Identification of Accident Factors on Highway Segments: A Method and Applications

T. CHIRA-CHAVALA and KING K. MAK

ABSTRACT

An algorithm was developed to provide traffic engineers a means to identify factors or combinations of factors that cause accident overrepresentation at a given highway location relative to some average. The output from this algorithm can be used for further site investigation to develop accident countermeasures that may be responsive to the problems at the site. It also provides quantitative measures of the degree of overrepresentation for each factor at that site. Available mainframe computer programs to facilitate the computation involved are given. An example of application of this algorithm to a highway location in Texas is described in full detail. Comparison of the output from this full algorithm with that from an automated microcomputer program developed at the Texas Transportation Institute is also discussed.

The objective of this paper is to develop an algorithm to identify accident factors that are overrepresented at a selected roadway location relative to some "average," as well as to determine the magnitude of the overrepresentation. The output from such an analysis can thus be used as input for

1. Developing specific accident countermeasures for that location; and
2. Establishing priority locations within a city, county, or district where accident remedies may be more urgently needed.

INTRODUCTION

Traffic engineers and planners are often faced with the tasks of having to identify roadway locations that are deemed accident hazardous and to determine effective remedies to alleviate the problems. The task of identifying locations with high accident history has been greatly facilitated by available computerized accident data, roadway inventory files, and computer programs to isolate highway segments that show high accident rates, frequency, and severity.

To determine appropriate accident countermeasures, the causes of the problems at the site must first be understood and identified. Once this is accomplished, engineers or accident investigators can conduct in-depth site investigations and then develop appropriate remedies. Procedures to reliably determine causative accident factors and their magnitude of overrepresentation at a given site have not been developed or well documented.

Accidents are complex phenomena, and the problems at different locations are likely to be different or site specific. When properly conducted, analyses of past accident records for a specific site of interest can help illuminate possible causes of accidents at that location.

Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843.

CONCEPTUAL BASIS

The procedure developed here is aimed at identifying, from computerized accident data, characteristics of accidents (factors) or combinations of factors that are overrepresented at the site relative to some average. This average may be accidents on a similar road class within a city, county, or district. The algorithm is based on the principles of discrete-multivariate models and is capable of simultaneously analyzing a number of potential variables. In this way, both independent (or main) effects and interactions of these variables with one another can be systematically determined. Effects due to confounding variables, which may jeopardize the results, can therefore be minimized or avoided.

The number of accidents on a given segment of highway 2 to 3 mi long is not likely to be very large to permit a simultaneous analysis of an unlimited number of variables of interest. Fortunately, most accident variables usually correlate and interact with one another in such a way that only some, but not all, will be required in the analysis. This procedure therefore incorporates a variable-selection step that takes place before the model estimation. The two-stage algorithm is fully described in the following section.

ALGORITHM

The algorithm consists of two stages: variable selection and modeling. Variable selection involves the systematic selection of variables that may be significantly overrepresented at the site and elimination of those that are statistically nonsignificant. Only the significant variables are further analyzed at the modeling stage. The modeling involves determining specific factors or combinations of factors that are overrepresented at the site, as well as the magnitude of such overrepresentation. Figure 1 shows a flow chart for the entire algorithm.

Variable Selection

This is a sequential procedure based on two measures of statistical association in contingency table

