Automated Analysis of High-Accident Locations

KING K. MAK, T. CHIRA-CHAVALA, and BARBARA A. HILGER

ABSTRACT

A procedure was developed to identify high-accident locations on urban freeways, to analyze the accident experience at these locations, and to determine and evaluate appropriate remedial measures. The procedure consists of (a) a mainframe computer program to identify and rank highway sections by number of injury and fatal accidents per 100 million vehicle miles of travel, (b) a microcomputer program to identify factors overrepresented in accident occurrence at these locations relative to the average for similar highways in the area, (c) a multidisciplinary approach to identify accident causative factors and to devise appropriate remedial measures, and (d) evaluation of remedial measures actually implemented. The procedure is currently being field tested.

Identification of high-accident locations and associated accident causative factors as well as determination and evaluation of appropriate remedial measures at these sites are continuing functions of transportation engineers. This process is time consuming and tedious, requiring extensive compilation and analysis of accident data. Computerized accident data have long been used to identify high-accident locations, but the analyses of accident data to identify causative factors have not been as well developed or automated.

A study is being conducted by the Texas Transportation Institute (TTI) for the Texas State Department of Highways and Public Transportation (SDHPT) to develop a procedure to aid engineers in performing this task in a more systematic and efficient manner. Although the procedure is designed for use with urban Interstate highways and urban non-Interstate freeways, it can easily be modified for use with other highway types. The major components of the procedure are as follows:

1. A mainframe computer program to rank highway sections by using accident rate,
2. A microcomputer program to analyze accident data at selected high-accident locations,
3. A multidisciplinary approach to identify accident causative factors and to devise appropriate remedial measures, and
4. Evaluation of remedial measures actually implemented.

Only the first three steps of the procedure are reported in this paper, with emphasis on the microcomputer program for automated analysis of accidents.

The key steps for the two computer programs and their interactions are illustrated in the schematic diagram as shown in Figure 1. Brief descriptions of the two computer programs are presented as follows.

WINDOW PROGRAM

A mainframe computer program previously developed by TTI for the Texas SDHPT, known as the "WINDOW" program, is used to determine the accident frequency/rate of highway segments and to rank the segments according to the accident frequency/rate. The program utilizes a "window," that is, a highway segment of specified length, which is then moved along the highway network in 0.1 mi increments. For each window, the accident frequency/rate is calculated and compared to that of other windows. Those windows with the highest accident frequency/rate are identified.

The WINDOW program was designed with numerous built-in options to accommodate user-specified inputs, including

1. Years of accident data (1 to 5);
2. Accident selection (subsetting) criteria, for example, county, highway type, accident type, accident severity;
3. Length of window (0.1 to 10 mi);
4. Ranking by accident frequency or rate; and
5. Output format, for example, number of roadway segments to be ranked, reports to be generated.

For this specific application, the latest 3 years of accident data are used. The accidents are subset by county (only one county is studied each time); highway type (urban Interstate highways and urban non-Interstate freeways); accident type (excluding construction zone accidents); and accident severity (injury and fatal accidents only, excluding property-damage-only accidents). A 2-mi long window is used and the roadway segments are ranked by accident rate per 100 million vehicle miles of travel.

Construction zone accidents are excluded from consideration because traffic operating conditions, and hence the accident characteristics, are very different in construction zones when compared to normal highway conditions. The determination of accident frequency/rate is based on injury and fatal accidents only in an attempt to include accident severity in the identification of high-accident locations. Also, this will minimize the impact of differing accident reporting thresholds between various law enforcement agencies within the study area (i.e., county). Some large urban police departments in Texas have adapted the policy of reporting only injury and fatal accidents as opposed to the statewide reporting threshold of injury accidents or accidents involving more than $250 in property damages. It should be noted, however, that all accidents, including property-damage-only accidents, are used in the accident analysis of the procedure.

