TECHNICAL REPORT STANDARD PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/LA-92/248		

4. Title and Subtitle	5. Report Date
Wet Weather Highway Accident Analysis and Skid Resistance Data Management System	June 1992
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
R. C. McIlhenny, K. S. Lee, Y. S. Chen	248 (Volume I)

9. Performing Organization Name and Address	10. Work Unit No.
Industrial Engineering Department Louisiana State University Baton Rouge, LA 70803-6409	11. Contract or Grant No. 90-4SS

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Louisiana Transportation Research Center 4101 Gourrier Avenue Baton Rouge, LA 70808	Final Report (Volume I) 2-1-90 Thru 6-30-92
	14. Sponsoring Agency Code HPR No. 0010(15)

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation Federal Highway Administration.

16. Abstract

The objectives and scope of this research are to establish an effective methodology for wet weather accident analysis and to develop a database management system to facilitate information processing and storage for the accident analysis process, skid resistance testing, and other related tasks. The methodology employed consists of four phases: review and documentation of current LDOTD and LTRC procedures, engineering and statistical review of literature and procedures in the area of accident analysis, identification and recommendation of improvements which may facilitate data management and recovery, and design and development of a new computer information system based on recommendations defined in the third task. An effective wet weather accident analysis, testing, and database management system that allows only needed locations to be identified, tested, and reported is implemented.

Volume II of this report consists of the data base management systems Users manual.

Volume III of this report consists of data base management systems Reference manual.

17. Key Words	18. Distribution Statement
wet weather accidents, skid resistance testing, data management, accident data analysis	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	157	

DOT F1700.7 (1-92)

WET WEATHER HIGHWAY ACCIDENT ANALYSIS AND SKID RESISTANCE DATA MANAGEMENT SYSTEM
(Volume I)

by

R. C. McIlhenny, Ph.D.
Associate Professor of Industrial Engineering
Louisiana State University
Baton Rouge, LA 70803

K. S. Lee, Ph.D., P.E.
Associate Professor of Industrial Engineering
Louisiana State University
Baton Rouge, LA 70803

Y. S. Chen, Ph.D.
Associate Professor of Quantitative Business Analysis
Louisiana State University
Baton Rouge, LA 70803

conducted for

LOUISIANA DEPARTMENT OF TRANSPORTATION AND DEVELOPMENT
LOUISIANA TRANSPORTATION RESEARCH CENTER

in cooperation with
U.S. Department of Transportation
FEDERAL HIGHWAY ADMINISTRATION

The contents of this report 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 Louisiana Transportation Research Center, the Louisiana Department of Transportation and Development or the Federal Highway Administration. This report does not constitute a standard, specification or regulation.

MAY 1992

ACKNOWLEDGEMENTS

The support of Louisiana Transportation Research Center, the Federal Highway Administration and the Louisiana Department of Transportation and Development is highly appreciated.

The research program was proposed and initiated by Dr. June Park, former Assistant Professor of Quantitative Business Analysis at LSU, and by Dr. Kwan S. Lee, Principal Co-Investigator and Associate Professor of Industrial and Manufacturing Systems Engineering at LSU. The time and effort spent by Dr. Park during the initial proposal writing phase of this research project is gratefully acknowledged.

Special thanks are due to Professor Lawrence Mann, Ex-Director of LTRC, Professor Peter Stopher, Director of LTRC, Mr. William H. Temple, Director of Research at LTRC, and Mr. Steven L. Cumbaa, Administrator of Special Studies Research at LTRC, Messers Glenn Chustz, Tom Richardson and Ted Stockwell at DOTD, David Broussard at LTRC for their efficient collaboration to this study.

Special appreciations are due to Professor Julia L. Higle at University of Arizona, Tucson, for her advice. Appreciations also extend to present and former research assistants, Srikanth S. Nagarajan, Alok Satyawadi, Gyana R. Parija, V. Sudhakar, Kasthuri Rangan and Vivek Goel for their technical contributions.

