Model of Relationship Between Interstate Crash Occurrence and Geometrics

Exploratory Insights from Random Parameter Negative Binomial Approach

Narayan S. Venkataraman, Gudmundur F. Ulfarsson, Venky Shankar, Junseok Oh, and Minho Park

This paper proposes the use of a random parameter negative binomial (NB) model for the analysis of crash counts. With the use of a 9-year, continuous panel of histories of total crash frequencies on Interstate highways in Washington State for 1999 to 2007, a random parameter NB model was estimated to account for parameter correlations, panel effects that contributed to intrasegment temporal variations, and between-site effects. Interstate geometric variables, such as lighting type proportions by length, shoulder width proportions, lane cross-section proportions, and curvature variables, were used in the model specification. Curvature variables included the number of horizontal curves in a segment, the number of vertical curves in a segment, the shortest horizontal curve in a segment length, the largest degree of curvature in a segment, the smallest vertical curve gradient, and the largest vertical curve gradient in a segment. Segments were analyzed at the interchange and the noninterchange levels. A total of 1,153 directional segments of the seven Washington State Interstates were analyzed. The analysis yielded a statistical model of crash frequency on the basis of 10,377 observations. Several curvature effects were found to be random, which meant that they varied from segment to segment. Although, for example, the numbers of horizontal and vertical curves in a segment were fixed-parameter effects, the largest degree of curvature, as well as the smallest and largest vertical curve gradient variables, were random parameters. The logarithm of average daily travel and the median and point lighting proportions were also found to be random parameters. These results suggested that segment-specific insights into crash frequency occurrence could be improved for appropriate design policy and prioritization.

Researchers worldwide have used a variety of statistical models to explain crash occurrence. Count models, such as the negative binomial (NB), have been used extensively until recently for crash modeling purposes because of the well-known problem of overdispersion (*1–4*). Another study by Shankar, Milton, and Mannering (*5*) suggests that traditional applications of the NB distributions do not address the possibility that more than one underlying process may be affecting crash frequency likelihoods. They proposed the use of the zero-inflated Poisson and the zero-inflated NB models, both of which seek to account for the preponderance of zero-crash observations, often found in crash frequency data. There have been counterarguments to the use of zero-inflated models in studies (*6*). The evolution of count modeling since these studies were done has resulted in a recent emphasis on random parameter approaches, multivariate approaches, and other approaches that involve Bayesian methods (*7–10*). Lord and Mannering overviewed these approaches (*11*). Nevertheless, the promise of these recent methods has not been fully exploited. In particular, random parameter approaches offer substantial promise in the identification of heterogeneity that is segment-specific. Heterogeneity, in conventional approaches such as the NB, has been exploited through the quadratic overdispersion form for various modeling objectives. A few studies, for example, have specifically explored the relationship between highway geometrics, traffic-related elements, and crash frequencies. Miaou (*3*) and Shankar, Mannering, and Barfield (*12*) were among the first to attempt to accomplish this in their studies. They used Poisson and NB to evaluate effects of roadway geometrics and environmental factors on rural crash frequency, and also modeled the frequencies of specific collision types. Poch and Mannering used the NB to identify left-turn volume and total approaching volume as highly significant determinants of crash frequency at intersections (*13*). The early consensus was that, as a mix of Poisson and gamma distributions, NB models could account for the site-specific variations (*14, 15*). This, however, rested on the treatment of heterogeneity as a fixed effect across segments. The issue of heterogeneity continues to be a substantial methodological issue in the field of traffic safety. The objective of this study was to shed further light on the effects of heterogeneity at the segment level in a divided Interstate freeway setting. In particular, a data set with Interstate geometrics and a 9-year panel of crash counts for the period 1999 to 2007 was used to evaluate heterogeneity. The rest of this paper is organized as follows. A brief review of recent studies that relate to random parameter models is provided. It is followed by a discussion of the methodology, the modeling results from the estimation of the Interstate data set, and conclusions and recommendations.

N. S. Venkataraman and G. F. Ulfarsson, Department of Civil and Environmental Engineering, University of Iceland, VR-II, Hjardarhagi 6, 101 Reykjavik, Iceland. V. Shankar, J. Oh, and M. Park, Department of Civil and Environmental Engineering, Pennsylvania State University, 226C Sackett Building, University Park, PA 16802. Corresponding author: V. Shankar, shankarv@engr.psu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2236, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 41–48.
DOI: 10.3141/2236-05

RELEVANT MODELING LITERATURE

The major limitation of fixed parameter NB models is that they cannot incorporate time variation or segment-specific effects, which results in an underestimation of standard errors in the regression coefficients and in subsequently inflated *t*-ratios. Some studies have

attempted to fix this problem by introducing a trend variable in the crash model (*15*). If heterogeneity is present because of panel effects, however, this method still will not capture all the unobserved heterogeneity. To account for this issue, Hausman et al. examined fixed- and random effects NB models for panel count data (*16*). The fixed-effects NB model does not allow for section-specific variation, but the random effects NB model allows for randomly distributed, section-specific variation (*15*). In a similar study, researchers were able to identify 11 specific variables that significantly affected the safety of intersections through the random effects NB model (*17*).

