Models for Predicting Accidents on Two-Lane Rural Highways

Prianka N. Seneviratne, Koti R. Kalakota, and Prabhakar Attaluri,
Utah State University

Empirical and theoretical models have been used to describe the expected change in accidents at a given location with varied geometric elements. Using the Highway Safety Information System (HSIS) data for a selected two-lane rural corridor in northern Utah, the differences in the functional forms, significance of explanatory variables, and predictive accuracies of previous models are examined. Three separate models are calibrated using the interactive regression approach. The first model, defined as the tangent model, was able to explain more than 90 percent of the variation in accident rate on tangents. The second model, developed to estimate expected accident rate on curve sections, explained 50 percent. The third (corridor) model, calibrated with curve and tangent data combined, explained more than 90 percent of the variation. Section length was found to be the most significant variable in all three models. The other geometric variables, such as degree of curvature and shoulder width, explain at most 10 percent of the variation. Models calibrated in the present case are reasonably valid in the same environment over short periods of time. However, previous models were found to be unsuitable for predicting accidents in the present case.

As highway travel grows and funding for new road capacity dwindles, two-lane highway networks in most countries will be expected to play a more important role. Many road segments will require upgrading to meet future traffic volumes, changing road composition, and diverse vehicle and driver characteristics. Consequently, questions arise about the forms of upgrading required, specifically, what geometric elements of the highways should be upgraded and to what extent.

The answer is complex because, even after four decades, practitioners and researchers are still uncertain of the interactions among drivers, vehicles, and the road environment. Moreover, the emerging knowledge of the intricate relationship accidents have with geometric and human factor variables is fragmented because of inadequate coordination among researchers. The lack of uniform data collection and recording procedures is an additional problem. These issues must be resolved, and reliable sources of information should be developed. Even so, much of the evidence to date suggests that microlevel changes to roadway alignment add to geometric inconsistencies, which in turn contribute to most of the accidents not related to driver error.

Notwithstanding the cost, geometric changes that can improve consistency are perhaps the easiest upgrades to undertake. For this reason, it is important to determine the changes that can produce the maximum reduction in accidents. This may be accomplished partly by examining the correlation between various roadway geometric elements and traffic accidents.

A study to verify the influence of geometric variables on accidents on two-lane rural highways was initiated,

the primary objective of which was to derive one or more mathematical models to explain the variation in accident rate as a function of a set of geometric variables. Subsequently, the validity of model parameters over time and space was investigated. To meet the above objectives, the following procedure was adopted:

- Examine and categorize previous accident prediction models on the basis of underlying modeling principles,
- Identify statistically significant variables within the study corridor,
- Determine the appropriate form or forms for models,
- Test the validity and predictive accuracy of models, and
- Compare calibrated model parameters and coefficients with parameters of models from other studies to determine the potential for model transfer.

Background

The relationship between highway design parameters and accidents has been widely discussed and researched for almost 40 years (*1–3*). Explanatory variables and assumptions about the functional form of models have changed little over the years, as evidenced by the significance of shoulder width, section length, traffic volume, and horizontal curvature in the established relationships. However, statistical methods and advanced computer software have enabled recent researchers to establish better relationships than those noted by early researchers such as Baldwin (2).

Of the models that have been proposed over the years, Zegeer's (*3*) two models are perhaps the most cited and best known among highway agencies. One of these describes accidents as a function of the cross-sectional variables, and the other describes accidents in relation to curve geometry. Both are user-friendly models, but their validity over time and space is not well documented. In terms of specific models for rural two-lane highways, Gupta and Jain (*4*), and Cleveland and Kitamura (*5*) arrived at some early conclusions. Both of these studies considered moderate-volume highways, but the predictive accuracy of the models was comparatively low. The latter authors analyzed the data by dividing them into three volume groups and established more credible relationships. Neuman (*6*) examined the relation between accidents and curve geometry on two-lane highways. Despite the large sample of 3,557 sites containing 13,545 crashes from four states, their regression model was able to explain less than 20 percent of the variation. The subsequent discriminant analysis performed on the same data set proved more fruitful because it enabled the identification of geometric elements that increase the potential of accidents. The drawback of the latter approach is the subjectivity of the definition of discriminating variables. Datta et al. (*7*) included several operational variables in the list of nongeometric surrogates, but evidently only the speed differential was found to have any significant influence on accidents.

