Transportation Planning and Technology, 1990, Vol. 15, pp. 41–58
Reprints available directly from the Publisher
Photocopying permitted by license only

ESTIMATING TRUCK ACCIDENT RATE AND INVOLVEMENTS USING LINEAR AND POISSON REGRESSION MODELS

SARATH C. JOSHUA

Bellomo-McGee Inc., Vienna, VA, USA

and

NICHOLAS J. GARBER

Civil Engineering Department,
University of Virginia, Charlottesville, VA, 22901, USA

(Received September 26, 1989)

During the past few years, vehicle miles of travel (VMT) for large trucks having six or more wheels in contact with the road and having a gross weight greater than 10,000 lbs, have been steadily increasing. This has resulted in an increased interaction between large trucks and other vehicles which is being manifested in an increasing rate of fatal accidents involving large trucks. For example, between 1982 and 1984, fatal accident rates for 100 million VMT for large trucks in Virginia, increased by 54% while the rate for other vehicles (passenger cars, vans and pickups) remained approximately constant at less than a 0.3% increase. In order to arrest this increasing trend in large truck fatal accident rates, it is necessary that appropriate counter measures be developed. For this to be done, however, it is necessary to identify those factors that are associated with large truck accidents. A recent analysis of large truck accident data in Virginia, indicated that certain traffic and highway geometric characteristics may be associated with the occurrence of large truck accidents. In order to identify these specific characteristics and determine the extent to which they influence accident rates, a study was conducted with the primary objective of establishing relationships between large truck accidents, and traffic and roadway geometric variables. This paper presents mathematical relationships obtained through multiple linear and Poisson Regression Analyses, relating the number of truck involved accidents per year at a section of highway with traffic and geometric variables. These models indicate that the slope change rate (absolute curve of slope changes in the vertical direction divided by the highway segment), the average daily traffic, the percent of trucks and the difference in speed between trucks and non-trucks influence the number of truck involved accidents at a given stretch of highway.

KEY WORDS: Truck accidents, highway geometry, traffic characteristics.

INTRODUCTION

Although the overall accident involvement rate for large trucks (trucks having six or more wheels in contact with the roadway and having a gross weight greater than 10,000 lb) has been less than that for other vehicles (passenger cars, vans and pickups), the large truck fatal accident rate has been significantly higher.[1] Recent trends also indicate that the number of vehicle miles travelled (VMT) by large trucks is increasing at a higher rate than that for other vehicles, which most likely will result in an increase in the annual number of fatal accidents involving large trucks. This problem is made worse by the fact that average truck sizes and weights are also increasing as wider and larger trucks are now permitted on the interstate and designated primary highways, due to the passage of the Surface Transportation Act (STAA) of 1982. It is reasonable to assume that as truck sizes increase and roadway geometrics remain constant, there will be an increasing likelihood of incompatibilities between the two, which will result in increasing accident rates for large trucks.

In order to arrest this trend in large truck accident rates, it is essential that appropriate counter measures be developed. Unfortunately, this cannot be done without sufficient knowledge of those factors that significantly affect accident rates of large trucks and to what extent each of these factors contribute to large truck accidents.

A recent analysis of large truck accident rates in Virginia indicated that certain traffic and highway geometric characteristics may be associated with large truck accidents.[1] There is however limited knowledge on the quantitative relationship between the probability of occurrence of a large truck and these traffic and geometric variables. A study was therefore carried out with the overall objective of identifying those traffic and geometric variables that are significantly associated to large truck accidents. All large trucks were considered in one group and not categorized into single units, tractor-trailers or twin trailers, as preliminary analysis did not indicate any significant difference in the causal factors among the different types of trucks.

In this analysis, mathematical relationships were developed for estimating the expected number of accidents for a given set of significant traffic and geometric variables. The Akaike's Information criterion was used to select the combination of variables that will best predict the expected number of truck involved accidents.

METHODOLOGY

Data Source

The Virginia Department of Transportation (VDOT) computerized data files for all accidents on Virginia highways between 1984 and 1986 were used as the source for the accident and traffic data used in the study. Data on geometric variables were however obtained directly at the study sites during the study.

Accident data base. Subfiles on large truck involved accidents were created from the VDOT main files. These subfiles contained information on the location of each accident in terms of city/county, route numbers and section numbers. These were used to actually identify the exact stretch of highway within which the accident occurred. Other information stored in these subfiles included the type of collision, severity of the accident, the major associated causal factor, the highway type and the average daily traffic (ADT).

Sites selection. The sites selected for in depth study were identified from two sources. The first source was the subfiles created from the VDOT files and the second was the Virginia Department of State Police. By systematically sorting the accident data in the subfiles by location, sites at which a large number of truck related accidents occurred were identified as possible candidate test sites. The Department of State Police was also asked to provide a list of sites which are prone to large truck accidents. The selection of sites with high truck accidents on the study sites was based on the premise that these sites will exhibit traffic and geometric characteristics that are associated with large truck accidents. One final set of study sites was then selected using the following criteria:

- Length of highway segment restricted to a maximum of two miles.
- Majority of accidents should be of the mobility type.
- A consistent pattern of accident occurrence should exist.
- Feasibility of Data Collection.

The fourth criterion was necessary, as sometimes temporary maintenance or rehabilitation at a site may result in certain traffic characteristics, such as speed, which are significantly different from those at normal times. A minimum of 15 sites was then selected from each of these following highway categories:

- Interstate Highways,
- Four or More Lanes Divided Primary Highways.
- Four or More Lanes Undivided Primary Highways.
- Two-lane Primary Highways.