In the area of classical models that incorporate randomness into parameters of count data, Milton, Shankar, and Mannering showed the first documented application in the field of traffic safety (*18*). They examined the contribution of geometric and traffic variables to the occurrence of severity proportions at the segment level. They illustrated the flexibility of a mixed logit structure by modeling the proportion of severity in relation to five major reported types, namely, property damage only, possible injury, evident injury, disabling injury, and fatality. The flexibility of the structure arose from the fact that several geometric variables were found to influence the severity proportion in a nonfixed manner. Individual, segment-specific variation in variable influence was found to be statistically significant. The study also showed that segment effects need not always be unidirectional, with the exception of length and average daily traffic (ADT). It was empirically determined, for example, that the frequency of vertical grade changes per mile was negatively associated with severities in some segments and positively associated with a severities increase in others. The study showed that incorporation of randomness into parameter variation is a useful approach to uncover segment-specific heterogeneity, although more studies need to be conducted in a variety of empirical contexts to corroborate this finding into a generalizable concept. Whether the variation occurs in a continuous form (e.g., normal distribution) or in a discrete form (classes of variation), it remains an area for application of rich combinations of random parameter models. The following studies are leading examples of the importance of accounting for segment-specific heterogeneity. Anastasopoulos and Mannering used random parameter NB regression to account and correct for heterogeneity, which can arise from a number of factors that relate to road geometrics, pavement, and traffic characteristics, driver behavior, vehicle types, socioeconomic factors, variations in police recording crashes, time, and other unobserved factors (*9*). Through their use of vehicle crash data, their findings provided factors related to pavement condition and quality, which were found to significantly influence vehicle crash occurrences, including the effects of friction, the international roughness index, pavement rutting, and pavement condition rating. Geometric factors and their effect on vehicle crash frequencies, road segment length, median types and width, number of ramps and bridges, horizontal and vertical curves, and shoulder widths were all found to be statistically significant. The traffic variables of annual ADT and the percent of combination trucks in the traffic stream were both found to have significant impacts on crash frequencies. El-Basyouny and Sayed used covariates to represent segment length, annual ADT, crosswalk density, business land use, and unsignalized intersection density. They found the number of lanes between signals to significantly influence crash frequencies (*8*). Their approach led to new insights into how the covariates affect crash frequencies, and to account for heterogeneity as a result of unobserved road geometrics, traffic characteristics, environmental factors, and driver behavior. Inclusion of corridor effects in the mean function could explain enough variation that some of the model covariates would be rendered nonsignificant, and the inference would be affected as well. Malyshkina et al. (*10*) used Markov switching count-data models to provide a superior statistical fit for crash frequencies relative to standard zero-inflated models. The estimation helped in detailed study of individual roadway segments, which can lead to a better understanding of why some segments are considerably safer than others during some time periods. Their study concluded that two-state Markov switching count-data models are likely to be a better alternative to zero-inflated models to account for the excess of zeros often observed in crash-frequency data. Washington et al. (*19*), Ulfarsson and Shankar (*20*), and Sittikariya and Shankar (*21*) underscored the importance of heterogeneity as a result of time-varying effects.

METHODOLOGY

A generalized representation of the conditional density function for crash counts for i segments with t years of observation is as follows:

$$P\left(y_{it}\middle|x_{it},\beta_{it}\right)=g\left(y_{it},\beta'_{it}x_{it}\right) \qquad \text{for } i=1,2,\ldots,1{,}153;\ t=1,\ldots,9 \tag{1}$$

The $g(.)$ is a density function that relates to the mean crash occurrence rate in terms of the vector of estimable parameters β'_{it} and vector of observed variables y_{it} and x_{it}. The mean rate is the lambda variable that is estimated in conventional NB models, wherein lambda is used to characterize the density function $P(.)$. The variables in Equation 1 are subscripted with both i and t to represent yearly observations of crash count y and a vector of regional, directional, lighting, geometric, and traffic attributes x. The total number of segments consistently observed was 1,153 for both directions of travel, and the total number of years of observation was 9. This structure assumed a balanced panel, which need not be the case if, in some years, data were incomplete, or if part of them were missing. Geometric variables may change over time, as may lighting variables and traffic volumes. If, however, the subscripts in β'_{it} are removed so as to operate as β', then the parameter effect is restricted to being the same across all segments across all years. This representation usually is used in nonrandom parameter models, such as the classical NB. That the subscript notation can be applied to ancillary parameters, such as the overdispersion effect as well to an unsubscripted overdispersion effect, implies that unobserved heterogeneity is the same across all segments across all years. One that varies over segments and years usually results in a random effects type of structure.

The main characteristic of a random parameter NB model is that β is not fixed. Because the x's are constrained to the observed data set, the modeling objective is to maximize the information available from the x's. As such, insights from the model should be limited to the range of values observed in the x's. One possible way to achieve this maximization objective is to vary the β across segments, rather than fix them to be constant across segments and years. Assume, for example, that normal distribution for β would be a way to introduce a continuous variation in parameter randomness. The likelihood function then is modified in the sense that it is based on probabilities that are calculated from varying β's instead of fixed β's such that

$$\beta_{it}=\beta+\Delta z_i+\Gamma v_{it} \tag{2}$$

where the second term indicates heterogeneity in the mean of the parameter as a function of variables in z_i, and the third term is a random deviation from the mean. This is a general form of describing parameter heterogeneity. In particular, the vector delta includes parameters for exogenous variables that are found to influence the beta vector systematically. If heterogeneity in parameter means is not explicitly modeled as a function of exogenous variables, then the second term in Equation 2 disappears. All heterogeneity is modeled

by using the third term as a distribution specific to the type of *x* variable used. If *x* is a dummy, then the second term is modeled as a uniform distribution; if *x* is continuous, a normal or lognormal distribution may be used. The third-term parameter vector gamma represents the lower triangular matrix, which includes the diagonals to represent both the scale and covariance matrix of the random parameters in the model estimated in this study. In this study, correlation between random parameters was allowed. In the more general form, if the second term is estimated and includes a vector *z*, then variables that enter *z* are preferably different from those that enter *x*. The subscripts are such that heterogeneity in the means explained by a systematic combination of variables in z_i indicates segment-specific variations in the parameter mean, whereas the unobserved heterogeneity denoted by ν_{it} is assumed to exist at the yearly level for each segment. The likelihood contribution of the *i*th segment to the sample likelihood is conditioned on the unobserved heterogeneity ν_{it} and denoted by

$$P(y_{i1}, \ldots, y_{i5} | x_{it}, z_{it}, \nu_{it}) = \prod_{t=1}^{9} g(y_{it}, \beta'_{it} x_{it}) \quad (3)$$