Traffic volume and other roadway-related data are obtained from the computerized roadway inventory

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

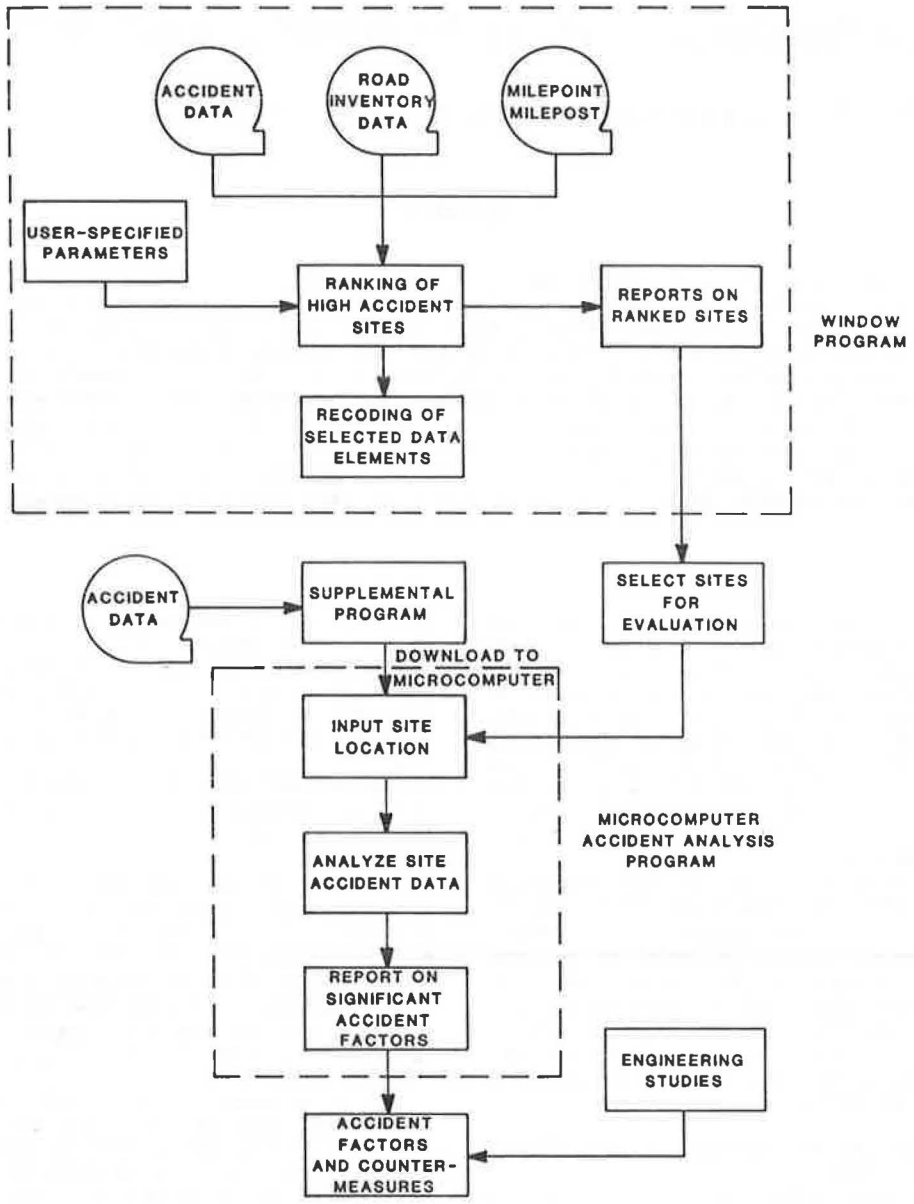

FIGURE 1 Schematic diagram illustrating key steps for the WINDOW program and the automated accident analysis programs.

file. The milepoint-milepost equivalency file establishes a track in going-down-the-highway order. The window is then moved along this track and takes snapshots every 0.1 mi to find the most hazardous locations with the highest accident rates.

The WINDOW program then outputs a user-specified number of 2-mi highway sections ranked by accident rate, as given in Table 1. The section length of 2 mi is selected arbitrarily and can be changed as appropriate. Other reports can also be generated, such as listing of highway sections sorted by highway number and accident counts by 0.1 milepoints.

