Relationship Between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach

SHAW-PIN MIAOU, PATRICIA S. HU, TOMMY WRIGHT, AJAY K. RATHI, AND STACY C. DAVIS

A Poisson regression model is proposed to establish empirical relationships between truck accidents and key highway geometric design variables. For a particular road section, the number of trucks involved in accidents over 1 year was assumed to be Poisson-distributed. The Poisson rate was related to the road section's geometric, traffic, and other explanatory variables (or covariates) by a loglinear function, which ensures that the rate is always nonnegative. The primary data source used was the Highway Safety Information System (HSIS), administered by FHWA. Highway geometric and traffic data for rural Interstate highways and the associated truck accidents in one HSIS state from 1985 to 1987 were used to illustrate the proposed model. The maximum likelihood method was used to estimate the model coefficients. The final model suggested that annual average daily traffic per lane, horizontal curvature, and vertical grade were significantly correlated with truck accident involvement rate but that shoulder width had comparably less correlation. Goodness-of-fit test statistics indicated that extra variation (or overdispersion) existed in the developed Poisson model, which was most likely due to the uncertainties in truck exposure data and omitted variables in the model. This suggests that better quality in truck exposure data and additional covariates could probably improve the current model. Subsequent analyses suggested, however, that this overdispersion did not change the conclusions about the relationships between truck accidents and the examined geometric and traffic variables.

The passage of the 1982 Surface Transportation Assistance Act preempted more-restrictive state vehicle size and weight limits and has allowed longer and wider trucks to travel on a designated national highway network (*1*). Furthermore, states are expected to provide reasonable access beyond the national network to truck terminals and service facilities. Safety questions, such as whether the current highway design is adequate to serve these larger trucks and which highway geometric conditions pose the most serious safety problems for large trucks, are of primary concern. These questions can be better addressed if truck accident involvement rates, defined as the number of trucks involved in highway accidents per truck miles traveled, can be accurately estimated for different truck types under different highway geometric conditions.

S.-P. Miaou, P. S. Hu, A. K. Rathi, and S. C. Davis, Energy Division, Center for Transportation Analysis, Oak Ridge National Laboratory, P.O. Box 2008, MS 6366, Oak Ridge, Tenn. 37831. T. Wright, Mathematical Sciences Section, Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, P.O. Box 2008, MS 6367, Oak Ridge, Tenn. 37831.

The occurrences of vehicle accidents have long been recognized as complex events involving the interactions of many factors. Previous attempts to establish relationships between vehicle accidents and highway geometric design variables have had mixed results, and no specific relationships are widely accepted. In addition, the relationships have typically been studied through conventional linear regression models. These models are, however, known to have several undesirable statistical properties in describing discrete random events such as vehicle accidents. (A brief review of relevant studies is given later.) The objective of this paper is to present a potential statistical framework for establishing relationships between truck accidents and key highway geometric design variables. The types of trucks of interest are large trucks with gross vehicle weight ratings of 10,000 lb or more. The specific truck safety questions that this proposed model framework is intended to address include:

1. Given a section of highway, how safe is it for large trucks in terms of accident involvement rate and accident probability?
2. Given a set of highway geometric design elements, which elements are relatively more critical to the safety performance of large trucks?
3. What reduction in large-truck accident involvement rates can be expected from various improvements in highway geometric design?

In this study, accident probability refers to the probability of observing y vehicles involved in accidents during a period of time, where $y = 0, 1, 2, 3, \ldots$.

This paper is organized as follows: statistical models used in recent studies to establish relationships between vehicle accidents and highway geometric design are briefly reviewed; a statistical framework for establishing such relationships is presented; and the data used in this study and the associated statistics are described. Then the results are summarized and directions for future work are suggested.

LITERATURE REVIEW

The empirical relationships between vehicle accidents and highway geometric design variables, such as horizontal cur-

vature, vertical grade, lane width, and shoulder width, have been addressed in many studies. For example, *NCHRP Report 197* contains a summary of research performed until 1978, including a synopsis of major findings from more than 400 reports and publications (*2*). More recent studies include those by Zegeer et al. (*3*) and Joshua and Garber (*4,5*). Unfortunately, most of the studies did not distinguish accident rates between trucks and other types of vehicles. Among those studies that did focus on truck accidents, accident rates were often reported at the aggregated national or state level for a particular roadway class, not in terms of specific highway geometries (*1,6*). Joshua and Garber's study is, to the best of our knowledge, the first study that actually focused on quantifying the effect of highway geometric design variables on truck accidents statistically. At present, we are unaware of any statistical investigations that have established a relationship by both truck configuration and accident severity type. An up-to-date literature review on large-truck accidents and their relationships to various roadway, driver, vehicle, traffic, and environmental characteristics as well as safety implications of various truck configurations relevant to highway geometric design, is contained in a paper by Miaou et al. (*7*).

The need to establish relationships between truck accidents and geometric design and the frustration among researchers to find such relationships were properly described by Harwood et al. (*8*):

> The data . . . clearly illustrate the effect of two key variables related to hazardous materials routing—roadway type and area type—on truck accident rate. An attempt was made to determine the relationship between two traffic volume factors (AADT [annual average daily traffic] and percent trucks) and truck accident rate, but no consistent results were obtained. Consideration of the effects of additional geometric variables . . . on truck accident rates . . . would be desirable. . . . However, it should be recognized that the development of reliable relationships between geometric features and accidents is a difficult statistical task. Previous attempts . . . have had mixed results and no set of geometric-accident relationships is widely accepted.