In addition, control sites with similar highway configurations were identified either upstream or downstream of each study site and included in the investigation.

Traffic and geomtric data. At each selected site, actual field data were collected on the following:

- Twenty-four Hour Vehicular Volume.
- Twenty-four Hour Vehicular Classification.
- Twenty-four Hour Vehicular Speeds.
- Speed Samples for Trucks and Non-trucks.
- Speed Limit or Advisory Maximum Speed.
- Highway Type, Number of Lanes, Lane and Shoulder Widths.
- Vertical and Horizontal Alignments.

Streeter Amet traffic counters were used to obtain traffic data such as speeds and vehicular volumes. Pneumatic tubes were used to collect volume data when average daily traffic (ADT) was less than 15,000, but for volumes higher than 15,000, the nearest permanent induction loops were used, as pneumatic tubes are damaged frequently when ADTs are greater than 15,000.

The individual speeds of large trucks and other vehicles were obtained by using a type of radar speed detector that emitted radar signals in the form of pulses rather than continuously. This considerably reduced the chances of detection by radar detectors and facilitated the discrete collection of speed data at each site. All of the traffic data were collected on work days, excluding Mondays and Fridays, so as to eliminate the influence of weekend traffic.

Geometric data such as number of lanes and shoulder widths were obtained by direct measurements at each site. The data on other geometric characteristics such as radii of horizontal curves and grade lengths and percent (slope) were obtained by using the Slopemaster Ball Bank Indicator Data Acquisition Unit (DAU). The length and percent of each grade were obtained directly as output from the DAU, but it was not possible to obtain radii of horizontal curves directly. A special procedure for obtaining this information was developed using the DAU. This technique used two angular readings of the ball bank indicator, corresponding to two different speeds of the test vehicle at each horizontal curve. The radius of each curve was then determined from a model which was developed for use in the procedure. A detailed description of the model is given in Reference 1.

Surrogates of Roadway Alignment

The sites considered in this study were not spot locations but segments of highways of one to two miles long. It was therefore necessary to use variables that describe the geometric characteristics along the length of each segment. Three surrogate measures of horizontal and vertical alignments were therefore used in this project. The

Curvature Change Rate (CCR) is used widely in Germany to describe horizontal alignment of a roadway.[2,3] The other two, Slope Change Rate (SCR) which is analogous to CCR, and Absolute Mean Slope (AMS) are new definitions. These two measures are directly proportional to the degree of variation in the vertical alignment. The Curvature Change Rate is defined as the absolute sum of the angular changes in horizontal alignment divided by the length of the highway segment, and it is given as:

$$\mathrm{CCR} = \left[\sum_{i=1}^{n} \left|\frac{L_i}{R_i}\right| + \sum_{j=1}^{n} \left|\frac{L_s}{2R_i}\right|\right](57.3)(5280)/L \text{ deg/mile} \qquad (1)$$

where L_i = length of circular curve i at a given highway segment (see Figure 1).
L_s = length of transition curve s at a given highway segment (see Figure 1).
R_i = radius of circular curve i at a given highway segment (see Figure 1).
L = total length of highway segment.

Similarly the slope change rate is given as:

$$\mathrm{SCR} = \left[\sum_{j=1}^{k} |G_{j+1} - G_j|\right](5280)\ L \text{ percent per mile} \qquad (2)$$

where G_j = slope of jth grade, percent (see Figure 2).
L_j = length of jth grade, ft. (see Figure 2).
L = length of entire segment, ft. (see Figure 2).
$l_{j,j+1}$ = length of curve between jth and $(j + 1)$th grades, ft. (see Figure 2).

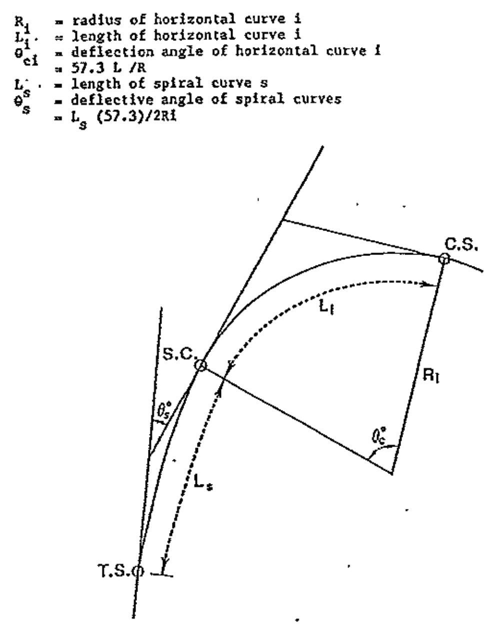

Figure 1 Components of highway horizontal alignment.

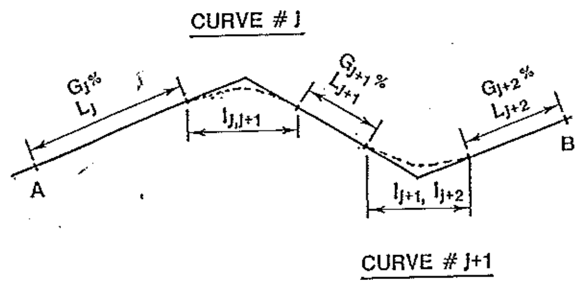

Figure 2 Components of highway vertical alignment.

The absolute mean slope is defined as the sum of the absolute grade changes in the vertical alignment divided by the length of the highway segment. it is given as:

$$\text{AMS} = \sum_{j+1}^{k} \left|\frac{G_j + G_{j+1}}{2}\right| l_{j,j+1} + \sum_{j=1}^{k} |G_j L_j| / L \text{ percent per mile} \tag{3}$$

where G_j = slope of jth grade, percent (see Figure 2).
L_j = length of jth grade, ft. (see Figure 2).
$l_{j,j+1}$ = length of curve between jth and $(j + 1)$th grades, ft. (see Figure 2).