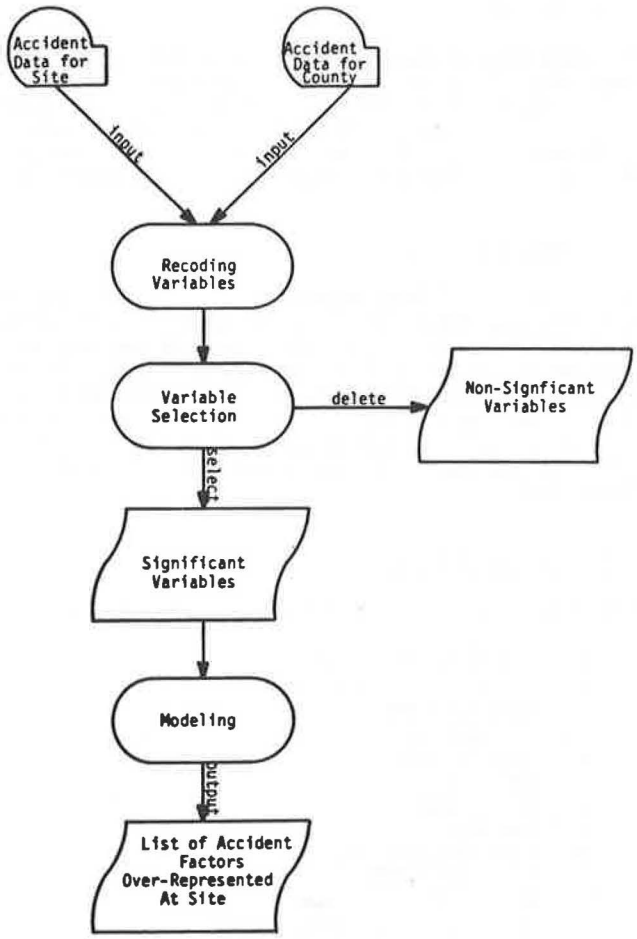

FIGURE 1 Flow chart for the algorithm.

analyses: Q_T and Q_{CMH} (1). In each step of the variable selection, one independent variable that is the most significant is selected after examining the effects of all (unselected) variables on accident overrepresentation at a site. The dependent variable for variable-selection analyses is site/average. The null hypothesis associated with the tests of Q_T and Q_{CMH} can be stated as follows (1):

> H_O: for each level of the independent variables, the accidents are distributed at random between the site and the county (assuming county as the comparison average) for all levels of the covariable(s).

The variable selection algorithm follows these steps (2):

1. A two-way contingency table of accident frequency is formed between the dependent variable (site/county) and each of the potential independent variables. A Pearson chi-square is then computed. The variable selected is one that has the highest value of chi-square per degree of freedom.
2. For each of the variables not yet selected, a contingency table of accident frequency is formed among this variable, the dependent variable, and the variable previously selected. Q_T is calculated, which reflects both the main effect of this variable, as well as its interaction with the previously selected variable. The variable selected in this step is the one that has the highest Q_T value per degree of freedom. On the other hand, those variables showing statistically nonsignificant Q_T values will be eliminated from further analysis.

In this context, Q_T expresses the extent of "total association" of the variable with the dependent variable, having accounted for the previously selected variable. The derivation of Q_T is given by Landis et al. (1) and its formula is given as

$$Q_T = \sum_{h=1}^{q} G_h' \, [\mathrm{Var}\,\{G_h \mid H_O\}]^{-1} G_h \qquad (1)$$

where

$h = 1, 2, \ldots, q$, is the levels of the previously selected variable(s);
G_h = a matrix of the differences between observed and expected frequencies under H_O; and
G_h' = a transposed matrix of G_h.

The degrees of freedom for Q_T is $q(s - 1)(r - 1)$, where s and r are the levels of the independent variable under investigation and the dependent variable, respectively.

3. The variable selection process then continues in this manner until completion. After the first few steps of the variable selection, however, the cell frequencies of the contingency table may thin out considerably, and the degrees of freedom may increase so rapidly that Q_T may become less effective. In this situation, Q_{CMH} will be used as a selection criteria instead. Q_{CMH} is not as sensitive to small cell size as is Q_T, and its test of significance is based on degrees of freedom of $(s - 1)(r - 1)$. Q_{CMH} is capable of capturing a weak but consistent effect of a variable although it does not reflect the total contribution, which includes interactions with other variables as does Q_T (2).

The derivation of Q_{CMH} is also reported by Landis et al. (1) and its formula is given as

$$Q_{CMH} = G' \, [\mathrm{Var}\,\{G \mid H_O\}]^{-1} G \qquad (2)$$

where $G = \sum_{h=1}^{q} G_h$.

The variable selected using Q_{CMH} is one that has the highest Q_{CMH} per degree of freedom.

The variable selection process is completed when either of the following is met: (a) the list of potential variables is exhausted, or (b) the data thins out so much that neither Q_T nor Q_{CMH} are appropriate.

Modeling

Accident characteristics or combinations of these characteristics that are overrepresented at a site relative to the county can be isolated and the magnitude of their overrepresentation quantified using the following model estimation technique.

