Injury Accident Prediction Models for Signalized Intersections

Michael Yiu-Kuen Lau and Adolf D. May, Jr.

An intuitive methodology for developing accident prediction models for signalized intersections based on the Traffic Accident Surveillance and Analysis System (TASAS) in California is illustrated. A fairly new grouping and classifying technique called Classification and Regression Trees (CART) was used as a building block for developing injury accident models. The proposed methodology is a three-level procedure with a "tree" structure for easy interpretation and application. Macroscopic-type models for injury accidents per year are derived, and the following factors have been found to be significant: traffic intensity, proportion of cross street traffic, intersection type, signal type, number of lanes on cross streets and main streets, and left turn arrangements. On the basis of the results, it is also apparent that the models derived from the proposed methodology and TASAS provide more intuition and flexibility than the existing models used in California and other models derived from both site observations and accident record systems.

The purpose of this paper, which is based on a project funded by the California Department of Transportation (Caltrans) and the FHWA, is to develop accident prediction models for signalized intersections on the basis of intersection characteristics such as geometric design elements, traffic control measures, traffic demand patterns, environmental factors, and accident history. The data base for the study was derived from the Traffic Accident Surveillance and Analysis System (TASAS) maintained by Caltrans and the California State Highway Patrol (*1*). One of the themes of study is the investigation of whether prediction models can be developed on the basis of existing accident record systems. Major investments of time and effort have been made in these systems, and they also require continual efforts by state and federal agencies for their maintenance.

Because it was not the aim of the study to collect additional data for model development in addition to the information contained in TASAS, it was assumed that macroscopic models that use existing data would suffice. The term "macroscopic models" refers to models that are not derived on the basis of detailed information and do not require such information, including turning movement counts, headway distributions, and so forth—details that are not usually available in accident record systems. Furthermore, classification of accidents for this kind of model is confined to injury, fatal, and property damage only (PDO) accidents. The advantages include easy comprehension and simple translation into monetary terms, which are required by most economic and feasibility analyses. Disadvantages include inadequacy in reflecting the process of collisions and the concept of conflicts.

The issue of systematically grouping entities (such as intersections) with similar accident patterns on the basis of different control, design, and demand features of the entities has not been examined critically in previous studies. A powerful and systematic tool for grouping or classifying entities can be a very important building block in developing accident prediction models because lack of homogeneity within subgroups classified by nonsystematic methods could introduce biases in prediction. A fairly new grouping and classifying technique called Classification and Regression Trees (CART), developed by Breiman at the University of California, Berkeley (*2*), was used in the study for this purpose.

This paper describes a methodology and specific techniques for developing prediction models for injury accidents. The proposed methodology is based on the results and comments of the pilot study, working paper, and review process, as described elsewhere (*3*). Injury accidents here include accidents in which the people involved have injuries ranging from very slight to serious. It is believed that the reporting level for injury accidents is about 80–90 percent, whereas the reporting level for PDO accidents is believed to be about 50 percent or less. There is also a difference in reporting levels among different jurisdictions. In contrast, fatal accidents are rare events, and only 0.5 percent of all intersection accidents are fatal. In view of these facts, models for injury accidents were developed as a first step for the project. The application of our three-level prediction procedure to some 2,500 signalized intersections that are under the control of the state of California and contained in TASAS is illustrated.

DATA BASE: TASAS

The data base used in the study is basically a simplified version of the TASAS system, containing information on about 2,500 signalized intersections and 122,000 accidents that occurred at these intersections from 1979 through 1985. Some of the intersections have shorter reporting periods. This is a reflection of changes in design, control, and so on.

Intersection-Related Characteristics

There are 2,498 signalized intersections in the data base, 95 percent of which are located in urban areas. Some of their characteristics are briefly discussed in the following paragraphs. Detailed information may be found in Table 1 and elsewhere (*3*).

Institute of Transportation Studies, Civil Engineering Department, University of California, Berkeley, Calif. 94720.