To obtain the marginal density, the conditional density is integrated with respect to the density function of ν_{it} as follows:

$$L_i = \int_0^{\infty} h(\nu_{it}) \prod_{t=1}^{9} g(y_{it}, \beta'_{it} x_{it})\, d\nu_{it} \quad (4)$$

The likelihood for segment *i* is, as a result, a nonclosed form problem, because there is a distributional assumption for β_{it}. This integral can be approximated, however, through simulation by using the relationship

$$E_{\nu_{it}}\left[L_i | \nu_{it}, t = 1, \ldots, 9\right] \gg \frac{1}{R} \sum_{r=1}^{R} L_{ir} | \nu_{itr} \quad (5)$$

For the entire sample then,

$$\log L = \sum_{i=1}^{1,153} \log L_i \quad (6)$$

and by evaluating the first- and second-order conditions with respect to parameters, estimates of the effects of variables associated with segment crash counts can be determined. The above-mentioned procedure is generally called the method of simulated maximum likelihood and relies heavily in its approximation accuracy of the actual likelihood on the number of draws *R*. Draws are taken from a distribution such as the normal (for continuous variables) or uniform (for binary variables) and rejected until a draw from the desired region is obtained. This process, however, may require a large number of draws, if the desired region lies close to the tails of the normal distribution. To speed up the estimation process, quasi-random sequences, such as Halton draws, are used. The Halton sequence arises on the basis of a set of values of simulation draws, which involves sequences that are efficiently spread over the unit interval in which draws are generated. For each simulation, it is desirable to ensure that the sequence of draws generated is the same for each segment; furthermore, it is desirable to maintain the sequence for multiple runs.

With the above-described simulated maximum likelihood method, two types of random parameter NB structures were considered for the Washington State Interstate data set. They differed on the relationship between heterogeneity in year *t* and year *t*-1. In an autocorrelated structure, there was a one-lag assumption (AR(1), $\nu_{it} = W\nu_{i,t-1} + \mu_{it}$, where *W* was a diagonal matrix of coefficient-specific autocorrelation coefficients, and μ_{it} satisfied the earlier specification for ν_{it}.) In theory, it is entirely possible that the lag may differ across segments; the one-lag model is therefore to be considered a benchmark model that assumes a restrictive, nonzero correlation structure. It was determined that an unstructured correlation model appeared as a plausible alternative. Although they pose convergence problems, such structures require evaluation across multiple states in terms of being considered viable alternatives to the one-lag, correlated, random parameter NB. The random effects model (*15, 17*) is a special case of the random parameter NB in which the only randomness is in the constant term. The above-mentioned generalized structure for the random parameter NB can be relaxed to accommodate sub-vectors of parameters that are to be fixed, as opposed to those that are random and correlated.

DATA DESCRIPTION

The panel data in this research consisted of 9 years (1997 to 2007) of crash counts, geometric, and traffic volume data for Washington State's Interstate system. Figure 1 shows an Interstate and national highway map for Washington State. The seven Interstates, namely, I-5, I-90, I-405, I-82, I-182, I-205, and I-705 were modeled in this study. The remaining seven national highways were not used in this study, because data were not comprehensively available. It can be expected, however, that the use of data from the national highways can yield richer insights on heterogeneity because of variations by functional class and environmental interactions. Environmental effects have been found to be significant contributors to crash occurrence likelihoods (*22*). Another reason that the national highway system may yield more insight is because of the level of segmentation used in this study. Because the study involves segmentation at the interchange and noninterchange levels, interchange density (and therefore segment length) can be a key heterogeneity factor in the influence of geometrics on crash occurrence. A total of 1,153 directional segments were assembled on the basis of interchange and noninterchange definitions. Interchange segments were defined by the farthest merge and diverge ramp limits for each direction. As such, a single interchange can have different limits by direction. Noninterchange segments were defined as continuous travel segments between two interchanges. Over a continuous period of 9 years, the 1,153 segments yielded a balanced panel of 10,377 observations.

Table 1 shows the descriptive statistics for key variables used in the estimation of the correlated, random parameter NB of total annual crash frequency. More than 100 variables were measured by visual scanning and integration with electronic information from the Washington Department of Transportation databases. The variable list included lighting type, lane cross sections, shoulder widths, horizontal curve parameters, and vertical curve parameters.

