PERGAMON

Accident Analysis and Prevention 34 (2002) 357–365

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Effects of road geometry and traffic volumes on rural roadway accident rates

Matthew G. Karlaftis *, Ioannis Golias

Department of Transportation Planning and Engineering, Faculty of Civil Engineering, National Technical University of Athens, 5 Iroon Polytechnion Street, 157 93 Zografon, Athens, Greece

Received 16 May 2000; received in revised form 1 February 2001; accepted 28 February 2001

Abstract

This paper revisits the question of the relationship between rural road geometric characteristics, accident rates and their prediction, using a rigorous non-parametric statistical methodology known as hierarchical tree-based regression. The goal of this paper is twofold; first, it develops a methodology that quantitatively assesses the effects of various highway geometric characteristics on accident rates and, second, it provides a straightforward, yet fundamentally and mathematically sound way of predicting accident rates on rural roads. The results show that although the importance of isolated variables differs between two-lane and multilane roads, 'geometric design' variables and 'pavement condition' variables are the two most important factors affecting accident rates. Further, the methodology used in this paper allows for the explicit *prediction* of accident rates for given highway sections, as soon as the profile of a road section is given.

Keywords: Accident rates; Rural roads; Hierarchical tree based regression

1. Introduction

Road safety modelling has attracted considerable research interest in the past four decades because of its wide variety of applications and important practical implications. Public agencies, such as State Departments of Transportation, may be interested in identifying accident-prone areas to promote safety treatments. Transportation engineers may be interested in identifying those factors (traffic, geometric, etc.) that influence accident frequency and severity to improve roadway design and provide a safer driving environment.

The very high cost of highway accidents paid by societies around the world makes highway safety improvement an important objective of transportation engineering. Highway safety specialists can influence traffic safety either through means such as road rules, law enforcement, and education, or by applying local traffic control and geometry improvements. An overwhelming majority of previous studies have indicated that improvements to highway design could produce significant reductions in the number of crashes. Recognizing this, the Federal Highway Administration (FHWA) promotes safety and accident investigation by encouraging States to pursue the development of Safety Management Systems (SMS). And, although SMSs are not Federally required as of 1996, most States continue to work on their development, suggesting the need for improving on existing empirical models for accident measurement.

Following a long line of studies concerned with identifying major factors contributing to highway accidents, this paper revisits the problem of the relationship between rural road geometric characteristics, accident rates and their prediction, using a rigorous non-parametric statistical methodology known as hierarchical tree-based regression (HTBR).[1] The goal of this paper is not only to develop a methodology that quantitatively assesses the effects of various rural road geomet-

* Corresponding author. Tel.: +30-1-6711203; fax: +30-1-7721327.

E-mail address: mgk@central.ntua.gr (M.G. Karlaftis).

[1] For information on the process of functional road classification the reader should refer to US DOT (1969).

PII: S0001-4575(01)00033-1

ric characteristics on accident rates, but also to provide a straightforward, yet fundamentally and mathematically sound way of *predicting* accident rates. The ability to predict accident rates is very important to transportation planners and engineers, because it can help in identifying hazardous locations, sites which require treatment, as well as spots where deviations (either higher or lower rates) from expected (predicted) warrants further examination. The remainder of the paper is organized as follows. The next section provides some background necessary for the development of the methodology used in this paper. Following this, the data and methodology that were used, the estimation results, and examine the effects of various geometric characteristics on accident rates are presented and discussed. The final section of the paper summarizes the findings and offers some concluding remarks.

2. Background

Much literature exists that addresses the problem of accident rate estimation, and the identification of the various factors affecting this rate. Joshua and Garber (1990) used multiple linear and Poisson regression to estimate truck accident rates using traffic and geometric independent variables. Jones and Whitfield (1991) used Poisson regression with data from Seattle to identify the daily characteristics (traffic, weather, etc.) that may influence the number of traffic accidents. Miaou et al. (1992) used Poisson regression on traffic data from 8779 miles of roadway from the Highway Safety Information System (HSIS) to establish quantitative relationships between truck accident rates and highway geometric characteristics. Their results indicate that surrogate measures for mean absolute curvature (for horizontal alignment) and mean absolute grade (for vertical alignment) are the most important variables for accident rate estimation.