ABSTRACT

The objectives and scope of this research are to establish an effective methodology for wet weather accident analysis and to develop a database management system to facilitate information processing and storage for the accident analysis process, skid resistance testing, and other related tasks. The methodology employed consists of four phases: review and documentation of current LDOTD and LTRC procedures, engineering and statistical review of literature and procedures in the area of accident analysis, identification and recommendation of improvements which may facilitate data management and recovery, and design and development of a new computer information system based on recommendations defined in the third task. An effective wet weather accident analysis, testing, and database management system that allows only needed locations to be identified, tested, and reported is implemented.

IMPLEMENTATION STATEMENT

The results of this research will cut costs by eliminating unnecessary testing and will save lives through the eventual reduction of accidents; these results are embodied in a user friendly system using the SAS package running under the TSO environment. The system has been implemented on IBM 3090 Machine. The software requirement for this package includes SAS/STATS, SAS/AF and SAS/SQL. Programs in SAS/STATS have made use of SAS Version 5.18 and programs in SAS/AF and SAS/SQL have made use of SAS Version 6.06. This system can be implemented in any environment as long as the above mentioned requirements are met. These packages are leased on a yearly basis by SAS Institute based in North Carolina. SAS institute can be contacted at an address mentioned at the bottom of this page. It is recommended that, before installing the system, the user should have a copy of each of the manuals of the packages mentioned above. The manual for using the database system is presented in Volume III of the report.

SAS Institute Inc.
SAS Circle, Box 8000
Cary, NC 27512-8000

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

PROBLEM STATEMENT

Louisiana's Skid Accident Reduction Program is intended to identify and take corrective actions on portions of the highway system which pose an inadequate level of frictional "skid" resistance. This program identifies possible problem areas, i.e., "abnormal" sections, through an analysis of accident reports; tests the identified sections for skid resistance according to ASTM E-274; maintains a computer file system of skid test results for each section; and transmits the results to the Highway Needs, Priorities and Program Engineer for possible inclusion in the Department's construction or maintenance programs. It is felt that in many instances, the current system fails to identify the sections of highway with the highest probability of inadequate skid resistance, and therefore not all sections that need to be tested are correctly identified. Additional problems exist in the logistics of testing and in skid test data storage, maintenance and retrieval.

BACKGROUND OF THE RESEARCH

Accidents occur for various reasons. Some of these numerous factors are represented in a skidding accident analysis model shown in Hankins (1971). This model pinpoints the primary factors in a skidding accident as: (1) highway design, (2) environmental factors such as physical and meteorological environments, (3) driver personalities, and (4) vehicle defects. Due to the numerous

secondary factors involved, a simple analysis scheme would not be sufficient to determine the exact reason for an individual accident, and the hazardous nature of a highway location cannot be ascertained easily.

Although Hankins' model (1971) indicates a variety of factors involved in an accident, the only factor that lies in the hands of the highway agency is the design and condition of the road. Of particular interest in this regard is the number of accidents which may be caused by the slipperiness of roads when they are wet, snowy or muddy. This slipperiness is caused by a loss of frictional force (or the decrease of skid resistance) between the tires of a moving vehicle and the pavement surface due to the presence of water or ice.

A study in 1980 (National Transportation Safety Board, 1980) showed that 13.41% of fatal accidents in Louisiana occurred on wet pavements. In the same study, the number of fatal accidents occurring in Louisiana on wet pavements was above the national average.

A frequency analysis of accident data was performed using the accident data for Louisiana from 1984-1988:

Table 1 : Frequency distribution of accident by surface condition condition of pavement in Louisiana for the years 1984-88

Conditions	Frequency	Percentage
Dry	195393	75.68
Rainy	59659	23.11
Snowy/icy	1052	0.41
Muddy	161	0.06
Other	856	0.33
Missing	1032	0.41

The table indicates that the percentage of wet accidents (a wet accident is defined as an accident which occurs under rainy, snowy/icy or muddy surface conditions) in the time period of analysis was 23.58% while the parishes were, on an average, wet only 6% of the time (see also Tables A1-A8 in Appendix A for a year-by-year analysis of wet accident percentages).