The user then selects specific locations for evaluation from the list of high-accident locations generated from the WINDOW program. For evaluation purposes, minor changes can be made in the beginning and ending milepoints of the locations to coincide with identifiable landmarks, such as interchanges and bridge structures. These changes, if necessary, are accommodated by the microcomputer program before analysis of the accident data. Each of the high-accident locations selected is then analyzed individually using the microcomputer accident analysis program.

A supplemental mainframe computer program is used to create an accident data file from the state master accident data file for use with the microcomputer accident analysis program. The data file includes all accidents within the study area (which is a county for the purpose of this study) that meets the subsetting criteria used with the WINDOW program, except for accident severity (i.e., property-damage-only accidents are also included in the data file).

Because storage space is limited on the microcomputer, only selected data elements are included in the output data file, a list of which is given in Table 2. Also, many of the data elements are recoded to fewer levels for use with the microcomputer accident analysis program. The subsetting and recoding of the data elements are also handled by the supplemental computer program. The output acci-

TABLE 1 Example Output from WINDOW Program

Rank	Highway District	Highway	Beginning Milepoint: County	Beginning Milepoint: Control Section	Beginning Milepoint: MPT	Ending Milepoint: County	Ending Milepoint: Control Section	Ending Milepoint: MPT	Accidents	Rate (accidents/ 100 MVM)	Fatal Accidents	Fatalities	Injury Accidents	Injuries	PDO Accidents
1	12	US 0059	Harris	177-7	4.5	Harris	177-7	6.5	38	718.49	2	2	36	54	0
2	12	US 0059	Harris	177-11	5.1	Harris	177-11	7.1	282	343.23	12	13	270	377	0
3	12	SH 0146	Harris	389-5	1.1	Harris	389-5	3.1	54	333.06	2	2	52	98	0
4	12	SH 0225	Harris	502-1	7.1	Harris	502-1	10.7	93	319.69	3	3	90	125	0
5	12	US 0059	Harris	27-13	6.2	Harris	27-13	8.2	391	262.24	4	4	387	526	0
6	12	US 0059	Harris	177-11	2.3	Harris	177-11	4.3	197	230.26	8	9	189	283	0
7	12	US 0059	Harris	27-13	8.4	Harris	27-13	10.4	307	227.02	5	5	302	418	0
8	12	SH 0225	Harris	502-1	11.1	Harris	502-1	13.1	54	224.84	4	6	50	78	0
9	12	SH 0225	Harris	502-1	14.4	Harris	502-1	16.4	39	207.07	2	2	37	55	0
10	12	US 0059	Harris	177-11	7.5	Harris	27-13	1.0	187	190.74	5	5	182	263	0
11	12	US 0059	Harris	177-7	9.5	Harris	177-11	1.7	141	190.67	3	3	138	192	0
12	12	US 0059	Harris	27-13	4.1	Harris	27-13	6.1	249	181.05	7	7	242	337	0
13	12	SH 0146	Harris	389-5	3.2	Harris	389-5	5.2	18	170.17	1	1	17	31	0
14	12	US 0059	Harris	27-13	2.0	Harris	27-13	4.0	207	167.05	3	3	204	287	0
15	12	SH 0146	Harris	389-12	9.7	Harris	389-5	0.6	31	161.77	2	2	29	43	0
16	12	SH 0225	Harris	502-1	1.2	Harris	502-1	3.2	104	146.95	6	6	98	139	0
17	12	US 0059	Harris	27-13	12.0	Harris	27-13	14.0	93	114.46	4	5	89	130	0
18	12	US 0059	Harris	177-7	7.1	Harris	177-7	9.1	83	107.21	3	3	80	133	0
19	12	SH 0225	Harris	502-1	3.4	Harris	502-1	5.4	54	105.52	3	5	51	85	0

Note: 1980-1982 Texas on-system accidents—non-Interstate urban freeway. Rank 30, 2-mi segments, main lane Harris County. Subset excludes PDO and construction accidents. Segments sorted by rank for rate.