DEVELOPMENT OF MODELS

The main objective of this task was to develop mathematical relationships that can be used to predict the expected number of large truck accidents during a year at a given segment of highway, from a given set of independent traffic and geometric variables. Multiple linear regression and Poisson regression models are two of the several types of models investigated in the study and are reported here. The independent variables considered were:

Roadway Geometry:

a) number of lanes
b) lane width (LNWD)
c) shoulder width (SHLDWD)
d) Curvature Change Rate (CCR)
e) Absolute Mean Slope (AMS)
f) Slope Change Rate (SCR)
g) segment length (SEGLEN)

Traffic Variables:

a) Average Annual Daily Traffic (AADT)
b) mean speed (all vehicles)
c) speed variance (all vehicles)
d) mean speed (trucks)
e) speed variance (trucks)

f) mean speed (non-trucks)
g) speed variance (non-trucks)
h) percent of large trucks (TPERCNT)
i) difference in mean speed between trucks and non-trucks

In developing the models, all the selected sites were grouped by roadway configurations and traffic volumes into the following environments:

Environment I	Primary Highways (undivided—four lane and two lane) AADT < 15,000
Environment II	Primary Highways (divided—four lane) AADT ≤ 15,000
Environment III	Interstate/Primary Highways (divided—four lane) AADT > 15,000

Environment I includes all primary highways with undivided carriageways carrying AADT volumes less than 15,000 and all two-lane highways. Environment II includes all four-lane primary highways with divided carriageways that carry an AADT volume less or equal to 15,000. Environment III includes all highway environments with AADT greater than 15,000. All of the highways that fall within this range of AADT have divided carriageways and are generally superior in design.

Criteria for Selecting Best Models

A major problem in statistical modeling is the identification of the best model to fit the observed data, particularly when the underlying relationship between the dependent variable and the independent variables is unknown. Several model selection procedures, such as Adjusted Multiple Correlation, Mallows C_p, prediction Sum of Squares (PRESS), etc. have been developed, but one that is becoming increasingly popular is the Akaike's Information Criteria (AIC). This procedure identifies significant independent variables, and determines the best model without the necessity to stipulate a level of significance. AIC, which was developed by Akaike[4] is an information-theoretic criterion for the identification of an optimal model. This criterion selects the optimal model taking into account the model complexity.

An objective measure of the distance between the true model and the hypothesized model is Boltzmann's generalized entropy or the negentropy. This measure is also known as the Kullback-Liebler information quantity.[5] Model selection using AIC is based on the concept of entropy maximization or in other words the minimization of the negentropy. In AIC, the Kullback-Liebler information quantity is estimated by the mean loglikelihood for the model. A complete explanation of the derivation of this criteria and its extensions is given by Bozdogan.[6] For the purpose of this analysis the criterion is described in its final form as:

If $\{M_k : k \in K\}$ is a set of competing models indexed by $k = 1, 2, \ldots, K$. Then, the criterion AIC is given by,

$$\mathrm{AIC}(k) = 2 \log L(\hat{\theta}_k) + 2k \tag{4}$$

where $\mathrm{Log}\ L\ (\theta_k) = \mathrm{Log}_e$ [maximized likelihood]
k = number of free parameters in the model

AIC is minimized to choose the model M_k over the set of competing models.

The first term in the above equation is a measure of the badness of fit or bias, when the maximum likelihood estimates of the parameters are used. The second term, $2k$, is a measure of the complexity of the model and compensates for the first term.

Using this criterion, the model yielding the minimum AIC was selected as the best model. The resulting model is the one with least complexity and highest information gained.

Multiple Linear Regression Model

This is the simplest case of generalized linear models, and can be expressed as,

$$Y = X\beta + e \tag{5}$$

where Y is an $N \times 1$ response vector
X is an $N \times p$ matrix of explanatory variables
β is a $p \times 1$ vector of parameters
e is an $N \times 1$ random vector whose elements are independent identically and normally distributed
i.e. $e_i \approx N(O, \sigma^2)$ for $i = 1, \ldots, N$

AIC derivation for multiple regression models. A least square estimation of parameters was carried out using procedures available in the statistical software package SAS. The parameters, $\beta_0, \beta_1, \ldots, \beta_p$ were estimated through the following minimization.

Minimize $S(\beta_o, \beta_1, \ldots, \beta_p) =$

$$\sum_{i=1}^{n} [y_i - E\{y_i/x_{i1}, x_{i2}, \ldots, x_{in}\}]^2 \tag{6}$$

$$= \sum_{i=1}^{n} \{y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\}^2 \tag{7}$$

In matrix form,

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \tag{8}$$

The vector of partial derivatives of S(β) with respect to β elements is,

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta = 0 \tag{9}$$

which gives

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{10}$$

Assuming that,

$$y_i \approx N(\mathbf{x}_i\beta, \sigma^2) \tag{11}$$

$$\varepsilon_i \approx N(0, \sigma^2) \tag{12}$$

for $i = 1, \ldots, n$

The likelihood function of the dependent variable observation vector Y is,

$$L(\beta, \sigma^2) = F(Y, \beta, \sigma^2) \tag{13}$$

$$= [1/(2\pi\sigma^2)^{n/2}] \exp [(-1/2\sigma^2)(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)] \tag{14}$$

the likelihood function is,

$$L(\beta, \sigma^2) = \log L\,(\beta, \sigma^2) \tag{15}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \tag{16}$$

