000
MaxTtSpeed	0.024	0.004	0.004	0.001	0.015	0.004
FlsLight	-	-	-	-	-0.057	0.000
XwBells	-0.001	0.248	-0.001	0.000	-0.002	0.041
NHwyTrSngl	0.997	0.000	0.174	0.000	0.900	0.040
TwoWyTrfc	-	-	0.101	0.004	-	-
Rural	-0.477	0.037	-	-	-0.600	0.048
SpG50	0.517	0.018	0.787	0.071	-	-
Log(theta)	0.517	0.018	0.500	0.000	0.587	0.000
Zero-inflation model coefficients (binomial with logit link)						
(Intercept)	1.200	0.015	1.324	0.005	1.474	0.005
TotalTrains	-0.278	0.005	-0.214	0.004	-0.211	0.040
Theta	3.198		2.190		2.170	
Log-likelihood	-219.978		-207.204		-204.74	
AIC	502.74		512.70		527.14	
BIC	512.872		528.807		547.800	

However, it is interesting to note that the models performed well when missing data was imputed using LOCF imputation, as evident from AIC and BIC values. The best overall fitness of the ZINB model was observed for the dataset in which the missing values were imputed using KNN imputation. Additionally, when other machine learning-based datasets were fitted using ZINB, they also showed better fitness performances compared to statistical imputations. Figure 3 presents a visual comparison of AIC and BIC values derived from ZINB models based on datasets imputed using statistical and machine learning-based imputation techniques.

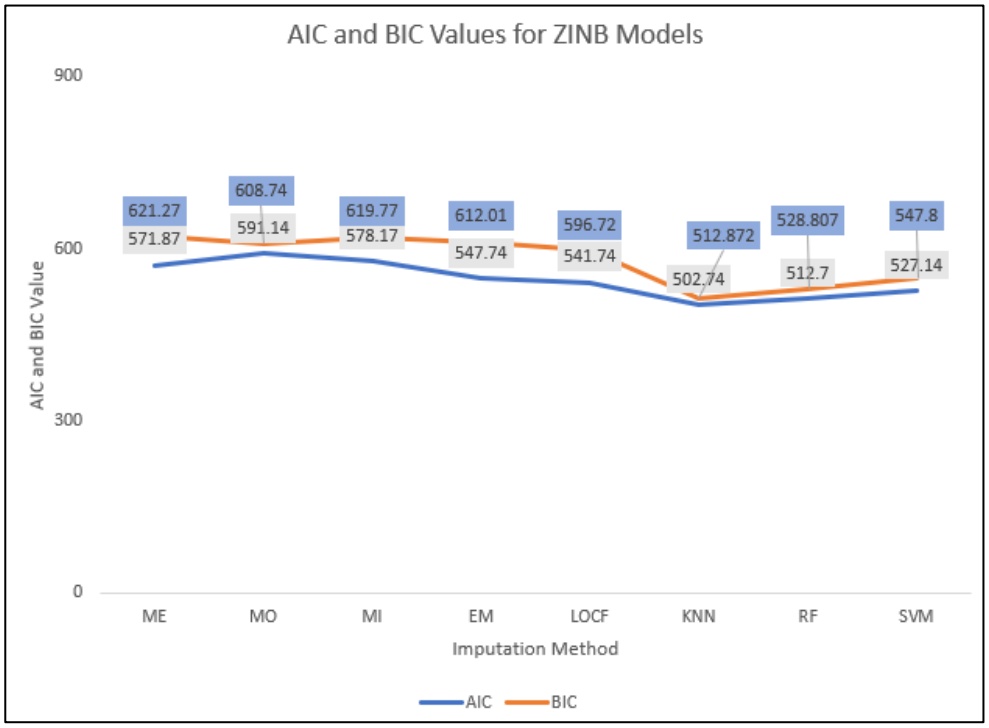

Figure 3: AIC and BIC values obtained from ZINB models based on imputed eight datasets

7 CONCLUSIONS

In this study, a comparison of statistical and machine learning data imputation methods was carried out to evaluate the fitness of crash prediction models for HRGCs. For this purpose, randomly selected inventory and crash datasets (2018-2022) involving 2000 HRGCs were obtained from the national FRA HRGC database. These two datasets were aggregated to combine information on the past crash history and inventory details for every crossing. Using the aggregated dataset, a detailed investigation of missing values was conducted by examining heat maps and estimating the percentage of missing data for each variable in the dataset. Furthermore, eight different statistical and machine learning methods were applied to impute the missing values in the dataset. These methods included mean, median, mode, EM, LOCF, KNN, RF, and SVM imputations. The process revealed that after missing value imputation, the means and standard deviations of the variables started to differ. Based on these eight imputed datasets, zero-inflated negative binomial models were applied, and the models were compared based on fitness parameters such as AIC and BIC. The modeling revealed that overall, there wasn't a significant difference between the statistically imputed data's ZINB model performance. However, the machine-learning imputed data, particularly the KNN imputed dataset, provided a better-fitted ZINB model. This study underscores the importance of using high-quality and missing-value-free datasets for HRGC crash prediction modeling. With better datasets, the reliability of crash prediction models can be improved, significantly aiding correct decision-making regarding resource allocation and safety interventions at HRGCs. Moreover, this study advocates the application of machine learning-based data imputation techniques to elevate data quality, thereby enhancing crash prediction modeling for HRGCs. Notwithstanding the absence of significant distinctions in model performances between imputed data based on KNN, RF, and SVM imputation, the findings unequivocally indicated superior model performance for datasets in which missing values were imputed through machine-learning approaches when compared to classical statistical methods.