Indeed, vehicle accidents are complex processes involving the interactions of many factors, including not only the road, but also the vehicle, the drivers, the traffic, and the environment (e.g., weather and lighting conditions). To establish such a relationship, the analysis requires good accident, traffic, and highway geometric data, as well as good truck travel (or exposure) information. In part, lack of accurate information on vehicle miles traveled by truck configuration and inadequate accident reporting have made the evaluation of truck safety performance on national highways nearly unattainable. A TRB Special Report was prepared specifically to address these data problems and has proposed a national monitoring system (NMS) for truck safety (*9*). Unfortunately, the data collection plan adopted in NMS did not address the highway geometric aspect of truck safety issues.

The rest of this section presents a discussion on statistical models that have been used for establishing the relationships between vehicle accidents and highway geometric design variables. We then summarize the characteristics of the problems and outline the desired capabilities of a candidate model for establishing such relationships.

Vehicle Accident Models

Multiple linear regression models have been used frequently in establishing vehicle accidents–geometric design relationships, as summarized in *NCHRP Report 197* (*2*). The regression models have the following general form:

$$\frac{y_i}{v_i} = x_i'\beta + \varepsilon_i \qquad (i = 1, 2, \ldots, n) \tag{1}$$

where

i = road section index;
y_i = number of vehicles involved in accidents during a time period;
v_i = total vehicle miles traveled;
x_i = vector of explanatory variables (or covariates) associated with the section, such as AADT, horizontal curvature, vertical grade, and shoulder width;
β = vector of regression coefficients to be estimated; and
ε_i = zero mean model residual for road section i.

Vehicle miles traveled, v_i, is usually estimated as 365 × AADT_i × ℓ_i × (number of years under consideration), where AADT_i and ℓ_i are the AADT (in number of vehicles) and the section length (in miles) of section i, respectively. It was also suggested that the dependent variable y_i/v_i be log-transformed whenever appropriate. The ordinary least-squares (OLS) method was typically used to estimate the regression coefficients, although the weighted least-squares (WLS) method was occasionally used. A recent study that used the multiple linear regression model was reported by Zegeer et al. (*3*), in which the WLS method was used for coefficient estimation. For the 10,900 curved road sections they studied, 12, 123 vehicle accidents were reported over a 5-year period. However, those accidents occurred on only 44.3 percent of the road sections, so 55.7 percent of the road sections had no observed accidents in the 5 years.

There are several statistical properties of the conventional multiple linear regressions that are considered undesirable in establishing the relationships between vehicle accidents and highway geometric design. These undesirable properties relate mainly to the underlying distributional assumption of a conventional multiple linear regression model, some of which have been discussed by Jovanis and Chang (*10*). The following are some examples:

- For a given road section, the number of vehicles involved in accidents are random discrete events that take nonnegative integer values: 0, 1, 2, 3, . . ., each of which has some probability of being observed. The use of a continuous distribution such as normal distribution to model accident events is at best an approximation to a truly discrete process. Furthermore, the occurrences of vehicle accidents are sporadic in nature. In most studies of this kind, the analyst is faced with a problem of dealing with a large number of road sections that have no accidents during the observed period. Zegeer et al.'s study is a good example (*3*). This suggests that for several years most of the road sections considered would have a much higher probability of being observed with no accidents than with more than one accident. In other words, the underlying distri-

bution of the occurrences of vehicle accidents on most of the road sections is positively (or rightly) skewed. Normal distribution is not a good approximation under this condition.

- There are other inferential assumptions of multiple linear regression in Equation 1 that are probably too restrictive for this type of study—for example, the residuals of the model, ε_i, are assumed to be uncorrelated with the explanatory variables, x_i. Other limitations of using Equation 1 include the following: (a) it may occasionally predict negative accident involvement rates, and (b) it does not provide a clear linkage between accident involvement rate and accident probability. That is, for an estimated accident involvement rate from the regression model, it is difficult to compute the probability of observing y vehicles involved in accidents on a particular road section during a period of time.

In contrast to multiple linear regression models, the Poisson regression models are widely used for modeling accident and mortality data in epidemiology. It is only in a recent study by Joshua and Garber (*4,5*) that the model was introduced to establish the relationships between truck accidents and highway geometric design. A limitation of using the Poisson regression model, which is well-known in the statistical literature (*11,12*), is that the variance of the data is restrained to be equal to the mean. In this study, we used a Poisson regression model with a different model structure than that used by Joshua and Garber and considered the consequences of the limitation on Poisson distribution.

Although the proposed model in this paper is also in the Poisson context, it does not have two limitations that we have observed for the Joshua and Garber model. We will briefly discuss these limitations using their first model [Equation 28 in the work by Joshua and Garber (*4*)] as an example. First, one can show that the Joshua and Garber model will always give a small prediction of truck accidents for relatively leveled highway sections no matter what other variables are being included in the model. In other words, regardless of the AADT and the percentage of trucks for a given road section, as long as that road section is relatively flat, the predicted number of accidents for that section based on their model will be small. Second, it can also be shown that the Joshua and Garber model suggests that increases in AADT or length for a given road section, while holding other variables constant, will lead to a decrease in the predicted truck accident rate for that road section. This is contrary to what one would expect.