The maximum likelihood estimates $\hat{\beta}$ of β is obtained by the following partial derivatives,

$$\frac{d}{d\beta} L(\beta, \sigma^2) = 0 \tag{17}$$

$$\frac{d}{d\sigma^2} L(\beta, \sigma^2) = 0 \tag{18}$$

Resulting in,

$$\sigma^2 = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)/n = \mathrm{ee}'/n = \mathrm{SSE}/n \tag{19}$$

Now the maximized loglikelihood becomes,

$$\log L(\hat{\theta}_k) = l(\hat{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{\mathrm{SSE}}{n} - \frac{n}{2} \tag{20}$$

But,

$$\mathrm{AIC} = -2\log(\hat{\theta}_k) + 2k \tag{21}$$

which gives

$$\mathrm{AIC} = n\log 2\pi + n\log\left\{\frac{\mathrm{SSE}}{n}\right\} + n + 2k \tag{22}$$

where $k = (p + 1) + l$
$p + 1$ = number of β coefficients estimated
l = estimated parameters, σ^2

Identification of best multiple linear regressions models. Using AIC as the criterion for model selection, the best model was the model yielding the minimum AIC value. The best subset of singleton variables was first determined by using a backward elimination procedure. The model that included all variables and its AIC at the next lower level (i.e., with the number of variables reduced by one) was then determined. At each level (i.e., number of variables) of the variable selection process, all the terms of the expression for AIC remains the same except for SSE. Therefore, the best submodel would be the model yielding the minimum SSE, which is also the model with the maximum R square at that level. Using PROC RSQUARE of SAS, the best subset of variables at each level was thus determined. The AICs for all levels are then compared with each other, and the model with the minimum AIC is then selected as the best model.

Poisson Regression Model

The occurrence of highway accidents can be described by the non-stationary Poisson Process.[7,8] According to the basic assumption of Poisson processes, we assume that the numbers of accidents occurring within each observed time interval (one year, in our model) are independent with the expectation defined as in the following equation

$$E(y_{ij}) = f(x_i, \beta) \qquad (23)$$
$$i = 1, \ldots, n$$
$$j = 1, \ldots, m_i$$

where $x_i = x_{i1}, x_{i2}, \ldots, x_{ip-1}$ is the ith set of values for the $p - 1$ independent variables.

m_i = number of replications of ith experimental condition
$\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})$ is a p-dimensional vector of parameters
Y_{ij} = a particular realization of the experiment

It is assumed that a general form of the model exists and $f(x_i, \beta)$ is a differential function of β. The experiment yields n values of the independent variables, where n is supposed to be sufficiently greater than p to ensure the estimability of the β parameters.

The probability of k accidents occurring during each interval (one year) can be represented as:

$$P_i(k) = \frac{e^{-y_i} y_i^k}{k!}$$

In the estimation procedure we will determine the parameter vector β where,

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \ldots x_i^{\beta_i} + \varepsilon \qquad (24)$$

y = expected accident frequency per year
x_i = traffic and highway geometry variables

Three different methods can be used to estimate the parameters of the Poisson regression model. These are the maximum likelihood principle, weighted least squares analysis and the minimum chi-square estimation. The weighted least squares method of parameter estimation was used in this analysis, as it is more efficient in terms of computing time.

AIC derivation Poisson regression model. The AIC derivation for least squares estimation of the parameters of this model is basically the same as that for the Linear Model described earlier.

Identification of the best Poisson regression model. The model investigated was the multiplicative model described by Eq. (24) for the expected accident involvement. The parameter estimation was carried out in two steps. First, the model was linearized through a logarithmic transformation and parameters were estimated treating the transformed function as a linear model. Using the backward procedure described earlier, the best model at each level was then determined based on the AIC values. Then, starting out with these parameter estimates and the corresponding models, a more precise estimation of the non-linear model was carried out using the PROC NLIN procedure in SAS. This produced least-squares estimates of the parameters obtained through the Marquardt iterative method, where the residuals are regressed on to the partial derivatives of the model with respect to the parameters until the iteration converges.

RESULTS

Multiple Linear Regression Models

Linear regression models were developed for the annual truck accident involvement rate (TRATE) (i.e., accidents per 100 million VMT) as a linear function of highway and traffic related variables. The results obtained are given for the three different environments described earlier.

Environment I. Table I gives the best model for TRATE at each level for this environment. It can be seen that the model with eight variables has the minimum AIC of 288.46 and is therefore the best model. This model is given as:

$$\begin{aligned} \text{TRATE} &= 289.233 - 27.249\ (\text{LNWD}) + 6.6912\ (\text{SHLDWD}) \\ &- 26.413\ (\text{AMS}) + 1.9328\ (\text{SCR}) + 0.0269\ (\text{CCR}) \\ &+ 0.00492\ (\text{ADT}) - 1.3822\ (\text{TPERCNT}) \\ &+ 0.2362\ (\text{SPDIFSQ}) \\ R^2 &= 0.6832 \end{aligned} \tag{25}$$

Environment II. The best model at each level is shown in Table II, which indicates that the best overall model is that consisting of three variables, with an AIC of 245.21 and is given as:

$$\begin{aligned} \text{TRATE} &= -12.196 + 0.0231\ (\text{CCR}) + 0.00077\ (\text{ADT}) \\ &+ 0.6444\ (\text{TPERCNT}) \\ R^2 &= 0.2187 \end{aligned} \tag{26}$$

Table I Linear models for TRATE—Environment I

Level	SSE	AIC	Variables
1	5,251.91	307.89	ADT
2	5,079.73	308.59	SCR, ADT
3	4,851.08	308.79	SCR, ADT, TPERCNT
4	4,475.27	307.65	AMS, SCR, ADT, SPDIFSQ
5	3,454.94	299.56	SHLDWD, SCR, ADT, TPERCNT, SPDIFSQ
6	2,532.25	289.44	LNWD, SHLDWD, AMS, SCR, ADT, SPDIFSQ
7	2,369.44	288.85	LNWD, SHLDWD, AMS, SCR, CCR, ADT, SPDIFSQ
8	2,228.90	288.46†	LNWD, SHLDWD, AMS, SCR, CCR, ADT, SPDIFSQ, TPERCNT

† Identifies best model.