TABLE 1 FACTORS, LEVELS, AND PERCENTAGES OF INTERSECTION CHARACTERISTICS

Factor	Percentage	Levels (Notes)
Terrain	0.9	1 (mountainous)
	17.7	2 (rolling)
	81.4	3 (flat)
Design speed (mph) (SPEED)	5.0	1 (less than 30)
	1.4	2 (30–34)
	8.9	3 (35–39)
	5.3	4 (40–44)
	13.1	5 (45–49)
	12.0	6 (50–54)
	5.6	7 (55–59)
	26.8	8 (60–64)
	21.8	9 (greater than 65)
Rural/urban (RORU)	4.4	1 (rural)
	95.6	2 (urban)
Inside/outside city (IORO)	87.0	1 (inside)
	13.0	2 (outside)
Intersection type (ITYPE)	72.5	1 (four-legged)
	2.8	2 [multilegged (>4)]
	3.9	3 (offset)
	18.2	4 ("T")
	1.0	5 ("Y")
	1.5	6 (other)
Control type (CTYPE)	22.7	1 (pretimed two-phase)
	4.2	2 (pretimed multiphase)
	10.2	3 (semi–traffic actuated two-phase)
	7.2	4 (semi–traffic actuated multiphase)
	11.1	5 (full traffic actuated two-phase)
	44.6	6 (full traffic actuated multiphase)
Lighting type (LIGHT)	0.8	1 (no lighting)
	98.8	2 (lighted)
Main-line signal mast arm (MSM)	8.3	1 (no mast arm)
	91.2	2 (signal on mast arm)
	0.5	3 (missing)
Main-line left turn channelization (MLT)	42.3	1 (curbed median left turn channelization)
	20.2	2 (no left turn channelization)
	36.5	3 (painted left turn channelization)
	0.6	4 (raised bars)
	0.4	5 (missing)
Main-line right turn channelization (MRT)	80.1	1 (no free right turns)
	19.7	2 (provision for free right turns)
	0.2	3 (missing)
Main-line traffic flow (MTF)	4.0	1 (two-way traffic, no left turns permitted)
	89.0	2 (two-way traffic, left turns permitted)
	0.8	3 (two-way traffic, left turns restricted during peak hours)
	5.2	4 (one-way traffic)
	1.0	5 (other)
Main-line number of lanes (MNL)	9.2	2 (two lanes)
	3.9	3 (three lanes)
	66.9	4 (four lanes)
	2.2	5 (five lanes)
	16.5	6 (six lanes)
	0.2	7 (seven lanes)
	0.7	8 (eight lanes)
Main-line ADT (MADT)		(packed numeric 999999)
Cross street signal mast arm (XSM)	46.3	1 (no mast arm)
	53.2	2 (signal on mast arm)
	0.5	3 (missing)
Cross street left turn channelization (XLT)	13.8	1 (curbed median left turn channelization)
	58.7	2 (no left turn channelization)
	26.7	3 (painted left turn channelization)
	0.2	4 (raised bars)
	0.6	5 (missing)
Cross street right turn channelization (XRT)	78.0	1 (no free right turns)
	21.6	2 (provision for free right turns)
	0.4	3 (missing)
Cross street traffic flow (XTF)	2.2	1 (two-way traffic, no left turn permitted)
	90.5	2 (two-way traffic, left turn permitted)
	0.2	3 (two-way traffic, left turn restricted during peak hours)
	6.4	4 (one-way traffic)
	0.7	5 (other)
Cross street number of lanes (XNL)	57.4	2 (two lanes)
	6.0	3 (three lanes)
	32.1	4 (four lanes)
	0.7	5 (five lanes)
	2.3	6 (six lanes)
	0.0	7 (seven lanes)
	0.1	8 (eight lanes)
Cross street ADT (XADT)		(packed numeric 999999)
Median indicator	23.4	1 (undivided)
	76.2	2 (divided or independent) (alignment)
IADT		[= XADT/(XADT + MADT)]

There are ~72 percent four-legged and ~18 percent "T" intersections. The remainder are multilegged, offset, or "Y" intersections. Two thirds (~66 percent) of the intersections have four lanes on main streets, and ~57 percent have only two lanes on side streets. For left turn arrangements on main streets, 42 percent have curb median left turn channelization, and 36 have painted left turn channelization. On side streets, 14 percent have curb median left turn channelization, and 26 percent have painted left turn channelization. About half have design speeds of less than 55 mph. Some 44 percent of the intersections are controlled by multiphase, fully actuated traffic controllers, and 22 percent are controlled by two-phase, pretimed traffic controllers. About 10 percent have two-phase, semi-traffic-actuated controllers, and 11 percent have two-phase, fully actuated controllers. On main streets, 89 percent have two-way traffic with left turn permitted, and 4 percent have two-way traffic with no left turns permitted. About 5 percent have one-way traffic. The percentages for side streets are 90 percent two-way with left turn, 2 percent two-way without left turn, and 6 percent one-way. The median average daily traffic (ADT) on main streets is ~27,000, and the maximum is ~72,000. The median ADT on side streets is ~17,500, and the highest ADT is ~45,000.

Accident-Related Characteristics

There are ~122,000 accident records in the data base. These are accidents that occurred within 250 ft of the 2,498 intersections, from 1979 to 1985.

The majority of the accidents (76 percent) occurred in clear weather conditions, and only 12 percent of them occurred in cloudy conditions and 8 percent of them in rainy conditions. It was also recorded that 35 percent of the accidents were rear end collisions and 31 percent were broadside collisions. Only 2.4 percent of them were automobile-pedestrian collisions. The percentage of fatal accidents with one or more people killed was only ~0.5, and the percentage of injury accidents was ~38. The remaining were PDO accidents. About 75 percent of the accident reports noted that the parties involved had not been drinking, and only ~5 percent of them reported that those who were involved were under the influence of alcohol.

METHODOLOGY

Methodology is a systematic study of the subject in question, including such tasks as definition of objectives, selection of measures of effectiveness, generation of solutions, refinement of solutions, selection of models, and so on. A flowchart showing the proposed methodology used here, based on this principle, is shown in Figure 1, and the details of the procedure are discussed in the sections that follow.

The study of accidents has been regarded as a controversial and indefinite field by many researchers because occurrences of accidents are highly stochastic in nature. Because of this, a fairly long record of accident history would be required for any meaningful study. A long accident history also has its disadvantages. The question of changes in both basic and operating characteristics of entities and in methods of collecting and recording accidents during the period must be addressed. Furthermore, accident data and records are usually regarded as "incomplete" because useful information for model building, such as vehicle condition and design, driver's characteristics and behavior, and so on, is usually not contained in such systems.

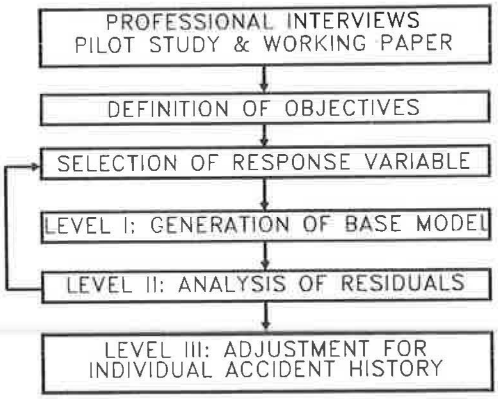

PROFESSIONAL INTERVIEWS
PILOT STUDY & WORKING PAPER
DEFINITION OF OBJECTIVES
SELECTION OF RESPONSE VARIABLE
LEVEL I: GENERATION OF BASE MODEL
LEVEL II: ANALYSIS OF RESIDUALS
LEVEL III: ADJUSTMENT FOR
INDIVIDUAL ACCIDENT HISTORY

FIGURE 1 The proposed methodology.