As shown in Table 1, the total annual crash frequency had a mean of 10.84 and a standard deviation of 18.9, with a maximum of 388 crashes. The percentage of observations with more than 10 crashes per year was 30.79%. The mean of directional, per lane ADT was 6,529.78 with a maximum of 22,111.89 and a standard deviation of 8,391.396. Directional, per lane ADT was a weighted (by length) measure of ADT to address cases where cross sections changed, or ADT itself changed within a segment. In the estimated model, this measure was used in logarithmic form, which indicated that it was interactive with Interstate geometrics in a multiplicative manner. Length was used in a logarithmic form as well, which indicated its multiplicative effect. (The logarithmic forms of ADT and length

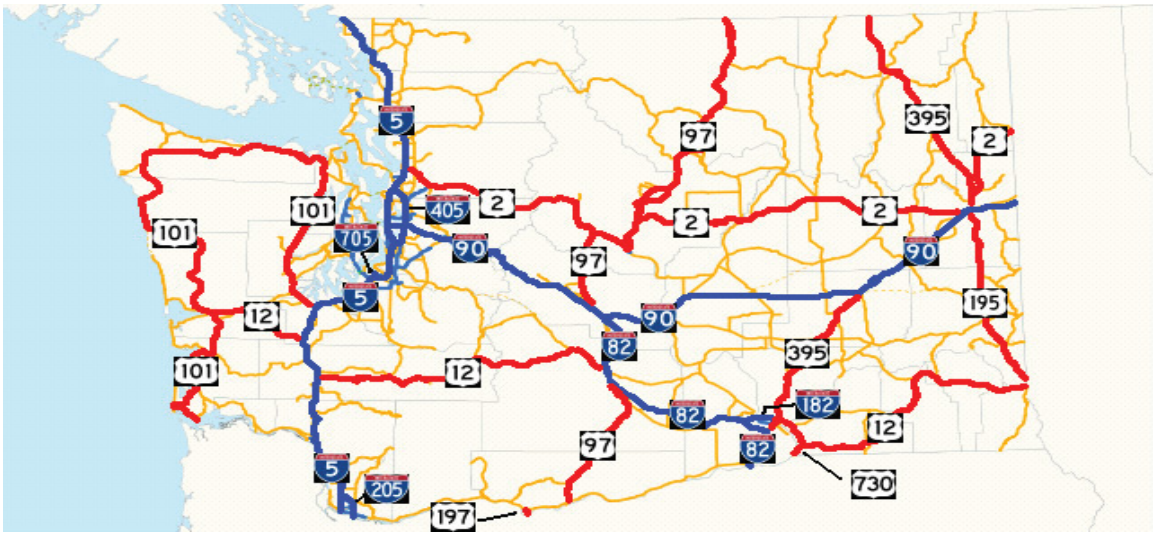

FIGURE 1 Washington State Interstate and highway map.

yielded statistically better fits as well.) In the current data set, the average segment length was 1.326 mi, with a maximum of 20.38 mi. The entire Interstate network was scanned visually for fixed roadway lighting and was measured in terms of proportion of segment length with a particular type of lighting. The following categories were observed: no lighting, median side lighting, right-side lighting, both-side lighting, point lighting, and tunnel lighting. The median roadway lighting proportion had a mean of 0.12 and a standard deviation of 0.287, while the right-side roadway lighting had a mean of 0.097 and a standard deviation of 0.258. In addition to these popular types, both-side roadway lighting had a mean of 0.005 and a standard deviation of 0.061. Point roadway lighting had a mean of 0.011 and a standard deviation of 0.039. No roadway lighting had a mean of 0.78 and a standard deviation of 0.0359. Segments with less than three-lane cross sections were few; two-lane length proportions had a mean of 0.52 and a standard deviation of 0.494. Three-lane segments had a proportion mean of 0.34 and a standard deviation of 0.462, while four-lane and five-or-more-lane segments had proportion means of 0.13 and 0.002, respectively.

Shoulder widths were measured on the Interstate network in increments of a foot, on the basis of available data from Washington DOT. An assessment of the shoulder width densities indicated that point densities were significant for the following groups: ≤2 ft, 3 to 4 ft, 5 to 9 ft, 10 ft, and >10 ft. From a modeling standpoint, the lack of sufficient density at the 1-ft increment for shoulder widths created estimation problems. By grouping shoulder widths in accordance with the clustering of densities, as mentioned above, a stable, estimable specification was feasible. All shoulder width proportions ranged from 0 to 1. The 2-ft, left-shoulder-width segment proportion had a mean of 0.17; the 3- to 4-ft, left-shoulder-width proportion had a mean of 0.199; the 5- to 9-ft, left-shoulder-width proportion had a mean of 0.132; the 10-ft, left-shoulder-width proportion had a mean of 0.472; and the more-than-10-ft, left-shoulder-width proportion had a mean of 0.019.

The 2-ft, right-shoulder-width segment proportion had a mean of 0.179; the 3- to 4-ft, right-shoulder-width proportion had a mean of 0.2065; the 5- to 9-ft, right-shoulder-width proportion had a mean of 0.119; the 10-ft, left-shoulder-width proportion had a mean of 0.463; and the more than 10-ft, right-shoulder-width proportion had a mean of 0.029. The number of horizontal curves in a segment had a mean of 1.805, with a maximum frequency of 37. (Incidentally, 1.45% of all segments had more than 10 horizontal curves.) Shortest horizontal curve-in-segment length in miles had a mean of 0.182 and a maximum value of 1.22, while the longest horizontal curve-in-segment length had a mean of 0.275 and a maximum value of 2.40. The shortest horizontal curve-in-segment length was defined as the length of the shortest horizontal curve in the case of more than one horizontal curve in a segment. Likewise, the longest curve-in-segment length was the length of the longest horizontal curve in the segment. If the segment had exactly one horizontal curve, the shortest and longest values would be identical. The smallest horizontal-curve-in-segment central angle had a mean of 12.18° and a maximum of 98.43°, while the largest horizontal-curve-in-segment central angle had a maximum of 111.294° and a mean of 21.062°.