In a study of approximately seven thousand miles of roadway logs in Utah, Mohamedshah et al. (1993) used linear regression to predict truck accident involvement rate per mile per year, based on average Average Annual Daily Traffic (AADT) and truck AADT per lane, shoulder width, horizontal curvature, and vertical gradient. The results suggest that truck involvement rate increases with AADT and truck AADT, degree of curvature and gradient. Hadi et al. (1993), using data from the Florida Department of Transportation's Roadway Characteristics Inventory (RCI) system, estimated negative binomial (NB) regression for accident rates on various types of rural and urban highways with different traffic levels. Their results suggest that higher AADT levels and the presence of intersections are associated with higher crash frequency, while wider lanes and shoulders are effective in reducing crash rates. In that paper, the authors also provide an extensive review of earlier findings relating accident rates and geometric characteristics.

More recently, Ivan and O'Mara (1997), using NB regression on 1991–1993 data from the Traffic Accident Surveillance Report of Connecticut found that annual average daily traffic was a critical accident prediction variable, while geometric design variables and speed differential measures were not found to be effective predictors of accident rates. Karlaftis and Tarko (1998), based on a county accident data set from Indiana, estimated macroscopic accident models that attempt to explicitly control for cross-section heterogeneity in NB regression that may otherwise seriously bias the resulting estimates and invalidate statistical tests. Data collected from the States of Minnesota and Washington on rural two-lane highways, estimated accident models for segments and three-legged and four-legged intersections stop- controlled on the minor legs. Independent variables for their models included traffic, horizontal and vertical alignments, lane and shoulder widths, roadside hazard rating, channelization, and number of driveways. Results imply that segment accidents depend significantly on most of the roadway variables collected, while intersection accidents depend primarily on traffic.

This brief review of some of the existing literature suggests that a variety of traffic and design elements such as AADT, cross-section design, horizontal alignment, roadside features, access control, pavement conditions, speed limit, lane width (LW), and median width, affect accident rates. And, most of these results have been based on multiple linear or Poisson and NB regression models.

Much of the early work in the empirical analysis of accident data was done with the use of multiple linear regression models. As the literature has repeatedly pointed out, these models suffer from several methodological limitations and practical inconsistencies in the case of accident modelling (Lerman and Gonzales, 1980). To overcome these limitations, several authors used Poisson regression models that are a reasonable alternative for events that occur randomly and independently over time. Despite its advantages, Poisson regression assumes equality of the variance and mean of the dependent variable. This restriction (which, when violated, leads to invalid t-tests of the parameter estimates), can be overcome with the use of NB regression which allows the variance of the dependent variable to be larger than the mean. As a result, most of the recent literature has used NB regression models to evaluate accident data.

But, while NB regression has been instrumental in overcoming most of the problems associated with models involving count data, it still remains a parametric procedure requiring the functional form of the model to be specified in advance, it is not invariant with respect to monotone transformation of the variables, it is easily and significantly influenced by outliers, it does not handle well discrete independent variables with more than two levels, and it is adversely affected by multcollinearity among independent variables (Hadi et al., 1993; Mohamedshah et al., 1993; Tarko et al., 1996; Karlaftis and Tarko, 1998). It is likely, for example, that while the accident models have been correctly specified, multicollinearity has inflated the variance of some of the independent variables coefficient estimates, leading to lower t-statistic values and to coefficients that are not significant and/or are counter-intuitive.

In this paper a methodology which attempts to recognize the existence of the above mentioned problems and develop a framework to account for them is introduced. This methodology, known as HTBR or as Binary Recursive Partitioning (BRT) (Breiman et al., 1984), can be of assistance in overcoming some of the problems associated with multiple linear and NB regression. It should be noted that besides overcoming the above, rather theoretical problems, the proposed methodology has three additional strengths. First, it allows for straightforward and quantitative assessment of the effect of various rural road geometric characteristics on accident rates; second, it allows for the quick estimation of predicted accident rates for a given rural road section; and, third, it is easily amenable to 'if-then' statements for incorporation in expert systems which have become increasingly popular and useful in safety management. The strengths and weaknesses of the proposed methodology are demonstrated using Indiana State Police Accident Information records and Indiana Department of Transportation's Road Inventory database. The combined database includes five years (1991–1995) of crashes on Indiana rural roads, along with the geometric and traffic characteristics for these roads.

3. Data and methodology

3.1. The data

The data used in this paper concern rural roads and come from two sources: the Road Inventory database, from the Indiana Department of Transportation (INDOT), and the Accident Information Record form the Indiana State Police. The first database contains a list of road sections and various traffic and geometric characteristics for those sections. The second database contains a description of the location and type of accidents that occurred on Indiana's roads. Combining these two yields a database that contains five years (1991–1995) of accident data for Indiana along with the traffic and geometric characteristics for the location of each accident.