For illustration, assume that three independent variables were selected by the variable selection process. The following procedure applies, which remains unchanged for any number of the independent variables:

1. Accident frequency for county only is cross-classified by these three variables.
2. A log-linear model that best describes this county contingency table is estimated. A log-linear model expresses the cell probabilities as a function

of significant main effects of and interactions among the variables. Such a model can generally be expressed as

$$\ln(P_{ijk}) = u + u_1 + u_2 + u_3 + U_{12} + u_{13} + \ldots \quad (3)$$

where

P_{ijk} = estimated cell probability for the (i,j,k)th cell,
u = overall mean,
u_1 = main effect of variable 1,
u_2 = main effect of variable 2,
u_3 = main effect of variable 3, and
u_{12} = interaction between variables 1 and 2, and so on.

The goodness-of-fit for a log-linear model is a likelihood ratio statistic (3):

$$X^2 = -2 \sum \text{(observed) log (estimated/observed)} \quad (4)$$

3. A contingency table of expected accident frequency for the site, if sites and county are not different, is constructed. The following formula is used to compute the expected cell counts for the site. This contingency table is cross-classified by the same variables as those in Step 1.

$$E_{ijk} = N \times P_{ijk} \quad (5)$$

where N is the total number of accidents at the site.

4. For each cell of the site contingency table, compute the overrepresentation indicator or the Freeman-Tukey deviate (3):

$$Z_{ijk} = (X_{ijk})^{1/2} + (X_{ijk} + 1)^{1/2} - (4E_{ijk} + 1)^{1/2} \quad (6)$$

where X_{ijk} is the observed accident frequency of the (i,j,k)th cell for the site. The overrepresentation indicator (Z_{ijk}) reflects the extent to which the actual observed number of accidents in any one cell of the site contingency table differs from the expected number of accidents in that cell, if the site is indeed no different from the county. A large positive value of Z_{ijk} indicates that the observed number of accidents at the site is higher than expected, and therefore an overrepresentation is indicated for that cell. A negative value of Z_{ijk} indicates that the author did not observe as many accidents as expected at the site for that cell. When the observed and the expected number of accidents are similar, Z_{ijk} will be a small positive number that is less than 1.

One property of this indicator that is particularly useful here is that its magnitude is a function of both (a) the extent to which the observed accident frequency differs from the expected frequency and (b) the cell size. That is, a larger value of either (a) or (b) will result in a larger positive Z_{ijk}. Therefore, a cell that has higher accident counts will display a higher overrepresentation ranking indicator than a cell that has lower accident counts, even if both indicate identical percent differences between observed and expected frequencies.

5. The cells that show Z_{ijk} values larger than, for example, 1.5 are listed. These cells represent combinations of accident factors that are found to be significantly overrepresented at the site relative to the county. The selection of 1.5 as the cutoff point is based on the fact that when the observed and the expected frequencies are similar, Z_{ijk} will be less than 1. Its value will be even smaller for increased cell size.

APPLICATIONS

The algorithm developed in the Algorithm section has been applied to an analysis of accident data at a number of sites on urban Interstate and urban non-Interstate systems in Texas. As an example for a case study, the analysis carried out for one of these sites is fully described in the next paragraph.

Description of Site

The site is a 2.4-mi segment of a U.S. highway in San Antonio, Texas. It is a six-lane divided urban freeway, full access control, with both straight and sharp-curve segments. From 1980 through 1982, 254 accidents were reported between milepoints 23.8 and 26.1. These were fatal, injury, or property-damage-only accidents. The comparison average for this site are the accidents on all urban U.S. highways in Bexar County.

Independent Variables

The following 12 variables were initially considered:

1. Degree of curvature
2. Weather and surface condition
3. Accident time
4. Accident type
5. Vehicle type
6. Severity
7. Driver age
8. Speeding
9. Driving while intoxicated (DWI)
10. Light condition
11. Vehicle damage scale
12. Driver license status

The levels of these variables are given in Table 1. Two different levels for these variables were defined: one for variable selection purposes and the other (more detailed) for modeling. This was to minimize the number of overly small cells that would adversely affect variable selection more than modeling. The first five variables are important because they are directly applicable to developing traffic engineering-related countermeasures; they are called primary variables. The other seven variables are mostly driver related and they help to further illuminate the causes of accidents; they are called secondary variables. Their usefulness in traffic engineering-related countermeasures is probably more limited.

To ensure that the five primary variables will have a good chance of being evaluated before the data hopelessly thin out, the variable selection will be applied to these variables first. Once this is done, the secondary variables will then be examined.

Result of Variable Selection

The variable selection yielded the following outcome step by step:

Step 1: Table 2 gives the values of the Pearson chi-square, X_p^2, obtained for the five primary variables evaluated: degree of curvature, weather and surface condition, accident time, accident type, and vehicle type. Degree of curvature, which shows the largest X_p^2 per degree of freedom, was selected.