Approach Objectives

Because of the problems mentioned previously, the proposed three-level estimating procedure has been derived with the following objectives in mind:

- To develop an approach that is intuitive and yet is reflective of the stochastic nature of accident occurrences;
- To develop a staged approach that can allow users to appreciate the importance and consequences of the process, in contrast to some "black box" approaches; and
- To develop a systematic approach that is capable of extracting some patterns from the fairly "incomplete" data sets in most accident record systems.

Selection of Response Variable

Selection of a response variable is a very important step in the process because this variable determines, to a large extent, the final model. The selection of the preferred model may hinge on the choice of the response variable. This variable, which is also commonly known as the dependent variable, is a measure of performance of the system, for example, the risk level of an intersection.

In this paper, only injury accidents are addressed because of the different reporting levels of PDO accidents and rarity of fatal accidents. The first task is to find an appropriate derivative of injury accidents for comparison and evaluation purposes because common sense indicates that it is not reasonable to compare an intersection with 1 accident in 1 year with another intersection with 1 accident in 10 years. For this purpose, normalization by time is a logical step. The next matter is a further normalization by exposure measures such as traffic intensity or traffic intensity-distance. This is a common practice; however, simple reasoning indicates that the attitude of "why bother" should prevail, if possible. As a result, the number of injury accidents per year was initially selected as our response variable. It was found that no further transformation was required because this response variable appeared to be adequate.

Generation of the Base Model

Instead of putting some form of the traffic intensity variable in the denominator of the response variable, a base model was built, with injury accidents per year as the response variable and traffic intensity as the predictor variable. One of the advantages of this approach is that it allows researchers to see the relationship between the two variables in an undistorted manner, such as a scatter plot. On the basis of the untransformed information in a graph, different functional forms to model the relationship between the two variables can be tried. Estimates of the parameters can be obtained by techniques such as least squares, maximum likelihood, and so on. The base model so obtained is called a Level I prediction in this paper.

Analysis of Residuals and Grouping of Intersections by CART

Further details, such as design, control, degree of conflicts, and environmental features of the intersections, are also considered to be major factors affecting safety at intersections. The importance of these factors can sometimes be reflected in the large variations between injury accidents and traffic intensity found in most scatter plots. One of the approaches that can be used is to analyze the residuals of the base model on the basis of other characteristics of the intersections. In other words, those intersections with similar characteristics that have higher or lower accident records than other intersections in general would be grouped together. The residual is defined as the difference between the observed value and the value predicted by the base model.

The next question is how many groups should be selected to represent high– or low–accident risk intersections. Extreme solutions include one and *n* groups, where *n* is the number of intersections in the data base. With a single group, the model is equivalent to the base model and is therefore not interesting, because some understanding of the design factors that tend to affect the safety of intersections should be provided. If there are *n* groups, it might be possible that the given characteristics of the intersections cannot be used to produce a grouping that reflects similar accident patterns. Also, if *n* groups are used, there is no way to identify those intersections that are "out of line" for accident surveillance purposes. Engineering judgment and a technique to group intersections with error measures would therefore be important to the process of getting an optimal group size.

The CART (Classification and Regression Trees) program can be used to analyze the residuals of the base model and then group intersections with similar accident patterns. The refinement of estimates by grouping intersections on the basis of other intersection characteristics is called the Level II prediction in this paper.

Adjustment by Accident History

The materials just covered refer to estimates of accident prediction for groups of intersections. In other words, all intersections within a certain group will have an equal estimate. It can be argued that the grouping made or models derived were only based on information that is available in the list of predictor variables, which may not contain all the factors involved. The kind of information that cannot be captured by the models or variations within groups can be found easily in the accident history of individual intersections. As a result, the accident history of individual intersections could become a very valuable source of information that reflects the safety level of individual intersections.

Level I and Level II predictions refer to group estimations or predictions, whereas the Level III prediction is aimed at the concept of safety estimation at individual intersections. The idea of linear combination of group estimate and individual accident history, as proposed by Hauer and Persaud (*4*), was used in the study. A detailed discussion on Level III prediction and its application to the proposed methodology is presented later in this paper. This staged procedure could allow flexibility to users that have different input requirements. At the same time, it gives them an opportunity to observe the evolution of their estimates.

CLASSIFICATION AND REGRESSION TREES (CART)

CART is a computer program and technique developed by Breiman et al. to classify and group entities on the basis of a set of measurements or characteristics, using tree methodology (*2*). CART is particularly useful to this study because the data set has high dimensionality, a mixture of data types (continuous and categorical), and (perhaps) lack of homogeneity. Lack of homogeneity refers to different relationships that hold between variables in different parts of the measurement space. Most important of all, the tree structure output provides information regarding the main factors and interactions between factors that are important in predicting accidents in a form that is easily interpreted and understood. CART differs from other tree structured programs in the pruning and estimation process, that is, in the process of "growing an honest tree" (*2*, Chapter 8).