TABLE 2 List of Primary and Secondary Variables

Variable	Level
Primary	
Accident type	Single vehicle (fixed object)
	Other
	Multivehicle
	Rear-end
	Sideswipe
	Other
Accident time	Weekday, rush hour
	Weekday, nonrush hour
	Weekend, daytime
	Evening/night
Weather/surface condition	Adverse
	Not adverse
Degree of curve	Straight
	Less than 4 degrees
	Greater than 4 degrees
Vehicle type	Passenger car
	Pickup truck/van
	Truck/bus
Secondary	
Accident severity	Fatal and injury
	Property damage only
Driver age	Under 21
	21 to 55
	Over 55
Speeding	Yes
	No
DWI or DW drugs	Yes
	No
Driver license status	Out-of-state or military
	In-state

dent data file is then downloaded onto the microcomputer.

MICROCOMPUTER ACCIDENT ANALYSIS PROGRAM

The microcomputer accident analysis program (MAAP) is designed to provide users a list of accident factors and their interactions that are significantly overrepresented at the location under consideration in comparison to an average. The program is written in turbo-pascal for use with IBM PC-XT or compatible microcomputers with MS-DOS version 2.1 or above. The program has more than 2,300 lines of code and requires 150K of memory. A minimum configuration of 256K memory and a hard disk drive is required to use the program.

The accident analysis methodology is based on the simple concept of overrepresentation. The assumptions are that certain accident characteristics (factors) or combinations of factors, or both, are overrepresented at a high-accident location when compared to the average of similar highway types within the study area (note that a different baseline of comparison can be used as appropriate for other applications), and that these overrepresented accident factors and/or combinations of factors are indicative of accident causative factors at the high-accident location.

The accident analysis is based on a discrete-multivariate algorithm. A two-staged procedure is used: variable selection and modeling. The first stage selects a set of significant variables or factors for further analysis in the second stage. This intermediate step is required because the number of variables that can be simultaneously analyzed in the modeling stage is restricted by the number of accidents at a given site. It is therefore desirable to reduce the number of variables to only those that are statistically significant to minimize the problem of insufficient sample size in the modeling process.

The algorithm for the entire analysis, variable selection and modeling, is completely automated. Users' intervention at any of the intermediate steps is not required. Once a site is specified by the user, the algorithm will start with the variable selection process and automatically proceed to modeling at the end of variable selection. The output of overrepresented accident factors for that site is then printed.

Variable Selection

The purpose of the variable selection process is to narrow down the list of 13 potential variables to only those with significant influence on accident overrepresentation at the high-accident sites. The significant variables are then analyzed in the modeling process while the nonsignificant variables are eliminated from further consideration.

The 10 variables, as given in Table 2, are categorized as either primary or secondary. The primary variables (1 through 5) are considered to be more important because they are directly applicable to

the development of traffic engineering-related countermeasures. The secondary variables (5 through 10) contain mostly driver-related factors and are useful for law enforcement-related countermeasures.

A step-by-step description of the algorithm is presented as follows:

1. Each of the primary variables is cross-classified with the dependent variable (i.e., site versus average) to form a two-way table with accident counts as entries in the cells. Pearson chi-square statistic is calculated for each of these tables. The variable with the smallest p-value (i.e., highest level of significance) is then selected in this initial step.

2. For each of the remaining primary variables, a three-way contingency table is formed among this variable, the dependent variable, and the variable selected in Step 1. A statistic, Q_T, is then calculated (1-3), which reflects both the main effect of this variable and its interaction with the previously selected variable. The variable with the smallest p-value for the Q_T statistic is then selected as the second variable. Also, variables with nonsignificant p-values in the Q_T statistic are eliminated from further consideration.

3. The process in Step 2 is repeated for the remaining primary variables, with the addition of one more selected variable at each step. The process will continue until all primary variables have been either selected or eliminated, or until the data are exhausted. In other words, the data may have thinned out so much that the sample size for a large number of cells in the contingency table becomes too sparse for proper analysis. In such a case, the last entered significant primary variable is dropped and the process as described in Step 2 is repeated with each of the sparse variables. If the Q_T statistic is significant, the sparse variable will be included in the modeling process. If the Q_T statistic is not significant or if the data remain sparse, the variable will be dropped from further consideration.