A formidable effort was made by Cleveland et al. (*8*) to examine the influence of geometric and traffic variables by bundling them into compatible groups. A total of 21 models were tested with different combinations of the variables within the bundles. These models were able to explain between 30 and 75 percent of the variation in accidents, and the most significant variable was found to be traffic volume. The influence of geometric elements was noted to be relatively insignificant.

The traffic conflict technique drew considerable attention when it was first introduced. Instead of geometric variables, the technique used conflicts that were defined as a function of traffic volume or some surrogate variable. A similar concept was advanced by Reinfurt et al. (*9*) as an alternative to accident prediction models. They proposed that surrogates such as center line and edge line encroachment rates when related to degree of curvature demonstrate the sensitivity of curve design to accident potential. These authors showed that both rates are linearly related to degree of curvature at values greater than *5* degrees, and they suggested that accident potential increases with edge line encroachments. However, earlier studies by Datta et al. (*7*) and Terhune and Parker (*10*) demonstrated no robust relations between such nonoperational variables and accidents.

More recently, new statistical modeling techniques have emerged in the traffic safety arena. The generalized linear modeling (GLM) approach and Poisson modeling approach are noteworthy. Both these techniques have been used to overcome random variations (or inherent uncertainty) that distort the response of a dependent variable to changes in an independent variable. Roine and Kulmala (*11*) are perhaps the first to use GLM in modeling two-lane highway accidents. They employed the popular generalized linear interactive modeling (GLIM) package for their work.

Study Corridor

A 53-mi corridor of Highway 89/91 from Brigham City to Bear Lake in northeastern Utah was selected for calibrating and validating a set of new models as well as testing the transferability of previous models. This corridor is close to the research laboratory, which made the verification of geometric data relatively easy. More-

over, it was classified as two-lane rural highway by the state department of transportation, had vertical grades ranging from 2 to 8 percent and horizontal curves varying from 2 to 32 degrees, and had an annual average daily traffic of less than 1,000 vehicles. The data were made available by FHWA from their HSIS records for Utah and were contained in four separate files: accidents, curve, roadway, and grade.

For the purposes of these analyses, the curve, roadway, and grade files were combined into one file which was used as the roadway inventory file. The file containing accident records (from 1987 to 1990) was combined with the roadway inventory file, so the road characteristics at each accident site were available in one file. This was a tedious task since each accident is recorded as a separate entry in the HSIS data base, and a given 1-mi segment could have up to 10 accidents of different types in a given year. The accidents in each segment in a given year had to be aggregated first then reentered in a new column so that they corresponded with the roadway and traffic characteristic of that segment. Another problem with the data was that the same characters were used to describe more than one variable. Furthermore, the variables were sometimes defined by characters and other times by numbers or blank spaces.

Two separate data bases were created from the above data base. One contained all the tangent sections, and the other contained all the curve sections. The variables in each of these data bases and a short description of their relationship with accident rates are given below.

Section Length (L)

The distributions of the length (L) of all tangents and horizontal curves (in meters) are given in Table 1. Approximately 70 percent of the tangent sections in the corridor are less than 300 m. More than 90 percent of the curves are less than 300 m, and the curved segments make up approximately 52 percent of the entire study corridor.

One important feature that was missing from the data is information on spirals. There was no way of knowing if the curves were preceded by spirals. But since terrain within the corridor does not permit long spirals, errors due to this in distinction could be regarded as minimal.

The annual average accident rate (AAAR) for the tangents and curves is also shown in Table 1. On the tangents, the highest AAAR was observed in sections longer than 1500 m, and the lowest was observed in sections less than 150 m. On curved sections, the highest AAAR was observed in sections between 600 and 900 m.

Degree of Curvature (D)

Since tangent sections in the HSIS base are identified as having zero degrees of curvature, it permitted detailed analyses of accidents on curves. The AAAR as a function of degree of curvature in the study corridor is shown in Figure 1. Contrary to previous findings (*4,7*), more than 40 percent of accidents on curves occurred in 4- to 8-degree curves. In terms of total accidents in the corridor, less than 15 percent of the accidents occurred in these curves.