The number of vertical curves in a segment had a maximum frequency of 30 and a mean of 3.12; and 5.22% of all segments had more than 10 vertical curves. The smallest vertical curve-in-segment gradient had a maximum of 7.43%, with a mean of 1.231%, while the largest vertical curve-in-segment gradient had a maximum of 10% and a mean of 2.777%. The upshot of using variables in a random parameter model was that, given the variable types (e.g., shortest and longest horizontal curves, smallest and largest central angles, smallest and largest vertical curve gradients) it was expected that these variables would capture segment-specific heterogeneity with greater resolution than a conventional count-of-curves variable. Secondly, any heterogeneity not captured by these variables would be captured by the segment-specific variations in the marginal effects through the random parameter specification. In this sense, the marginal effects represented by the conditional means of the geometric variables were potentially closer to the true effects.

MODELING RESULTS

Table 2 provides the modeling results of the NB model with correlated random parameters. The off-diagonal matrix elements from the Cholesky decomposition were not provided because of space constraints; all off-diagonal elements, however, were determined to be significant. The ADT-minimum gradient parameter correlation and point-lighting-proportion-minimum gradient parameter correlation

TABLE 1 Summary Statistics of Crash, Geometric, and Traffic Variables for Interstate Segments in Washington State

Variable	Mean	SD	Min.	Max.
Number of Crashes per Year	10.84	18.917	0	388
Geometric Characteristics				
Portion of segment with median roadway lighting	0.12	0.287	0	1
Portion of segment with right-side roadway lighting	0.097	0.258	0	1
Portion of segment with both-side roadway lighting	0.005	0.061	0	1
Portion of segment with point roadway lighting	0.011	0.039	0	0.75
Portion of segment with no roadway lighting	0.78	0.359	0	1
Portion of segment with two lanes	0.52	0.494	0	1
Portion of segment with three lanes	0.34	0.462	0	1
Portion of segment with four lanes	0.13	0.331	0	1
Portion of segment with five or more lanes	0.002	0.043	0	1
2-ft left-shoulder-width segment proportion	0.17	0.336	0	1
3- to 4-ft left-shoulder-width proportion	0.199	0.484	0	1
5- to 9-ft left-shoulder-width proportion	0.1321	0.597	0	1
10-ft left-shoulder-width proportion	0.472	0.478	0	1
>10-ft left-shoulder-width proportion	0.0187	0.262	0	1
2-ft right-shoulder-width segment proportion	0.179	0.341	0	1
3- to 4-ft right-shoulder-width proportion	0.2065	0.489	0	1
5- to 9-ft right-shoulder-width proportion	0.119	0.560	0	1
10-ft right-shoulder-width proportion	0.463	0.477	0	1
10-ft right-shoulder-width proportion	0.02855	0.434	0	1
Number of horizontal curves in segment[a]	1.805	2.458	0	37
Shortest horizontal curve-in-segment length (miles)	0.182	0.193	0	1.219
Longest horizontal curve-in-segment length (miles)	0.275	0.277	0	2.402
Smallest horizontal curve-in-segment central angle (degrees)	12.176	14.690	0	98.426
Largest horizontal curve-in-segment central angle (degrees)	21.062	21.001	0	111.294
Number of vertical curves[b]	3.12	3.210	0	30
Smallest absolute vertical curve gradient (%)	1.23	1.385	0	7.43
Largest absolute vertical curve gradient (%)	2.78	2.075	0	10.00
Other Segment Characteristics				
Segment length (miles)	1.326	1.753	0.1	20.380
ADT per lane[c]	6,529.78	8,391.40	458.23	22,111.89

NOTE: Crash characteristics are based on 9-year panel from 1999 to 2007; SD = standard deviation; min. = minimum; max. = maximum.
[a]Represents a curve presence count.
[b]Includes curves not fully subsumed in segment.
[c]Weighted ADT value. If one or more values exist in a segment because of change in cross section or change in total ADT, weighted value represents composite contribution by length proportion.

were the weakest, with estimated *t*-statistics of 3.016 and 3.356, respectively. It appears that the unstructured correlation matrix for parameters yielded the best likelihood. The statistical model presented shows improvement over the baseline fixed parameter NB, with an improvement in likelihood from –29,146.43 to –26,311.44 (see Table 3). It is premature to make a general statement as to whether a correlated random parameter model will yield likelihood improvements of the type noted in this study until future studies across multiple states are able to determine trends in random parameter modeling likelihoods. Out of the 21 geometric and traffic variables used in the model, six variables had random parameters that were significant, which meant that the standard deviations of the means were significant. The six variables were logarithm of ADT, point and median lighting proportions, largest degree of horizontal curvature in segment, and smallest and largest vertical curve gradients in segment. Fixed parameter variables included logarithm of length, no lighting and right lighting proportions, all shoulder width proportions, number of horizontal curves in segment, number of vertical curves in segment, and shortest horizontal curve-in-segment length. A detailed discussion of the variables mentioned above in terms of associated marginal effects is provided in the following section.

