On the need to address fixed-parameter issues before applying random parameters: A simulation-based study

Numan Ahmad [a, b], Tanmoy Bhowmik [c], Vikash V. Gayah [b, *], Naveen Eluru [c]

[a] National Institute of Transportation
National University of Sciences and Technology, Risalpur, Pakistan

[b] Department of Civil and Environmental Engineering
The Pennsylvania State University
231 Sackett Building
University Park, PA 16802, U.S.

[c] Department of Civil, Environmental and Construction Engineering
University of Central Florida
12800 Pegasus Drive, Room 301D
Orlando, Florida 32816, U.S.

* Corresponding Author: Tel. phone: +1 814-865-4014

Email Addresses: numan.ahmad@nit.nust.edu.pk; nma5753@psu.edu (N. Ahmad), tanmoy78@Knights.ucf.edu (T. Bhowmik), gayah@engr.psu.edu (V. Gayah), naveen.eluru@ucf.edu (N. Eluru)

ABSTRACT

Count regression models have been applied to model expected crash frequency at individual roadway locations. Random parameters have been increasingly integrated into these models to account for unobserved heterogeneity. However, the introduction of random parameters might also mask issues in the model specification, leading to inaccurate relationships and model interpretation. Two of these specification-related issues are: 1) not considering the appropriate functional form of explanatory variables; and, 2) ignoring the best set of significant explanatory variables. To better examine the need for careful model specification, this study uses synthetic data to demonstrate that the consideration of random parameters does not address the two model specification issues identified. The results from the simulation study illustrate that (a) model specification issues cannot be circumvented by random parameters alone and (b) random parameter models including the exhaustive set of explanatory variables available offer significant model improvements.

Keywords

Random Parameter Negative Binomial Regression; Model specification; Unobserved heterogeneity; Synthetic Data; Simulation-based statistical analysis

1. Introduction

In the road safety literature, count data models are typically used to model crash frequency on individual roadway facilities (roadway segments, intersections, and interchanges) during a specific interval of time. Such models assume a functional form for how traffic volume and roadway, or roadside features relate to crash frequency. While the model specifications identify the important factors, the crash process is substantially complex and thus several factors affecting crash frequency typically remain unknown to an analyst. Drawing on burgeoning work in econometric model development, researchers have proposed frameworks that allow for flexible parameter distributions across individual observations by integrating random parameters into the model (Anastasopoulos and Mannering 2009, Park 2013, Mannering et al. 2016, Qin et al. 2018, Wali et al. 2018, Mannering et al. 2020, Alnawmasi and Mannering 2022, Feknssa et al. 2023). Similar to random parameter models, latent class (semi-parametric) models have also been applied which take the unobserved heterogeneity into account by classifying observations into a set of X different classes (Greene and Hensher, 2003, Vij et al., 2013, Mannering and Bhat, 2014, Mannering et al., 2016, Mahmud et al., 2023). The approaches directly tackle the unobserved heterogeneity present in the data by assuming correlation exists between observed parameters and the unobserved features that cannot be captured. In particular, the emergence and widespread adoption of random parameters has enhanced the safety modeling literature.

In recent years, there has been a significant surge in the application of random parameters to crash frequency models. The surge can be attributed to widespread availability of econometric model tools in open source and proprietary software. However, in some cases, there is growing emphasis on including random parameters in the models without carefully evaluating the impact of independent variables available in the dataset. In some cases, the introduction of random parameters might actually be capturing misspecification of the crash frequency models. Thus, as a first step, it is critical to ensure that the model is properly specified before incorporating random parameters. For example, researchers have focused on examining the underlying form of crash frequency models and offered functional forms of crash exposure variables (usually annual average daily traffic (AADT)) such as traditional forms with linear formulation (Venkataraman et al. 2011, Gooch et al. 2016, Khattak et al. 2020), Hoerl forms (Hauer 2015, Wang et al. 2020), and flexible linear forms (Gayah and Donnell 2021, Eluru and Gayah 2022). The reader would note that incorporating additional observed variables in the model will not account for unobserved heterogeneity. Additional steps – such as the introduction of random parameters, latent classes, or Markov switching processes – are needed. However, failure to properly specify the model with over-reliance on random parameters to circumvent misspecification issues is not the right solution.

This paper serves as a cautionary tale to demonstrate the importance of proper model specification, and the appropriate role of random parameter models. A simulation-based framework is employed in which crash data are generated synthetically according to some known underlying process. This is done to ensure full control over the crash data generation process; use of empirical data would always be subject to omitted variable bias as there are countless factors that would influence crash frequency. Using this synthetic data, various model (mis)specifications are used to demonstrate conditions under which randomness in parameters might appear in the presence of misspecification. These misspecifications include: 1) not considering the appropriate functional form of explanatory variables; and 2) ignoring the best set of significant explanatory variables. The results confirm that analysis should first and foremost focus on model specification and not rely on random parameters as a "crutch" for proper model specification. In this way, the true power of random parameters can be better harnessed to capture unobserved heterogeneity that might be present.

The rest of this paper is organized as follows. The next section discusses the methods implemented in this study which include a) modeling methodology for the fixed parameter and random parameter negative binomial models and different functional forms b) procedure for generating synthetic data and c) experimental design of the study. The methods section is followed by the results and discussion section which discusses the key findings from the study followed by the conclusions.

2. Methods

This section describes the methods used in this paper. Section 2.1 "Models" include the description of the negative binomial and random parameter negative binomial model used to predict crash frequency. The different functional forms of the models considered as also described in Section 2.1 (see Section 2.1.3 "Different functional forms in crash frequency models"). Section 2.2 "Synthetic Data Generation" describes the synthetic data generation process that is applied for this study. Finally, Section 2.3 "Experimental Design" presents the experimental design as well as discusses the procedure for comparing the values of true parameters with their corresponding recovered values for both the fixed parameter and random parameter negative binomial models with each of the three functional forms.

2.1. Models

2.1.1. Fixed parameter negative binomial model

Considering the discrete and non-negative nature of crash frequency, Poisson or negative binomial regression can be applied to estimate the expected crash frequency on roadway segments. One of the key limitations of Poisson regression is that it is based on the assumption that the mean and variance are equal. The negative binomial regression model relaxes the abovementioned assumption by allowing the variance to exceed the mean (Var $(Y_i) > \mathrm{E(Y)}$). For the negative binomial regression model, the general functional form can be given below:

$$\ln(\lambda_i) = \beta X_i + \varepsilon_i \,, \tag{1}$$

where λ_i indicates the expected crash frequency on a particular road segment i during a specific interval of time; β indicates a vector of estimable parameters of key independent variables X_i associated with crash frequency; and ε_i indicates an error term assumed to follow a gamma distribution. For the negative binomial model, the mean-variance relationship can be given below:

$$\mathrm{Var}\,(y_i) = E(y_i) + \varphi E(y_i)^2 \,, \tag{2}$$

where y_i refers to the observed crash frequency occurred on a particular roadway segment i; λ_i refers to the expected crash frequency on a particular roadway segment i, and φ refers to the overdispersion parameter that comes from the negative binomial model. In the negative binomial model, the assumed probability distribution can be given below:

$$\mathrm{f}\,(y_i) = \frac{\Gamma(\varphi+y_i)}{\Gamma(\varphi)y_i!}\left(\frac{\lambda_i}{\varphi+\lambda_i}\right)^{y_i}\left(\frac{\varphi}{\varphi+\lambda_i}\right)^{\varphi}, \tag{3}$$

where Γ and φ refer to the gamma function and overdispersion parameter, respectively. Following the maximum likelihood estimation method, both coefficients (β) and the overdispersion parameter (φ) can be estimated using the likelihood function as given below:

$$L_i = \prod_{i=1}^{N} P(y_i) \,, \tag{4}$$

where Equation (4) shows the likelihood that the observed crash frequency can be observed conditional that the expected crash frequency is estimated using parameters $\beta_j$and φ_j obtained from the model. It should be noted that values for the two parameters need to be selected in a way that the likelihood is maximized and the model with the best fit to data is obtained.