Concept and Theory

The concept involves partitioning the intersections by a sequence of binary splits into terminal nodes, as shown in Figure 2. In each terminal node t, the predicted response value $y(t)$ is a constant. Because the predictor $d(X)$ is constant over each terminal node, the tree can be thought of as a histogram estimate of the regression surface, as shown in Figure 3. In developing predictive models by a tree technique with a learning sample, L, the following three issues for determining a tree predictor should be addressed:

- A way to select a split at every intermediate node,
- A rule for determining when a node is terminal, and
- A rule for assigning a value $y(t)$ to every terminal node t.

A detailed discussion on these subjects can be found in Breiman's work (*2*). To put it in simpler terms, the issue of the node assignment rule is the easiest to resolve, and it can be shown that the average of y_n, the observed value of the response variable, in terminal node t could be used to represent the group.

A regression tree is formed by iteratively splitting nodes, and one way to select a split at every intermediate node is to

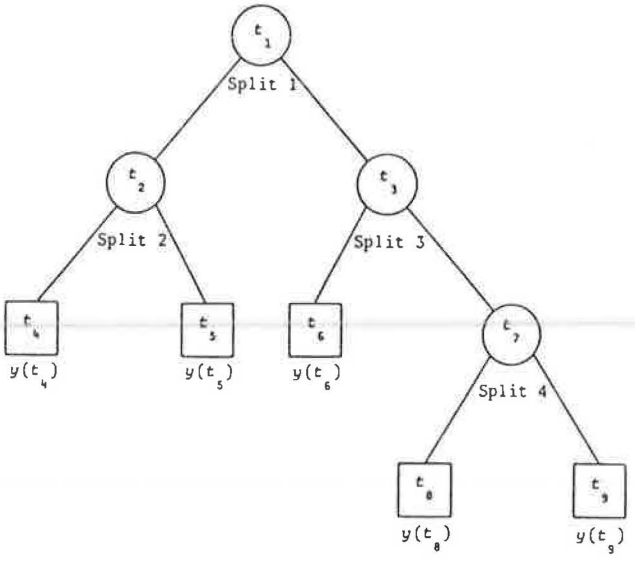


FIGURE 2 Partitioning of intersections.

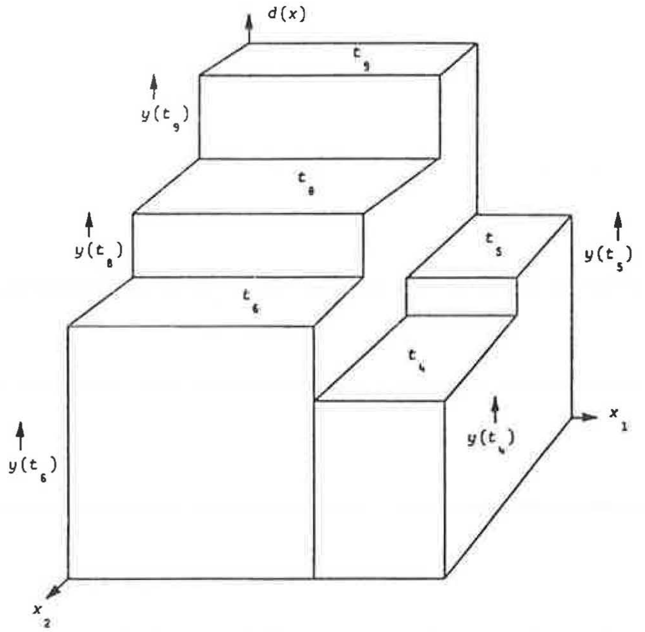


FIGURE 3 Histogram estimate at nodes.

maximize the decrease in the resubstitution error measure. The best split of t is the split that minimizes the weighted variance

$$P_L S^2(t_L) + P_R S^2(t_R) \tag{1}$$

P_L and P_R are the proportions and $S^2(t_L)$ and $S^2(t_R)$ are sample standard deviations of the cases in t that go left and right, respectively. Furthermore, this value is also smaller than $S^2(t)$, which is the sample variance of the y_n values in node t. In essence, the best split of t here corresponds to the most important factor that should be selected at t for splitting purposes.

The rule for determining whether a node is a terminal node is related to the issue of pruning of trees and can also be influenced by some options specified by users, such as minimum node size, tree selection rule, and so on.

Notes on the CART Program and Its Applications

The CART program is written in standard Fortran and can be run on most computers in both interactive and batch modes. One of the most important options that is available is the estimation method. This is the method specified by users to calculate the accuracy of models. There are three methods available in the CART program: resubstitution, test sample, and cross-validation.

Resubstitution uses all the cases in the learning sample, L, to construct predictor $d(X)$ and to estimate its error $R^*(d)$. For the resubstitution estimate, the following equation could be used:

$$R(d) = (1/N) \sum_n [y_n - d(X_n)]^2 \tag{2}$$

The problem with the resubstitution estimate is that it is computed with the same data used to construct the predictor $d(X)$ instead of an independent sample. This estimate therefore tends to give an overly optimistic picture of the accuracy of the predictor. Trees built by this method are sometimes called "exploratory trees." This is a rapid method of growing exploratory trees, and it can be used to explore a range of parameters, the effects of adding or deleting variables during preliminary stages, or both.

The second method is test sample estimation. This method randomly divides the cases, L, into two sets, L_1 and L_2. Only the cases in L_1 are used to construct predictor $d(X)$. The cases in L_2 are used to estimate $R^*(d)$ by the following equation:

$$R^{ts}(d) = (1/N_2) \sum_{(X_n,\, y_n) \in L_2} [y_n - d(X_n)]^2 \tag{3}$$

Care should be taken so that cases in L_1 are independent of cases in L_2. The drawback of this method is the reduction in effective sample size; thus the test sample estimation method is only recommended for problems with a large sample size.