Step 2: Table 3 gives the values of Q_T for the

TABLE 1 Levels of Potential Independent Variables

Variable	Level for Variable Selection	Level for Modeling
Degree of curvature	Straight	Straight
	Curve	Less than 2 degrees
		Greater than 2 degrees
Weather/surface condition	Dry	Dry
	Wet	Wet
Accident time	Weekday, rush hour	Weekday, rush hour
	Weekday, non-rush hour; or weekend, day	Weekday, non-rush hour Weekend, day
	Evening or night	Evening or night
Accident type	Single-vehicle	Single-vehicle: barrier/object
	Multivehicle	Single-vehicle: other
	Rear-end, sideswipe	Multivehicle: rear-end sideswipe angle, head-on
	Other	
Vehicle type	Passenger cars only	Passenger cars only
	At least one pickup or van	At least one pickup or van
	At least one heavy truck or bus	At least one heavy truck or bus
Severity	Fatal or injury	Fatal
	Property damage only	Injury Property damage only
Driver age	At least one over 55 or at least one under 21	At least one over 55 At least one under 21
	21 to 55	21 to 55
Speeding	At least one speeding	At least one speeding
	No speeding	No speeding
Driving while intoxicated (DWI)	At least one DWI	At least one DWI
	No DWI	No DWI
Light condition	Daylight	Daylight
	Other	Other
Vehicle damage scale	Front	Front
	Rear	Rear
	Side	Side
	Other	Other
Driver license status	At least one out of state	At least one out of state
	Other	Other

TABLE 2 Result of Variable Selection: Step 1

Independent Variable	X_p^2	Degrees of Freedom	p-value
Degree of curvature	228.0	1	0
Weather/surface condition	31.2	1	0
Accident time	8.8	2	0.012
Accident type	14.0	2	0.001
Vehicle type	1.6	2	0.437

TABLE 3 Result of Variable Selection: Step 2

Variable	Q_T	Degrees of Freedom	p-value
Curvature x surface condition	25.5	2	0
Curvature x accident time	17.9	4	0.001
Curvature x accident type	12.7	4	0.013
Curvature x vehicle type	1.3	4	0.856[a]

[a]Eliminated.

four primary variables not selected in Step 1. Weather and surface condition, which shows the largest Q_T per degree of freedom, was selected. On the other hand, vehicle type with a nonsignificant Q_T value was eliminated from further analysis.

Step 3: Table 4 gives the values of Q_T for the primary variables not yet selected or eliminated. Q_{CMH} values were also computed for these variables. Although the Q_T values for both variables were not highly significant, the Q_{CMH} value for accident time was.

TABLE 4 Result of Variable Selection: Step 3

Variable	Q_T	Degrees of Freedom	p-value	Q_{CMH}	Degrees of Freedom	p-value
Curvature x surface x accident time	16.9	8	0.032	11.8	2	.003
Curvature x surface x accident type	15.2	8	0.055	8.7	2	.013

This indicates that accident time had consistent main effect on accident overrepresentation at the site. This variable was, therefore, selected. Accident type was retained because its Q_{CMH} showed a p-value of 0.013, indicating that this variable may have consistent (though relatively weak) main effect.

Step 4: A cross-classification of accident frequency by degree of curvature, weather and surface condition, accident time, and accident type, resulted in 32 percent of the cells having fewer than four accidents. Accident type was therefore regarded as a "sparse" variable, and no variable selection analysis was performed for this variable. Sparse variables are those associated with too many small (less than four accidents) or empty cells to warrant meaningful variable-selection analyses. Such variables have neither been selected nor eliminated as significant variables because of the sample size limitation. As an option, they can be further investigated in the modeling stage.

Step 5: Having exhausted the primary-variable list, the selection process continued with the secondary variables. Each secondary variable was first cross-classified with the dependent variable and all those variables already selected. Only severity, driver age, speeding, and driver license status showed cells with reasonable sample size to justify variable-selection analyses. All other secondary

TABLE 5 Result of Variable Selection: Step 4

Variable	Q_T	Degrees of Freedom	p-value	Q_{CMH}	Degrees of Freedom	p-value
Selected variables x speeding	22.4	12	.034	1.16	1	.281
Selected variables x driver age	14.0	12	.302[a]	0.78	1	.378
Selected variables x severity	11.2	12	.511[a]	0.43	1	.510
Selected variables x license status	9.7	12	.641[a]	2.33	1	.127

[a]Eliminated.

variables were sparse. The values of Q_T and Q_{CMH} for severity, driver age, speeding, and driver license status are given in Table 5. Of these, only the Q_T for speeding showed a p-value less than 0.05; all four showed nonsignificant Q_{CMH}s. Therefore, speeding was selected while the other three variables were eliminated from further analysis.