2.1.2. Random parameter negative binomial model

According to the road safety literature, in addition to the observed factors, there could be some unobserved factors (which could be either available but not used in the model or could be missing in the data) that could have a significant influence on the crash frequency. Studies suggest using random parameter negative binomial regression to account for the effects of such unobserved factors on crash frequency on roadway segments (Dong et al. 2014, Wali et al. 2018, Tang et al. 2019, Huo et al. 2020). A general form for an RPNB model can be given below:

$$\beta_{i,j} = \overline{\beta_j} + \varphi_{i,j} \tag{5}$$

where, $\beta_{i,j}$ indicates a unique parameter estimated for the jth independent variable for a segment i; $\overline{\beta_j}$ refers to the mean parameter for a particular independent variable; whereas $\varphi_{i,j}$ represents the error terms related to the jth parameter for observation i with a random distribution. The expected crash frequency for the random parameter negative binomial model can be computed as below:

$$\lambda_i|\varphi_i = \exp\left(\beta X_i + \varepsilon_i\right) \tag{6}$$

The log-likelihood function for a random parameter negative binomial model can be given as below:

$$LL = \sum_{\forall_i} ln \int_{\varphi_i}^{i} g(\varphi_i)\, P(n_i|\varphi_i) d\varphi_i \tag{7}$$

where, $g(.)$ refers to the probability density function of φ_i, $P(n_i|\varphi_i)$ indicates the Poisson probability of an observation "roadway segment" (i) to have crashes (n_i) conditioned on φ_i. Given the computational complexity associated with the numerical integration of the random parameter negative binomial model with no closed-form expression, a simulation-based maximum likelihood method is used to maximize the log-likelihood function using Halton draws.

2.1.3. Different functional forms in crash frequency models

This section provides a description of three common potential functional forms that are considered in crash frequency models. These are used to demonstrate how failure to properly specify these relationships may manifest as randomness in the model. It should be noted that these three forms are just examples and not meant to serve as an exhaustive list, as these relationships may take numerous forms in practice. Thus, analysts should be prepared to test many forms when developing a crash frequency model. The three functional forms serve as the observed component of the modeling exercise and can be employed as fixed parameter models. At the same time, these functional forms can be overlaid with random parameters to accommodate unobserved heterogeneity. As the focus of our current research is on interplay the role of observed and unobserved parameters, we employ three functional forms to eliminate any specific functional form related bias toward observed or unobserved effects. The exact mathematical details of the functional forms follow.

2.1.3.1. Traditional functional form

The Highway Safety Manual (Part 2010) and relevant studies (Fitzpatrick et al. 2010, Venkataraman et al. 2011, Bauer and Harwood 2013, Gooch et al. 2016, Gayah and Donnell 2021) have mostly applied the traditional (trad) functional form for the expected crash frequency:

$$N_{trad} = e^{\beta_0} x\, L^{\beta_1}\, x\, AADT^{\beta_2} x\, e^{(\beta_3 X_3 + \beta_4 X_4 + --- + \beta_n X_n)}, \quad (8)$$

where N_{trad} refers to the expected crash frequency computed via model with the traditional functional form, $(\beta_0, \beta_1, \dots, \beta_n)$ refers to a vector of estimable parameters, $(X_3, X_4, \dots, X_n)$ refers to a vector of the roadway or roadside features, L refers to the segment length, and $AADT$ refers to the annual average daily traffic on a specific roadway segment. One of the key advantages of the traditional functional form is that β_2 provides a constant slope over the range of AADT values indicating that change in expected crash frequency due to a unit increase in AADT is constant. While using the traditional functional form could be simple and straightforward, studies suggest that this could not be a reasonable approach. For instance, the impact of a unit increase in the AADT on the expected crash frequency could be different in different ranges of AADT values (Shankar et al. 1998, Ulfarsson and Shankar 2003). Further, these approaches completely ignore the presence of unobserved heterogeneity.

Studies have used the random parameter version of the traditional negative binomial model which allows the coefficient of AADT to vary across the individual roadway segments due to site-specific unobserved factors (Anastasopoulos and Mannering 2009, Venkataraman et al. 2011). Still, the elasticity of the crash frequency for the AADT (refers to the percent change in the predicted crash counts or frequency due to one percent change in AADT) is constant for the roadway segment as per the random parameter models. Furthermore, the random parameter models are simulation-based and could be extensive in terms of computation. Some studies have accounted for continuing change in the relation of expected crash frequency and AADT using sigmoid functions, but such methods were based on parametric techniques like neural networks which could have lower interpretability and could not suit the crash predictive models in the Highway Safety Manual appropriately (Kononov et al. 2011). To account for the varying relationships between the expected crash frequency and AADT, some studies have used the Hoerl form as discussed below.

2.1.3.2. Hoerl functional form

In the Hoerl form, the expected crash frequency is allowed to vary as a function of traffic volume which is accommodated by including an additional term in Eq. (8) that also treats AADT as a traditional variable in the model:

$$N_{Hoerl} = e^{\beta_0} x\, L^{\beta_1}\, x\, AADT^{\beta_2} x\, e^{(\beta_3 AADT + \beta_4 X_4 + --- + \beta_n X_n)}, \quad (9)$$

While the Hoerl form can account for the varying relationship between the expected crash frequency and AADT, still a more flexible functional form is needed which is consistent with the traditional functional form and can have the potential to capture the varying relationships between the expected crash frequency and traffic volume that could exist in the different ranges of the traffic volume. To cope with the need for the functional form which is consistent with the traditional form but can account for varying relationships between the expected crash frequency and traffic volume, studies have introduced alternative functional forms with a flexible form of AADTs as discussed below.

2.1.3.3. Eluru and Gayah functional form

Studies have proposed alternative functional forms which capture the relationship in the data in a better manner with a reasonable computational cost (Gayah and Donnell 2021, Eluru and Gayah 2022). While the two functional forms including Gayah and Donnell and Eluru and Gayah forms capture the relationships between the expected crash frequency and exposure in flexible ways, the latter better accommodates the addition of random parameters; more details are provided in (Eluru and Gayah 2022). Since this study focuses on estimating both fixed parameter and random parameter negative binomial models, the Eluru and Gayah functional form is used as a flexible version of the traditional functional form. The general form of the Eluru and Gayah negative binomial model can be given below (Eluru and Gayah 2022):

$$N_{EG} = \beta_{AADT} \ln(AADT_i) + \delta_1 AADT_{inc1} + \delta_2 AADT_{inc2} + \gamma z_i + \varepsilon_i \quad (10)$$

where N_{EG} refers to the expected crash frequency computed via the crash frequency model with the Eluru and Gayah functional form, β_{AADT} indicates the parameter to be estimated for $\ln(AADT_i)$ for roadway segment i, δ_1 and δ_2 refer to the parameters to be estimated for the newly created variables which capture the changes to the effects of the AADT variable "$\ln(AADT_i)$"; γ indicates a vector of parameters to be estimated for a set of other explanatory variables "z_i" related to a roadway segment i, and ε_i refers to the error terms with a gamma distribution having a mean and variance of 1 and α, respectively. It should be noted that the independent variables which are newly added include $AADT_{i\_inc1}$ and $AADT_{i\_inc2}$ and are defined as below (Eluru and Gayah 2022):

$$AADT_{i\_inc1} = Max[0, \ln(AADT_i) - \ln(T_1)] \quad (11)$$

$$AADT_{i\_inc2} = Max[0, \ln(AADT_i) - \ln(T_2)] \quad (12)$$

T_1 and T_2 refers to the $\ln(AADT_i)$ threshold points at which the slope is expected to vary. For a detailed discussion about the Eluru and Gayah functional form, please refer to (Eluru and Gayah 2022). To provide a general understanding of the Eluru and Gayah functional form, Equation (10) shows a general equation of the Eluru and Gayah functional form which includes two threshold points (T_1 and T_2) at each of which the slope (effect of AADT on crash frequency) may start to change. To modify Equation (10) for the one threshold framework for the Eluru and Gayah form, the term ($\delta_2 AADT_{inc2}$) should to be excluded. Furthermore, in the one threshold framework, Equation (11) alone serves the purpose where Equation (12) is no more relevant.

2.2. Synthetic Data Generation

Synthetic data are generated and used in this study to compare model performance under a variety of conditions. . While synthetic data cannot account for the complex nature of crash generation that occurs in reality, any empirical data would always be subject to some amount of unobserved heterogeneity as all safety-influencing features could never be fully obtained and captured in this model. Thus, the use of synthetic data is desirable in this case because it allows randomness to be introduced systematically and only when necessary to better compare the performance models without and with random parameters. Two data generation processes are used. The first is to generate independent variables in the model and the second to generate dependent variables (crash frequencies). Detailed discussion and examples about synthetic data generation can also be found in (Abowd and Lane 2004, Abowd and Woodcock 2004, Reiter

2005a, Drechsler and Reiter 2010, Kinney et al. 2011). One of the recent studies (Reiter 2005b) has proposed non-parametric methods to generate synthetic data using classification and regression tree methods. Some studies also suggest using alternative machine learning methods including support vector machine, random forest, and bagging for synthetic data generation (Caiola and Reiter 2010, Drechsler and Reiter 2011).