Model Capabilities

The characteristics of the problem a researcher will face in establishing empirical relationships between vehicle accidents and highway geometric design can be summarized as follows:

- Vehicle accidents are complex interactions involving many factors. Many of these variables will never be available for individual road sections. Therefore, in developing empirical models, one should recognize the fact that, no matter how many covariates one manages to include, there are always some variables that will be excluded, especially those qualitative types of variables.
- The occurrences of vehicle accidents are sporadic and discrete random events.
- Road sections differ not only in geometric features and traffic conditions but also in vehicle exposure.
- Vehicle accidents and exposure data are both subject to sampling and nonsampling errors. Not all accidents are reported, especially minor property damage accidents. Also, vehicle exposure data come primarily from FHWA's Highway Performance Monitoring System (HPMS) (*13*), which is a sampling-based system.

These characteristics suggest that a potential model for establishing such relationships for trucks should be a probabilistic model capable of

- Addressing safety questions in terms of both accident involvement rate and accident probability,
- Predicting "nonnegative" accident involvement rates,
- Taking into account the differences in truck exposure across road sections,
- Giving proper statistical weights to a great portion of road sections with no observed truck accidents,
- Providing inferential statistics that allow the evaluation of model uncertainties due to the uncertainties of truck exposure data and possible omitted variables in the model, and
- Handling different roadway classes, truck configurations, and accident severity types.

POISSON REGRESSION MODEL

Model Formulation

Consider a set of n highway sections of a particular roadway type, say, rural Interstate. Let Y_i be a random variable representing the number of trucks involved in accidents on highway section i during a period of, say, 1 year. Furthermore, assume that the amount of truck travel (or exposure) on this highway section, V_i, is also a random variable, estimated through a highway sampling system, such as HPMS. Associated with each highway section i, there is a $k \times 1$ covariate vector, denoted by $x_i = (x_{i1} = 1, x_{i2}, \ldots, x_{ik})'$, describing its geometric characteristics, traffic conditions, and other relevant attributes. Given V_i and x_i, truck accident involvements Y_i, $i = 1, 2, \ldots, n$ are postulated to be independent, and each is Poisson-distributed as

$$p(Y_i = y_i|\Lambda_i = \lambda_i, V_i = v_i, x_i) = \frac{(\lambda_i v_i)^{y_i} e^{-\lambda_i v_i}}{y_i!}$$

$$(i = 1, 2, \ldots, n; y_i = 0, 1, 2, \ldots,) \quad (2)$$

where Λ_i (>0) is the truck accident involvement rate on highway section i, and it is expected to vary from one highway section to another, depending on its covariates x_i. For each highway section i, the Poisson model implies that the conditional mean is equal to the conditional variance:

$$E(Y_i|\Lambda_i = \lambda_i, V_i = v_i, x_i) = \mathrm{Var}(Y_i|\Lambda_i = \lambda_i, V_i = v_i, x_i) = \lambda_i v_i \quad (3)$$

and is proportional to truck exposure v_i for a given truck accident involvement rate λ_i. The definition and properties of the Poisson process are well-known and will not be repeated here [see elsewhere (*14*)].

To establish a relationship between truck accident involvement rate and highway geometric and traffic variables, the following exponential form is used:

$$\Lambda_i = \exp(x_i'\beta + \varepsilon_i) = \lambda_i \exp(\varepsilon_i) \quad (4)$$

where β is a $k \times 1$ coefficient vector and ε_i is a specification error due, for instance, to omitted variables. (Note that higher-order and interaction terms of covariates can be included in Equation 4 without difficulties whenever appropriate.) This particular loglinear relationship ensures that the truck accident involvement rate is always nonnegative. Also, the specification error, ε_i, admits the fact that the functional relationship is at best an approximation to the true relationship. This type of functional relationship has been widely employed in statistical literature and found to be very flexible in fitting different types of count data (*12,14–16*).

If x_i and V_i are given with no (or negligible) uncertainties and Λ_i is assumed to be a constant (i.e., $\varepsilon_i = 0$, for all i), then Equation 2 becomes a classical Poisson regression model. The uncertainties in V_i and Λ_i introduce extra variations (or overdispersion) in the Poisson model (*12,17*). The consequences of ignoring the extra variations in the Poisson regression are that the maximum likelihood estimates (MLEs) of the regression coefficients, β, under the classical Poisson model, are still consistent; however, the variances of the estimated coefficients would tend to be underestimated. In other words, we may overstate the significance levels of the estimated coefficients (*11,18*).

Throughout this study, we used the classical Poisson regression, assuming that truck exposure V_i and covariates x_i were observed without error and that truck accident involvement rate, Λ_i, was a constant for each road section i. The potential underestimation of coefficient variance, because of overdispersion in the Poisson regression model, was corrected using an estimate of overdispersion suggested by Wedderburn (*19*) [and elsewhere (*20*)].