For smaller sample sizes, the V-fold cross-validation method is recommended. The cases in L are randomly divided into V subsets of as nearly equal size as possible. For each v, $v = 1, \ldots, V$, apply the same construction procedure to the learning sample $L - L_v$, so that the predictor $d^{(v)}(X)$ is obtained. Then the estimate can be computed as follows:

$$R^{cv}(d) = (1/N) \sum_v \sum_{(X_n,\, y_n) \in L_v} [y_n - d^{(v)}(X_n)]^2 \tag{4}$$

Cross-validation is the recommended procedure because it provides "honest" trees. It is preferred over test sample procedure because every case is used in constructing the predictor and calculating error estimates.

Options are also available for selecting different sizes of optimal trees. These options are explained in detail in the work of Breiman et al. (2).

ACCIDENT PREDICTION MODELS

It is generally true that the accuracy of prediction models depends on the details of the information base on which the models are built. This relationship should not depend on the type and complexity of the models or techniques used. In other words, it should be possible to improve the accuracy of a model by feeding more information into the process. This pattern was also found in the current study. Evidence is presented in the examples given at the end of this section. The initial parts of the section describe the application of the proposed procedure.

Level I: Model Based on Traffic Intensity

It has been demonstrated in this data set that traffic intensity was indeed the most important single factor in predicting injury accidents. A base model was built on this fact. In this paper, traffic intensity is expressed as the total number of vehicles (in millions) entering an intersection per year, calculated from average daily traffic on cross and main streets. A scatter plot with the average number of injury accidents per year (IACCYR) on the vertical axis and millions of vehicles entering an intersection from all legs per year (MVYR) on the horizontal axis is shown in Figure 4. There are only 2,488 intersections or points in the figure because 10 intersections had incomplete information.

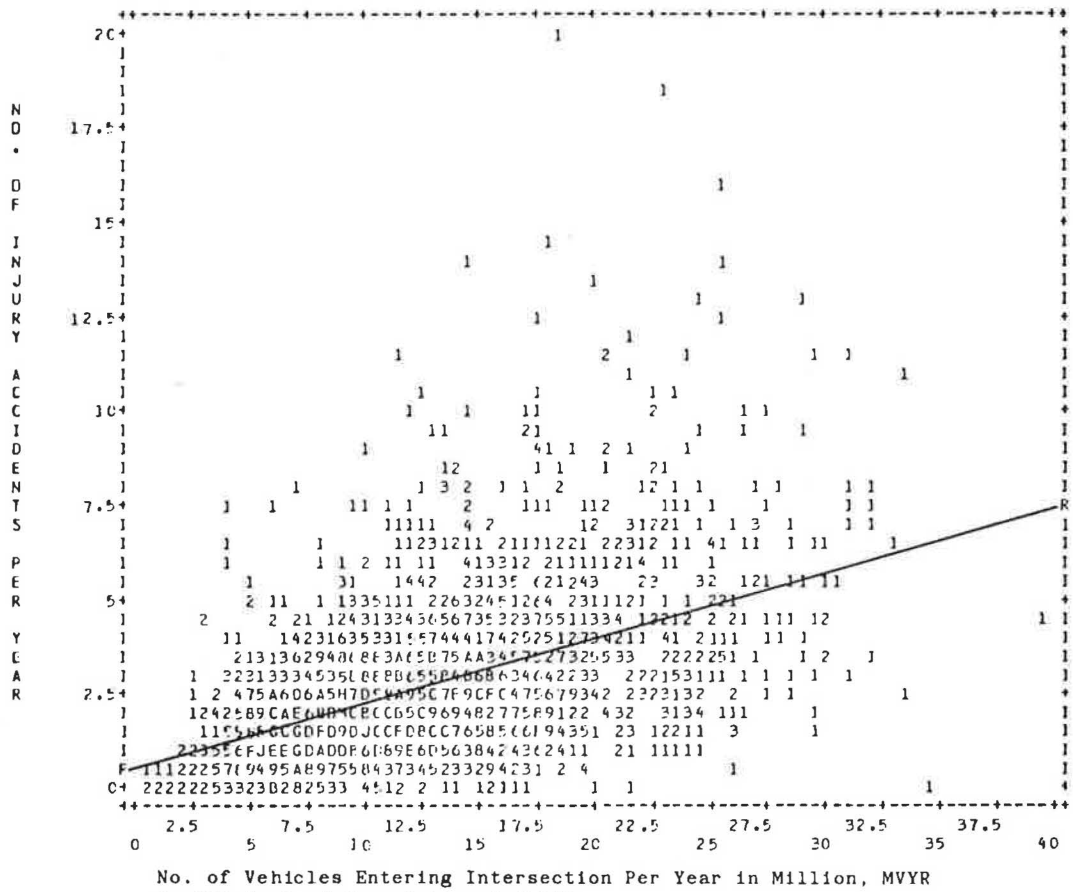
Different functional forms, including power and logarithm transformations, have been used to fit the data, and a straight line relationship was finally selected because it was as good as any other form. Estimates of slope and intercept were then obtained by regression analysis, as shown in Figure 4. Other estimation methods could have been used but were considered unnecessary. The following equation was selected as the base model for injury accidents per year:

$$\text{FIACCYR} = 0.61856 + 0.16911 * \text{MVYR} \quad (5)$$

The question of nonzero intercept was also discussed, and it was decided that no adjustment should be made, simply because such an adjustment would introduce biases in the estimation process and the difference would be minimal. However, when MVYR is less than 3.0, a warning to users should probably be made. On the other hand, no consistent biases (overestimation or underestimation) were found in the estimation process.