Vertical Grade (G)

The study corridor traverses two major canyons, Sardine and Logan, in northern Utah. Consequently, there are many segments where a horizontal curve is connected directly or through a short section to a vertical curve. These sections are not easily identifiable from the

TABLE 1 Distribution of Lengths and Accident Rates on Tangents and Curves

Section	Tangent Sections		Curve Sections	
Length (m)	% Sections	AAAR	% Sections	AAAR
0-150	49.00	0.1429	48.30	0.1973
150-300	21.00	0.4000	42.50	0.3107
300-450	7.00	0.4800	6.50	0.8000
450-600	9.80	0.8571	1.30	1.5000
600-900	2.80	1.0500	1.40	3.8000
900-1500	2.10	0.7330	0.00	0.0000
>1500	8.30	9.5200	0.00	0.0000

AAAR = Average Annual Accident Rate (accidents per million vehicles)

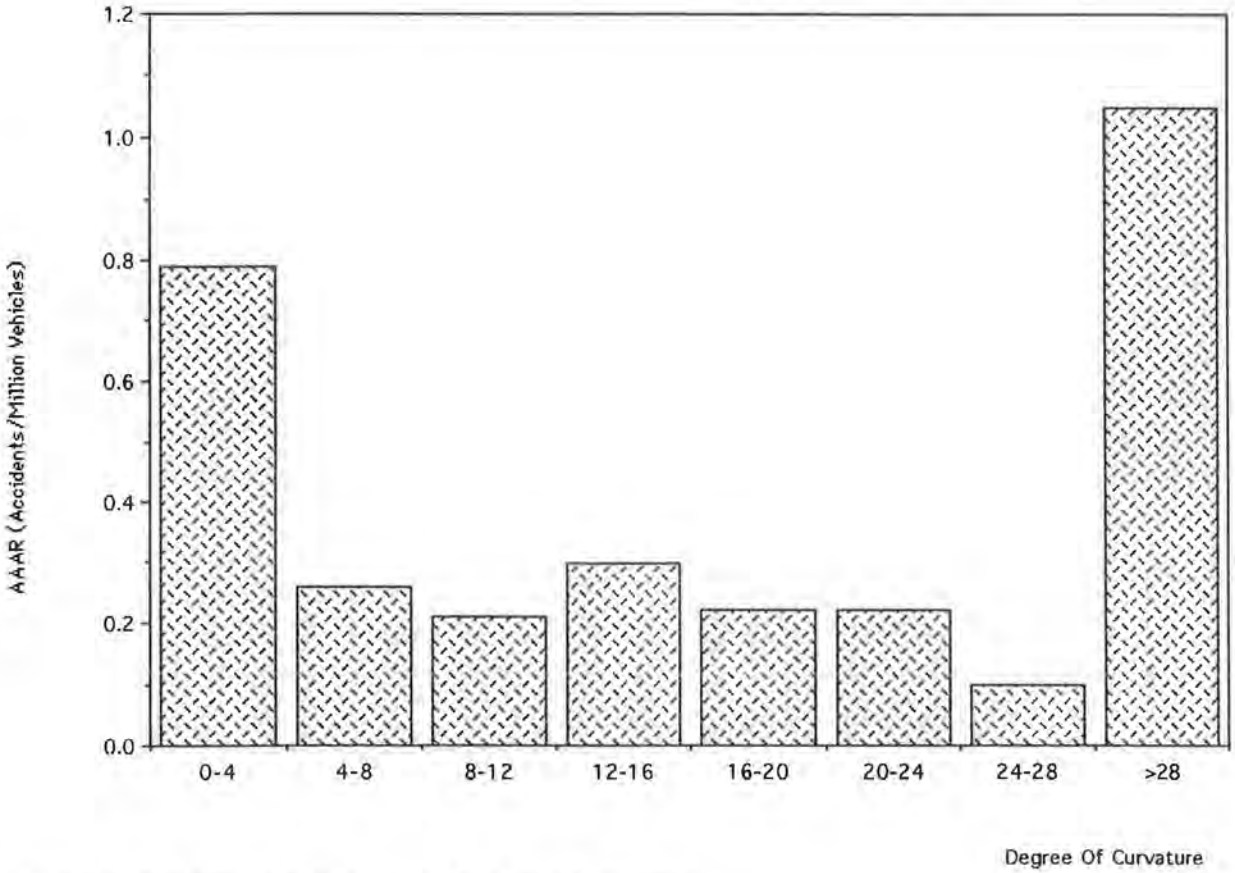

FIGURE 1 Distribution of accident rates by degree of curvature.