The availability of such data allows for inferences to be drawn on the effects of traffic and geometric characteristics on highway accidents. Further, to avoid the possibility of heterogeneity among roads with different number of lanes and based on previous findings in the literature (Hadi et al., 1993; Mohamedshah et al., 1993; Karlaftis and Tarko, 1998), road sections were grouped into two main categories: rural two-lane and rural multilane. The variables available for model estimation appear in Table 1.

3.2. The methodology

As previously mentioned, NB regression has accounted for most of the theoretical issues in count data research. Nevertheless, there still remain a number of issues that have not been addressed (Hadi et al., 1993; Mohamedshah et al., 1993; Tarko et al., 1996; Karlaftis and Tarko, 1998). First, NB regression, much like multiple linear and Poisson regression, is a parametric procedure requiring for the functional form of the model to be known in advance. Second, it is easily and significantly affected by outliers. Third, it cannot handle missing data well. Fourth, it does not treat satisfactorily discrete variables with more than two levels. Fifth, it does not deal well with multicollinear independent variables.

HTBR is a tree-structured non-parametric data analysis methodology that was first used in the 1960s in the medical and the social sciences (Morgan and Sonquist, 1963). An extensive review of the methods used to estimate the regression trees and their applications can be found in Breiman et al. (1984). HTBR is technically binary, because parent nodes are always split into exactly two child nodes, and is recursive because the process can be repeated by treating each child node as a parent. In essence, the HTBR algorithm proceeds by iteratively asking the following two questions: (i) which of the independent variables available should be selected for the model to obtain the maximum reduction in the variability of the response (dependent variable); and (ii) which value of the selected independent variable (discrete or continuous) results in the maximum reduction in the variability of the response. These two steps are repeated using a numerical search procedure until a desirable end-condition is met. In mathematical terms, deviance D is initially defined as[2]

[2] In this section only the essential parts of the HTBR methodology formulation that may be of interest to the reader are presented. Readers interested in the details of the formulation are encouraged to refer to Breiman et al. (1984) for an in-depth treatment, or Washington and Wolf (1996) and Washington et al. (1996) for a presentation of the methodology in the context of engineering applications. The discussion of HTBR presented in this paper is based on Washington and Wolf (1996).

$$D_a = \sum_{l=1}^{L} (y_{ia} - \bar{x}_a)^2, \qquad (1)$$

where D_a is the total deviance of a variable y at node a, or the sum of squared error (SSE) at the node, y_{ia} is the observation on dependent variable y in node a and is the mean of L observations in node a.

A split of the observations can be found at node a on a value of an independent variable x_1 that results in two branches and corresponding nodes b and c, each containing M and N of the original L observation ($M + N = L$). The goal of HTBR is to find the variable x_1 at its optimum split (i) so that the reduction in deviance is maximized, or more formally when

$$\Delta_{(\forall x)} = \text{maximum}. \qquad (2)$$

The maximum reduction occurs at some $x_{1(i)}$ (independent variable x_1 at value i). When the data are split at this value of x, the remaining two samples have much smaller variance of y than the original data set. Numerical search procedures are employed to maximize Eq. (2).

The HTBR methodology has several attractive technical properties: it is non-parametric and does not require specification of a functional form; it does not require variables to be selected in advance since it uses a stepwise method to determine optimal splitting rules; its results are invariant with respect to monotone transformations of the independent variables; it can handle data sets with complex (non-homegeneous) structure; it is extremely robust to the effects of outliers; it can use any combination of categorical and qualitative (discrete) variables; and, it is not affected by multicollinearity between the independent variables. Further, and as it pertains to this research, HTBR can straightforwardly yield predictions for the 'dependent' variable (y), incorporating the optimal splitting rules in an 'if-then' series of statements, making the incorporation of the results in an expert system rather simple.

4. HTBR model estimation and interpretation

As previously mentioned, HTBR partitions the data into relatively homogeneous (low standard deviation) terminal nodes, and it takes the mean value observed in each node as its predicted value. In general, HTBR models can be fairly complex and detailed, and therefore difficult to illustrate mathematically. Nevertheless, the methodology lends itself to graphical 'tree' like representations well.