2.2.1. Generation of independent variables in synthetic data

For independent variables, the "synthpop" package in R software is used to create set of the independent variables . To generate a synthetic version of the roadway and traffic variables, the synthpop package is used to replicates observed data from two-lane rural roads in Engineering District 3 of Pennsylvania between 2005 and 2012. The "synthpop" method selects a given independent variable to be synthesized and randomly generates values using a random sampling method with replacement from the actual or observed data (Nowok et al. 2016). Next, a parametric or non-parametric method can be applied to generate the values of a second, third, fourth, etc. independent variable while using the first one, first two, and first three synthetic independent variables as predictors in the machine learning methods, respectively. In the present study, a non-parametric classification and regression tree method is applied to synthesize the variables where the type of the classification and regression tree method including regression and classification is decided based on the variable type continuous, binary, or categorical etc.

The original data included various independent variables as given below:

- Annual average daily traffic (AADT)
- Segment length in mile (length_mi)
- Density of access points "number of access points per mile" (acc_den)
- Degree of horizontal curvature per mile (d_seg_mi)
- Presence of rumble strips along outer shoulder (sh_rs)

It should be noted that sh_rs is a binary variable while all other independent variables are continuous. Using the synthpop method, the variable "sh_rs" was first synthesized while randomly sampling its values from the actual data with replacement. Next, the synthpop was used to predict (synthesize) the d_seg_mi variable, using the sh_rs variable as the only predictor in the regression tree. The two variables including sh_rs and d_seg_mi are then used as predictor variables to predict the acc_den variable. Next, the values of the length_mi were predicted using the regression tree while using the aforementioned three variables (already synthesized) as predictor variables. Finally, all the pre-synthesized four variables are used to predict the values of AADT using the regression tree.

2.2.2. Generation of dependent variables (crash frequencies) using equations with different functional forms in synthetic data

After generating the set of independent variables used in the study, the next step is to generate the synthetic values for the observed annual crash frequency for each observation. In this process, the expected crash frequency was generated via pre-described equations. Then, the negative binomial distribution was applied with random number generation to randomly assign a crash frequency to individual roadway segments.

Three different functional forms were considered, based on the forms described in the METHODS (Different functional forms in crash frequency models) section. The fixed parameter versions of these models are:

$$N_{trad} = e^{-4.5} x\ AADT^{0.8}\ x\ length\_mi\ x\ e^{(-0.8\ x\ sh\_rs\ +\ 0.02\ x\ acc\_den + 0.004\ x\ \deg\ \_seg\_mi)} \quad (13)$$

$$N_{Hoerl} = e^{-4.95}\ x\ AADT^{0.9}\ x\ length\_mi\ x\ e^{(-0.0001 x AADT - 0.8\ x\ sh\_rs\ +\ 0.02\ x\ acc\_den\ +0.004 x\ \deg\ \_seg\_mi)} \quad (14)$$

$$N_{EG} = e^{-5.25}\ x\ AADT^{1}\ x\ length_{mi}\ x\ e^{(-0.8\ x\ sh\_rs + 0.02\ x\ acc\_den + 0.004\ x\ d\_seg\_mi + 2.1\ x(AADT>1900) - 0.5x(AADT>1900)\ x\ ln(AADT))} \quad (15)$$

This provides three synthetic crash frequencies associated with each observation in the synthetic dataset.

Random parameter versions of these models are also considered to account for when unobserved heterogeneity is present and significant. In this case, specific parameter coefficients in Equations (13-15) are also randomly generated following a given distribution when estimating the expected crash frequency associated with a segment. This is then applied to the random crash frequency generation process. In this way, models are able to be estimated in cases when unobserved heterogeneity is not present (or not significant) and when it is.

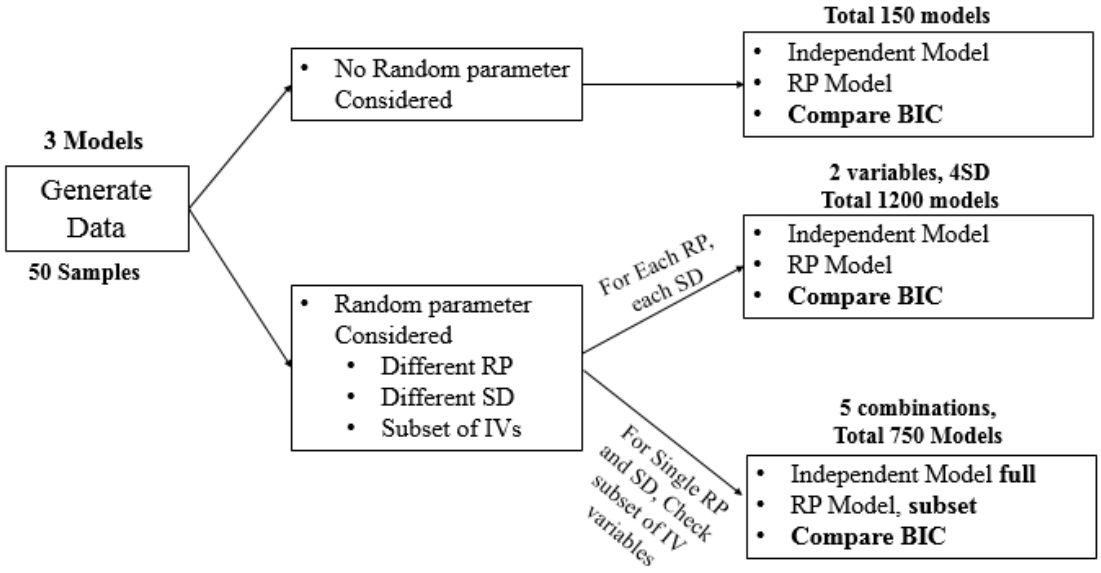
2.3. Experimental Design

The experimental design of the study is shown in Figure 1 and can be briefly summarized below:

1. A total of 100,000 (100K) unique observations were generated using the proposed data generation approach.
2. A total of 50 synthetic samples, each including 10,000 (10K) observations, are randomly selected from the synthetic data of 100K observations. To compute the expected crash frequencies for each observation in a given specific random sample, the following two methods are used:
    a. Fixed parameter models using Equations (13-15) which consider no randomness in the data.
    b. Random parameter models using Equations (13-15) where one or more of the coefficients in the models are assumed to be drawn from some known distribution for each observation.
        i. For simplicity, only two independent variables are treated as potentially having random parameters: AADT and sh_rs.
        ii. Various values of randomness (standard deviations associated with 5%, 8%, 12%, and 17% of the mean coefficient values) are considered. A normal distribution was assumed and used for the distribution of random parameters in this data generation process.
3. No randomness and one-time randomness (standard deviation associated with 12% of the coefficient values) for the AADT variable are considered. The key motivation behind this part of the analysis is to understand the impact of independent variables on the performance of random parameter negative binomial models.
4. Using each of the expected crash frequencies estimated above, annual crash frequencies for each observation are randomly determined using the negative binomial distribution.
5. Fixed parameter and random parameter models are then estimated using the synthetically generated independent variables and annual crash frequencies for each observation. Models are estimated

considering both that all relevant independent variables are included in the model and that one or more independent variables are excluded and with and without the proper functional form considered.

As shown in Figure 1, a total of 150 fixed parameter negative binomial models (50 data sets x 3 functional forms for each) and 150 random parameter negative binomial models are estimated. Corresponding models of a given functional form are compared based on their BIC values. When randomness is incorporated into the dependent variable generation (Step 2b (ii) above), each model is generated 8 times (2 potential random variables x 4 levels of randomness in each) leading to a total of 1200 fixed parameter negative binomial models (50 data sets x 8 x 3 functional forms for each) and 1200 random parameter negative binomial models. Referring to Step 3 above, five models are generated per dataset including five different combinations of independent variables (but not using all of the five independent variables at a time) leading to 250 fixed parameter negative binomial models (50 data sets x 5 combinations of independent variables) and 250 random parameter negative binomial models for each functional form (total of 750 fixed parameter and 750 random parameter negative binomial models for the three functional forms).