It is the responsibility of the highway agency to reduce the risk of wet pavement accidents. Such risks may be reduced by a comprehensive maintenance and rehabilitation program which calls for thorough inspection of suspected locations and subsequent correction of identified problem associated with these locations.

Unfortunately, most highway agencies do not have sufficient resources to carry out this type of extensive maintenance of rehabilitation year but have to select the top few suspected locations. As discussed earlier, the factors involved in an accident are so many that the exact hazardous nature of a location cannot be ascertained easily. Therefore, the problem is the determination of a comprehensive, justified analysis scheme to select the top few suspected wet hazardous locations every year for further inspection and analysis.

A highway accident analysis system is the total set of procedures for storing, maintaining, retrieving, and analyzing information related to highway accidents [Zeeger, 1982]. These procedures can be divided into two tasks, the accident analysis and the database management of accident, roadway and traffic records. The accident analysis in this research is limited to the

task of identifying the problem location. This identification is very important for maintenance of highways, especially in setting up the priority for the resurfacing of the roadway. To maximize the efficiency of the identification process, reference methods and length of spots or sections, and fixed vs. floating segment lengths were also studied. Database management includes merging or interfacing data files, recovering missing information, processing accident and other data files and reporting.

Prior to the discussion of accident analysis methods, three types of road elements commonly used are defined here:

(1) **Intersections (or junctions):** These are usually defined as the locations where two or more road segments intersect or cross over. Intersections are potentially more hazardous because of the various cross traffic flows. These cross traffic flows are commonly defined to be locations within a 0.1 mile radius of intersecting road segments.

(2) **Sections:** These are sufficiently long road segments with no cross traffic. They are usually of variable or fixed lengths. Some states classify them to be of variable length primarily because of different road elements such as bridges and changing geographical patterns.

(3) **Spots:** These are small road segments with usually high traffic accident potential. They are locations with radii of approximately 0.1 mile and have a total number of accidents greater than or equal to 2.

Hazardous Location Identification Methods

Extensive research has been done in the area of accident analysis, and a variety of methods have been proposed over the years. Some of the most popular methods for identifying hazardous locations are briefly discussed below along with their advantages and disadvantages.

(a) Accident Frequency Method

In this method, the locations are listed in descending order of the total number of accidents, and locations that meet or exceed a predefined accident criterion are identified. This method is simple and most popular in states where section lengths are fixed. The method, however, has several disadvantages. In particular, this method does not take into account differing section lengths and differing traffic volumes, that is, average daily traffic (henceforth ADT) among sections. Many states utilize this method as a preliminary accident file search and then apply another method such as rate-quality control to rank locations for further analysis [Zeeger, 1982].

(b) Accident Rate Method

The accident rate method consists of simply dividing the accident frequency at a location by the vehicle exposure to determine the number of accidents per million vehicles at intersections and other spots (generally defined as 0.3-mile segments or less). For highway sections, the accident rate is computed in terms of accidents per million (or hundred million) vehicle-miles of travel. This method is currently employed in the

state of Louisiana. According to Zeeger (1982), this is an improvement over the accident frequency method as it considers the differing vehicle exposures (i.e. differing ADT).

Several problems can arise in merging the data from the volume file for purposes of rate calculation. The most significant problem involves the overlap of the intersection itself. When conducting a computer search for long sections, problems can also occur, even if cross-street volumes are available in the traffic volume file.

(c) Frequency Rate Method

The frequency rate method is used for identifying locations based on both accident numbers and rates. Usually, this method is applied by selecting a sample of locations that meet the accident frequency criterion and then ranking the selected locations by accident rate. However, some agencies identify locations by rate and then rank them by frequency. For some agencies, a location must meet or exceed both a minimum number of accidents and a minimum accident rate, to be considered for further analysis.