The model shown in Fig. 1 is the result of the HTBR methodology applied to crashes on rural two-lane roads. Interpreting the tree, both for explanatory and predictive purposes, is rather straightforward. The top of the tree, or root node, shows that the first optimal split for crashes on rural two-lane roads occurs on

Table 1
Independent variables available for model estimation

Variable	Symbol	Type	Description
Section length	L	Continuous	Length of the road section were an accident occurred
Number of lanes	NoL	Count	Number of moving traffic lanes in the section
Lane widths	LW	Continuous	Widths of the northbound, southbound, and average lane widths
Shoulder widths	SW	Continuous	The widths of the left, right, inside, and outside shoulders
Median width	MW	Continuous	Width of the median (or 0 if median not available)
Shoulder type	ST	Qualitative	Dummy variables for type of shoulder (paved, earth, stabilized)
Pavement type	PT	Binary	The variables takes the value of 1 if the road surface is bituminous concrete, sheet or rock asphalt, and 0 otherwise
Concrete pavement	CP	Binary	The variable takes the value of 1 if the road surface is Portland concrete cement and 0 otherwise
Median type	MT	Qualitative	The variable takes the value of 0 if there is no median, 1 for grass or sod, 2 for bituminous concrete, and 3 for non-mountable barrier median
Turn lanes	TL	Binary	These variables indicate the presence of left, right, left and right, and continuous turn lanes
Number of curbs	NoC	Count	The number of curbs on the road section (0, 1, 2)
Number of park lanes	NoPL	Count	The number of parking lanes on the section (0, 1, 2)
Friction	FR	Continuous	Coefficient of wet sliding (skidding) FR at 40 mph between a wet pavement surface and a standard tire
Pavement Serviceability Index	SI	Qualitative	Takes the value of 0 for dirt and gravel roads, 1 for very poor, 2 for deteriorated, 3 for fair, 4 for good, and 5 for very good pavements
Access control	A	Qualitative	Takes the value of 1 for no access control, 2 for partial access control, and 3 for full access control
AADT	AADT	Continuous	Annual Average Daily Traffic