Table II Linear models for TRATE—Environment II

Level	SSE	AIC	Variables
1	295.01	250.93	SCR
2	274.10	248.96	CCR, TPERCNT
3	246.38	245.21†	CCR, ADT, TPERCNT
4	243.30	246.53	CCR, ADT, TPERCNT, SPDIFSQ
5	241.22	248.06	AMS, SCR, CCR, ADT, TPERCNT
6	239.75	249.73	AMS, SCR, CCR, ADT, TPERCNT, SPDIFSQ
7	239.15	251.60	LNWD, AMS, SCR, CCR, ADT, TPERCNT, SPDIFSQ
8	239.08	253.58	LNWD, SHLDWD, AMS, SCR, CCR, ADT, TPERCNT, SPDIFSQ

† Identifies best model.

Table III Linear models for TRATE—Environment III

Level	SSE	AIC	Variables
1	1,470.14	871.12	TPERCNT
2	1,414.37	866.39†	CCR, TPERCNT
3	1,408.07	867.61	CCR, ADT, TPERCNT
4	1,402.29	868.89	CCR, ADT, TPERCNT, SPDIFSQ
5	1,398.58	870.43	AMS, SCR, CCR, TPERCNT, SPDIFSQ
6	1,396.51	872.18	AMS, SCR, CCR, ADT, TPERCNT, SPDIFSQ

† Identifies best model.

Environment III. Table III shows that the best modél for this environment is that consisting of only two variables and is given as:

$$\begin{aligned} \text{TRATE} &= 5.416 + 0.1119\ (\text{CCR}) - 0.1580\ (\text{TPERCNT}) \\ R^2 &= 0.2317 \end{aligned} \tag{27}$$

Using R^2 as a simple measure to compare how well each model represented the variation in data, only the model for Environment I, with an R^2 value of 0.68 demands any consideration. This model implies that an increase in SCR, CCR, ADT, SHLDWD or SPDIFSQ increases the TRATE while an increase in LNWD, AMS and TPERCNT decreased in the TRATE. An increase in the variable SHLDWD resulting in an increase in TRATE is contrary to expectations. However, this may be due to observed correlation between ADT and SHLDWD. Also there is no logical explanation for an increase in either AMS or TPERCNT resulting in a decrease in TRATE. Because of this and the low R^2 values obtained it can be concluded that linear models do not adequately describe the relationship between TRATE and the traffic and geometric variables.

Poisson Regression Models

Environment I. The best non-linear model at each level is shown in Table IV. The best overall model for this environment is the model at the third level, with variables SCR, ADT, TPERCNT. The parameters for these variables were then used as the starting values for the iterative least squares estimation. The resulting model is:

$$\text{TINVOL} = .015237\ (\text{SCR})^{0.0577}\ (\text{ADT})^{0.5024}\ (\text{TPERCNT})^{0.5731} \tag{28}$$

$$\text{AIC} = 62.06$$

Table IV Poisson regression model for Environment I

Level	SSE	AIC	Log variables
1	10.29	64.71	LNWD
2	9.14	62.09	LNWD, AMS
3	8.67	62.06†	SCR, ADT, TPERCNT
4	8.36	62.60	LNWD, SHLDWD, AMS, SEGLEN
5	7.88	62.32	LNWD, SHLDWD, SEGLEN, TPERCNT, SPDIFSQ
6	7.80	63.94	LNWD, SHLDWD, CCR, SEGLEN, TPERCNT, SPDIFSQ
7	7.77	65.76	AMS, SCR, CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ
8	7.70	67.41	LNWD, SHLDWD, AMS, CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ
9	7.67	69.28	LNWD, SHLDWD, AMS, SCR, CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ

† Identifies best model.

This model indicates that ADT, SCR and TPERCNT are variables that are significant in predicting the expected number of truck accident involvements. The low AIC value obtained for this model indicates that it is a good predictor for the number of truck involvements. Intuitively, one would expect to see the variables, SEGLEN and LNWD, be included in the final model. SEGLEN was however not included in this model as all of the sites in this environment were about two miles in length. The lane width (LNWD) is also not included as the analysis showed some correlation between ADT and LNWD.

Figures 3 and 4 show how the number of truck involvements vary with the truck percentage (TP) and SCR for this model. The number of involvements shown is the expected total for one year. These figures indicate that according to this model the number of truck involvements is almost doubled as the percentage of trucks increases from 5 to 15%.