The independent variables that were found to be significant from the variable selections were

- Degree of curvature
- Weather and surface condition
- Accident time
- Speeding

The independent variables that were sparse were

- Accident type
- Light condition
- Vehicle damage scale
- DWI

The independent variables that were found to be nonsignificant and thus eliminated from further analysis were

- Vehicle type
- Driver age
- Severity
- Driver license status

Modeling Result

The modeling result is described next, step by step:

1. A contingency table of accident frequency for county, cross-classified by the selected variables (a) degree of curvature, (b) weather and surface condition, (c) accident time, and (d) speeding is given in Table 6.

TABLE 6 Number of Accidents for County

Curvature (V4)	Condition (V3)	Time (V2)	Speeding (V1) Yes	Speeding (V1) No
Straight	Dry	Weekday, rush hour	42	82
		Weekday, non-rush	25	59
		Weekend, day	7	21
		Evening or night	90	179
	Wet	Weekday, rush hour	15	17
		Weekday, non-rush	9	10
		Weekend, day	6	7
		Evening or night	40	28
Less than 2 degrees	Dry	Weekday, rush hour	3	6
		Weekday, non-rush	3	6
		Weekend, day	2	3
		Evening or night	10	21
	Wet	Weekday, rush hour	0	2
		Weekday, non-rush	0	2
		Weekend, day	2	0
		Evening or night	2	6
Greater than 2 degrees	Dry	Weekday, rush hour	0	4
		Weekday, non-rush	0	3
		Weekend, day	0	1
		Evening or night	9	9
	Wet	Weekday, rush hour	0	2
		Weekday, non-rush	2	1
		Weekend, day	0	2
		Evening or night	3	4

2. A log-linear model that best describes the data in Table 6 was found to be

$$\ln (P_{ijkl}) = u + u_2 + u_4 + u_{13} \quad (7)$$

This model states that the probability of accidents for the county is influenced by the main effects of accident time and degree of curvature, as well as the interaction between speeding and weather and surface condition. The chi-square goodness-of-fit test was 44.14 for 39 degrees of freedom (a p-value of 0.263), indicating a very good fit.

3. If the site and the county were similar in accident characteristics, the expected number of accidents at the site, cross-classified by degree of curvature, weather and surface condition, accident time, and speeding, could be obtained (using Equations 5 and 7) as given in Table 7. Also given in Table 7 are the actually observed number of accidents at the site.

4. The magnitude of accident characteristics that was overrepresented at the site relative to the county was computed by using Equation 6. The result is given in Table 8. Only those characteristics associated with the overrepresentation indicator (z) of greater than 1.5 and the observed accidents of at least 7, are given in Table 7. These are cells that show a significantly higher than expected number of observed accidents at the site. Thus accident overrepresentation is indicated by these cells. The following findings can be drawn from Table 8:

(a) The curve section with curvature greater than 2 degrees was a major cause of accident overrepresentation, almost regardless of time of day, weather and surface condition, or presence or absence of speeding.

(b) The combination of this sharp curve, wet conditions, and speeding was particularly serious in causing accident overrepresentation, as indicated by consistently high values of the overrepresentation indicator (z).

(c) Accidents on this sharp curve were found to be especially overrepresented in the evening or at night (very high values of overrepresentation indicator are given).

(d) To a lesser extent than evening and nighttime, accidents on this sharp curve tended to be overrepresented during rush hours of week days.

NOTE: The accident overrepresentation analysis can conclude at this point or the analyst may choose to further examine the four sparse variables. As mentioned earlier, sparse variables may not necessarily be nonsignificant. Their statistical significance was not tested because of the sample size limitation. Analyses in the modeling stage to incorporate these sparse variables can help illuminate causes of accident overrepresentation at the site even further, if these variables are indeed significant. Sparse variables can be analyzed in the modeling stage by one of two methods:

1. Replacement of the last independent variable selected with each of the sparse variables, and Steps 1 through 5 are repeated. Because there were four sparse variables, four modeling analyses would be required.
2. Incorporation of each of the four sparse variables into a modeling analysis together with the four independent variables already selected, and Steps 1 through 5 are repeated. Again, four analyses would be required for the four sparse variables.