4. After all primary variables have been evaluated, the selection process is continued for the secondary variables. The process described in Steps 2 and 3 are repeated until all the secondary variables are either selected or eliminated, including the sparse variables.

An intermediate program output, which summarizes the results of the variable selection process, is provided. Each variable is listed as significant, sparse but significant, or nonsignificant. Only variables found to be significant, or sparse but significant, are evaluated in the modeling process.

Modeling

The purpose of the modeling process is to identify and to isolate combinations of levels within the significant variables that contribute to accident overrepresentation at the high-accident location, relative to the average. A step-by-step description of the modeling algorithm is presented as follows:

1. A contingency table on accident frequency (or counts) for the county is created, including all the significant primary and secondary variables previously identified, but excluding those sparse variables that are significant. The cell probabilities for all the cells in the contingency table are then computed. There are a number of ways that these cell probabilities can be obtained (1). The method chosen for this microcomputer program is as follows. For the (i,j,k)th cell, the cell probability, P_{ijk}, is determined by dividing the accident count in the cell (Y_{ijk}) by the overall total ($\sum_{ijk} Y_{ijk}$), that is

$$P_{ijk} = Y_{ijk} / \sum_{ijk} Y_{ijk}$$

The subscripts i, j, and k denote the levels of the selected significant variables.

2. A contingency table for the expected accident frequency of the site under evaluation, E_{ijk}, is then computed based on the cell probabilities of the county determined under Step 1, that is,

$$E_{ijk} = N \times P_{ijk}$$

where N is the total number of accidents for the site under evaluation.

3. Cell residuals are then computed by comparing the actual or observed accident frequencies at the site under evaluation, X_{ijk}, to the expected accident frequencies, E_{ijk}, determined under Step 2. The Freeman-Tukey residuals (4), Z_{ijk}, are then calculated for all the cells of the contingency table:

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

Those cells with Z_{ijk} greater than +1.5 are considered to be significantly overrepresented; that is, the observed accident counts are significantly higher than expected frequency based on the county-wide average. The value of +1.5 is chosen arbitrarily and can be changed as appropriate. These cells are then printed out in descending order of magnitude for the Z_{ijk}'s.

4. This modeling process, as described in Steps 1 through 3, is then repeated for each of those variables that are sparse but significant. Recall that these sparse variables are tested without the last entered significant variable. Thus, the last entered significant variable is also excluded in the modeling process for the sparse variables.

Program Output

The output from the program is illustrated using a study site in San Antonio, Texas. The study site is on a six-lane divided U.S. highway with full access control. The end points of the study site have been adjusted to coincide with interchanges, and the total section length is 2.4 mi. A total of 254 accidents were reported at this site in the 3-year period from 1980 to 1982. Results from the variable selection process are as follows:

Selected	Sparse	Rejected
	Primary Variables	
Degree of Curve	Accident Type	Vehicle Type
Weather/Surface Condition		
Accident Time		
	Secondary Variables	
Speeding	DWI Involvement	Accident Severity
	Driver License Status	Driver Age

The results obtained from the modeling process are summarized in Figure 2. The first four variables (from left to right): degree of curve, weather/surface condition, accident time, and speeding, are those identified as statistically significant on accident overrepresentation and selected by the variable selection algorithm. These significant variables were analyzed first.