Model Estimation and Diagnostic Checking

Regression coefficients, β, were estimated using the MLE procedure. The detailed derivation of the MLEs and the corresponding covariance matrix is omitted from this paper but can be found in work by Miaou et al. (*7*). The MLE was obtained by maximizing the log-likelihood function, $L(\beta)$, with respect to the coefficient β using a nonlinear optimization technique called the Davidon-Fletcher-Powell algorithm (*21*). In determining whether a specific variable should be included in Equation 4, we first checked to see if the estimated coefficient of the variable had the expected sign, and then we examined whether its t-statistic was greater than 1.96 (or 1.645 for a lower α-level). In addition, we used Akaike's information criterion (AIC) for model selection. Models with smaller AIC values are preferred. Bozdogan's article is an excellent reference on the theory and application of AIC criterion (*22*).

To help assess the overall goodness-of-fit of the proposed model, we considered two statistics: Pearson's chi-square statistic (X^2) and likelihood ratio statistic (G^2) (*23*). The basic idea of both statistics is to compare the observed frequency with the expected frequency based on the model. In this particular study, the observed and the expected frequencies refer to the observed and the expected number of trucks involved in accidents [y_i and $\exp(x_i'\hat{\beta})v_i$] on each road section, respectively. However, X^2- and G^2-statistics are usually poorly approximated by chi-square distribution when a large number of y_i are zero (*23*).

To reconcile this small frequency problem, instead of comparing frequencies for each individual road section, we consider a group of road sections as a comparing unit. First, each covariate is categorized into a number of subintervals; road sections are then cross-classified according to the values of their covariates. In other words, we generate a multidimensional "contingency table" in such a way that each road section is assigned to one of the cells in the table according to the values of their covariates. For road sections that are assigned to the same cell, their observed and expected numbers of trucks involved in accidents and truck exposure are added up respectively to produce the corresponding cell values. Note that a cell with no realized truck exposure is called a structural zero and is treated as if it does not exist.

To construct the multidimensional table, the strategy used in this study to choose the cutoff points of each covariate was to avoid creating too many cells with very low estimated frequencies. The X^2- and G^2-statistics were then computed by comparing the observed and the expected frequencies in each cell of the multidimensional table. The estimated model may be inadequate when X^2 and G^2 are greater than a reference point, $\chi^2_{0.05}$ (df $= H - p$), where df is the degrees of freedom, H is the total number of cells with truck exposure greater than zero, and p is the number of coefficients considered in the model. However, it should be noted that there are many possible reasons for a model to fail the tests, including the overdispersion problem discussed earlier.

In sum, a selected model should have (a) the expected signs in all estimated coefficients, (b) low AIC value and (c) high t-statistics for model coefficients. For the model to be useful in practice, the signs of the coefficients were given the highest consideration. At the same time, covariates with signs in coefficients contrary to expectation should be checked and further investigated. If higher-order terms in the model are found to be statistically significant, they should be considered but carefully checked.

DATA AND SUMMARY STATISTICS

Data Source

Data from the Highway Safety Information System (HSIS), an accident data base developed by the Highway Safety Research Center (HSRC) of the University of North Carolina for FHWA, were employed for developing relationships between truck accidents and key highway geometric design variables. The HSIS currently contains information on five states. Of these states, one in the Midwest was considered to be the state that has the most complete information on highway geo-

metric design. In addition, this state was also the only HSIS state with a "historical" road inventory file in which year-to-year changes on highway geometric design are recorded. Thus, accidents in a given year can be matched to the road inventory information of the same period. For these reasons, road and accident data from this HSIS state were chosen for illustration in this paper.

Detailed descriptions of the HSIS data base, including data quality, are available in a guidebook prepared by the HSRC and in work by Miaou et al. (*7*). At the time of this study, the data base of this state is maintained on an annual basis from 1985 to 1987. Data are stored in six files: roadlog, horizontal curvature, vertical grade, accident, vehicle, and occupant files. Thus, these files had to be linked before any analysis could be performed. Key variables used in linking these files were the route numbers and milepoints at which accidents occurred and the route numbers and milepoints at which a road section, a curve, and a grade began and ended.

Each record on the road inventory file represented a homogeneous section in terms of its road characteristics, such as number of lanes, lane width, shoulder width, median type and width, and AADT. Thus, for example, once the lane width changes on a particular road section, this section, along with its neighboring sections, was redelineated to reflect the change. It should be noticed that each road section in the roadlog file was not necessarily homogeneous in terms of its horizontal curvature and vertical grade. On the other hand, each road section in the horizontal curvature and vertical grade files was homogeneous in terms of its horizontal curvature and vertical grade, respectively, but not necessarily in terms of other road characteristics.

Rural Interstate highways and the associated large-truck accidents from 1985 to 1987 were extracted to illustrate the proposed model. Horizontal curvature and vertical grade files were only available for 1987. Since these two highway geometric elements usually change very little over the years, we used the 1987 horizontal curvature and vertical grade data for all 3 years under consideration.

Data and Statistics

The time period considered in this study was 1 year, which means that the same road section, even if nothing had changed, was considered as three independent sections—one for each year from 1985 to 1987. Therefore, there were 1,644 road sections of rural Interstate, which constituted 8,779.21 lane-mi of roadway, during the 3-year period. Data for each year contained roughly one-third of the total sections and lane miles. The section lengths varied from 0.01 to 14.9 mi—with an average of 1.35 mi. Simple descriptive statistics of these 1,644 road sections and the associated truck accidents, truck exposure, traffic, and key highway geometric design variables are given in Table 1. The following is a detailed discussion of these data.