RP, SD, and IVs refer to the random parameter, standard deviation (for random parameter), and independent variables respectively.

Figure 1. Experimental Framework

The parameter retrieval capability of the models is assessed by comparing the mean values of parameters for the independent variables in the models with their corresponding true parameters. The true parameter of a specific independent variable refers to its value used in Eq.13/Eq.14/Eq.15 that was used to compute the expected crash frequency. The key motivation behind comparing the mean parameter with true parameter for the independent variables was to see how credible the estimation results for the models are based on the synthetic data sets. One of the most recent studies provides detailed guidance about the retrieval of parameters in the models estimated using simulated data (Bhowmik et al. 2021). Since a specific model is estimated numerous times using multiple data sets (as shown in Figure 1), the mean parameter value is computed for a specific independent variable which is then compared with its true value used to

compute the crash frequency. The difference between the two values can be presented in terms of absolute percentage bias which can be computed below:

$$\text{Absolute percentage bias} = \frac{(\text{true parameter} - \text{mean estimate})}{\text{true parameter}} * 100 \quad (17)$$

3. Results and Discussions

3.1. Actual versus Synthetic Data

As previously mentioned, crash data (N = 22,488) from 2005 to 2012 for engineering district 3 in Pennsylvania were used to generate a synthetic dataset with 100,000 observations. A comparison of the distribution of the key independent variables in the actual and synthetic datasets is provided in Figure 2. As shown, the distribution of all the independent variables in the synthetic data seems reasonable compared to the actual (observed) data.

The correlations between the independent variables in the synthetic data were also computed and then compared with the corresponding values in the actual data; see Table 1. Again, similar correlations between the variables in the two datasets can be noticed which justifies that the synthetic dataset is reasonable compared to the actual data.

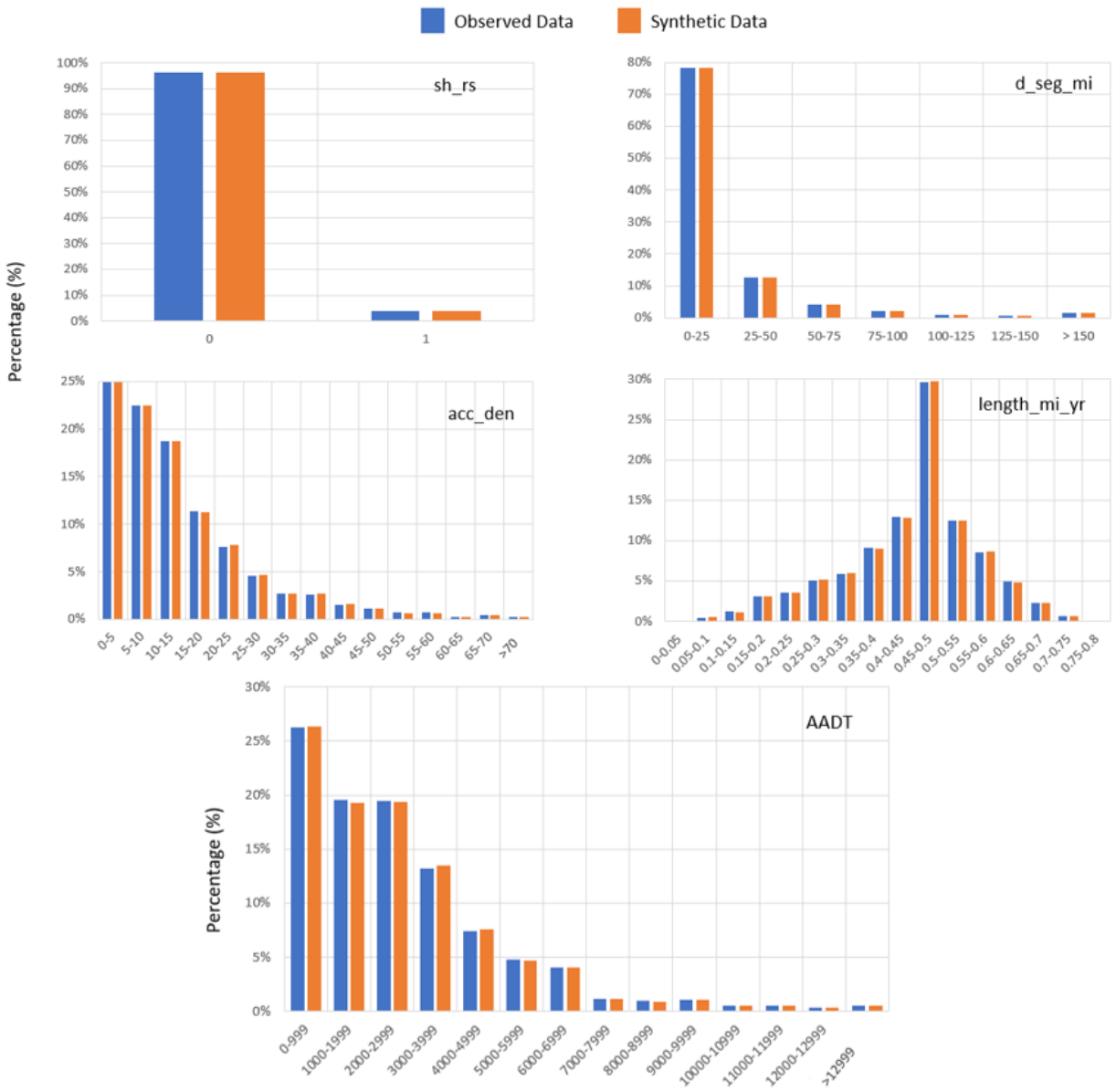


Figure 2. Percentage-wise Distribution of Variables in Observed (Actual) Versus Synthetic Data

Table 1. Descriptive Statistics of Independent Variables in Actual Data versus Synthetic Data

Correlation between Actual Independent Variables (N = 22,488)					
	AADT	length_mi	sh_rs	acc_den	d_seg_mi
AADT	1	---	---	---	---
length_mi	-0.100	1	---	---	---
sh_rs	0.233	-0.002	1	---	---
acc_den	0.318	-0.202	-0.043	1	---
d_seg_mi	-0.195	-0.071	-0.044	-0.006	1
Correlation between Synthetic Independent Variables (N = 100k)					
	AADT	length_mi	sh_rs	acc_den	d_seg_mi
AADT	1	---	---	---	---
length_mi	-0.093	1	---	---	---
sh_rs	0.242	-0.004	1	---	---
acc_den	0.315	-0.204	-0.044	1	
d_seg_mi	-0.199	-0.078	-0.046	-0.007	1

Table 2 provides comparisons between the actual crash frequency and the synthetic crash frequencies considering different functional forms and random parameters with various standard deviations for the two independent variables (AADT and presence of should rumble strips). It can be noticed that both mean and standard deviations of the actual crashes are a bit off from the synthetic crash frequencies computed under different scenarios; however, it was expected as the predicted crashes were computed assuming parameters in the respective models and were not based on models estimated using the actual data.

Table 2. Comparison of Actual versus Synthetic Crash Frequencies

Distribution of Synthetic Crash Frequency Using Different Functional Forms and Standard Deviations of Random Parameters				
Crash Frequency	**Mean**	**SD**	**Minimum**	**Maximum**
Actual Data	0.481	0.866	0	16
Fixed Parameter Models				
Synthetic Data	**Mean**	**SD**	**Minimum**	**Maximum**
Traditional	3.842	5.998	0	150
Hoerl	3.630	4.836	0	150
Eluru and Gayah	8.034	16.131	0	150
Random Parameter Models				
AADT as Random Parameter with 5% Standard Deviation				
Traditional	3.854	6.008	0	150
Hoerl	3.641	4.868	0	150
Eluru and Gayah	8.052	16.194	0	150
AADT as Random Parameter with 8% Standard Deviation				
Traditional	3.944	6.235	0	150
Hoerl	3.728	5.145	0	150
Eluru and Gayah	8.230	16.797	0	150
AADT as Random Parameter with 12% Standard Deviation				
Traditional	4.152	6.996	0	150
Hoerl	3.934	5.700	0	150
Eluru and Gayah	8.595	17.994	0	150
AADT as Random Parameter with 17% Standard Deviation				
Traditional	4.493	8.246	0	150
Hoerl	4.236	6.678	0	150
Eluru and Gayah	9.147	19.741	0	150
Shoulder Rumble Strip as Random Parameter with 5% Standard Deviation				
Traditional	3.890	6.152	0	150
Hoerl	3.665	4.917	0	150
Eluru and Gayah	8.128	16.434	0	150
Shoulder Rumble Strip as Random Parameter with 8% Standard Deviation				
Traditional	3.910	6.243	0	150
Hoerl	3.683	5.005	0	150
Eluru and Gayah	8.169	16.581	0	150
Shoulder Rumble Strip as Random Parameter with 12% Standard Deviation				
Traditional	3.929	6.386	0	150
Hoerl	3.710	5.203	0	150
Eluru and Gayah	8.230	16.829	0	150

3.2. Modeling Results

This section compares model results obtained when using fixed parameter and random parameter models under various conditions. As discussed in the methodology section, prior to comparing the results, the study first assessed the parameter retrieval capabilities for both fixed parameter and random parameter negative binomial with the three functional forms. To conserve space, we are restricting ourselves from providing the parameter retrieval results for data generated with AADT random parameter for 5% standard deviation only. However, the results were quite consistent across all the synthetic datasets including synthetic crash frequencies computed with different values of standard deviation for the AADT. The values presented in Table 3 clearly illustrate that the proposed model system recovers the parameters extremely well as indicated by the smaller absolute percentage bias values.