INTERPRETATION OF MODEL PARAMETERS

The logarithms of length and ADT variables are positively signed, consistent with the expectation of higher crash frequencies with longer lengths and higher ADTs. In addition to their strong statistical significance, these variables retained the strongest elasticities, after accounting for panel correlation. The elasticity of length was approximately

TABLE 2 Model Results for Correlated Random Parameter NB Estimation of Total Annual Crash Frequency on Interstate Segments in Washington State

Variable	Constant Parameter		Random Parameter				
	Mean	*t*-Statistic	Distribution	Mean	*t*-Statistic	Standard Deviation	*t*-Statistic
Exposure and Context							
Constant	–7.258	–65.024	NA	NA	NA	NA	NA
Logarithm of length of segment in miles	0.709	86.513	NA	NA	NA	NA	NA
Logarithm of ADT	NA	NA	Normal	0.916	95.475	0.074	80.441
Interchange indicator (1 if segment is an interchange segment; 0 otherwise)	–0.030	–3.130	NA	NA	NA	NA	NA
Urban segment indicator (1 if segment is in an urban location; 0 otherwise)	0.107	8.491	NA	NA	NA	NA	NA
Lighting Type by Length Proportion							
Point lighting segment proportion	NA	NA	Normal	0.582	4.144	2.873	19.805
Median continuous segment proportion	NA	NA	Normal	0.989	17.785	0.767	53.285
No lighting segment proportion	0.753	13.378	NA	NA	NA	NA	NA
Right lighting segment proportion	0.874	15.867	NA	NA	NA	NA	NA
Number of Lanes by Length Proportion							
Three-lane cross-section segment proportion	0.447	29.515	NA	NA	NA	NA	NA
Four-lane cross-section segment proportion	0.998	59.258	NA	NA	NA	NA	NA
Five-lane cross-section segment proportion	0.770	8.649	NA	NA	NA	NA	NA
Left Shoulder Width by Length Proportion							
3- to 4-ft left-shoulder-width proportion	–0.284	–13.315	NA	NA	NA	NA	NA
5- to 9-ft left-shoulder-width proportion	–0.429	–24.616	NA	NA	NA	NA	NA
10-ft left-shoulder-width proportion	–0.530	–35.212	NA	NA	NA	NA	NA
Right Shoulder Width by Length Proportion							
3- to 4-ft right-shoulder-width proportion	–0.201	–9.293	NA	NA	NA	NA	NA
5- to 9-ft right-shoulder-width proportion	–0.377	–21.598	NA	NA	NA	NA	NA
10-ft right-shoulder-width proportion	–0.408	–27.445	NA	NA	NA	NA	NA
Horizontal Curvature							
Number of horizontal curves in segment	0.043	18.316	NA	NA	NA	NA	NA
Shortest horizontal curve-in-segment length in miles	–0.179	–5.674	NA	NA	NA	NA	NA
Largest degree of curvature in segment	NA	NA	Normal	0.032	12.862	0.110	42.504
Vertical Curvature							
Smallest vertical curve gradient in segment	NA	NA	Normal	0.011	2.364	0.140	38.098
Largest vertical curve gradient in segment	NA	NA	Normal	0.010	3.165	0.014	12.154
Number of vertical curves	0.016	6.337	NA	NA	NA	NA	NA
Scale parameter for overdispersion	18.085	33.177	NA	NA	NA	NA	NA

NOTE: Number of observations = 10,377. NA = not applicable. Parameter values are fixed across observations for variables with constant parameters or defined by distributions for variables with random parameters.

TABLE 3 Comparison of Model Fit Statistics

	Constant Parameter NB	Random Parameter NB
Log likelihood	–29,146.43	26,311.44
Akaike information criterion	5.622	5.080
Bayesian information criterion	5.640	5.112
Hannan-Quinn information criterion	5.630	5.091

0.709, which indicated that a 1% increase in length contributed to a 0.709% increase in crash occurrence. By comparison, ADT had an elasticity of approximately 0.916 at the conditional mean, with the consideration that it was a random parameter. On the basis of the conditional mean, a 1% increase in ADT should contribute to a 0.916% increase in total annual crash frequency on Washington State Interstates. Because the ADT parameter was random, its marginal effect could vary from segment to segment. The construct of the correlated model ensured that up to 1,153 marginal effects could occur. The interchange effect was to result in a roughly 3% decrease in expected crash frequency. Although this finding may appear counterproductive, once the urban rural effect is controlled for, the offsetting magnitude

is reasonable. An urban segment, for example, was expected to experience a roughly 10.2% increase in crash frequency compared with rural segments. Overall, the net effect at urban interchanges might be a marginal increase in crash frequencies as a result.

Lighting proportion effects on crash occurrence need to be interpreted by lighting type proportion by length. Lighting type proportion by length varied from 0 to 1 for all segments, and 1 represented 100% occupancy of the segment of a particular lighting type. The point lighting proportion variable had a mean of 0.582 and a standard deviation of 2.873. This outcome suggests that, in 41.97% of the segments, an increase in point lighting proportion by length would be associated with fewer crash frequencies, whereas in 58.03% of segments, it would tend to be associated with higher crash frequencies. Median continuous lighting proportion appeared to have a larger counterproductive effect on crash occurrence; nearly 90.12% of segments were expected to have higher crash frequencies with an increase in median lighting proportion, whereas 9.88% of segments were expected to have fewer crashes with an increase in median lighting proportion by length. Right-side lighting and no lighting were also expected to be associated with higher crash occurrences because of the increase in proportion by length. A 1% increase, for example, in right-side lighting proportion was associated with a 0.08% increase in crash frequencies, whereas an increase in the no lighting proportion had a much higher elasticity of 0.59%.