Accident Data

During the 3-year period, 933 large trucks were involved in accidents on the rural Interstate highways, regardless of truck configuration and accident severity type. With the total estimated to be 1,057.54 million truck-mi (MTM), the overall truck accident involvement rate was 0.88 truck involvements per MTM. Out of these 933 large trucks, the relative splits by single-unit and combination trucks were 109 and 824, respectively. These accidents occurred on only 32.2 percent of the road sections. The maximum number of trucks involved in accidents being observed on an individual road section was 10. On average, each section in the rural Interstate had 0.57 trucks involved in accidents per year.

Truck Exposure Data

For each highway section, AADT, truck percentage, and section length were available from the roadlog file. For each road section i, truck exposure was computed as v_i = 365

TABLE 1 Mean and Standard Deviations of Characteristics of 1,644 and 5,105 Road Sections

		1,644 Sections		5,105 Sections	
Accidents, Exposure, and Covariates	Measure	Mean	Standard Deviation	Mean	Standard Deviation
No. of Trucks Involved in Accidents		0.5675	1.0854	0.1832	0.5585
Truck Exposure (million truck miles)		0.6433	0.8019	0.2072	0.3514
AADT/Lane (in 1000's of vehicles)		1.8943	1.6509	1.6640	1.3774
Horizontal Curvature (degrees/100 ft arc)	CD			1.0108	2.0655
	CCR	4.5664	33.1867		
	MAC	0.4409	0.7530		
	MC	1.1484	1.9144		
Vertical Grade (percent)	VG			2.1736	1.6039
	GCR	2.7305	5.1859		
	MAG	1.9623	1.1955		
	MG	2.5262	1.5027		
Shoulder Width: Inside & Outside (ft)		13.7518	1.0930	13.7877	1.0103
Section Length (miles)		1.3489	1.5226	0.4343	0.6916

$\times$ AADT$_i$ $\times$ ($T\%_i$/100) $\times$ ℓ_i, where $T\%_i$ is the percentage of trucks (e.g., 15) and ℓ_i is the length of road section i. Note that AADT$_i$ $\times$ ($T\%_i$/100) is the "truck AADT" of road section i. Thus, truck exposure is related to truck AADT and length of the road section. As indicated in Equation 3, the proposed Poisson model has the property that the expected number of trucks involved in accidents on a road section is proportional to the truck exposure of that road section.

Traffic Variable

AADT is typically used to indicate traffic condition or congestion level of a road section. Because the number of lanes varies from one road section to another, particularly in urban areas, in this study we generalized this variable by considering AADT per lane. This traffic variable represents the average density of vehicle flow on the road in an average day. Conceptually, the higher the vehicle density, the greater the chance for a truck to be involved in a conflicting position with other vehicles when negotiating its way through the road section. For example, consider two road sections, i and j, with identical geometric design and the same truck exposure (i.e., $v_i = v_j$), but Section i is more congested than Section j: (AADT/lane)$_i$ > (AADT/lane)$_j$; one would expect to observe more trucks involved in accidents on Section i than that on Section j. (Of course, this example is valid only if all other conditions, such as environment and driver factors, are the same between the two road sections.)

Highway Geometric Design Variables

The highway geometric design variables available from the selected HSIS state included (a) lane width, (b) paved shoulder width, (c) median width and type, (d) horizontal curvature, and (e) vertical grade. Because all of the road sections were coded as having 12-ft lane width, we were unable to distinguish the effects of different lane widths on truck accident involvement rate in this study. In addition, most road sections were divided. The highway geometric design variables used in the model follow.

Shoulder Width Paved shoulder widths were recorded separately for inside (or left) and outside (or right) shoulders in a given direction. In this study, because the inside and outside shoulder widths are highly correlated, we considered total shoulder width (i.e., the sum of inside and outside shoulder widths). Furthermore, we considered 20 ft to be an "ideal" shoulder width in that it practically adds an additional lane on each side of the road. Based on this consideration, we defined a variable called "deviation from ideal shoulder width," which is the total shoulder width short of the ideal shoulder width. In other words, for a particular road section i deviation from ideal shoulder width, denoted by SWD$_i$, was defined as SWD$_i$ = max{0, 20 − SW$_i$}, where SW$_i$ is the total inside and outside shoulder width of Section i.

Horizontal Curvature and Vertical Grade Horizontal curvatures and vertical grades were coded in degrees per 100 ft arc and percent, respectively. In addition, positive values indicated "right turn" and "upgrade" whereas negative values indicated "left turn" and "downgrade." As indicated earlier, each road section in the road inventory file was relatively homogeneous in terms of general road characteristics and traffic conditions, but not necessarily in horizontal curvature and vertical grade. Therefore, each road section in the road inventory file may have contained more than one horizontal curvature or vertical grade. Two ways of resolving this problem were considered. One way was to create surrogate measures to characterize the curvature and grade conditions along the length of a road section. Another way was to disaggregate those road sections with multiple curvatures and grades into smaller subsections in such a way that each subsection contains a unique set of horizontal curvature and vertical grade. The former was considered less direct from the engineering point of view and it may be difficult for design engineers to incorporate these measures into their current practice, but the second method was considerably easier to interpret in a design context. However, it should be mentioned that because the location of an accident is often estimated and occasionally it is roughly assigned to the nearest milepost of the route on which it occurred, assigning vehicle accidents to very short road sections is more susceptible to locational error than assigning to longer road sections. In this study, we used both approaches for comparison purposes.