Level II: Models Based on Intersection Characteristics

The residuals of the base model, as shown in Equation 5, are analyzed by CART, and the factors and levels that are available and used in the analyses are shown in Table 1. These factors were selected from the TASAS system after consultation with some of the Caltrans practicing engineers who will be using the models. Fortunately, missing values represent only a very small proportion of the data set and are treated as separate levels in the analysis for simplicity reasons. An additional factor (IADT), called the index of conflict, was also created for the



Coefficient of Correlation=0.47498

Slope=0.1691 (t value=26.9), Intercept=0.6186 (t value=6.9)

FIGURE 4 Scatter plot and regression result of base model.

study because turning movement counts were not available in TASAS. This index of conflict is calculated as the proportion of cross street traffic to total traffic. IADT turned out to be a very important factor in predicting intersection accidents, as shown in Figure 5, which shows a ninefold cross-validation tree. The tree, with nine terminal nodes and a relative error of 0.90, was selected by the 0.3 standard error rule. A relative error of 0.90 implies that CART is able to provide a further reduction in error of about 10 percent. This also shows that the 2,488 intersections can be divided into nine groups with similar accident characteristics on the basis of the tree structure, as shown in the same figure.

Group 1 is characterized by intersections that have an index of conflict of less than or equal to 0.065, that is, 6.5 percent of all entering traffic is on the cross street. A simple classification rule in this situation appears to indicate that when the level of conflicts is low, no other factor except traffic intensity is significant in affecting the safety of intersections. There are 492 intersections in this group, with an average residual of about −0.73 and a standard deviation of about 1.5. The average negative residual of −0.73 here means that 0.73 injury accidents should be deducted from the base model derived from Equation 5. In simple terms, intersections in this category generally have a lower risk.

Group 2 is characterized by intersections that have an index of conflict larger than 0.065 and are either "T," "Y," or multilegged intersections. Because less than 3 percent of the intersections in the data set are multilegged, it can be assumed that this group represents those intersections that have fewer legs. This could be why they have lower accident statistics in comparison with intersections with more legs, such as four-legged and offset intersections. As shown in the intermediate node, these latter intersections have an average residual of 0.36, compared with −0.62 in Group 2.

Group 3 has an average residual of 0.05. Intersections in Group 3 have narrow cross streets (less than or equal to a total of three lanes), and they tend to have smaller residuals than the remaining groups (4 to 9).

In comparison with the intermediate node corresponding to intersections with pretimed controllers and two-phase fully actuated controllers, Group 4 has a smaller average residual than the other groups. It could be argued from a safety viewpoint that it might be worthwhile to spend additional money on installing actuated controllers.

Group 6 is similar to Group 5, with the exception that the index of conflict is greater than 0.475. Both groups have pretimed signal controllers. It is obvious that Group 6 has a much higher accident potential than Group 5. This difference is due to a higher level of conflict.

Groups 7 and 8 are similar to groups 5 and 6, with the exception that all of them have two-phase fully actuated controllers. Group 7 includes intersections that have 5 lanes or fewer on main streets, but Group 8 includes only those intersections that have 6 or more lanes on main streets. Group 8 has a higher residual (3.56) than Group 7, and this appears to indicate that very wide intersections might require more than two phases to accommodate the turning movements that occur in these intersections.

Group 9 has an average residual of 4.82, which appears to indicate that left turn prohibitions during peak hours may not increase risk.

From the results, it can be seen that the groupings and their estimates are not unreasonable and that their characteristics could provide additional insights to engineers who are designing or redesigning signalized intersections. These models contain as many reasonable factors as other models obtained by regression analysis for similar projects in which field observations and accident record systems were used (*5*).

Level III: Models Based on Information that Includes Individual Accident History

It is obvious that no matter how complicated the models are, there are likely to be some unique intersection features that cannot be captured. One method of compensating for this is to employ a concept of combining estimates on the basis of the accident history of the intersection and estimates for the group in which the intersection belongs. Hauer and Persaud (*4*) have derived an estimate, Z, on the basis of a linear combination of the two results of two approaches for predicting the safety of an individual intersection. The estimate is obtained from the following equation:

$$Z = aE\{m\} + (1 - a)x \qquad (6)$$

where

$$a = (1 + \mathrm{Var}\{m\}/E\{m\})^{-1}$$

and where m = expected accident statistics and x = accident count. Hauer and Persaud also suggested that the sample mean $\bar{x}$ could be used to estimate $E\{m\}$ and that sample standard deviation(s) could be used to estimate $\mathrm{Var}\{m\}$ by the following equations:

$$E\{m\} = E\{x\} \qquad (7)$$

$$\mathrm{Var}\{m\} = (s^2 - \bar{x}) \qquad (8)$$

This combination technique does not appear to be very powerful for predicting future accidents if there is a long accident history with many accidents. However, the technique is particularly useful in predicting accidents at intersections with few events and short history. This combination technique would not be applicable to new intersections because a new intersection would not have any accident record or history.

Examples Illustrating the Procedures and Overall Significance of the Models

The procedure to be illustrated has a three-level structure, and users can terminate the analysis at the end of any level to suit their input requirements. Five intersections in the data set were arbitrarily selected for illustration purposes. The results, based on the procedures described in earlier sections, are tabulated in Table 2. The first eight rows, from MADT to MNL, are the characteristics of the intersections, described in Table 1. MVYR is the independent variable of Equation 5, and FIACCYR is the dependent variable of the same equation. FIACCYR is also the estimate for Level I prediction. IACCYR is the observed number of injury accidents per year that occurred at the intersection. $E\{x\}$ is the average or sample mean of the residuals of the subgroups based on grouping by CART.