The number of lanes proportion variable was to be interpreted with the two-lane cross section as the baseline. A two-lane by direction cross section typically was observed in rural areas. By comparison with this baseline, segments with three-lane directional cross sections were expected to have 56% higher crash frequencies; four-lane cross sections were expected to have roughly 171% higher crash frequencies, and five-lane sections were expected to have approximately 161% higher crash frequencies on an annual basis, when all other factors were controlled for. This was not inconsistent with expectation, because an increase in cross section is associated with an increase in volume as well as an increase in potential congestion-related effects. Of note was that, after controlling for both fixed and random parameter effects, the marginal impacts of four- and five-lane directional cross sections were fairly close, which indicated a potential plateau effect on traffic safety. Shoulder width effects on crash frequency were to be viewed with 2-ft shoulder widths as the baseline. In this respect, the marginal impact of 3- to 4-ft, left shoulder widths resulted in a 25% decrease in annual crash frequency, whereas 5- to 9-ft and 10-ft left shoulder widths were expected to result in a 35% to 41% decrease with respect to 2-ft, left shoulder widths. More than 10-ft, left shoulder widths were practically nonexistent. Right shoulder widths appeared to have a modest effect, compared with left shoulder width impacts in this regard. With respect to 2-ft, right shoulder widths as the baseline, the marginal impact of 3- to 4-ft right shoulder widths was to result in a 18% decrease in annual crash frequency, whereas 5-ft to 9-ft, and 10-ft, right shoulder widths were expected to result in a 31% to 34% decrease with respect to 2-ft, right shoulder widths. More than 10-ft, right-shoulder-width proportions were insignificant.

With respect to curvature variables, the discussion of fixed parameter effects is presented first. The shortest horizontal curve-in-segment variable, while significant, was expected to result in a 16% decrease for every 1 mi increase in length. Although the 1-mi increase in horizontal curve length may seem unrealistic, it provides a sense of the sensitivity of crash occurrence to a unit increase in the shortest curve in the segment. From a design standpoint, this appears to be a variable of some utility, when engineers attempt to control for factors that predispose a segment to an increase in crash occurrence. Variables that represented the number of horizontal and vertical curves were also fixed parameter effects. The marginal increase in crash occurrence as a result of the addition of a horizontal curve was expected to be about 4.44%. Vertical curves, although statistically significant, had weaker effects on crash occurrence. The increase in crash frequency with the addition of a vertical curve was expected to be about 1.57%.

Random parameter effects from curvature variables are now discussed. The largest degree of horizontal curve in the segment variable was associated with an increase in crash frequency in 61.55% of the segments, whereas in 38.45% of the segments, it was expected to be associated with a decrease. At the conditional mean, the average, marginal effect was to produce a 3.29% increase in expected crash frequency overall, when the largest degree of curvature was increased by 1°. The minimum, vertical curve gradient in a segment was associated with an expected increase in crash frequency in 52.85% of segments. In 47.15% of the segments, it was expected to result in a decrease in crash frequency on account of a gradient increase. At the conditional mean, on average, the effect of the minimum, vertical curve gradient was to result in an expected 1.07% increase in overall crash frequency. By comparison, maximum, vertical curve gradient effects were almost uniformly counterproductive on crash frequency. In 76.25% of segments, an increase in the maximum, vertical curve gradient (absolute value) was expected to result in an increase in annual crash frequency, with the conditional mean effect to be of 1% magnitude.

CONCLUSIONS AND RECOMMENDATIONS

The developed model is a prototype tool that can be applied to recommend measures for safety improvements on Interstates. Because this model provides detailed information on what factors contribute to crash occurrence, it can be highly useful in project prioritization by isolating out the sites that are most vulnerable and in need of immediate safety upgrades. Currently, the NB method is used by highway agencies, such as the Washington DOT, to predict crash occurrence and to determine location prioritization schemes. On the basis of these data, the sites were marked as high crash and low crash corridors. The biggest limitation, however, in making such predictions was the degree of uncertainty and randomness associated with crash occurrences and the decision-making process to classify locations. The NB process with random parameters can account for this uncertainty and randomness and thus allow for increases in the credibility of crash predictions and the identification of high crash locations. The proposed model, although exploratory, appears to have promise in its capacity to capture segment-specific heterogeneity. Heterogeneity is the key that underpins some parameters as random while others are plausibly fixed. It appears that lighting, ADT, and curvature variables, which are all important policy variables, have heterogeneity effects, because policy applications are heterogeneous especially through lighting and curvature variables.

The findings from the NB models indicate that lighting variables (e.g., point lighting, no lighting, and continuous lighting) have a counterproductive effect with respect to both side lighting as the baseline. This is a trade-off research question that merits a detailed evaluation in terms of energy and safety costs. It is estimated that crash occurrences can be expected to be higher when the both-side lighting proportion is not 100% by length. In the absence of such lighting, to account for societal costs through severity estimations can shed more light on the cost–benefit burden at key locations in terms of lighting policy.

Cross-sectional capacity effects appear to taper off in terms of safety impacts as lane provision increases from three to four directional lanes or more. This raises another trade-off question to be asked in further studies: the cost–benefit burden of providing for added

capacity as opposed to safety disbenefits. The impact of shoulder width variables appears to shed the most light in terms of low-cost improvements. The 5- to 9-ft shoulder width effect is close in magnitude to the 10-ft shoulder width effect, which suggests that shoulder widths can have an optimal safety effect in terms of cost–benefit in the 5- to 9-ft width range. This has implications in terms of right-of-way acquisitions and resulting trade-offs in terms of low-cost improvements and associated capacity opportunities. The trade, for example, of a 6-ft shoulder width for a 10-ft shoulder width to accommodate a high occupancy lane without additional right-of-way acquisition is a sample question that the proposed model can enable decision makers to address.

The proposed model also provides for decision insights into safety effects of particular geometries in segments for which new construction may be a restrictive option. The finding, for example, on the random parameter curvature variables indicates that it is beneficial to focus on the shortest horizontal curve in segment, while emphasis is placed on the minimization of the largest degree of curvature in the segment. Likewise, the finding from the vertical gradient variables indicates that policy caps on maximum gradients can be evaluated from a safety perspective as well. In concert with capacity evaluations on gradient caps, insights from the proposed model can complement policy guidelines for multidimensional design policy.