(d) Rate Quality Control Method

This method, proposed by Deacon et al. (1975), is very popular in most of the states in the United States. This method is based on the assumption that the accidents follow a Poisson distribution. The rate quality control method not only entails the calculation of the accident rate at each location, but also a statistical test to determine if that rate is significantly higher

than accident rates for other locations with similar characteristics.

(e) Accident Severity Method

The accident severity method is used to identify and/or rank locations based on the number of severe accidents at each location. Accident severity is defined by the National Safety Council and many states in the following categories: fatal accident, A-type injury (incapacitating) accident, B-type injury (nonincapacitating) accident, C-type injury (probable injury) accident, and PDO (property damage only) accident. One of the severity methods used for comparing highway locations is called the equivalent property damage only (EPDO) method. With this equation, each accident is classified by the most severe injury that occurred, and an accident is counted only once in the equation. Locations are ranked by their computed EPDO number.

(f) Bayesian Analysis

This method is based upon Bayesian theory using gamma distribution. In this method, the accident rate for a location is assumed to follow the Poisson distribution, and the regional accident rates are assumed to follow the gamma distribution which depicts the sum of multiple exponentially distributed random variables. When time between the occurrence of an accident follows the exponential distribution, the number of occurrence follows the Poisson distribution.

The previous methods of analysis usually use total annual accidents as the criterion for selection of hazardous locations.

This is shown to be incorrect in most cases [Hauer, 1986]. Locations whose accident rates were above the mean in a before_period tended to become lower in the after_period, and locations whose accident rates were below the mean in the before_period tended to increase in the after_period. This phenomenon, where a randomly large number of accidents for a certain site during a before_period is normally followed by a reduced number of accidents during a similar after_period, even when no measures have been implemented, is generally termed regression-to-the-mean. It is necessary to correct this bias for appropriate accident evaluation.

(g) Hazardous Roadway Features Inventory

The hazardous roadway features inventory is a method of identifying locations that do not necessarily exhibit a history of high accident experience, but may deserve consideration for improvement based on a potential for high accident frequency or severity. Such locations may be identified for several reasons, including that (1) they do not meet current design standards; or (2) they constitute an obvious hazard to traffic.

Hazards can be located by routine field inventories or by special studies conducted to locate a certain type of hazard. Dangerous roadway features should be routinely identified to prevent accidents. In Illinois, improvements of the state system are considered at locations having a potential for a large number of accidents due to substandard horizontal curves, improper super-elevation, Y-intersections, poor sight distance, etc.

(h) Other Methods

Other methods have been proposed based on regression plots between expected and observed number of accidents. McGuigan (1981, 1982) proposed that the difference between expected and observed number of accidents be the criterion used for ranking. This method has some drawbacks [Maher and Mountain, 1988] because the deviation of observed number of accidents from the expected value may also be caused by random errors.

Also, various other identification methods or combinations of methods are utilized by highway agencies. For example, besides applying the identification methods to a total accident database, many agencies identify specific locations that exhibit an abnormally large number of specific types of accidents. Michigan identifies locations with statistically high numbers of right-angle, rear-end, and left-turn accidents, etc. [Maleck, 1981].

In West Virginia, listings are routinely obtained and reviewed regarding locations with an excessive number of wet-weather, run-off-road, fatal, and night accidents. Also, a "Delta Accident Change" listing is used to analyze locations and produce a list of segments with an unusually high increase or decrease in accident experience as compared to previous years.

Database Management

The current database management system has the following problems:

(1) It is not user-friendly and is not amenable to analysis, tabulation, or graphics. Wet weather accident data should be integrated with all the skid data (including inventory, new materials, legal, special requests, etc.) into an integrated database, and a user friendly application system must be developed to support information retrieval and update on the database. In addition, a user friendly analysis and presentation system is needed. The database management system should provide on-line data entry that includes the capability of data entry from currently used field test data collection equipment.