Table 3. True Parameter versus Mean Parameters in Fixed Parameter and Random Parameter Negative Binomial Models

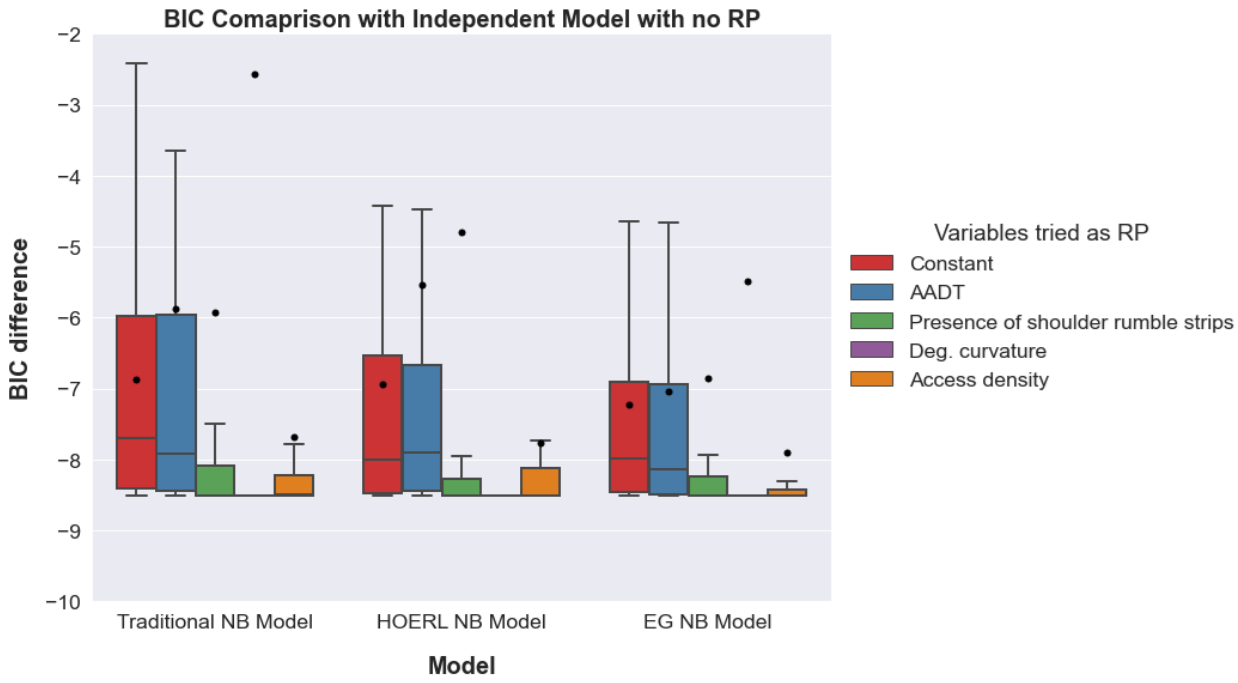
Model	Variables		Independent Model		RP Model	
		TP	MP	PB (%)	MP	PB (%)
Traditional NB	constant	-4.500	-4.550	1.117	-4.558	1.297
Traditional NB	ln(AADT)	0.800	0.808	0.977	0.809	1.094
Traditional NB	sh_rs	-0.800	-0.806	0.708	-0.805	0.586
Traditional NB	d_seg_mi	4.00E-03	3.81E-03	4.746	3.95E-03	1.326
Traditional NB	acc_den	2.00E-02	1.99E-02	0.663	1.96E-02	2.181
Traditional NB	Overdispersion Parameter	0.500	0.498	0.323	0.498	0.400
Hoerl NB	constant	-4.950	-4.882	1.376	-4.873	1.552
Hoerl NB	ln(AADT)	0.900	0.889	1.201	0.888	1.339
Hoerl NB	sh_rs	-0.800	-0.829	3.639	-0.792	0.942
Hoerl NB	d_seg_mi	4.00E-03	3.76E-03	5.882	3.92E-03	1.953
Hoerl NB	acc_den	2.00E-02	2.04E-02	2.034	2.02E-02	0.790
Hoerl NB	AADT	-1.00E-04	-9.58E-05	4.235	-1.95E-04	95.126
Hoerl NB	Overdispersion Parameter	0.400	0.402	0.512	0.389	2.800
EGNB	constant	-5.250	-5.216	0.652	-5.153	1.857
EGNB	ln(AADT)	1.000	0.998	0.175	0.991	0.877
EGNB	sh_rs	-0.800	-0.794	0.736	-0.800	0.033
EGNB	d_seg_mi	4.00E-03	3.75E-03	6.266	3.75E-03	6.229
EGNB	acc_den	0.020	0.019	2.956	0.019	3.784
EGNB	(AADT>1900)	2.100	2.159	2.826	2.037	2.991
EGNB	ln(AADT)*(AADT>1900)	-0.500	-0.511	2.194	-0.495	1.041
EGNB	Overdispersion Parameter	1.000	0.988	1.163	0.993	0.675

TP, MP, and PB refers to the true parameter, mean parameter, and percentage bias, respectively.

3.2.1. Comparison of Fixed Parameter versus Random Parameter Negative Binomial Models: Crash frequency computed with Fixed Parameter Negative Binomial Models

Figure 3 provides results that compare the fixed parameter and random parameter negative binomial models that were obtained when fixed parameter negative binomial models were used to compute the expected crash frequency for each observation. Values along the Y-axis show the difference in the mean values of BIC for the random parameter and fixed parameter negative binomial models in the corresponding figures for each of the three cases (traditional, Hoerl, and Eluru and Gayah functional forms). A positive difference in the mean BIC values along the Y-axis indicates that the random parameter negative binomial models perform better compared to their fixed parameter counterparts and vice versa (this applies to all relevant figures Figure 5-8). The different colors in Figure 3 refer to the situation when a constant or other specific independent variable is considered as the random parameter in the random parameter negative binomial models. The findings indicate that if there is no randomness (no random parameter is added to the equation of the fixed parameter model), the fixed parameter negative binomial models perform better compared to