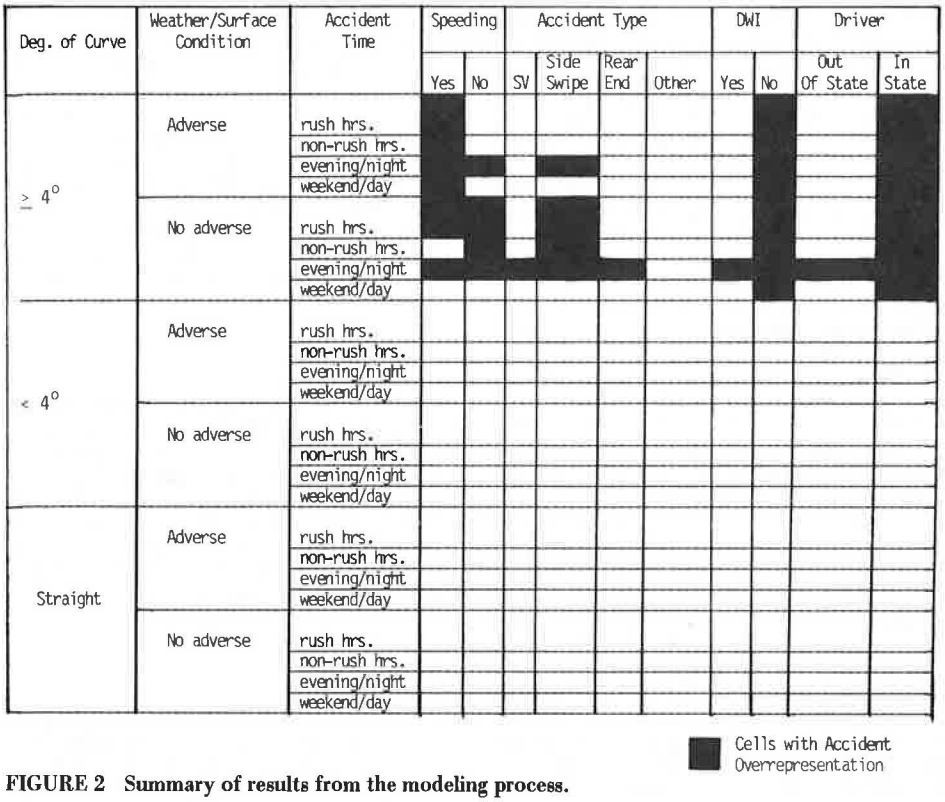

FIGURE 2 Summary of results from the modeling process.

The analysis was then repeated for each of those variables that are sparse but significant by replacing the significant variable that was selected last (i.e., speeding) with one of the sparse but significant variables. For example, accident type replaced speeding as the fourth variable and the analysis was repeated for the following variables: degree of curve, weather/surface condition, accident time, and accident type.

The analysis results indicate the following factors as causes of accident overrepresentation at this site relative to the average for the county:

1. Curve section with curvature greater than 2 degrees;
2. Combination of adverse (wet) weather/surface condition speeding on curve section;
3. Accidents are overrepresented during the time period of evening and night on curve section; and
4. Single vehicle accidents, especially those involving median barriers and rollovers, are overrepresented in the evening or at night on curve section as are sideswipes.

The accident analysis results were then combined with field observations and engineering studies to determine accident causative factors and applicable remedial countermeasures.

FIELD EVALUATION

It should be borne in mind that the results from the MAAP program are only indications of accident factors and combinations of factors that are significantly overrepresented at the location under evaluation. The program cannot and should not replace detailed field studies and sound engineering judgment in the effort to determine potential causative factors and possible remedial measures.

A multidisciplinary team approach is used for the field evaluation. The multidisciplinary team consists of an accident analyst, a traffic engineer, and an analyst with human factors or law enforcement expertise, or both, to provide a broad spectrum of expertise to the evaluation process. Results from the MAAP program and other available information, such as as-built plans, traffic counts, and so forth, are first analyzed to identify potential accident causative factors and remedial measures. The team then visits the location under evaluation to observe and assess the physical and traffic characteristics at the site and to identify potential problem areas and appropriate remedial measures. The site is also videotaped for future reference and further evaluation in the office.

Again using the San Antonio site as an illustrative example, the results of the accident analysis suggest that sharp horizontal curves, low skid resistance, speeding, and night visibility, are candidate accident causative factors. A review of the as-built plans and site visits confirm these potential problem areas.