An advantage of the second method is that the final analysis result of accident-overrepresentation causes would be more complete. However, the smaller cell size due to an additional variable may make the result less interesting for practical purposes.

COMPUTER SOFTWARE

The algorithm as described can be performed by using computer programs that are currently available. For variable selection, PARCAT (4), which is a mainframe program, can be used. For modeling, BMDP (5), ECTA (6), or any other standard log-linear model program will be satisfactory. In order to perform the analysis using these computer programs and to reach the final outcome given in Table 8, users are required to have sufficient familiarity with the statistics involved to make statistical decisions and to se-

TABLE 7 Expected and Observed Number of Accidents for Site

Curvature (V4)	Condition (V3)	Time (V2)	Speeding (V1) Yes		No	
Straight	Dry	Weekday, rush hour	(12.7)	7	(26.1)	9
		Weekday, non-rush	(8.8)	2	(18.2)	8
		Weekend, day	(3.7)	0	(7.7)	7
		Evening or night	(29.4)	8	(60.6)	11
	Wet	Weekday, rush hour	(5.3)	3	(5.4)	6
		Weekday, non-rush	(3.6)	4	(3.7)	6
		Weekend, day	(1.5)	6	(1.5)	3
		Evening or night	(12.2)	5	(12.5)	6
Less than 2 degrees	Dry	Weekday, rush hour	(1.3)	2	(2.8)	3
		Weekday, non-rush	(.9)	2	(1.9)	1
		Weekend, day	(.4)	0	(.8)	2
		Evening or night	(3.1)	1	(6.5)	7
	Wet	Weekday, rush hour	(.6)	1	(.6)	2
		Weekday, non-rush	(.4)	0	(.4)	0
		Weekend, day	(.2)	0	(.2)	1
		Evening or night	(1.3)	4	(1.3)	2
Greater than 2 degrees	Dry	Weekday, rush hour	(.8)	7	(1.6)	9
		Weekday, non-rush	(.6)	2	(.1)	11
		Weekend, day	(.2)	1	(.5)	6
		Evening or night	(1.8)	17	(3.8)	28
	Wet	Weekday, rush hour	(.3)	8	(.3)	1
		Weekday, non-rush	(.2)	11	(.2)	1
		Weekend, day	(.1)	7	(.1)	0
		Evening or night	(.8)	12	(.8)	9

Note: Numbers in parentheses are expected numbers of accidents.

TABLE 8 Accident Overrepresentation Indicators for Site

Curvature	Condition	Time	Speeding Yes	Speeding No
Straight	Dry	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
	Wet	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
Less than 2 degrees	Dry	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
	Wet	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
Greater than 2 degrees	Dry	Weekday, rush hour	3.43	3.44
		Weekday, non-rush	[a]	5.58
		Weekend, day	[a]	[a]
		Evening or night	5.50	6.66
	Wet	Weekday, rush hour	4.35	[a]
		Weekday, non-rush	5.44	[a]
		Weekend, day	4.26	[a]
		Evening or night	5.02	4.11

[a]Overrepresentation indicator (z) less than +1.50 or observed number of accidents less than 7.

lect, evaluate, and interpret data at every intermediate step of the analysis.

The Texas Transportation Institute (TTI) has developed an automated microcomputer program for the entire algorithm (7). The program is part of a study conducted for the Texas State Department of Highways and Public Transportation. Because the entire analysis procedure is fully automated, it does not require intervention by the users at any of the intermediate steps. Once the subset of accident and roadway data are specified by a user, the program will automatically initiate the variable-selection analysis, carry out the modeling, and finally report the accident overrepresentation factors for that site. This program was written in turbopascal for IBM PC-XT or compatible systems.

This automated microcomputer program was based on the algorithm described in this paper but was especially modified for use on microcomputers. The decision for such a simplification was made for practical reasons: manageable run-time and storage memory of microcomputers. The analysis output from the full algorithm and that output from the automated program are compared in the following section for the site mentioned in the Applications section.