HSIS files and are difficult to find without extensive field surveys or state department of transportation inventories. Therefore, no distinction was made between sections on the basis of this condition. An approximation was made when there was a change in grade within a section so that the average of the two end point grades become the grade variable.

Right Shoulder Width (SWR)

The nature of the terrain had determined the available right shoulder width in many of the sections through the canyons. None of the shoulders in the entire corridor were paved, and most segments had less than 1 m of shoulder. The distribution of the widths (in meters) and accidents is shown in Figure 2.

Traffic Volume (AADT)

Section traffic volumes are recorded in terms of (AADT) in the data files. However, the AADT changed little within the study corridor. The major difference was noted in the section through the city of Logan, which was eliminated to satisfy the rural condition. Because of this and because it is more of an operational variable, AADT was incorporated into the dependent variable as the exposure variable.

Data Analysis

Prior research and the present study objectives were used as a guide to select the explanatory variables. The task was not difficult because there were no other truly random geometric variables in the data base except those mentioned above. Others, such as lane width and left shoulder width, were either not relevant or invariable. Therefore, those listed above were considered, and the accident rate (accidents per million vehicles) averaged over 5 years (AAAR) was used as the dependent variable.

The nature of the relationship between each independent variable and the dependent variable was not clearly apparent from the individual plots. For example, Figure 1 demonstrates that the accident rate in the present case is not increasing with the degree of curvature