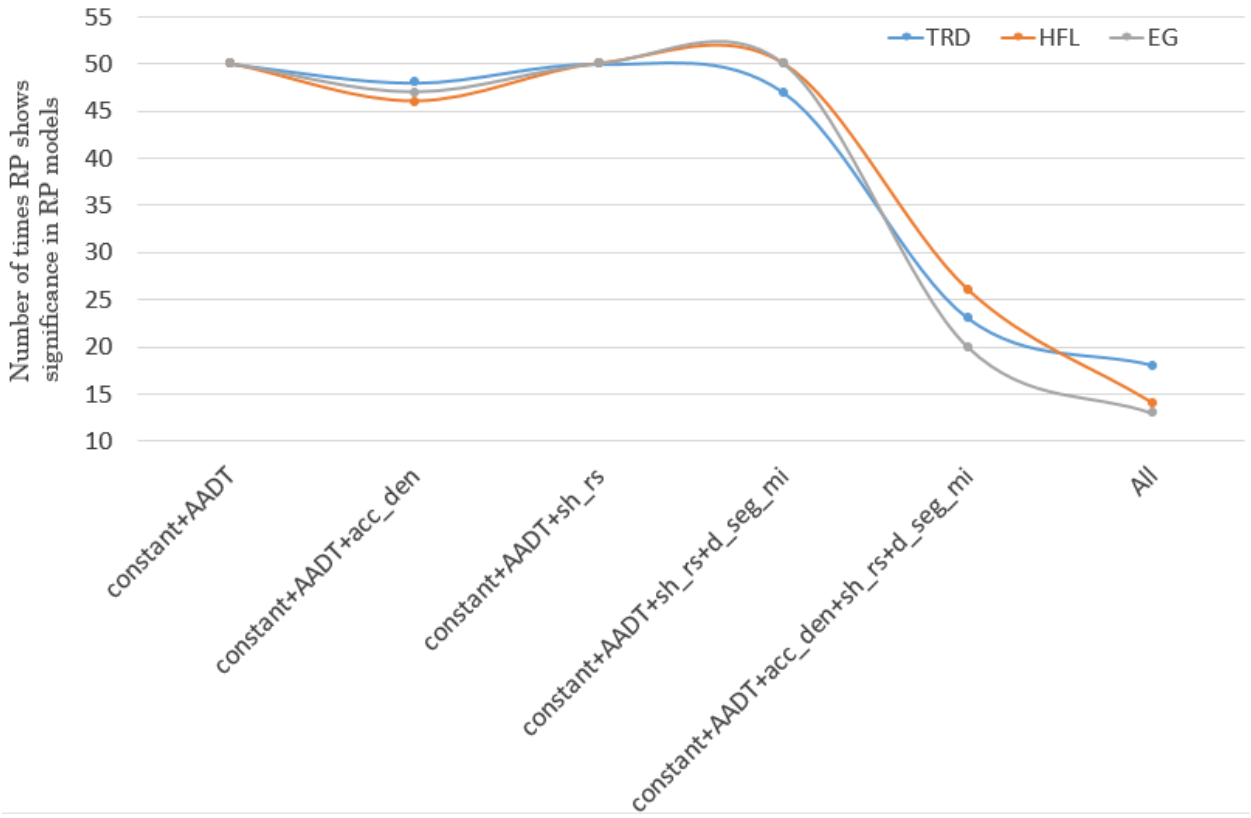
their random parameter counterparts based on the mean BIC values for all three functional forms. This is as expected since applying the random parameter models adds additional complexity and increases the number of parameters when estimated, but there is no additional benefit due to the lack of unobserved heterogeneity.



RP, NB, and EG refer to random parameter, negative binomial, and Eluru and Gayah, respectively.

Figure 3. Comparison of Random Parameter versus Fixed Parameter (Independent) Models based on Data with No Randomness

Figure 4 plots the number of times random parameters are statistically significant in a model even when no unobserved heterogeneity exists in the data. The results are plotted as a function of the number of independent variables considered in the random parameter model. Note that when there are few independent variables in the model, unobserved heterogeneity exists since relevant explanatory variables are not included. In this case, the random parameters are often significant and help account for this unobserved heterogeneity introduced by omitting an available and critical parameter. However, even when all six independent variables are considered in the random parameter models, random parameters are statistically significant in the random parameter models a considerable number of times. Specifically, random parameters are significant 12, 13, and 18 times (out of 50 times) for the models with traditional, Hoerl, and flexible forms respectively. However, in terms of the mean BIC values, the random parameter models do not outperform the fixed parameter despite the presence of statistically significant random parameter in the re-estimated random parameter models. These results suggest that random parameters might sometimes show up as significant in a model even when no unobserved heterogeneity exists.



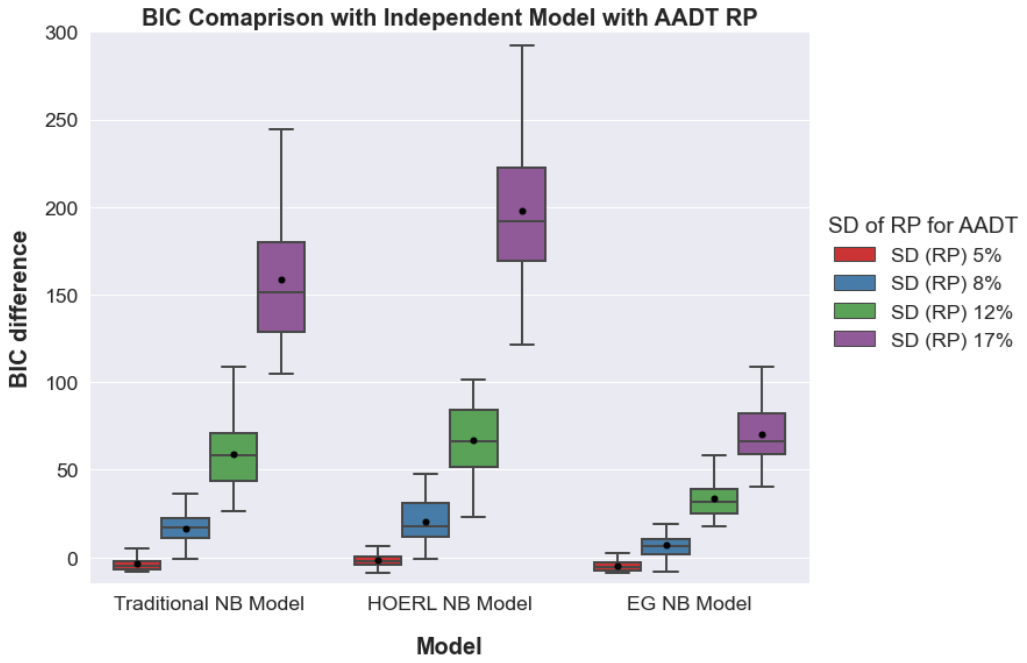
TRD, HFL, and EG refer to traditional, Hoerl, and Eluru and Gayah functional forms respectively.

Figure 4. Random Parameter versus Fixed Parameter Models based on Subset of Independent Variables

3.2.2. Comparison of Fixed Parameter versus Random Parameter Negative Binomial Models: Crash frequencies computed with Random Parameter Negative Binomial Models

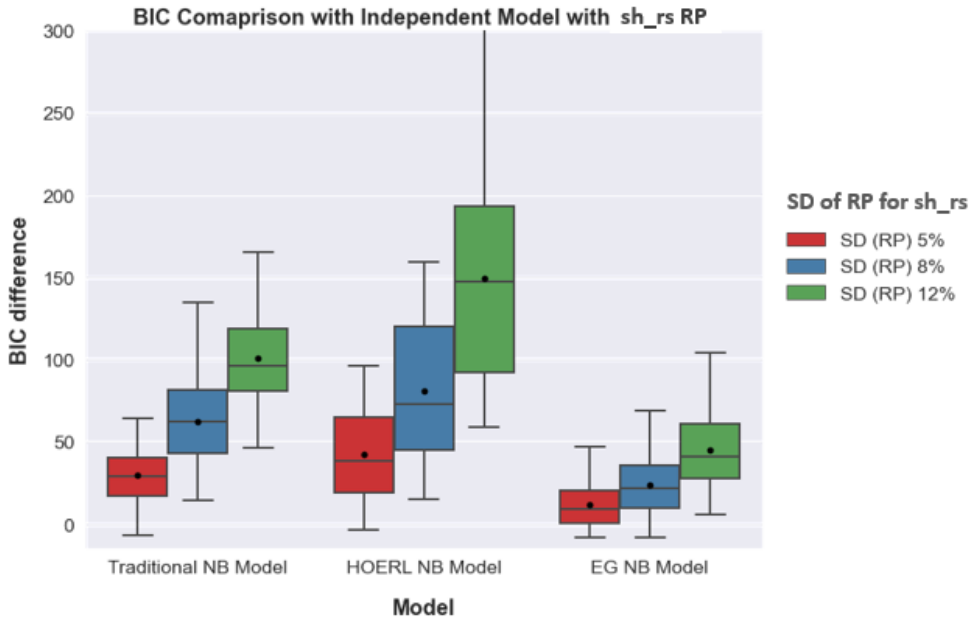
Figure 5 and Figure 6, respectively, compare the fixed parameter and random parameter negative binomial models that were obtained when random parameter models with different values of randomness (standard deviation based on the values/distribution of a specific random parameter) for the AADT and sh_rs variables were used to compute the expected crash frequency for each observation. The different colors in the two figures refer to different values of standard deviation for AADT and sh_rs used when these variables are added to the equations of fixed parameter models to compute the expected crash frequencies.