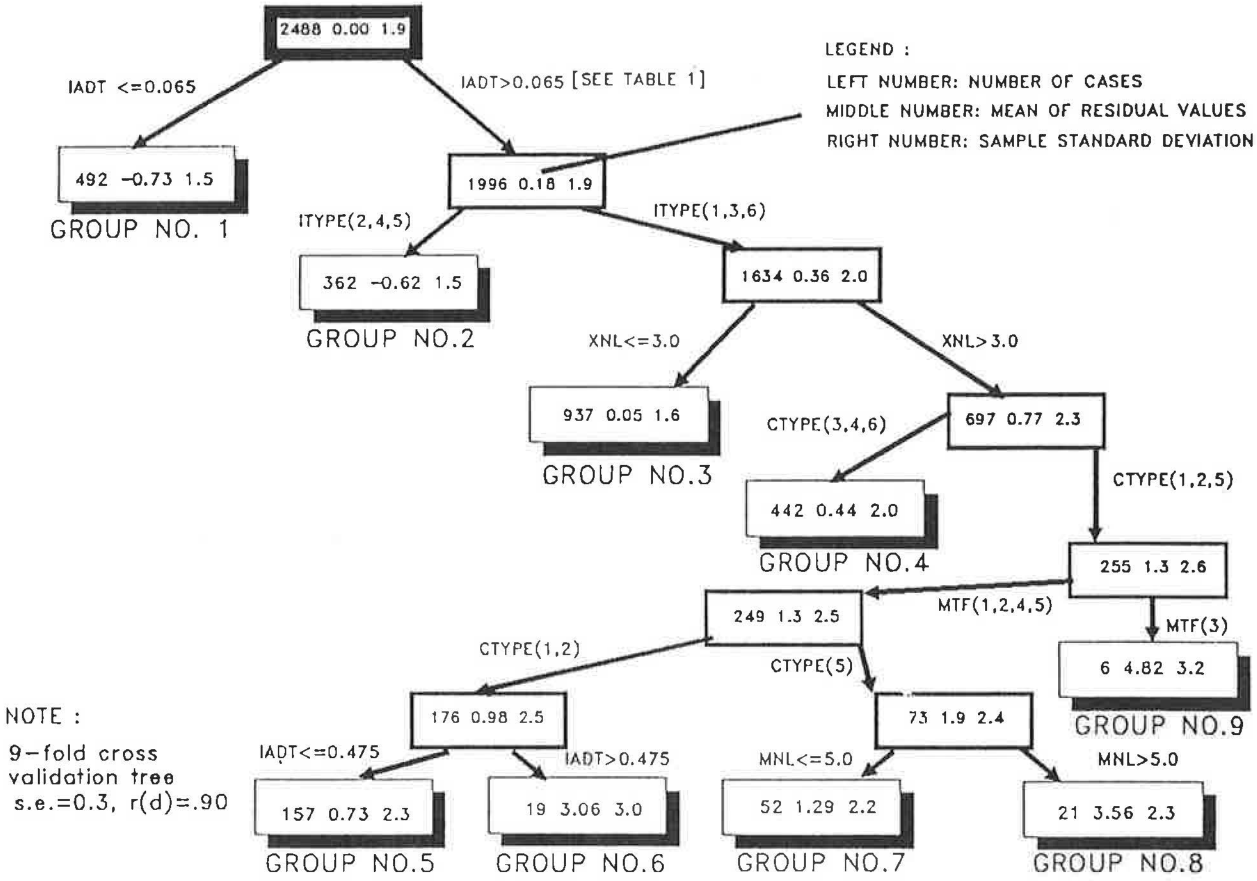


FIGURE 5 Grouping of intersections by CART.

CIACCYR is the estimate for Level II prediction, and it is calculated as the algebraic sum of $E\{x\}$ and FIACCYR. In other words, $E\{x\}$ by CART is the adjustment factor to account for different design or control features of a particular intersection.

For Level III prediction, the standard deviation(s) of the subgroups from CART and individual accident history would be required. The proportion or weight *(a)* to be used can be derived from Equation 6. The x in Equation 6 is calculated as the difference between the observed number of injury accidents (IACCYR) and FIACCYR. Finally, the new adjustment factor (Z) can be calculated with Equation 6, and the estimate for Level III prediction (HIACCYR) would be the sum of Z and FIACCYR.

It can be seen that the differences between the predicted and observed values generally improve from Level I to Level III. This improvement is not surprising because input requirements increase from Level I to Level III. It can be argued that examining the differences between the observed and predicted values for Level III predictions is not fruitful because they are adjusted according to the actual differences between observed and calculated values. Furthermore, it is not meaningful to compare the predicted values (especially Level III) with the observed values because the predicted values represent future accident potential. Out-of-sample testing would be more meaningful in this regard; however, this type of test might not be feasible due to the length of the record systems. A check on the accuracy of the models predicted by Levels I and II was also performed. This check could be considered redundant because the accuracy of these models was assured by the large t values of the estimates, as shown in Figure 4 for Level I, and by the cross-validated tree for Level II, as shown in Figure 5. A correlation of 0.475 between the actual and predicted values was found for Level I, and a correlation of 0.580 was found for Level II. A similar analysis for Level III was not carried out because Level III predictions represent *future* accident potential. The improvement from 0.475 to 0.580 might not appear to be large; however, it should be recalled that only existing information on road designs and traffic conditions found in TASAS was used, and factors such as driver's characteristics, vehicle design, weather, and so on were not employed in the analysis.