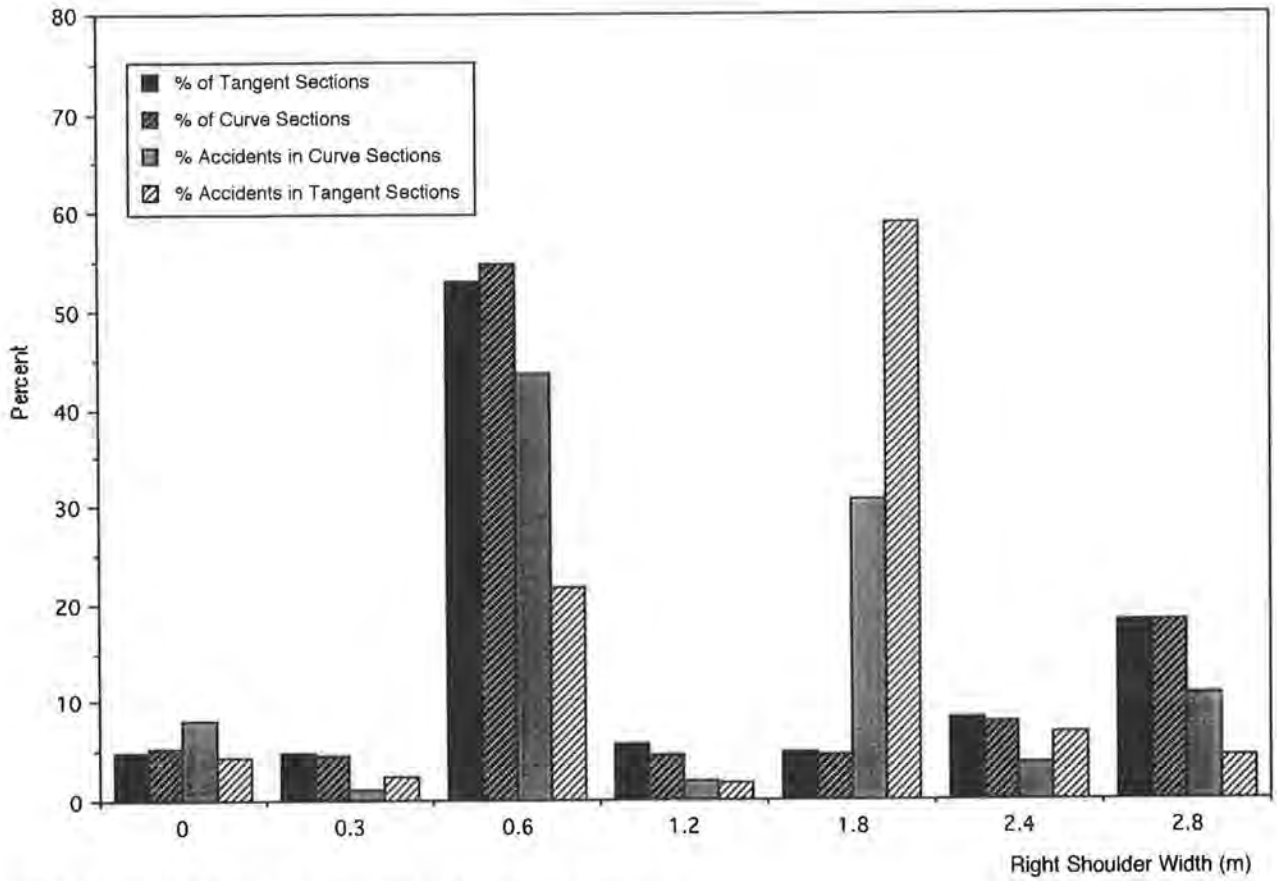

FIGURE 2 Distribution of right shoulder width and accidents.

that had been hypothesized by prior researchers. Moreover, the relationships noted in previous studies did not concur with one another. Therefore, it was important to test several different forms and combinations of variables. All these combinations of variables were tested for collinearity and correlation with the dependent variable (accident rate) in the case of tangents, curves, and the entire corridor.

Two modeling approaches were examined in the present study, generalized linear interactive modeling (GLIM) and interactive regression analysis (IRA). The GLIM procedure considers a predefined error structure for the dependent variable and a selected link between the linear predictor of the explanatory variables and the dependent variable. Accordingly, the occurrence of an accident is assumed to be a random variable (goodness-of-fit tests also confirmed the same) with an error structure of Poisson, and a link of log was selected. However, the results were not encouraging. Therefore, the second option, IRA, was chosen. Statgraphics software that permits IRA was used to perform the data analyses and model validation. This method is described in detail later.

Analysis of Variables

The study of the functional forms of previous models indicated that almost all models were calibrated with combinations of variables in which a single variable was used more than once. That contributed to illogical signs and insignificant coefficients, thereby decreasing the validity of the model. Care was taken to prevent such errors by examining the collinearity among the variables at each stage.

The explanatory variables to be included in the models were selected on the basis of their correlation with the dependent variable and the correlation among themselves. This exercise proved that with tangent, curve, and corridor, section length alone or section length in combination with another variable is highly correlated with accident rate. Nevertheless, all variables showing some degree of correlation were selected to be included in the first calibration phase.

After examining the results of the first phase, variables showing illogical signs and small correlation coefficients were omitted, and the models were recalibrated. This resulted in having the product of section length

and right-shoulder width in the tangent as well as the corridor model, and the product of degree of curvature and section length in the curve model.

Model Development

Formulation

In IRA (*12*), all or selected outliers in the data can be eliminated from the data set by selecting them off the screen until the best fit is observed. An outlier is any observation deviating significantly from the rest of the observations that is likely to considerably influence the model form. This increases errors in prediction and summary statistics.

The existence of an outlier in the data can be attributed to either of the two mechanisms briefly described below.

Mechanism 1

Data are from some heavy-tailed distribution, such as a *t*-distribution, in which the tailed values slowly tend to zero. Mathematically this can be expressed as

$$\Pr[X_n - X_{n-1} > 0] \geq \beta \quad \text{for all } n$$

$$\Pr[X_n / X_{n-1} > c] \geq \beta$$

where

X_n = test statistic with n degrees of freedom,
X_{n-1} = test statistic with $n-1$ degrees of freedom,
n = sample size, and
0, c = small value.