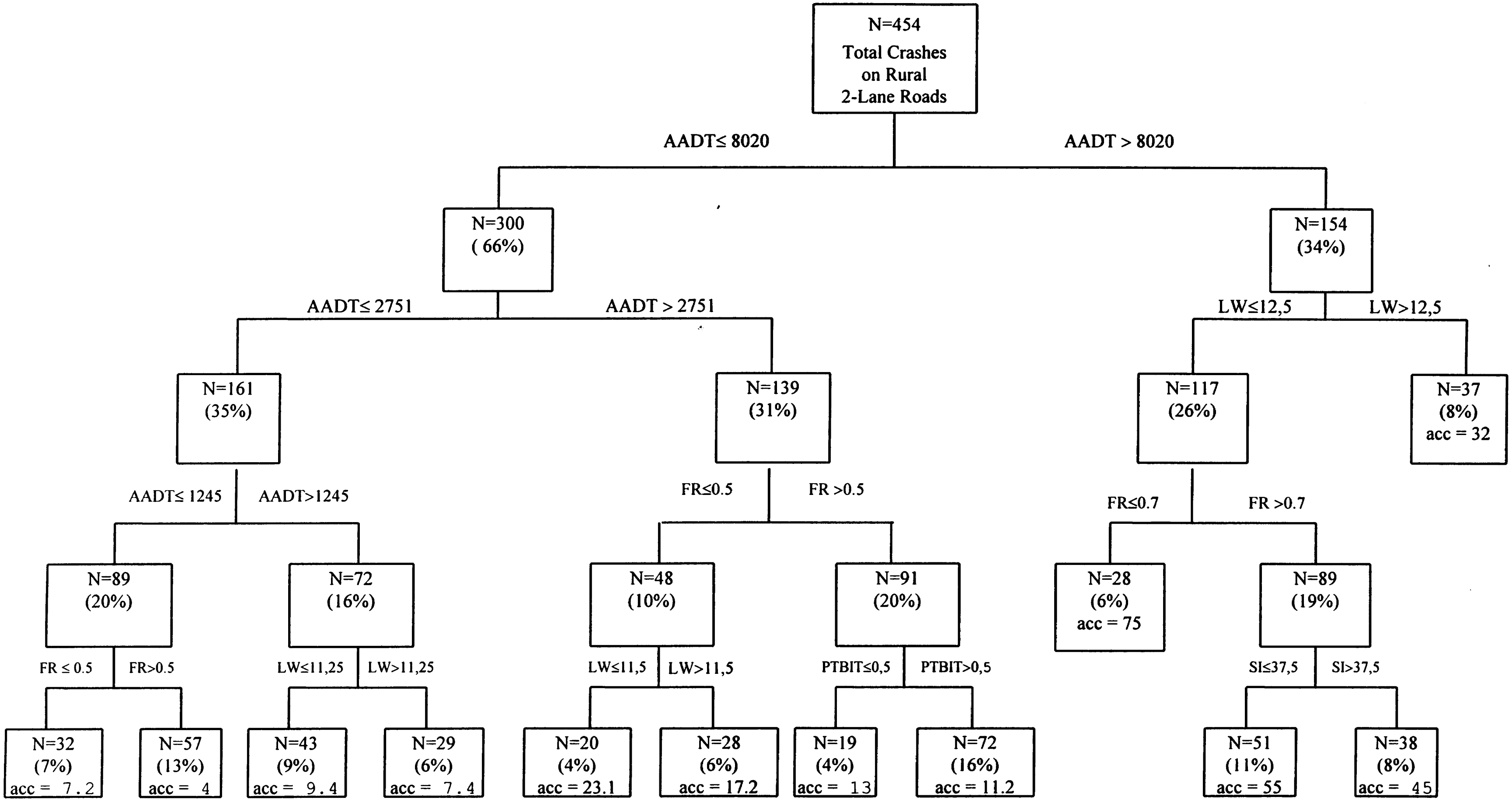

Fig. 1. Regression tree for accidents and geometric characteristics on rural two-lane roads.

AADT, sending cases (road sections) with less than or equal to 8020 to the left and all others to the right. In other words, the single best variable to explain the variability in total crashes on rural two-lane roads is AADT. Assume for the moment the interest is in rural roads with AADT larger than 8020. Conditional on this, the next best explanatory variable is LW. For LW less than or equal to 12.5 ft the road sections go to the left, where for LW larger than 12.5 ft the road sections go to the right forming what is called a terminal node, or leaf of the tree. For these road sections the tree predicts an average of 32 accidents (normalized on section length). The remaining splits, for the road sections with LW less than or equal to 12.0 ft, are made on Friction (FR) and Serviceability Index (SI). In general, an estimate on the number of accidents is obtained by continuing down the branches of the tree in similar fashion until a terminal node is reached. Recall that the estimate provided at terminal nodes is the mean of the sample at the node. This means that, since there is a number of observations that fall within the characteristics of a terminal node, the expected number of accidents is the mean of those observations. For example, there are 37 observations with $\mathrm{AADT} > 8020$ and $\mathrm{LW} > 12.5$, and their mean is 32 accidents.

More importantly, since transportation planners are very frequently interested in *predicting* accident rates for given highway sections, the profile of a road section can be examined, and the tree can be used to determine the prediction. For example, assume a planner wants to predict the expected number of accidents for a rural two-lane road section with AADT of 4000, FR of 0.5, and LW of 12.0 ft. Beginning at the root node (top of the tree), we branch left ($\mathrm{AADT} \leqslant 8020$), right ($\mathrm{AADT} > 2751$), left ($\mathrm{FR} \leqslant 0.5$), right ($\mathrm{LW} > 11.5$), to get the estimate of 17.2 crashes for that highway section.

It should be noted that, for the tree-structure, a X^2-test was used to evaluate the 'accuracy' of the predictions. Using a 'hold-out' sample of 120 randomly selected observations, the tree structure was estimated on the remaining 334 observations (for the rural two-lane road case). Then, using the 'if-then' rules yielded by the estimated tree, the accident rates for the 120 hold-out observations were estimated. At the 90% significance level for the X^2-test, the null hypothesis that the difference between the actual and predicted rates was zero could not be rejected.

Nevertheless, it should be noted that while the 'hold-out sample' method is a rather popular approach to validating the estimates yielded by the tree approach, it does have a shortcoming. Because both the estimation and validation samples are from the same general area (the State of Indiana), it is not surprising that their patterns are similar and hence the results of the model validation process are good. As such, it would be interesting to cross-validate the estimation results with data from a different area (but from rural roads nonetheless). In general of course, the process of randomly selecting a subsample for validation is the most frequently used technique.