COMPARISON OF ANALYSIS RESULTS FROM FULL ALGORITHM AND AUTOMATED MICROCOMPUTER PROGRAM

For the site presented here, there were no differences in the result of variable selection as far as the independent variables selected or eliminated or the order for which the variables were selected. For modeling, the factors or combination of factors that caused accident overrepresentation at the site were found to be the same for both algorithms. However, the magnitude of the accident overrepresentation indicators (z) from the two algorithms was slightly different. Table 9 gives the values of z obtained from the automated microcomputer program.

Generally, the result provided by the automated microcomputer version is not expected to be very different from that provided by the full algorithm unless it concerns "borderline" cases. These borderline cases may be variables indicating p-values close to 0.05 in the variable selection stage or the overrepresentation factors associated with values of the indicator (z) close to +1.5. For the purpose of accident countermeasures, these borderline cases are likely to be less interesting, and thus they may not affect subsequent actions taken by engineers.

One advantage of the full algorithm that has not been developed for the automated microcomputer version is that for a certain site, Step 2 of the modeling may yield the result such that the levels

TABLE 9 Accident Overrepresentation Indicators for Site (from Automated Program)

Curvature	Condition	Time	Speeding	
			Yes	No
Straight	Dry	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
	Wet	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
Less than 2 degrees	Dry	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
	Wet	Weekday, rush hour	[a]	[a]
		Weekday, non-rush	[a]	[a]
		Weekend, day	[a]	[a]
		Evening or night	[a]	[a]
Greater than 2 degrees	Dry	Weekday, rush hour	4.47	3.64
		Weekday, non-rush	[a]	4.54
		Weekend, day	[a]	[a]
		Evening or night	4.76	7.07
	Wet	Weekday, rush hour	4.83	[a]
		Weekday, non-rush	4.86	[a]
		Weekend, day	4.47	[a]
		Evening or night	4.83	3.64

[a]Overrepresentation indicator (z) less than +1.50 or observed number of accidents less than 7.

of some independent variables can be collapsed to form fewer levels. Collapsing the levels of variables, when statistically justified, is particularly desirable because it may allow additional variables to be incorporated in the modeling without sample size problems that may have arisen otherwise.

CONCLUSION

The algorithm reported here was developed to provide traffic engineers a powerful means to identify factors or combinations of factors that cause accident overrepresentation at a site relative to some average. In this way, engineers can use the output from this algorithm to develop remedial options that may be responsive to the problems at that site. Engineers and planners can also use the output to develop an areawide traffic safety improvement plan for the area's highway network by analyzing a number of potential sites and examining the values of accident overrepresentation indicator. Sites and/or factors that show higher values of such indicator may suggest more serious safety problems and a stronger need for safety improvement measures.

Currently, there are available mainframe computer programs to facilitate the statistical computation involved in carrying out the analysis. However, this requires the analysts to have sufficient knowledge of the statistical methods used. An automated microcomputer program has been developed by the Texas Transportation Institute for the Texas State Department of Highways and Public Transportation to eliminate this user requirement. Although the microcomputer program is a modified version of the full algorithm, its applications to date have suggested a comparable outcome to that provided by the full algorithm.

REFERENCES

1. J.R. Landis, E.R. Heyman, and G.G. Koch. Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Automotive Tests. International Statistical Review, Vol. 46, 1978, pp. 237-254.
2. J.E. Higgins and G.G. Koch. Variable Selection and Generalized Chi-Square Analyses of Categorical Data Applied to a Large Cross-Sectional Occupational Health Survey. International Statistical Review, Vol. 45, 1977, pp. 51-62.
3. Y.M.M. Bishop, S.W. Fienberg, and P.W. Holland. Discrete Multivariate Analysis: Theory and Practice. 5th ed. MIT Press, Cambridge, Mass., 1978.
4. J.R. Landis, M.M. Cooper, T. Kenedy, and G.G. Koch. A Computer Program for Testing Average Partial Association in Three-Way Contingency Tables (PARCAT). Biostatistics Technical Report 18. Department of Biostatistics, University of Michigan, Ann Arbor, May 1978.
5. BMDP Statistical Software, University of California Press, Berkeley (undated).
6. ECTA: A Program for the log-linear Analysis of Contingency Tables. The Population Studies Center, University of Michigan, Ann Arbor, Dec. 1975.
7. K.K. Mak, T. Chira-Chavala, and B.A. Hilger. Automated Analysis of High Accident Locations. Proc., National Conference on Microcomputers in Urban Transportation, San Diego, Calif., ASCE, New York, 1985.

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