Mechanism 2

Data can be broadly divided into two types of distributions, basic and contaminating. The latter contains more heavy tails (i.e., is highly skewed) than the former, yielding to contaminants or outliers. From these two distributions, a "heavy tailed distribution model" can be defined and any observation can be grouped into one of the two distributions using the model

$$f(.) = (1 - p)\, f_0(.) + p f_1(.)$$

where

$f(.)$ = any arbitrary observation,
$f_0(.)$ = basic distribution,
$f_1(.)$ = contaminating distribution, and
p = probability of belonging to a particular distribution

Several procedures are available to treat the observed outliers. The "rejection of the outlier" procedure is well known. Parameters are estimated by simply rejecting the outliers and calculating the arithmetic mean for the rest of the data. This procedure is well suited in case of multiple regression analysis, in which each variable will have its own distribution. In IRA, these outliers can be selected off the screen while the graphical fit of the least-squares regression line is observed.

Calibration

Each model was first calibrated with the entire data set corresponding to the type of road section it represents. Then the outliers (all the data points falling outside the 95 percent confidence interval) were removed, and the models were recalibrated. The relationships observed during the recalibration are as follows:

Tangent model:

$$AAAR = 0.12(L * SWR)$$

Curve model:

$$AAAR = 0.087(D * L)$$

Corridor model:

$$AAAR = 0.1224(L * SWR)$$

where

L = section length,
SWR = right shoulder width, and
D = degree of curvature.

Summary Statistics

The statistics given in Table 2 indicate that model validity and accuracy were computed following the calibration phase. These statistics are discussed briefly in the following section.

Multiple Correlation Coefficient (R^2)

According to the R^2-values, the tangent model explains 94 percent of the variation in the accident rate, the curve model 52 percent, and the corridor model 91 percent. R^2 indicates that with the exception of the curve model, the models have very good explanatory capabilities.

Standard Error

A regression model is as reliable as its parameters or coefficients. Ideally, the standard error of the coefficient

TABLE 2 Summary Statistics

Tangent Model
AAAR = 0.120(L*SWR)
N = 138, R^2 = 0.9416
Standard Error = 0.274
t-value = 46.9925
Root Mean Square Error = 14.85
Curve Model
AAAR = 0.087(D*L)
N = 133, R^2 = 0.5215
Standard Error = 0.238
t-value = 11.995
Root Mean Square Error = 13.57
Corridor Model
AAAR = 0.1224(L*SWR)
N = 292, R^2= 0.9113
Standard Error = 0.024
t-value = 54.68
Root Mean Square Error = 21.68

should be as small as possible or less than the model coefficients. The standard errors calculated were found to be 0.274, 0.238, and 0.024 for the tangent, curve, and corridor models, respectively.

Transferability

One objective of this study was to examine the transferability or applicability of the calibrated models over time and space. To test the temporal transferability, accident data from 1985 to 1989 were used to calibrate the models, and the 1990 accident data were retained to test their predictive accuracy. When the observed accident rates in 1990 were compared with the expected accident rates obtained using the calibrated models, the root mean square error (RMSE) values suggested a reasonably close fit in the case of tangent and corridor models. The curve model, on the other hand, was unable to predict as closely. As shown by the summary statistics in Table 2, the RMSE values for tangent, curve, and corridor models are 14.85, 13.58, and 21.68, respectively.

In spite of certain similarities in the explanatory variables, best-fit models in the present case are different from those derived in previous studies. Therefore, the most reasonable models for comparing the present case were those of Zegeer (*13,14*), which were developed as part of a national program with data from several states. Although these models were not of the same functional form, they are currently known to more state transportation officials than any others.

Attempts to calibrate these models with the present data were unsuccessful. Some input data, such as shoulder type and spiral information for the curve model, were unavailable. The number of accidents in the spe-

cific classes were also small. The R^2-value was found to be less than 12 percent, and the variables were insignificant. The predictions with the uncalibrated models were also inaccurate. For instance, estimates from both of Zegeer's models mostly lay outside the 95 percent confidence interval. Lack of knowledge about the predictive accuracy of the original models makes it difficult to say if this was to be expected.

Conclusions

This study showed that expected accidents of all types on two-lane rural highways can be accurately estimated. Moreover, it was found that the IRA approach followed here is more suitable than an ordinary linear regression type of approach. Rejecting the outliers contributing to model uncertainty is a more meaningful procedure than simply explaining the variation with a highly dispersed data set. However, disaggregating the data into tangents and curves did not improve model accuracy. Additionally, prior studies showed that disaggregation by type of accident considerably enhances accuracy.