Because of restrictions in available right-of-way and environmental impact concerns, the design speed of the highway was reduced from the typical 70 mph to 50 mph for the highway section under evaluation. Several sharp horizontal curves are present in the section, with high degrees of curvature. The curve at the beginning of the section is particularly troublesome. First, it is at the end of a long straight section with a downgrade approach. Also, it is a compound curve and the apex of the curve is not evident from the straight approach. Unfamiliar drivers could easily misjudge the sharpness of the curve and fail to respond properly.

Despite a reduction from 55 to 50 mph in the speed limit, speeding appears to be a problem at the site with a median speed of approximately 60 mph. Drivers are actually accelerating when they enter the curve because of the downgrade approach.

The concrete pavement surface is polished, but not slick. Also, the pavement surface is grooved and

the drainage appears good. However, under adverse weather or surface conditions, the demand for skid resistance may be fairly high at the sharp horizontal curves.

Night visibility at the site does not appear to be a problem. The section is lighted and well-delineated with raised pavement markers. Chevron panels have been erected on top of the concrete median barrier to better delineate the curve. Overrepresentation of accidents during evenings and nights may be attributable to other factors, such as increase in speed, alcohol involvement, and so forth.

After conferring with the SDHPT district personnel, a number of remedial measures have been implemented or planned for the site. First, an overhead warning sign with flashing beacons and accompanying advance curve warning sign were installed at the problem curve to forewarn drivers of the curve. The pavement surface was recently rotomilled to increase skid resistance and to improve drainage. Another planned countermeasure is the installation of transverse striping in an attempt to reduce the speed of traffic before it enters the curve. The effectiveness of these countermeasures will be evaluated as they are implemented.

Increased law enforcement at the site was also considered, but not implemented. Previous efforts in increased law enforcement at the site resulted in only temporary improvements. Also, the city police department has limited resources in terms of funding and manpower, and speed enforcement is not necessarily a high priority item. The Selective Traffic Enforcement Program (STEP) would be a good source of funding for this type of activity, but, unfortunately, the city does not participate in this program.

SUMMARY

Two computer programs developed by TTI for the Texas SDHPT have been reported in this paper. The first program, known as the WINDOW program, is designed for use on mainframe computers to identify and rank high-accident locations. This program has been fully operational for some time. An effort is currently underway to incorporate several minor changes into the program to improve its capabilities and flexibility.

The microcomputer program, MAAP, is being field tested with a small number of sites in Fort Worth, Houston, and San Antonio, Texas. A number of improvements are planned for the program and other changes may be identified from the field tests. Most of the planned improvements are in the areas of program output and reporting in an effort to make the program more user-friendly or to improve on the execution time. It is anticipated that the program will be ready for field operation some time in 1987.

Analysis results from these computer programs are then used with field evaluation and sound engineering judgment to determine candidate accident causative factors and remedial measures. This entire process provides a systematic and efficient means of analysis and evaluation in the effort to improve safety at identified high-accident locations.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support and assistance of the Texas State Department of Highways and Public Transportation and the guidance and direction of Herman Haenel, project manager for the study. Thanks are also due to numerous TTI staff, especially Robert Streckfus, Rebecca Yette, Edward Wever, Stephen Chan, and Mark Andrews, for their contributions to the programming and debugging of the computer programs at various stages.

REFERENCES

1. T. Chira-Chavala. An Algorithm for Identifying Accident Factors on Roadway Sections: A Discrete Multivariate Procedure. Report SAR-3. Accident Analysis Division, Texas Transportation Institute, Texas A&M University, College Station, May 1985.
2. J.R. Landis, E.R. Heyman, and G.G. Koch. Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. International Statistical Review, Vol. 46, 1978, pp. 237-254.
3. J.E. Higgins and G.G. Koch. Variable Selection and Generalized Chi-square Analysis of Categorical Data Applied to a Large Cross-Sectional Occupational Health Survey. International Statistical Review, Vol. 45, 1977, pp. 51-62.
4. Y.M.M. Bishop, S.E. Fienburg, and P.W. Holland. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, Mass., 1975, pp. 136-137.

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