CONCLUSIONS AND LIMITATIONS

The models derived here for injury accidents include factors such as traffic intensity, percentage of cross street traffic, intersection type, signal type, number of lanes on main and side streets, and left turn arrangements. However, only intersection type and traffic intensity are included in the current models used in California. The other factors used in the present models, such as rural or urban conditions and being inside or outside city limits, have not been found significant in this study. The models proposed in this paper have higher predictive power, and they also provide more insight for engineers who are designing or redesigning signalized intersections and evaluating alternative strategies.

TABLE 2 EXAMPLES ILLUSTRATING THE PROCEDURE

ID#	2574	2541	1794	934	290
MADT*	49000	37000	18100	20998	16000
XADT*	10000	301	3000	17111	1501
IADT*	0.169	0.008	0.142	0.449	0.086
ITYPE*	1	4	1	4	1
XNL*	4	2	2	4	2
CTYPE*	2	5	1	6	5
MTF*	2	2	2	2	2
MNL*	4	4	4	4	2
MVYR	21.53	13.61	7.70	13.91	6.39
IACCYR	5.86	1.33	3.71	0.00	2.00
LEVEL I(Equation 5.1)					
FIACCYR	4.26	2.92	1.92	2.97	1.70
LEVEL II(Figure 5.3)					
E(x)	0.73	-0.73	0.05	-0.62	0.05
Group no.	5	1	3	2	3
CIACCYR	4.99	2.19	1.97	2.35	1.75
LEVEL III(Equations 5.2-5.4)					
S	2.3	1.6	1.6	1.5	1.6
a	0.14	0.32	0.02	0.28	0.02
x	1.60	-1.59	1.79	-2.97	0.30
Z	1.48	-1.31	1.76	-2.32	0.30
HIACCYR	5.74	1.61	3.68	0.65	2.0

* Please refer to Table 1 for more details

LEGEND

ID#: Intersection Identity Number
MADT: Main Street ADT
XADT: Cross Street ADT
ITYPE: Intersection Type
XNL: Total Number of Lanes on Cross Street
CTYPE: Control(Signal) Type
MTF: Main Street Traffic Flow
MNL: Total Number of Lanes on Main Streets
MVYR: Total Number of Vehicles Entering Intersection per Year in Millions
IACCYR: Actual Number of Injury Accidents per Year
FIACCYR: Level I Forecasted Number of Injury Accidents per Year
CIACCYR: Level II Forecasted Number of Injury Accidents per Year
HIACCYR: Level III Forecasted Number of Injury Accidents per Year

As far as development of macroscopic models for injury accidents is concerned, it is concluded that the proposed methodology and TASAS are very suitable for this kind of study. The results obtained are not unreasonable and correspond with conventional wisdom. These factors would normally be expected to have a high degree of association with accident patterns. However, it could be argued that factors such as phases for left turn vehicles, provisions of left turn pockets, number of conflict points, and so on should also play an important role in the prediction process. Unfortunately, models with this kind of detail would require specific information on turning movement counts, conflict analyses, and other factors. Such information is not available in most accident record systems, including TASAS.

The proposed methodology follows the general pattern used in system analysis, thus allowing room for refinements and changes, if necessary. The three-level prediction procedure allows flexibility by having different levels of inputs. At the same time, this procedure gives users an opportunity to appreciate the evolution of their estimates. The selection of the response variable could also play an important role in the model development process. The use of number of injury accidents per year as the response variable has kept this analysis simple throughout. The use of traffic intensity as an exposure measure (similar to accident rate) can cause some problems, and this factor should be used as a predictor variable instead. For example, a reduction in accident rate with an increase in traffic flow has been found in some studies by regression analysis with accident rate (accidents per vehicle) as a dependent variable (*5*). Of course, accident rates could be used to summarize results at the end of the analysis, if this is found to be convenient.

The analysis of residuals of the base model by Classification and Regression Trees (CART) has proved to be a viable technique for building models based on data sets that have high dimensionality, a mixture of data types, and lack of homogeneity. Its tree structure output makes the interpretation of results easy. On the other hand, incorporating and interpreting interaction terms between factors can be quite time consuming in a regression analysis.

For future research, the following tasks would be useful in overcoming some limitations of the work completed to date. The models derived in this paper can be called macroscopic models, and there could be further attempts to analyze the accident aspects of the data base, for example, types of collision, time of occurrence, road condition, types of vehicle, driver's condition, and so on. On the other hand, the issues of underreporting of PDO accidents and the slight difference between serious injury accidents and fatal accidents should be considered and examined in future studies.

ACKNOWLEDGMENTS

We thankfully acknowledge that this research was sponsored by the California Department of Transportation and the Federal Highway Administration, U.S. Department of Transportation. Permission from these agencies to publish this paper is also acknowledged. We also wish to thank the following individuals for reviewing our paper and for their suggestions during the course of our research: E. L. Scott and L. Breiman, Statistics Department, University of California, Berkeley; R. N. Smith, L. Seamons, and H. Garfield, California Department of Transportation.

REFERENCES

1. *Manual of Traffic Surveillance and Analysis System.* 5 vols. California Department of Transportation, Sacramento, 1978.
2. L. Breiman et al. *Classification and Regression Trees (CART).* Wadsworth Publisher, Belmont, Calif., 1984.
3. M. Y. K. Lau and A. D. May. *Accident Prediction Models for Signalized Models.* Working Paper UCB-ITS-WP-86-10. ITS, University of California, Berkeley, 1986.
4. E. Hauer and B. N. Persaud. How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effect of Warning Devices. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987, pp. 131–140.
5. D. T. Silcock et al. *Relationships Between Accident Rates, Road Characteristics and Traffic on Two Urban Routes.* Research Report 40. TORG, Department of Civil Engineering, University of Newcastle-upon-Tyne, England, 1982.

The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the California Department of Transportation. This report does not constitute a standard, specification, or regulation.

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