Environment II. The best non-linear models for this environment, at different levels are shown in Table V. From these models the best overall model is the model at the fourth level, with variables SCR, ADT, TPERCNT and SEGLEN. The

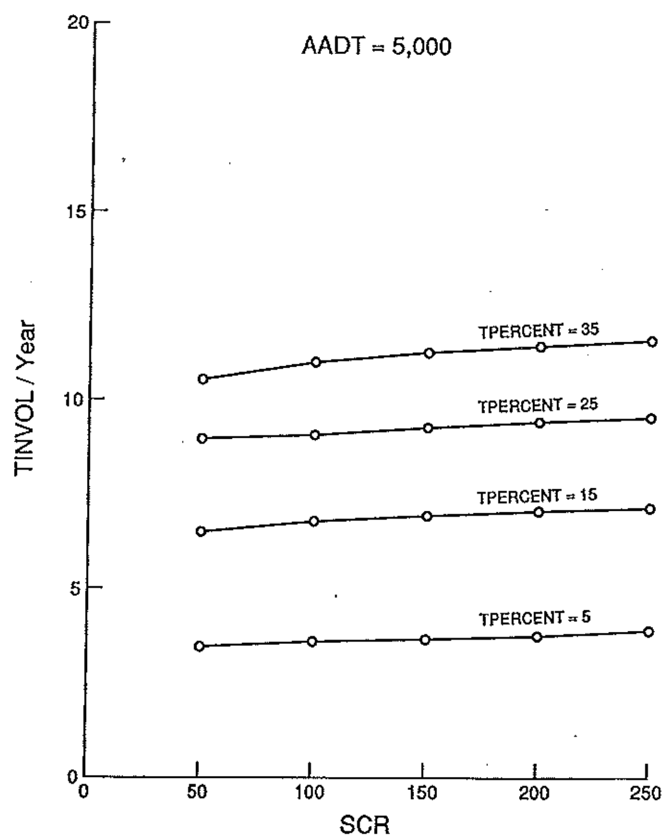

Figure 3 The effect of SCR and TPERCNT on TINVOL—Environment I (segment length = 2 miles) (AADT = 5000).

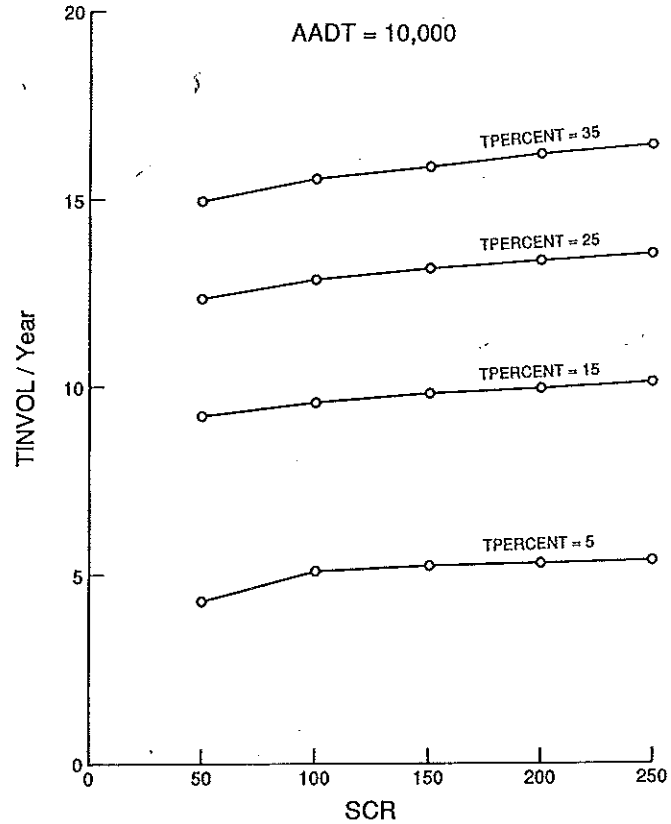

Figure 4 The effect of SCR and TPERCNT on TINVOL—Environment I (segment length = 2 miles) (AADT = 10,000).

Table V Poisson regression model for Environment II

Level	SSE	AIC	Log variables
1	15.57	92.10	ADT
2	13.09	84.71	ADT, TPERCNT
3	12.58	84.59	SCR, ADT, TPERCNT
4	11.90	83.56†	SCR, SEGLEN, ADT, TPERCNT
5	11.88	85.48	SCR, CCR, SEGLEN, ADT, TPERCNT
6	11.82	87.21	AMS, SCR, CCR, SEGLEN, ADT, TPERCNT
7	11.79	89.09	SHLDWD, AMS, SCR, CCR, SEGLEN, ADT, TPERCNT
8	11.72	90.77	LNWD, SHLDWD, AMS, SCR, CCR, SEGLEN, ADT, TPERCNT
9	11.64	92.40	LNWD, SHLDWD, AMS, SCR, CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ

† Identifies best model.

parameters for these variables were then used as the starting values for the iterative least squares estimation. The resulting model is:

$$\text{TINVOL} = 9 \times 10^{-8}\,(\text{SCR})^{0.0471}\,(\text{ADT})^{1.4358}\,(\text{TPERCNT})^{1.5232}\,(\text{SEGLEN})^{0.3826}$$
$$\text{AIC} = 83.56 \qquad (29)$$

This model indicates that SCR, ADT, TPERCVNT, and SEGLEN are the best descriptors of the TINVOL for a particular segment of highway.

Figures 5 and 6 show how the number of truck involvements vary with the truck percentage TPERCNT and SCR, using this model and a segment length of two miles. This model indicates a relationship between truck percentages, truck involvement and SCR similar to that for Environment I.