(2) The skid resistance test data is stored in the field on tape and transferred to the LTRC computer data files at completion of testing. Data is summarized by LTRC and submitted to the Highway Needs, Priorities and Programs Engineer. A historical analysis of abnormal sections or skid resistance test data is not currently performed. A historical database would be beneficial in that trends in skid resistance at a particular site can be observed, and additionally, if a particular site consistently appears on an abnormal list through the years, it can be flagged or highlighted to receive priority consideration for corrective action. The system should include the capability of maintaining historical data files by accident analysis period and location; integrating with construction and maintenance files to determine if an abnormal section has been corrected; ranking section test results by inadequacy of skid

resistance and importance of facility; and highlighting locations that have an adequate level of skid resistance, but where, as indicated by a continual appearance on the abnormal list, other hazardous conditions may exist. The database system must be of a flexible design which will readily integrate with LDOTD's proposed Pavement Management System as outlined in LTRC Research Report 195, "An Integrated Pavement Data Management and Feedback system (PAMS)."

Other related background

(a) Highway Classification Schemes

Appropriate groupings of highway classes or traffic level ranges or other logical subsets are required to be used in the traffic accident analysis when identifying "abnormal" sections.

The classification variables used by various states were reported by Zeeger (1982) as follows:

No. of Agencies	Classification Variables
6	None
8	Functional Classifications
5	Number of lanes
3	Interstate or other
3	Access control
1	Roadway width
1	Lane configuration
2	Others

Zeeger recommended guidelines on the use of classification schemes for the identification of high accident locations.

(b) Fixed versus Floating Segment

Highway spots and sections may be identified from a computerized accident file using either a fixed or floating

segment. Problems with the use of a fixed segment arise when a hazard exists near the boundary of two spots (such as at milepoint 0.6 in the example). In this case, some accidents would be reported in one spot and others in the adjacent spot. Thus, neither of the two spots would be identified as hazardous, and the high accident location would remain undetected. Zeeger (1982) stated that this situation can be partially prevented by using a floating segment length with which a search of the accident file is conducted as the segment length "floats" or moves sequentially by milepost or other reference numbers.

(c) Length of a Spot or a Section

An important consideration in the accurate identification of high accident locations is the selection of appropriate segment lengths for which accident data are to be accumulated. Segments are generally classified as either spots or sections. A spot is a short segment (usually defined as 0.3 mile or less) of highway used to identify problem "point" locations, such as short bridges, curves, intersections, and railroad crossings. A section is usually defined as a highway segment longer than 0.3 mile and is used to identify problems due to inadequate cross section, geometrics, pavement surface, a series of driveways, etc. Most state agencies define 0.1-mile segments, 0.3-mile segments, or intersections (within a certain number of feet) as spots [Deacon et al., 1975]. Deacon et al. also recommended guidelines for the selection of an appropriate spot or segment length.

SIGNIFICANCE OF THE RESEARCH

The current methodology employed by the LDOTD has several shortcomings. The simplicity of the current wet weather accident reporting procedure veils many imperfections in the process. For example, the current method does not take into account the drying time of the pavements after rainfall. The fact that the occurrence of an accident at a location is unpredictable has been ignored. Also, the current methodology assumes that the percentage of wet time at a single weather station is the same as the percentage of wet time in the entire parish. Again, the methodology does not take into account the relative humidity data and the solar radiation data in Louisiana for First Order Stations. Although these shortcomings make the current reporting process simple to operate, they induce a significant error factor.