The findings indicate that when randomness actual exists in the crash generation process, the random parameter negative binomial models show superior performance based on the mean BIC values compared to the fixed parameter counterparts for all the three functional forms. These findings make sense: if randomness (unobserved heterogeneity) in the data exists and is taken into account via estimating random parameter negative binomial models, the random parameter models perform better compared to their fixed parameter counterparts.



RP, SD, NB, and EG refer to random parameter, standard deviation, negative binomial, and Eluru and Gayah, respectively.

Figure 5. Comparison of Random Parameter versus Fixed Parameter Models based on Data with Randomness (due to AADT)



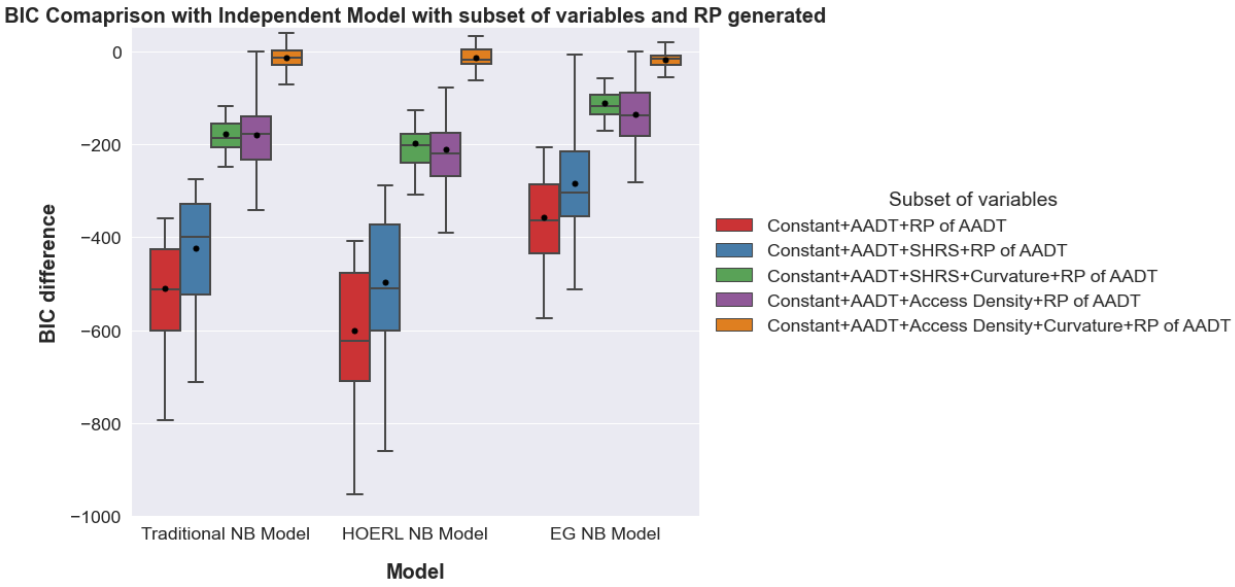
RP, SD, NB, and EG refer to random parameter, standard deviation, negative binomial, and Eluru and Gayah, respectively.

Figure 6. Comparison of Random Parameter versus Fixed Parameter Models based on Data with Randomness (due to sh_rs)

The above results lead to several key insights: 1) in the absence of structural randomness in the data generation, the fixed parameter negative binomial models perform better compared to random parameter negative binomial models;2) if randomness in the data exists, taking it into account via random parameter negative binomial models make the random parameter negative binomial models superior to their fixed parameter counterparts based on the mean BIC values for all three functional forms; 3) failure to incorporate key variables in the model with fixed impacts can manifest as randomness in other variables, even when such randomness does not exist. The reader would note that it is not expected that real world data will be free of unobserved factors. However, the exercise is conducted to illustrate how random parameter models cannot address all the challenges with data and analysts need to exercise caution in their adoption. To assess the role of considering randomness in data and independent variables with fixed parameters in the performance of random parameter negative binomial, please see the subsequent section.

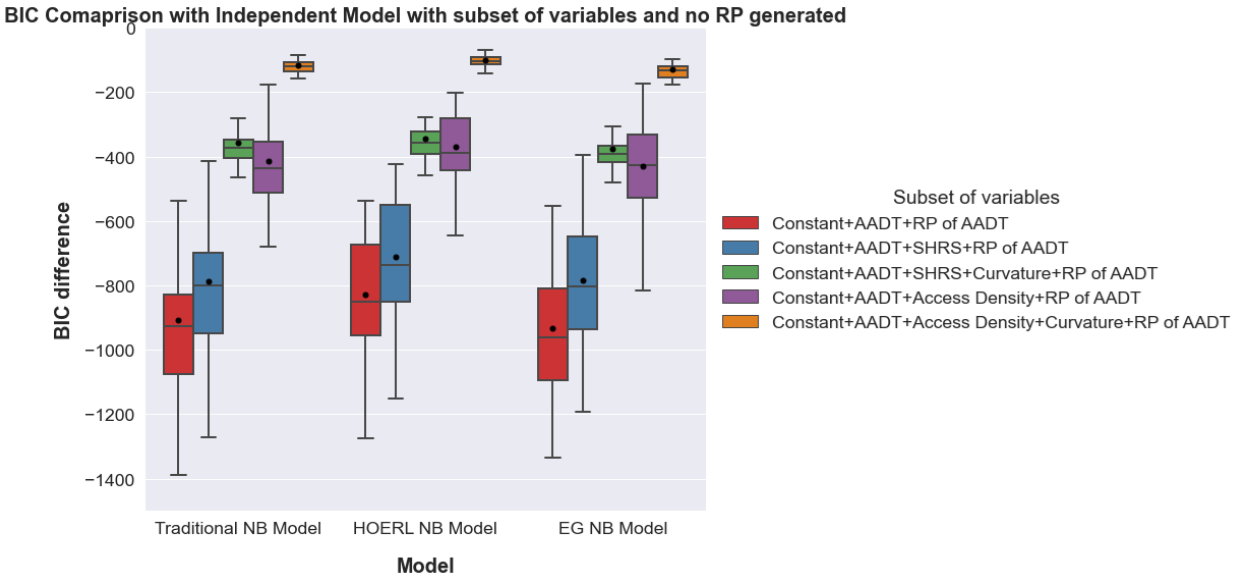
3.2.3. Comparison of Fixed Parameter versus Random Parameter Negative Binomial Models: Contribution of independent variables and randomness

Figure 7 illustrates the results which are used to compare the fixed parameter and random parameter negative binomial models obtained when random parameter models (with only 12% of standard deviation for AADT used as a random parameter) were applied to compute the expected crash frequency for each observation. The results indicate that if there is randomness in the data but the best combination of independent variables with fixed parameters was not used in estimating the random parameter negative binomial models, the fixed parameter negative binomial models may perform better compared to their random parameter counterparts based on the BIC value. This is critical finding highlights that missing important independent variables in the specification could significantly influence the performance of the random parameter negative binomial model. The result is particularly important to applied researchers. There is a focus on presenting models with large number of random parameters. The finding from our analysis should serve as a caution to any random parameter development. The consideration of random parameters should follow an exhaustive testing exercise with all relevant observed independent variables. The results suggest that it is better to develop models that have 6 fixed parameters (possibly with 6 independent variables) and 2 random parameters as opposed to developing a model with 4 fixed parameters and 4 random parameters. Figure 8 illustrates similar findings after comparing fixed parameter and random parameter negative binomial models obtained when fixed parameter models were used to compute the expected crash frequency.



RP, NB, and EG refer to random parameter, negative binomial, and Eluru and Gayah, respectively.

Figure 7. Comparison of Random Parameter versus Fixed Parameter Models based on Randomness and Subsets of Independent Variables



RP, NB, and EG refer to random parameter, negative binomial, and Eluru and Gayah, respectively.