Environment III. The best non-linear models for this environment, at each level is shown in Table VI. It can be seen that the AIC values for the models at the fourth and fifth levels are almost identical. The model at the fifth level was, however, selected as it included TPERCNT which has been shown to have some effect on TINVOL. The other variables are CCR, ADT, SEGLEN and SPDIFSQ. The parameters for these variables were then used as the starting values for the iterative least squares estimation. The resulting model is:

$$\text{TINVOL} = 0.001465\,(\text{CCR})^{0.0336}\,(\text{SEGLEN})^{0.3318}\,(\text{ADT})^{0.7086}\,(\text{TPERCNT})^{0.2064}\,(\text{SPDIFSQ})^{0.0475} \tag{30}$$

AIC = 407.48

This model indicates that in this environment the most critical variables are ADT,

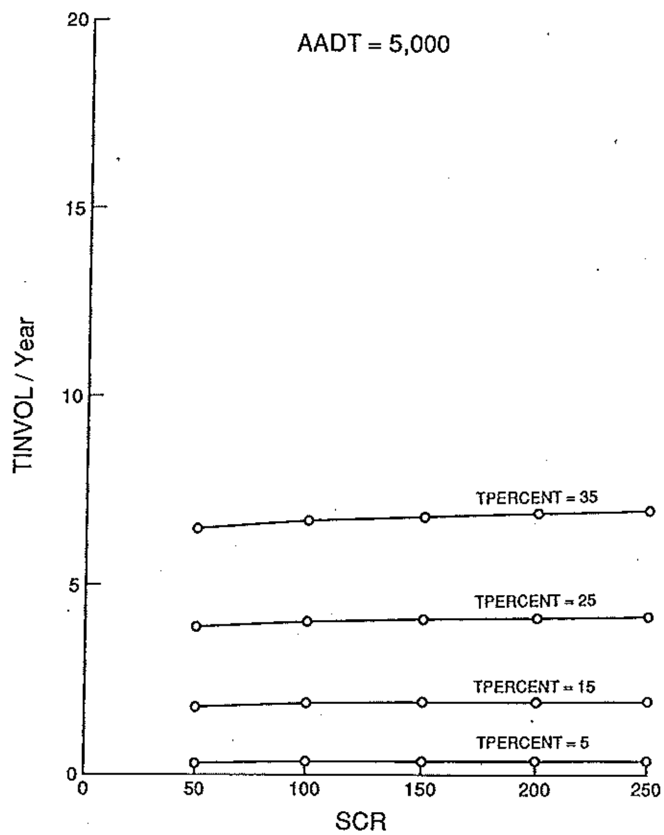

Figure 5 The effect of SCR and TPERCNT on TINVOL—Environment II (segment length = 2 miles) (AADT = 5000).

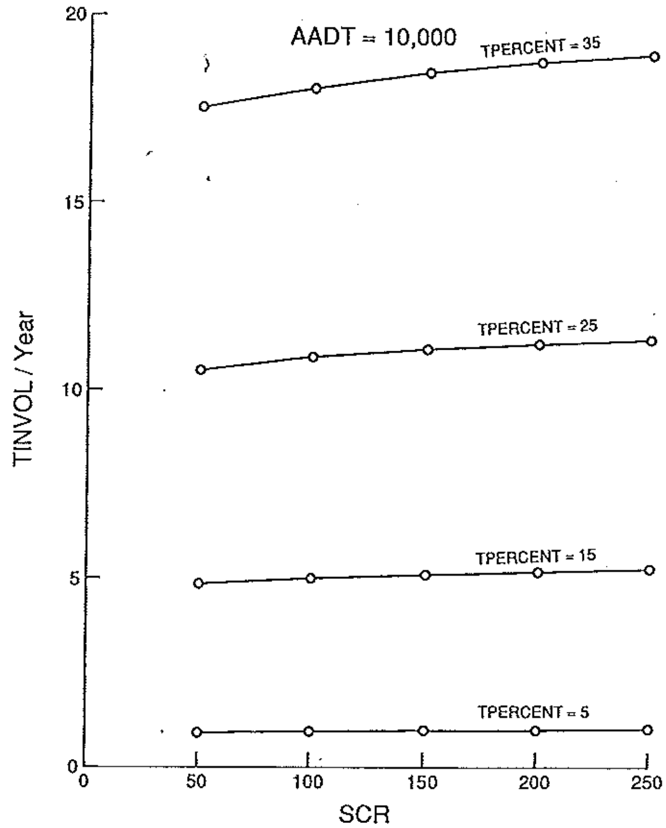

Figure 6 The effect of SCR and TPERCNT on TINVOL—Environment II (segment length = 2 miles) (AADT = 10,000).

CCR, and the Speed Difference (SPDIFSQ). The model indicates that the difference between the average speeds of trucks and non-trucks has some effect on large truck involvement in accidents on highways within this environment. Figures 7 and 8 show how the number of truck involvements vary with the CCR and Speed Difference. These figures indicate that both increasing speed difference and CCR tend to increase the expected number of truck accidents.

Table VI Poisson regression model for Environment III

Level	SSE	AIC	Log variables
1	107.23	415.55	ADT
2	101.59	408.15	SEGLEN, ADT
3	100.07	407.54	CCR, SEGLEN, ADT
4	98.89	407.47†	CCR, SEGLEN, ADT, SPDIFSQ
5	97.76	407.48	CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ
6	97.56	409.13	SCR, CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ
7	97.27	410.59	AMS, SCR, CCR, SEGLEN, ADT, TPERCNT, SPDIFSQ

† Identifies best model.