The results of the research will lead to significant improvements in accident analysis, precise accident location, trigger levels for response to the accident analysis, efficiency in skid resistance test programs, comprehensiveness of available accident and highway data, and ease of usage of the database system. These improvements apply directly to LDOTD and LTRC wet weather skid resistance analysis and testing programs. Additional benefits will be available to other LDOTD programs and activities in other state agencies which may use the accident database or events location strategy implemented here. The ultimate potential benefits will be a reduction in injuries, fatalities, and accident related costs for Louisiana drivers, and budget savings through

efficient and timely detection of high-hazard locations and prompt remedial actions.

OBJECTIVE OF THE RESEARCH

The operational objective of this research was twofold. First, the project sought to develop a new methodology which could identify abnormal sections effectively in conjunction with skid resistance tests. Second, the research aimed to develop a new computer information system, consisting of an integrated relational database and a user-friendly application system, which enables the wet weather accident analysis, skid resistance test, and other related tasks to be carried out correctly, consistently, and with minimal operational cost. The outcome of this research enables enhancement of the safety of the highway system.

SCOPE OF THE RESEARCH

The scope of the research includes four major tasks:

TASK 1:

1. Review current LDOTD Wet Weather Accident Reporting procedures (Cobol Programs).
2. Analyze the process of current data transfer from DPSACC to Master Accident File.
3. Review the process of current report generation.

TASK 2:

1. Review current literature in the area of 'Wet Weather Accident Analysis'.
2. Compare current LDOTD Accident Analysis Procedures with the reviewed literature.

TASK 3:

1. Recommend and implement a data recovery methodology to enhance the quality of the existing database.
2. Implement existing methods of wet accident analysis using SAS.
3. Correct errors and bias in classical statistical analysis schemes.
4. Implement recent Bayesian methods of analysis.
5. Compare methods developed using simulation techniques.

TASK 4:

1. Design and develop an information system consisting of 5 relational tables based on the analysis carried out in task 3.

2. Implement the designed information system and develop a user-friendly application system, taking into account, current and forthcoming LDOTD and LTRC enhancements.

CHAPTER 1

METHOD OF PROCEDURE

A wet weather highway accident analysis method consists of a task of counting the number of accidents for certain segments of the highway for statistical use for any period (usually one year) and identifying or predicting hazardous locations using the historical data of accidents. The former task requires the determination of a good highway classification scheme and assignment of accidents into a certain highway length such as a spot or a section, since the highway is continuous to a certain degree. The latter requires a method which would identify certain segments or locations of highway where accident potential is high. This requires us to look at the trend of the accidents occurrence.

For this research, a review and documentation of the current LDOTD and LTRC processes was undertaken to define, locate, and test abnormal sections. A literature survey was carried out at the same time to investigate and analyze existing procedures employed across the country in the area of accident analysis.

Before selecting a method which identifies the hazardous location; highway classification schemes, fixed versus floating segments, and length of a spot or a section were studied. These factors are required in the accident analysis when identifying hazardous sections and spots.

Then, alternatives for the accident analysis method were devised based on the engineering and statistical evaluation of

existing methods and the constraints identified through the investigation of the current LDOTD and LTRC practices. These alternatives were tested against actual accident and skid test data. After the selection of the identification method, the method as well as related data were analyzed to design a database management system.

1.1 LDOTD WET WEATHER HIGHWAY ACCIDENT ANALYSIS

The existing LDOTD wet weather highway accident analysis uses accident rate method (employing fixed segments of various lengths) and is supplemented by using the National Transportation Safety Board (NTSB) wet pavement accident index (1980). The wet weather accidents are read from the DOTDACC file which is the corrected and edited master accident file for each year. A program in Easytreive calculates the wet pavement accident index which is used in identifying hazardous locations. The wet pavement accident computation is a subset of the accident analysis schemes developed at LDOTD.

The wet pavement accident index used in the calculation is given by the following equation:

$$\text{Wet pavement accident index} = \frac{(\%\text{ wet accidents})/(\%\text{ dry accidents})}{(\%\text{ of wet time})/(\%\text{ of dry time})} \quad \ldots\ldots(1)$$

where the % of wet time is calculated from hourly precipitation data. The number of wet hours at a place is based on the NTSB

criterion of an hour being wet if it receives over 0.01 inch of rainfall. This index is used to rank locations which may be intersections, sections, or spots.