Figure 8. Comparison of Random Parameter versus Fixed Parameter Models based on No Randomness and Subsets of Independent Variables

4. Conclusions and limitations

This study applies a simulation-based statistical analysis to examine how proper model specification is critical for analysts to consider before applying random parameters to crash frequency models. A machine learning method is used to generate a synthetic dataset that matches closely to actual data observed in the field. Then, crash data is synthetically generated according to some known processes that the crash frequency models seek to recover, both ignoring and considering underlying unobserved heterogeneity via the introduction of random errors. Various fixed and random parameter models are estimated with improper and proper model specifications. The results suggest that failure to properly specify a model – either by omitting a critical variable or not incorporating the correct functional form – can cause parameters that may not necessarily be random to appear so. However, fixed parameters alone are not sufficient to account for systematic randomness that may exist in crash data, such as when unobserved heterogeneity is present. Since such unobserved heterogeneity is likely to always be present in real data, the implementation of random parameters is a critical and useful step in the development of crash frequency models, assuming that the model is first properly specified. The present study investigates the fixed parameter and random parameter negative binomial models from an estimation standpoint that relates to the role of model specification ignoring the significant fixed parameters in the random parameter model could lead to miss-specification. One of the key insights from the analyses is whether the randomness in the data generation exists or not, if the best set of significant variables with fixed parameters is not selected, including significant random parameter(s) does not improve the mean BIC values of the random parameter models compared to their fixed parameter counterparts. The results highlight how the consideration of random parameters alone cannot serve as a solution to mis-specification issues in model development. Thus, the consideration of random parameters should be implemented in concert with a well-defined model specification process (defined by considering different functional forms and all observed variables in model development).

While these findings provide useful insights, an interesting future research avenue is to check whether the misspecification could affect the out-of-sample prediction performance of the random parameter negative binomial models. Both the in-sample and out-of-sample predictions for the fixed parameter models are easily obtainable due to a consistent parameter estimates for all variables throughout observations. While obtaining the in-sample predictions for the random parameter models is still easier, it gets complicated when the aim is to obtain the out-of-sample predictions due to unavailability of observation-specific coefficients for random parameters in the random parameter negative binomial models. Most of the past studies have used the global mean approach to compute the out-of-sample predictions for the random parameter models (Wali et al. 2018, Tang et al. 2019). However, the out-of-sample predictions obtained via the global mean method could be flawed as it does not consider the variance due to observation-level coefficients. In this regard, some of the recent studies incorporate the mean and variance of the coefficients of random parameters when computing the out-of-sample predictions of the random parameter models (Hou et al. 2021, Xu et al. 2021, Hou et al. 2022). These studies reveal that if both mean and variance of the RPs are considered, the RP models can significantly outperform their fixed parameter counterparts. Considering the abovementioned points, the authors plan to extend their present analyses to understand how model specification relates to out-of-sample prediction performance of the random parameter negative binomial models as future research efforts.

6. References

Abowd, J.M., Lane, J., 2004. New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In Proceedings of the International Workshop on Privacy in Statistical Databases, pp. 282-289.

Abowd, J.M., Woodcock, S.D., 2004. Multiply-imputing confidential characteristics and file links in longitudinal linked data. In Proceedings of the International Workshop on Privacy in Statistical Databases, pp. 290-297.

Alnawmasi, N., & Mannering, F., 2022. The impact of higher speed limits on the frequency and severity of freeway crashes: Accounting for temporal shifts and unobserved heterogeneity. Analytic Methods in Accident Research, 34, 100205.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153-159.

Bauer, K.M., Harwood, D.W., 2013. Safety effects of horizontal curve and grade combinations on rural two-lane highways. Transportation Research Record 2398, 37-49.

Bhowmik, T., Yasmin, S., Eluru, N., 2021. A new econometric approach for modeling several count variables: A case study of crash frequency analysis by crash type and severity. Transportation Research Part B: Methodological 153, 172-203.

Caiola, G., Reiter, J.P., 2010. Random forests for generating partially synthetic, categorical data. Transactions on Data Privacy. 3 (1), 27-42.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis and Prevention 70, 320-329.

Drechsler, J., Reiter, J.P., 2010. Sampling with synthesis: A new approach for releasing public use census microdata. Journal of the American Statistical Association 105 (492), 1347-1357.

Drechsler, J., Reiter, J.P., 2011. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Computational Statistics and Data Analysis 55 (12), 3232-3243.

Eluru, N., Gayah, V.V., 2022. A note on estimating safety performance functions with a flexible specification of traffic volume. Accident Analysis and Prevention 167, 106571.

Feknssa, N., Venkataraman, N., Shankar, V., & Ghebrab, T., 2023. Unobserved heterogeneity in ramp crashes due to alignment, interchange geometry and truck volume: Insights from a random parameter model. Analytic Methods in Accident Research, 37, 100254.

Fitzpatrick, K., Lord, D., Park, B.-J., 2010. Horizontal curve accident modification factor with consideration of driveway density on rural four-lane highways in Texas. Journal of Transportation Engineering 136 (9), 827-835.

Gayah, V.V., Donnell, E.T., 2021. Estimating safety performance functions for two-lane rural roads using an alternative functional form for traffic volume. Accident Analysis and Prevention 157, 106173.

Gooch, J.P., Gayah, V.V., Donnell, E.T., 2016. Quantifying the safety effects of horizontal curves on two-way, two-lane rural roads. Accident Analysis and Prevention 92, 71-81.

Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: 18 contrasts with mixed logit. Transportation Research Part B: Methodological 37(8), 19 681–698.

Hauer, E., 2015. The art of regression modeling in road safety.

Hou, Q., Huo, X., Leng, J., Mannering, F., 2022. A note on out-of-sample prediction, marginal effects computations, and temporal testing with random parameters crash-injury severity models. Analytic Methods in Accident Research 33, 100191.

Hou, Q., Huo, X., Tarko, A.P., Leng, J., 2021. Comparative analysis of alternative random parameters count data models in highway safety. Analytic Methods in Accident Research 30, 100158.

Huo, X., Leng, J., Hou, Q., Zheng, L., Zhao, L., 2020. Assessing the explanatory and predictive performance of a random parameters count model with heterogeneity in means and variances. Accident Analysis and Prevention 147, 105759.

Khattak, A., Ahmad, N., Mohammadnazar, A., Mahdinia, I., Wali, B., Arvin, R., 2020. Highway Safety Manual Safety Performance Functions and Roadway Calibration Factors: Roadway segments phase 2, part 1. Tennessee. Department of Transportation.

Kinney, S.K., Reiter, J.P., Berger, J.O., 2011. Model selection when multiple imputation is used to protect confidentiality in public use data. Journal of Privacy and Confidentiality 2 (2).

Kononov, J., Lyon, C., Allery, B.K., 2011. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. Transportation Research Record 2236, 11-19.

Mannering, F. L., and Bhat, C. R. 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research, 1, 1-22.

Mannering, F. L., Shankar, V., Bhat, C. R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research, 11, 1-16.

Mannering, F., Bhat, C. R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research, 25, 100113.

Mahmud, A., Gayah, V. V., Paleti, R., 2023. Estimation of crash type frequency accounting for misclassification in crash data. Accident Analysis and Prevention, 184, 106998.

Nowok, B., Raab, G.M., Dibben, C., 2016. Synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software 74, 1-26.

Part, D., 2010. Highway safety manual. American Association of State Highway and Transportation Officials: Washington, DC, USA 19192.

Reiter, J.P., 2005a. Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society: Series A (Statistics in Society) 168 (1), 185-205.

Reiter, J.P., 2005b. Using CART to generate partially synthetic public use microdata. Journal of Official Statistics 21 (3), 441.

Shankar, V.N., Albin, R.B., Milton, J.C., Mannering, F.L., 1998. Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model. Transportation Research Record 1635, 44-48.

Tang, H., Gayah, V.V., Donnell, E.T., 2019. Evaluating the predictive power of an SPF for two-lane rural roads with random parameters on out-of-sample observations. Accident Analysis and Prevention 132, 105275.

Ulfarsson, G.F., Shankar, V.N., 2003. Accident count model based on multiyear cross-sectional roadway data with serial correlation. Transportation Research Record 1840, 193-197.

Venkataraman, N.S., Ulfarsson, G.F., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Exploratory insights from random parameter negative binomial approach. Transportation Research Record 2236, 41-48.

Vij, A., Carrel, A., Walker, J.L., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. Transportation Research Part A: Policy and Practice 54, 164–178.

Wali, B., Khattak, A.J., Waters, J., Chimba, D., Li, X., 2018. Development of safety performance functions: Incorporating unobserved heterogeneity and functional form analysis. Transportation Research Record 2672 (30), 9-20.

Wang, K., Zhao, S., Jackson, E., 2020. Investigating exposure measures and functional forms in urban and suburban intersection safety performance functions using generalized negative binomial-P model. Accident Analysis and Prevention 148, 105838.

Xu, P., Zhou, H., Wong, S., 2021. On random-parameter count models for out-of-sample crash prediction: Accounting for the variances of random-parameter distributions. Accident Analysis and Prevention 159, 106237.