Three surrogate measures for horizontal curvature and three surrogate measures for vertical grade were devised to characterize the horizontal and vertical alignments of each road section. On a particular road section i with length ℓ_i (in miles), assume that along the length of the section there are K curved subsections associated with it, indexed by $k = 1, 2, \ldots, K$. Each subsection k has length $\ell_{i,k}$ and curvature $\theta_{i,k}$ (which could be zero, positive, or negative). Similarly, we assumed that there were G different vertical graded subsections associated with a road section i, and each subsection had length $\ell_{i,g}$ and grade $\omega_{i,g}$ (which again could be zero, positive, or negative), where $g = 1, 2, \ldots, G$. These surrogate measures for a section i were defined as follows:

1. Horizontal curvature change rate (CCR) and vertical grade change rate (GCR):

$$\mathrm{CCR}_i = \sum_{k=1}^{K-1} |\theta_{i,k+1} - \theta_{i,k}|$$

$$\mathrm{GCR}_i = \sum_{g=1}^{G-1} |\omega_{i,g+1} - \omega_{i,g}| \tag{5}$$

If there is only one curvature (or grade), CCR (or GCR) is defined as zero.

2. Mean absolute horizontal curvature (MAC) and mean absolute vertical grade (MAG):

$$\mathrm{MAC}_i = \left(\sum_{k=1}^{K} \ell_{i,k}|\theta_{i,k}|\right) \Big/ \ell_i$$

$$\mathrm{MAG}_i = \left(\sum_{g=1}^{G} \ell_{i,g}|\omega_{i,g}|\right) \Big/ \ell_i \tag{6}$$

3. Maximum absolute horizontal curvature (MC) and maximum absolute vertical grade (MG):

$$MC_i = \max\{|\theta_{i,1}|,|\theta_{i,2}|, \ldots, |\theta_{i,K}|\}$$

$$MG_i = \max\{|\omega_{i,1}|,|\omega_{i,2}|, \ldots, |\omega_{i,G}|\} \quad (7)$$

All of these surrogate curvature measures are in the unit of degrees per 100 ft arc, and surrogate grade measures in percent. Similar measures were used by Joshua and Garber (*4*). Note that these surrogate measures are not unique (i.e., different combinations of curves and grades can result in the same values).

The second way to resolve the multiple curvatures and grades problem was to disaggregate road sections into smaller subsections so that each section had a unique set of curvature and grade measures. On the basis of this approach, 1,644 road sections were disaggregated into 5,105 subsections. Each subsection contained unique curvature and grade information. Two additional trucks involved in accidents were included because of the difference in assigning truck accidents to road sections when accidents occurred right at the cutoff point of two neighboring sections. With these redefined subsections, truck accidents occurred on 12.8 percent of the sections. In other words, 87.2 percent of the redefined road sections had no observed truck accidents during a year. Simple statistics for the redefined road sections are also given in Table 1. In our analysis, we used the absolute value of horizontal curvature (CD) and vertical grade (VG) on each subsection as the covariates.

RESULTS

Model Estimation and Selection

The proposed Poisson regression model was applied to develop empirical relationships between truck accidents and key highway geometric design variables described in the last section. The considered covariates include the following:

- x_{i1} = 1, representing a dummy intercept;
- x_{i2} = AADT per lane (thousands of vehicles);
- x_{i3} = horizontal curvature (degrees/100-ft arc);
- x_{i4} = vertical grade (%); and
- x_{i5} = deviation from ideal shoulder width (ft).

Several models were tested using different horizontal curvature and vertical grade measures. Table 2 presents the estimated coefficients (using maximum likelihood method) and their asymptotic *t*-statistics, AIC values, and negative log-likelihood functions $[-L(\hat{\beta})]$ of the tested models. Asymp-

TABLE 2 Estimated Coefficients of Tested Poisson Regression Models and Associated Statistics

Covariates & Statistics	Measure	Model 1	Model 2	Model 3	Model 4
x_{i1} = 1 (Dummy Intercept)		-14.4889 (-53.15)	-14.6804 (-55.96)	-14.6413 (-54.38)	-14.6833 (-55.67)
x_{i2} = AADT/Lane (in 1000's of vehicles)		0.069864 (3.77)	0.037828 (2.01)	0.076083 (4.14)	0.044691 (2.41)
x_{i3} = Horizontal Curvature (degrees/100 ft arc)	CD				0.172513 (8.96)
	CCR	0.000699 (0.42)			
	MAC		0.217495 (5.07)		
	MC			0.046771 (3.59)	
x_{i4} = Vertical Grade (percent)	VG				0.162218 (7.09)
	GCR	0.001249 (0.21)			
	MAG		0.222829 (7.91)		
	MG			0.091138 (3.70)	
x_{i5} = Deviation from Ideal Shoulder Width (ft)		0.064703 (1.48)	0.020024 (0.48)	0.036062 (0.83)	0.038589 (0.93)
$-L(\hat{\beta})$		1542.78	1492.57	1521.31	2222.95
AIC Value		3095.56	2995.14	3052.62	4455.90
Predicted vs. Observed Total Truck Involvements		932.35 933.00	932.88 933.00	932.99 933.00	931.85 935.00
Number of Road Sections		1644	1644	1644	5105

Values in parentheses are t-statistics of the coefficients above.

totic t-statistics were computed without the adjustment of possible overdispersion described earlier. Models 1, 2, and 3 used surrogate measures (i.e., CCR and GCR, MAC and MAG, MC and MG) for road sections with multiple horizontal curvatures and vertical grades, and Model 4 used the data for the disaggregated subsections that had a unique set of curvature and grade measures (i.e., CD and VG). Note that in Table 2, AIC values and log-likelihood functions are not comparable between Models 1 through 3 and Model 4 because the number of road sections (or sample sizes) used in developing the models was different.