In spite of being widely cited as a general model, the models by Zegeer et al. (3) were unable to explain more than 5 percent of the variation in accidents at the cross section or on curves. Of course, this may have been caused by the missing variables. But Zegeer et al. (3) demonstrated that curvature and section length are more significant among all the variables. The models should have been able to explain a larger part of the variance. Nevertheless, the three models developed in the present case predicted the 1990 accident rates with reasonable accuracy, indicating the potential for temporal transference.

The question of the validity of safety benefits from curve flattening is raised. Given the dominant significance of section length, the impact of curvature changes on accidents is likely to be diminished, particularly if curve flattening results in a significant elongation of the roadway. This question calls for further examination of the relationship between section length and degree of curvature.

There is an urgent need for more work in the accident-modeling area. If a reasonable level of uniformity can be achieved, parameter estimation and transferability will become practical. The use of modeling principles in transportation safety analysis will become commonplace. Ultimately, agencies responsible for transportation will be able to make informed decisions instead of educated guesses about the influence of geometric elements on traffic accidents.

References

1. Versace, J. Factor Analysis of Roadway and Accident Data. *HRB Bulletin 240*, HRB, National Research Council, Washington, D.C., 1960, pp. 24–32.
2. Baldwin, D. M. The Relationship of Highway Design to Traffic Accident Experience. *32nd Annual Convention Papers & Discussions*, AASHTO, Washington, D.C., 1946.
3. Zegeer, C. V., D. W. Reinfurt, T. Miller, W. W. Hunter, and E. Hamilton. *Cost-Effective Geometric Improvements for the Safety Upgrading of Horizontal Curves.* Report FHWA-RD-90-021. FHWA, U.S. Department of Transportation, May 1990.
4. Gupta, R. C., and R. Jain. *Effect of Certain Geometric Design Characteristics of Highways on Accident Rates for Two-Lane Roads in Connecticut.* Civil Engineering Department, University of Connecticut, Storrs, 1973.
5. Cleveland, D. E., and R. Kitamura. Microscopic Modelling of Two-Lane Rural Roadside Accidents. In *Transportation Research Record 681*, TRB, National Research Council, Washington, D.C., 1974, pp. 11–18.
6. Neuman, T. Accident Analysis for Highway Curves. In *Transportation Research Record 923*, TRB, National Research Council, Washington, D.C., 1983, pp. 65–69.
7. Datta, T. K., D. D. Perkins, J. I. Taylor, and H. T. Thompson. *Accident Surrogates for Use in Analyzing Highway Safety Hazards.* Report FHWA-RD-82-103-105. FHWA. United States Department of Transportation, 1983.
8. Cleveland, D. E., L. P. Kostynuik, and K.-L. Ting. Design and Safety on Moderate Volume Two-Lane Roads. In *Transportation Research Record 1026*, TRB, National Research Council, Washington, D.C., 1985, pp. 51–61.
9. Reinfurt, D. W., C. V. Zegeer, B. J. Shelton, and T. R. Neuman. Analysis of Vehicle Operation on Horizontal Curves. In *Transportation Research Record 1318*, TRB, National Research Council, Washington, D.C., 1991, pp. 43–50.
10. Terhune, K. W., and M. R. Parker. An Evaluation of Accident Surrogates for Safety Analysis of Rural Highways. Report FHWA-RD-86-127. FHWA, United States Department of Transportation, 1986.
11. Roine, M., and R. Kulmala. *Accident Prediction Models for Main Road and Rural Links on Single Carriage way Roads.* Research Note 1285. Technical Research Center of Finland, Espoo (in Finnish).
12. Hawkins, D. M. *Identification of Outliers.* Chapman and Hall, 1980.
13. Zegeer, C. V., J. Hunter, D. Reinfurt, L. Herf, and W. W. Hunter. *Safety Effects of Cross-Sectional Design for Two-Lane Roads.* Report FHWA/RD- 87/008. FHWA, United States Department of Transportation, Oct. 1987.
14. Zegeer, C. V., and J. A. Deacon. Effect of Lane Width, Shoulder Width, and Shoulder Type on Highway Safety. In *State of the Art Report 6*, National Research Council, Washington, D.C., 1987.