Looking closer at Fig. 1, it is clear that for lower flows the parameter that seems to be more important is the FR coefficient, while for higher flows LW seems to have the greatest importance. This seems to be justified by the fact that lower flows are related to higher speeds, which render the slippery of the roads a critical parameter. However, when flows are high the risk for an accident seems to stem mainly from the interaction of vehicles travelling at the same or opposite direction, rendering LW the more important factor.

Fig. 2 shows the results of the HTBR methodology applied to crashes on multilane rural roads. Interestingly, using again the X^2-test, the methodology yielded a 'simpler' tree. Its first optimal split occurs on AADT. Thus, it seems that AADT is the best variable to explain crash variability in multilane roads as well. What may also be noted is that for lower flows the existence of a median is an important factor while when it comes to higher flows the existence of access control seems to be the more important factor safety wise. Thus, vehicle interaction and vehicle maneuvering arrangements prove again to be important factors when traffic demand increases. For *predictive* purposes, the profile of a multilane road section can be examined similar to that of a two-lane road section and the tree can be used to determine the prediction. For example, assume a prediction is needed for the number of accidents on a rural multilane road section with AADT of 8300, and No Access Control. Beginning at the root node (top of the tree), we branch right ($\mathrm{AADT} > 6851$), right ($\mathrm{AADT} > 8075$), left ($A = 1$), to get the estimate of 27.1 crashes for that highway section.

It is interesting to note that some variables are selected more than once in the estimation process. For instance, taking all the left branches to the terminal node (leaf), AADT appears three times. Since one of the goals of HTBR is to develop a simple tree structure for data, relatively few variables will appear explicitly in the splitting criteria, and some highly important variables will appear more than once (such as AADT in this tree structure). While this could be taken to mean that the other variables are not important in understanding or predicting the dependent variable, an independent variable could be considered highly important even if it never appears as a primary node splitter. The software used in this paper (CART 1995) keeps track of surrogate splits in the tree-growing process, evaluating the contribution a variable makes in prediction by both primary and surrogate splits. That is, while the tree-structure can be used, as previously shown, for predictive purposes, a different measure called variable