Among the set of three surrogate measures for horizontal curvature and vertical grade, mean absolute curvature and vertical grade, MAC and MAG, (Model 2) performed better than the other two measures (Models 1 and 3) in terms of their ability to explain truck accident variations, which was indicated by lower negative log-likelihood function and AIC values. In addition, the coefficients estimated for MAC and MAG were in agreement with those obtained with CD and VG (Model 4) in terms of their values and signs. This suggests that surrogate measures MAC and MAG are probably more appropriate for establishing the model when disaggregating road sections into smaller subsections is not desirable. Overall, the estimated regression coefficients of AADT per lane, horizontal curvature, and vertical grade were all found to be highly significant in terms of their asymptotic t-statistics (when compared with a 1 percent α-level). Deviation from ideal shoulder width, on the other hand, was found to be insignificant at a 5 percent α-level.

Goodness-of-Fit Test and Adjusted t-Statistic

Model 4 was selected to conduct further goodness-of-fit tests. To compute the X^2- and G^2-statistics discussed earlier, road sections were first categorized by their covariates. In other words, multidimensional contingency tables cross-classified by the covariates were constructed. However, these two statistics were quite sensitive to the way covariates were divided into subintervals. Specifically, different categorization of the covariates could result in different contingency tables that had very different X^2- and G^2-statistics. We have examined, by trial and error, several categorizations of the covariates that were identified to be statistically significant in the selected model for generating the multidimensional contingency tables. An interesting feature we have found was that the generated tables were always very unsymmetrical, that is, the observed accident involvements concentrated on a small number of cells whereas the rest of the cells had very few observations. This was because most of the road sections had similar geometric features—for example, a great percentage of road sections were straight sections (i.e., curvature = 0) and a significant proportion of road sections had 2 percent grades. After trying several possible combinations of categorizing the covariates, we selected a contingency table that appeared to be most reasonable and did not contain too many cells with estimated frequency less than 1. The selected table had 112 cells in total, in which 9 cells had no realized truck travel and were ignored.

The X^2- and G^2-statistics were 160.35 and 150.33, respectively. The model failed the chi-square test at a 5 percent α-level, that is, X^2 and G^2 are greater than $\chi^2_{0.05}(\text{df} = 103 - 5 = 98) = 122.3$. This suggests that the overdispersion problem, due most likely to the uncertainties in truck exposure data and the omitted variables in Equation 4, is quite significant and that the current model can probably be improved by including additional covariates and by improving the accuracy of truck exposure data.

To check how the overdispersion affected the conclusions reached in the last subsection, the overdispersion parameter (τ) was estimated to be $X^2/(H - p)$, as suggested in Wedderburn (*19*). Better estimates of the t-statistics were derived by dividing the t-statistics obtained from the Poisson regression model by $\tau^{1/2}$ (= 1.28) (*23*,p.457). This adjustment did reduce the significance levels of the regression coefficients, but it did not alter the conclusions about the relationships between truck accidents and the examined geometric and traffic variables.

Example

To give an example of how truck accident involvement rate and accident probability can be computed from Model 4 for a rural Interstate highway section in the selected HSIS state, let us consider a hypothetical road section with the following characteristics:

- Lane width: 12 ft
- Section length: 1 mi
- Number of lanes: 4
- AADT/lane: 3,000 vehicles per lane
- Percentage trucks: 20
- Horizontal curvature: 3 degrees per 100-ft arc
- Vertical grade: 2 percent
- Shoulder width (left + right): 14 ft

Truck exposure in a year is first computed as

$$\begin{aligned} v &= 365 \times (\text{number of lanes} \times \text{AADT per lane}) \\ &\quad \times (\text{percentage trucks}/100) \times (\text{section length}) \\ &= 365 \times (4 \times 3{,}000) \times (20/100) \times 1 \\ &= 876{,}000 \text{ truck-mi} \end{aligned}$$

Based on the estimated model, the truck accident involvement rate is estimated as

$$\begin{aligned} \lambda &= \exp\{-14.6833 + 0.044691 \times 3 + 0.172513 \times 3 \\ &\quad + 0.162218 \times 2 + 0.038589 \times (20 - 14)\} \\ &= \exp\{-13.475718\} = 1.4047 \times 10^{-6} \end{aligned}$$

The expected number of trucks involved in accidents on this road section in 1 year is estimated at $E(y) = \lambda v = 1.4047 \times 10^{-6} \times 876{,}000 \approx 1.23$ trucks. The probability of observing y trucks involved in accidents on this particular road section in 1 year is then $p(y) = [(1.23)^y \exp(-1.23)]/y!$. For example, the probability of observing two trucks involved in accidents is $p(y = 2) = [(1.23)^2 \exp(-1.23)]/2! = 0.22$.