8 ACKNOWLEDGEMENTS

The authors would like to thank the Mid-America Transportation Center for providing resources to carry out this research.

REFERENCES

1. Abdulhafedh, A. (2016). Crash Frequency Analysis. JTTs, 04(06), 169-180. https://doi.org/10.4236/jtts.2016.64017
2. Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications, 27(1), e1873.
3. Anderson, M., Khattak, A. J., Farooq, M. U., Cecava, J., & Walker, C. (2020). Research on Weather Conditions and Their Relationship to Crashes (No. SPR-21 (20) M097). Nebraska. Department of Transportation.
4. Asgharpour, S., Javadinasr, M., Mohammadian, R., & Mohammadian, A. (2023). Missing Data Treatment in Crash Data: A Heuristic Optimization Weighting Approach. In International Conference on Transportation and Development 2023 (pp. 87-98).
5. Brod, D., Gillen, D., & Decisiontek, L. L. C. (2020). A new model for highway-rail grade crossing accident prediction and severity (No. DOT/FRA/ORD-20/40). United States. Department of Transportation. Federal Railroad Administration.
6. Deb, R., & Liew, A. W. C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. Information sciences, 339, 274-289.
7. Farooq, M. U., Ahmed, A., & Saeed, T. U. (2021). A statistical analysis of the correlates of compliance and defiance of seatbelt use. Transportation research part F: traffic psychology and behaviour, 77, 117-128. https://doi.org/10.1016/j.trf.2020.12.008.
8. Farooq, M. U., & Khattak, A. J. (2023). Investigating Highway–Rail Grade Crossing Inventory Data Quality's Role in Crash Model Estimation and Crash Prediction. Applied Sciences, 13(20), 11537. https://doi.org/10.3390/app132011537.
9. Farooq, M. U., & Khattak, A. J. (2023). A Heterogeneity-Based Temporal Stability Assessment of Pedestrian Crash Injury Severity Using an Aggregated Crash and Hospital Data Set (No. TRBAM-23-01089).
10. Farooq, M. U. (2023). The Effects of Inaccurate and Missing Highway-Rail Grade Crossing Inventory Data on Crash and Severity Model Estimation and Prediction (Doctoral dissertation, The University of Nebraska-Lincoln).
11. García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing, 72(7-9), 1483-1493.
12. Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2007). Analysis of binary outcomes with missing data: missing= smoking, last observation carried forward, and a little multiple imputation. Addiction, 102(10), 1564-1573.
13. Imprialou, M., & Quddus, M. (2019). Crash data quality for road safety research: Current state and future directions. Accident Analysis & Prevention, 130, 84-90.
14. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-115.
15. Jing, X., Luo, J., Wang, J., Zuo, G., & Wei, N. (2022). A Multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest. Water Resources Management, 36(4), 1159-1173.
16. Khattak, A., and M. U. Farooq. The Effects of Inaccurate and Missing Highway-Rail Grade Crossing Inventory Data on Crash Model Estimation and Crash Prediction. Presented at Transportation Research Board (TRB) 102nd Annual Meeting, Washington DC, 2023.

17. Khattak, A. J., M. U. Farooq, & A. Farhan. (2023). Motor Vehicle Drivers' Knowledge of Safely Traversing Highway-Rail Grade Crossings. Transportation Research Record. https://doi.org/10.1177/03611981231208902
18. Malarvizhi, R., & Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. Int. J. Eng. Res. Dev, 5(1), 5-7.
19. Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention, 26(4), 471-482.
20. Pantanowitz, A., & Marwala, T. (2009). Missing data imputation through the use of the random forest algorithm. In Advances in computational intelligence (pp. 53-62). Springer Berlin Heidelberg.
21. Pelckmans, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2005). Handling missing values in support vector machine classifiers. Neural Networks, 18(5-6), 684-692.
22. Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019, October). K-nearest neighbor (k-NN) based missing data imputation. In 2019 5th International Conference on Science in Information Technology (ICSITech) (pp. 83-88). IEEE.
23. Puri, A., & Gupta, M. (2017). Review on missing value imputation techniques in data mining. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2(7), 35-40.
24. Royston, P. (2004). Multiple imputation of missing values. The Stata Journal, 4(3), 227-241.
25. Sharma, A. K., & Landge, V. S. (2013). Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highway. International Journal of Advances in Engineering & Technology, 5(2), 292.
26. Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, 10(6), 363-377.
27. Templ, M., Kowarik, A., & Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis, 55(10), 2793-2806.
28. Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research, 16(3), 219-242.
29. Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J. S., Balis, U. J., … & Higgins, P. D. (2013). Comparison Of Imputation Methods For Missing Laboratory Data In Medicine. BMJ Open, 8(3), e002847. https://doi.org/10.1136/bmjopen-2013-002847
30. Woolley, S. B., Cardoni, A. A., & Goethe, J. W. (2009). Last-observation-carried-forward imputation method in clinical efficacy trials: review of 352 antidepressant studies. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 29(12), 1408-1416.
31. Ye, F., Wang, Y. (2018). Performance Evaluation Of Various Missing Data Treatments In Crash Severity Modeling. Transportation Research Record, 38(2672), 149-159. https://doi.org/10.1177/0361198118798485
32. Ye, F., Wang, Y. (2018). Performance Evaluation of Various Missing Data Treatments In Crash Severity Modeling. Transportation Research Record, 38(2672), 149-159. https://doi.org/10.1177/0361198118798485
33. Zhang, Y., & Liu, Y. (2009). Data imputation using least squares support vector machines in urban arterial streets. IEEE Signal Processing Letters, 16(5), 414-417.