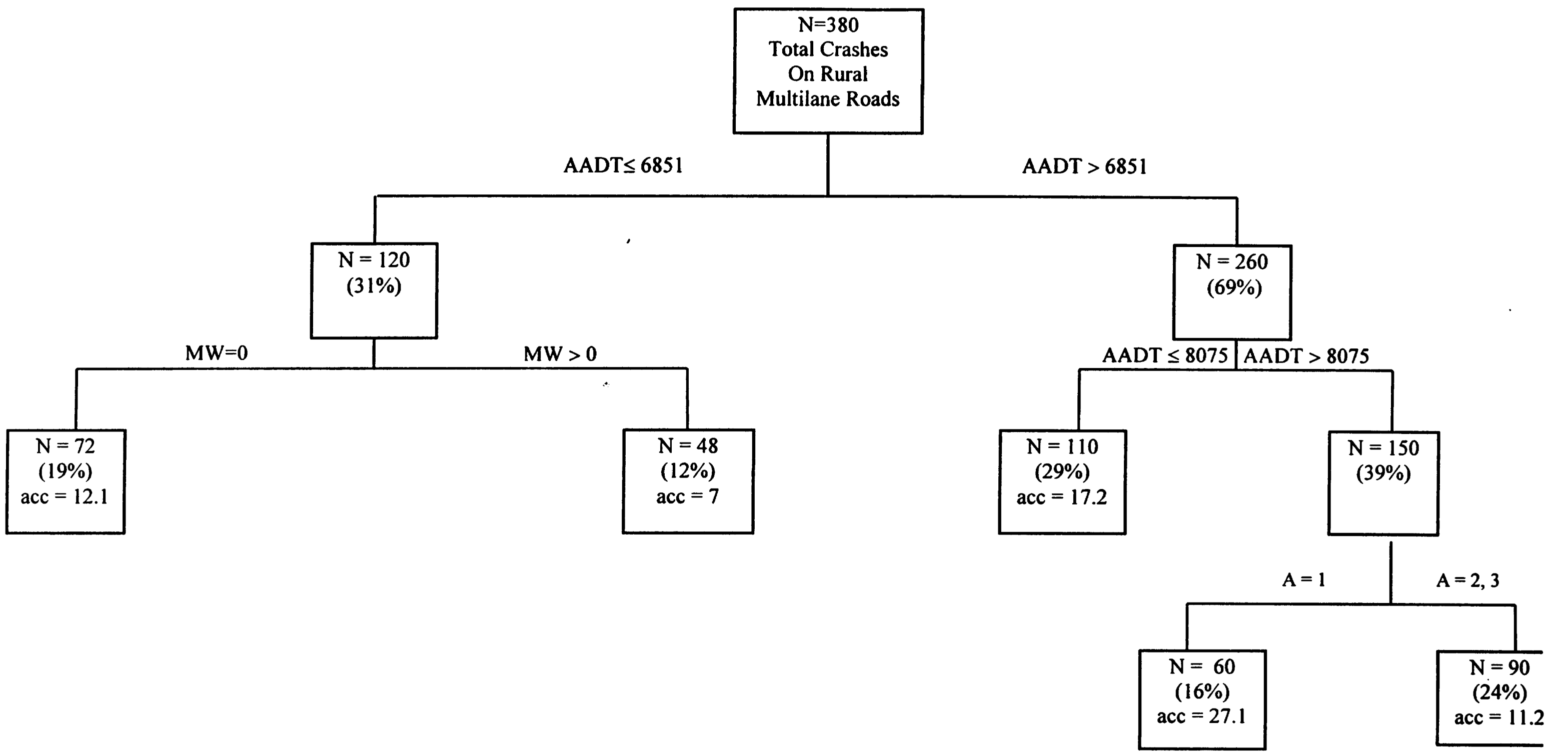

Fig. 2. Regression tree for accidents and geometric characteristics on rural multilane roads.

Table 2
Independent variable importance for crash rates (crashes normalized on highway section length)

Rural two-lane		Rural multilane	
Variable	Relative importance (%)	Variable	Relative importance (%)
AADT	100	AADT	100
Lane width	72	Median width	63
Serviceability index	59	Access control	59
Friction	32	Friction	25
Pavement type	30	Lane width	24
Access control	14	Serviceability index	21
		Pavement type	11

importance score should be used to estimate the importance of the effect of various geometric characteristics on accident rates.

To calculate a variable importance score, the software looks at the improvement measure attributable to each variable in its role as a surrogate to the primary split. The values of these improvements are summed over each node and totalled, and are then scaled relative to the 'best' performing variable. As a result, the variable with the highest sum of improvements is scored 100, and all other variables will have lower scores ranging downwards towards zero. The relative importance of the independent variables in explaining crash rates on various types of roadways appear in Table 2 (for crashes normalized on highway section length), and Table 3 (for crashes normalized on highway section length and AADT).

It is interesting to note the differences in the variables that 'explain' crashes on the two types of roadways. While AADT is overall the most important variable when crashes are normalized on section length (Table 2), the characteristics of subsequent importance vary for the two types of roadway. For the rural two-lane case, LW is the variable with the higher importance after AADT. It is obvious that the proximity of the opposing traffic streams renders the width of the lane an important factor for safety. The next more important variables – SI, FR and pavement type – are related to the road pavement conditions.

However, when it comes to multilane rural roads the variables with the higher importance after AADT are the existence of median width and of access control. The importance of these two factors seems to be justified mainly by the increased speeds on multilane rural roads. This fact renders the above two factors more important than LW and pavement condition variables, FR, SI and pavement type, which follow in importance (Table 2). Furthermore, it should be noted that when crashes are normalized on section length and AADT (Table 3), the variables of importance are similar to those of Table 2 (normalization on section length), the only new variable being the existence of a left turn lane, for both two-lane and multilane rural roads.

5. Discussion and conclusions

Much interest exists in the area of accident rate estimation, and the identification of the various factors affecting this rate. Much of the literature in this area has concentrated in identifying the factors affecting accident occurrence (accident rates), and secondarily in *predicting* them. The ability to predict accident rates is very important to transportation planners and engineers, because it can help in identifying hazardous locations, sites which require treatment, as well as spots where deviations (either higher or lower rates) from expected (predicted) levels warrants further examination. The aim of this paper was twofold. First, it developed a methodology that quantitatively assesses the effects of various highway characteristics on accident rates. Second, it provided a straightforward, yet fundamentally and mathematically sound way of *predicting* accident rates.

The methodology used in this paper, known as HTBR, has a number of both theoretical and applied advantages over multiple linear and NB regression that have been commonly used in accident rate research. It allows for the quantitative assessment of the effect of various geometric characteristics on accident rates. It allows for the quick estimation of predicted accident rates for a given highway section. Finally, it is easily amenable to 'if-then' statements for incorporation in expert systems, which have become increasingly popular and useful in safety management. The methodology was demonstrated using data from the Indiana State Police Accident Information records and the INDOT's Road Inventory database.