SUMMARY AND FUTURE RESEARCH

The Poisson regression model was proposed for establishing empirical relationships between truck accidents and key highway geometric design variables. Highway geometric and traffic data for rural Interstate highways and the associated truck accidents in one of the HSIS states from 1985 to 1987 were used to illustrate the proposed model. The estimated model suggested that AADT per lane, horizontal curvature, and vertical grade are significantly correlated with truck accident involvement rate, but that shoulder width has comparably less correlation. Goodness-of-fit test statistics indicated that extra variations (or overdispersion) existed in the developed Poisson model, which was most likely due to the uncertainties in truck exposure data and omitted variables in the model. This suggests that better quality in truck exposure data and additional covariates could probably improve the current model. The effect of correcting for the overdispersion was found to lower the significance level of the estimated Poisson regression coefficients. It, however, did not change the conclusions about the relationships between truck accidents and the examined traffic and highway geometric design variables.

An immediate extension of this study would be to apply the proposed Poisson regression model to other roadway types and to consider truck accidents by truck configuration and accident severity type. It would also be of practical interest to quantify the respective contribution of truck exposure data uncertainty and omitted variables to the overall overdispersion in the model. Other discrete distributions, such as negative binomial distribution, should also be explored. Finally, it would be interesting to study the sensitivity of the developed models to the uncertainties of accident location information.

ACKNOWLEDGMENTS

This research was sponsored by the Office of Safety and Traffic Operations Research and Development, FHWA, U.S. Department of Transportation. Many helpful discussions with Harry Lum, Jerry Reagan, and Justin True at FHWA and with our colleague Edward Frome at Oak Ridge National Laboratory are gratefully acknowledged.

REFERENCES

1. *Special Report 223: Providing Access for Large Trucks*. TRB, National Research Council, Washington, D.C., 1989.
2. Roy Jorgensen Associates, Inc. *NCHRP Report 197: Cost and Safety Effectiveness of Highway Design Elements*. TRB, National Research Council, Washington, D.C., 1978.
3. C. Zegeer, R. Stewart, D. Reinfurt, F. Council, T. Neuman, E. Hamilton, T. Miller, and W. Hunter. *Cost Effective Geometric Improvements for Safety Upgrading of Horizontal Curves*. University of North Carolina, Chapel Hill; FHWA, U.S. Department of Transportation, May 1990.
4. S. C. Joshua and N. J. Garber. Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models. *Transportation Planning and Technology*, Vol. 15, 1990, pp. 41–58.
5. N. J. Garber and S. C. Joshua. *Traffic and Geometric Characteristics Affecting the Involvement of Large Trucks in Accidents: Volume II*. Virginia Transportation Research Council, Charlottesville, 1991.
6. *Special Report 211: Twin Trailer Trucks*. TRB, National Research Council, Washington, D.C., 1986.
7. S.-P. Miaou, P. S. Hu, T. Wright, S. C. Davis, and A. K. Rathi. *Development of Relationships Between Truck Accidents and Highway Geometric Design: Phase I*. Technical memorandum. Oak Ridge National Laboratory, Tenn.; FHWA, U.S. Department of Transportation, Nov. 1991.
8. D. W. Harwood, J. G. Viner, and E. R. Russell. A Truck Accident Rate Model for Hazardous Materials Routing. In *Transportation Research Record 1264*, TRB, National Research Council, Washington, D.C., 1990.
9. *Special Report 228: Data Requirements for Monitoring Truck Safety*. TRB, National Research Council, Washington, D.C., 1990.
10. P. P. Jovanis and H. L. Chang. Modeling the Relationship of Accidents to Miles Traveled. In *Transportation Research Record 1068*, TRB, National Research Council, Washington, D.C., 1986, pp. 41–51.
11. D. R. Cox. Some Remarks on Overdispersion. *Biometrika*, Vol. 70, No. 1, 1983, pp. 269–274.
12. C. Dean and J. F. Lawless. Tests for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association*, Vol. 84, No. 406, June 1989, pp. 467–472.
13. *Highway Performance Monitoring System Field Manual*. FHWA, U.S. Department of Transportation, 1987.
14. D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Chapman and Hall, London, England, 1966.
15. P. K. Andersen and R. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, Vol. 10, 1982, pp. 1100–1120.
16. E. L. Frome, D. L. Cragle, and R. W. McLain. Poisson Regression Analysis of the Mortality Among a Cohort of World War II Nuclear Industry Workers. *Radiation Research*, Vol. 123, 1990, pp. 138–152.
17. A. C. Cameron and P. K. Trivedi. Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, Vol. 1, 1986, pp. 29–53.
18. A. C. Cameron and P. K. Trivedi. Regression-Based Tests for Overdispersion in the Poisson Model. *Journal of Econometrics*, Vol. 46, 1990, pp. 347–364.
19. R. W. M. Wedderburn. Quasi-Likelihood Functions, Generalized LInear Models, and the Gauss-Newton Method. *Biometrika*, Vol. 54, 1974, pp. 439–447.
20. P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, England, 1983.
21. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, England, 1986.
22. H. Bozdogan. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*, Vol. 52, No. 3, 1987, pp. 345–370.
23. A. Agresti. *Categorical Data Analysis*. John Wiley, New York, N.Y., 1990, pp. 246–250.

The opinions expressed in this paper are solely those of the authors.

Publication of this paper sponsored by Committee on Traffic Records and Accident Analysis.