Although the above-discussed conclusions show the promise of the proposed model, some caveats remain. This research study provides a means to model crash frequencies on the basis of data collected from seven major Interstates in Washington State. Information on weather and human factors, however, was not included in this model. Future research is recommended to include these dimensions into this model and test its robustness under the multiple criteria. Specifically, the weather information along the road (e.g., poor visibility, rain, snow) and human factors data (e.g., factors that range from road design and vehicle use to emotional and motivational determinants of behavior) for each segment should be used to further improve the model. This additional amount of detail will come at a cost, both in terms of information collection and model estimation capabilities, because additional resolution in heterogeneity in data will also make convergence a big issue in the model estimation. Electronic collation of the information and its aggregation to various segment levels are equally burdensome. As agencies move toward the era of complete electronic management of network attributes, the proposed model structure will start to gather appeal, mainly as a result of its flexibility and ability to capture micro-scale variations in the interaction between the physical attributes of the network and the dynamic attributes such as ADT. It is recommended that further research explore the interaction between dynamics, such as weather and human behavior, and the physical attributes, such as roadway geometrics.

ACKNOWLEDGMENTS

The authors thank the University of Iceland Research Fund for doctoral studies for financial support, the Washington Department of Transportation for its support of this research, and Vikas Sharma of Advantec Consulting Engineers for his contributions to the development of the study database. The authors also thank the anonymous referees for useful comments that greatly improved this paper.

REFERENCES

1. Engel, J. Models for Response Data Showing Extra-Poisson Variation. *Statistica Neerlandica,* Vol. 38, No. 3, 1984, pp. 159–167.
2. Lawless, J. Negative Binomial and Mixed Poisson Regression. *Canadian Journal of Statistics,* Vol. 15, No. 3, 1987, pp. 209–225.
3. Miaou, S. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis and Prevention,* Vol. 26, No. 4, 1994, pp. 471–482.
4. Maher, M. A New Bivariate Negative Binomial Model for Accident Frequencies. *Traffic Engineering and Control,* Vol. 32, No. 9, 1991, pp. 422–433.
5. Shankar, V., J. Milton, and F. Mannering. Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. *Accident Analysis and Prevention,* Vol. 29, No. 6, 1997, pp. 829–837.
6. Lord, D., S. Washington, and J. Ivan. Further Notes on the Application of Zero Inflated Models in Highway Safety. *Accident Analysis and Prevention,* Vol. 39, No. 1, 2007, pp. 53–57.
7. Ma, J., K. Kockelman, and P. Damien. Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity Using Bayesian Methods. *Accident Analysis and Prevention,* Vol. 40, No. 3, 2008, pp. 964–975.
8. El-Basyouny, K., and T. Sayed. Accident Prediction Models with Random Corridor Parameters. *Accident Analysis and Prevention,* Vol. 41, No. 5, 2009, pp. 1118–1123.
9. Anastasopoulos, P., and F. Mannering. A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention,* Vol. 41, No. 1, 2009, pp. 153–159.
10. Malyshkina, N., F. Mannering, and A. Tarko. Markov Switching Negative Binomial Models: An Application to Vehicle Accident Frequencies. *Accident Analysis and Prevention,* Vol. 41, No. 2, 2009, pp. 217–226.
11. Lord, D., and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. In *Transportation Research Part A,* Vol. 44, No. 5, 2010, pp. 291–305.
12. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention,* Vol. 27, No. 3, 1995, pp. 542–555.
13. Poch, M., and F. Mannering. Negative Binomial Analysis of Intersection-Accident Frequencies. *ASCE Journal of Transportation Engineering,* Volume 122, No. 2, 1996, pp. 105–113.
14. Kulmala, R. *Safety at Rural Three- and Four-Arm Junctions: Development and Application of Accident Prediction Models.* VTT Publication 233, Technical Research Center of Finland, Espoo, Finland, 1995.
15. Shankar, V. N., R. B. Albin, J. C. Milton, and F. L. Mannering. Evaluating Median Cross-Over Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using Random Effects Negative Binomial Model. In *Transportation Research Record 1635,* TRB, National Research Council, Washington D.C., 1998, pp. 44–48.
16. Hausman, J., B. Hall, and Z. Griliches. Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica,* Vol. 52, No. 4, 1984, pp. 909–938.
17. Chin, H., and M. Quddus. Applying the Random Effect Negative Binomial Model to Examine Traffic Accident Occurrence at Signalized Intersections. *Accident Analysis and Prevention,* Vol. 35, No. 2, 2003, pp. 253–259.
18. Milton, J., V. Shankar, and F. Mannering. Highway Accident Severities and the Mixed Logit Model: An Exploratory Empirical Analysis. *Accident Analysis and Prevention,* Vol. 40, No. 1, 2008, pp. 260–266.
19. Washington, S., M. Karlaftis, and F. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis.* Chapman Hall/CRC, Boca Raton, Fla., 2010.
20. Ulfarsson, G. F., and V. N. Shankar. Accident Count Model Based on Multiyear Cross-Sectional Roadway Data with Serial Correlation. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington D.C., 2003, pp. 193–197.
21. Sittikariya, S., and V. Shankar. *Modeling Heterogeneity: Traffic Accidents.* VDM-Verlag, Saarbrücken, Germany, 2009.
22. Shankar, V. N., S. Chayanan, S. Sittikariya, M.-B. Shyu, N. K. Juvva, and J. C. Milton. Marginal Impacts of Design, Traffic, Weather and Related Interactions on Roadside Crashes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1897,* Transportation Research Board of the National Academies, Washington D.C., 2004, pp. 156–163.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.