The results of the investigation of the roadway characteristics that affect accident rates are of interest. It is clear that for both rural two-lane and multilane roadways AADT is the most important variable. However, the factors of subsequent importance vary for each

Table 3
Independent variable importance for crash rates (crashes normalized on highway section length and AADT)

Rural two-lane		Rural multilane	
Variable	Relative importance (%)	Variable	Relative importance (%)
Lane width	100	Median width	100
Serviceability index	89	Access control	73
Pavement type	62	Friction	55
Friction	22	Lane width	25
Left turn	16	Serviceability index	19
		Left turn	16

case. Looking closely at the results of accident rates normalized on AADT (which cancels out the effect of AADT), it can be generally inferred that LW and pavement condition factors – SI, pavement type and FR – are the most important variables affecting crash rates for the two-lane case. The importance of LW seems to increase with higher flows. On the contrary, the importance of pavement condition factors seems to increase with lower flows due to higher speeds.

For rural multilane roads, with the effect of AADT cancelled out, median width and access control are the most important factors followed by pavement condition factors. It is worth mentioning that the importance of access control seems to increase with heavier traffic that renders vehicle maneuvering arrangements critical, while the existence of a median becomes more important in low flow conditions. Although the importance of isolated variables differs for the two roadway types it is obvious that 'geometric design' captured through LW and access control and 'pavement condition' captured through FR, SI and pavement type are, as expected, the two most important factors affecting accident rates.

The methodology used in this paper also allows the explicit prediction of accident rates for given highway sections. As soon as the profile of a road section is given, predictions regarding the expected accident rates can be obtained. In essence, when the AADT, LW, SI and FR of a road section are known, predictions can be obtained. Further, the 'if-then' rules for obtaining these predictions can be easily incorporated in an expert system that can automate the accident rate prediction effort. The work presented in this paper is part of the larger effort to tackle the problem of accident occurrence on the world's roadways. The extremely high cost of highway accidents paid by societies makes highway safety improvement maybe the most important objective of transportation engineering. This effort's eventual goal is to reduce injuries and fatalities due to highway design and maintenance deficiencies.

References

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees, Wadsworth International Group, Belmont, CA.

Hadi, M.A., Aruldhas, J., Chow, L.F., Wattleworth, J.A., 1993. Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression. Transportation Research Record, 1500, TRB, National Research Council, 169–177.

Jones, I.S., Whitfield, R.A., 1991. Predicting injury risk with new car assessment program crashworthiness ratings. Accident Analysis and Prevention 6 (20), 411–419.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvement using linear and poisson regression models. Transportation Planning and Technology 15, 41–58.

Karlaftis, M.G., Tarko, A., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30 (4), 425–433.

Lerman, S.R., Gonzales, S.L., 1980. Poisson regression analysis under alternate sampling strategies. Transportation Science 14 (4), 346–364.

Miaou, S.P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C., 1992. Relationship between truck accidents and highway geometric design: a poisson regression approach. Transportation Research Record 1376, 10–18.

Mohamedshah, Y.M., Paniati, J.F., Hobeika, A.G., 1993. Truck accident models for interstate and two-lane rural roads. Transportation Research Record 1407, 35–41.

Morgan, J.N., Sonquist, J.A., 1963. Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association 58, 415–434.

Ivan, J.N., O'Mara, P.J., 1997. Prediction of Traffic Accident Rates Using Poisson Regression, Presented in the 1997. Transportation Research Board Meeting, Washington, DC.

Tarko, A.P., Sinha, K.C., Farooq, O., 1996. A Methodology for Identifying Highway Safety Problem Areas, Presented in the 1997. Transportation Research Board Meeting, Washington, DC.

US DOT, 1969. National Highway Functional Classification Study Manual. Federal Highway Administration, Washington, DC.

Washington, S., Wolf, J., 1996. Hierarchical Tree-Based versus Ordinary Least Squares Linear Regression Models: Theory and Example Applied to Trip Generation. Presented in the 1996. Transportation Research Board Annual Meeting, Washington, D.C.

Washington, S., Wolf, J., Guensler, R., 1996. A Binary Recursive Partitioning Method for Modeling Hot-Stabilized Emissions from Motor Vehicles. Presented in the 1996. Transportation Research Board Annual Meeting, Washington, DC.