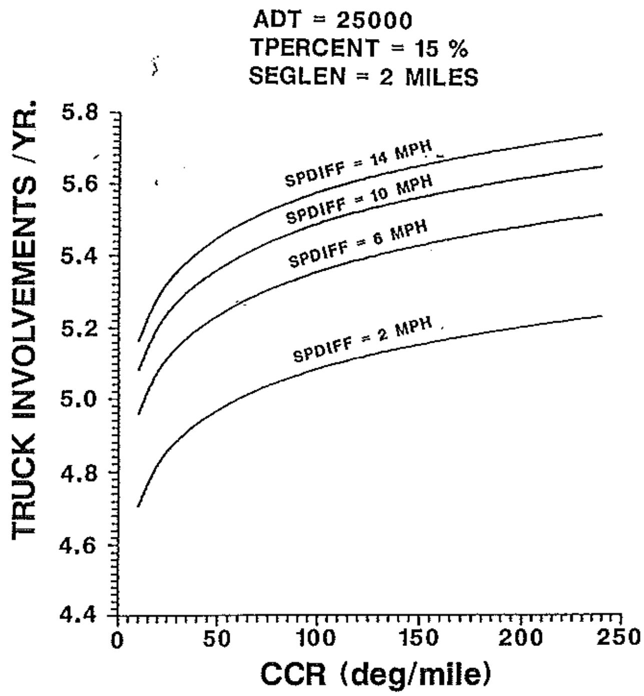

Figure 7 The effect of CCR and SPDIFF on TINVOL—Environment III (AADT = 25,000).

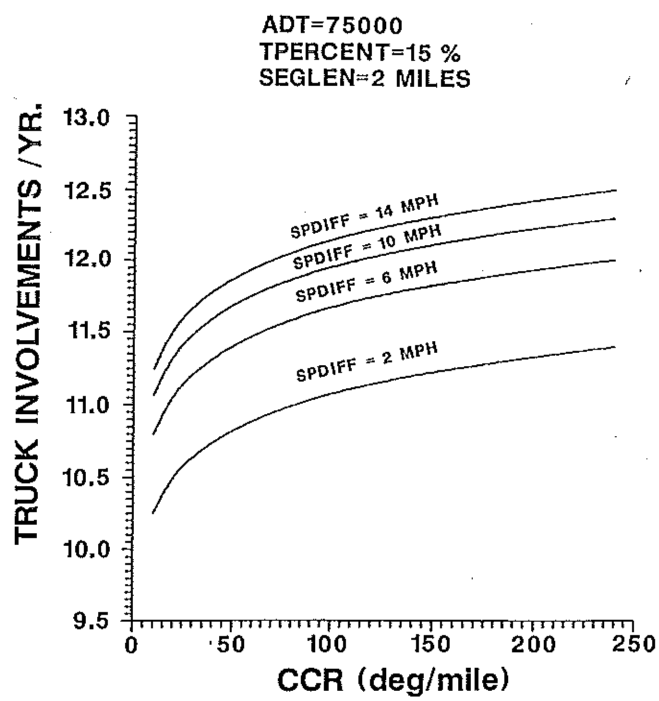

Figure 8 The effect of CCR and SPDIFF on TINVOL—Environment III (AADT = 75,000).

CONCLUSIONS

Based on the results of this study, the following conclusions are made.

Multiple Linear Regression Models

Multiple linear regression models do not seem to adequately describe the relationship between large truck involvement in accidents and associated traffic and geometric variables.

Poisson Regression Models

The Poisson models developed seem to adequately describe the relationship between large truck involvement in accidents and associated traffic and geometric variables.

Environment I

- The most significant traffic variables contributing to large truck involvement in accidents on highways within this environment are the ADT on the highway segment and the percentage of trucks.
- The most significant geometric variable is the slope change rate (SCR). Increasing SCR on a highway segment tends to increase the truck involvements, indicating that the vertical alignment is more critical for trucks in this environment.

Environment II

- The most significant traffic variables contributing to large truck accident causation are the ADT and the percentage of trucks on the highway segment.
- The most significant geometric variable is the SCR. Increasing SCR also increases the involvement of large trucks in accidents.

Environment III

- The most significant traffic variables contributing to large truck involvement in accidents to truck accidents are the ADT and the speed difference between trucks and non-trucks.
- The most significant geometric variable is the curvature change rate (CCR). Highway segments with high CCR tend to experience increased large truck involvements.

Measures for Complexity of Highway Alignment

The measures for the complexity of highway alignment, such as CCR, SCR and AMS introduced in this study have made a significant contribution to the quantification of geometric complexity. Without such a measure any work on the impact of highway geometry on safety or even operations can only be carried out for geometrically homogeneous highway segments. The role of these variables in the models developed, further verifies the relevance of such an approach.

References

1. N. J. Garber and Sarath Joshua, "Traffic and geometric characteristics affecting the involvement of large trucks in accidents." Virginia Transportation Research Council, Vol. II (1989).
2. R. Lamm and E. Choueiri, "Recommendations for evaluating horizontal design consistency." Transportation Research Record, Transportation Research Board, Washington DC (1987).
3. J. P. Leisch and J. E. Leisch, "New concepts in design speed applications." Transportation Research Record 631, Transportation Research Board, Washington DC (1977).
4. H. Akaike, "Information theory and an extension of the maximum likelihood priniciple." In B. V. Petrov and B. F. Csaki (Eds.), Second International Symposium on Information Theory. Academiai Kiado: Budapest, pp. 267–281 (1973).
5. S. Kullback and R. A. Liebler, "On information and sufficiency." *Annals of Mathematical Statistics*, 22, 79–86 (1951).
6. H. Bozdogan, "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions." *Psychometrika* 52(3), 345–370 (1987).
7. P. P. Jovanis and H. L. Chang, "Modeling the relationship of accidents to miles traveled." Transportation Research Record 1068, TRB, Washington DC (1986).
8. F. A. Haight, "Induced exposure." *Accident Analysis and Prevention* 5(2) (June, 1973).