The current LADOTD procedure to identify "abnormal" wet weather accident analysis comprises of the following steps:

(1) Determine statewide average for wet accidents / MVM for each highway class.
(2) If a subsection has more than 2 times the statewide average than retain.
(3) If 5 or more total wet + dry accidents occur within a subsection than retain.
(4) If 2 or more total wet accidents occur within a subsection than retain.
(5) If wet safety factor for a subsection is ≤ 0.67 then retain.

This current methodology has several disadvantages:

(1) It does not take into consideration the drying time of pavements after rainfall.
(2) It does not consider the fact that accident rates at a location are unpredictable.
(3) It assumes that the percentage of wet time in a parish is the percentage of wet time at a single weather station.
(4) It introduces a lot of redundancy in that each highway class and location type have to be run separately.
(5) It does not distinguish between the asphalt and concrete drying times.

(6) It does not consider the relative humidity and solar radiation data available in Louisiana for first order stations.

1.2 CLASSIFICATION, SEGMENTATION, AND LENGTH OF HIGHWAY SECTION FOR ACCIDENT ANALYSIS

Factors which are necessary for the accident analysis but are not dependent upon a specific location identification method were studied to improve the effectiveness of the accident analysis.

1.2.1 Highway Classification, Segmentation and Location Schemes

Existing highway classification schemes are standardized in Louisiana and are in accordance with the guidelines of the Traffic Engineering and Design Handbook. Therefore, changes in highway classification schemes are not necessary for this study. The spot has been redefined for the purposes of this study as a cluster of accidents within a radius of 0.1 mile not involving any intersection for the segmentation purpose.

In the current highway classification scheme, every location may be uniquely identified by the "control section" and "beginning control section log mile". Highway classes are identified as:

1. Rural 2-lane
2. Rural other
3. Rural multi-lane divided
4. Rural interstate
5. Urban 2-lane
6. Urban other
7. Urban multi-lane
8. Urban interstate

These 8 highway classes are quite different in traffic speeds, traffic volume, and road lengths. So, the analyses for all methods

are separate for each highway class. Highway numbers such as I10 (coded as I010 in the accident file) and I090, etc. are well known to any motorist.

"Control" and "section" numbers are uniquely defined in the control section data file which is then merged with the accident analysis file. Control is a 3 digit code while section is a 2 digit code. Parishes are numbered from 1 to 64, and, likewise, districts are number coded. "Beginning of control section log miles" are usually given to one-hundredths of a mile, such as 13.64, etc.

For reference, the guidelines on the use of classification schemes suggested by Zeeger (1982) are listed below:

(a) A distinction should be made between locations in rural and urban areas because of differences in accident patterns, frequencies, and severity.

(b) Further classification is desirable according to number of lanes and/or such factors as median separation and access control.

(c) Intersections should be distinguished, if possible, from other types of spots. Accident patterns at intersections are generally different from those at other spot locations, because exposure to traffic consists of vehicles entering the intersection on all approach legs. Many agencies report accident locations to the nearest intersection or with the distance to the nearest intersection. Some agencies define an intersection

accident as one that occurs within a specified distance from an intersection.

(d) The identification of spots or sections by highway class generally requires the interfacing of a roadway file with the accident file, as is the procedure in Michigan, West Virginia, and California.

(e) With most identification methods, the comparison of locations within similar groups is highly desirable. With the rate quality control method, for example, a major factor in the computation of the critical accident rate is the average rate for locations with similar characteristics. This includes locations both with and without accident experience. It should also be emphasized that the use of too many classification groups is also undesirable. If the number of classification groups is large, the number of sites per group will be small, and few or no locations will be identified within each group as having accident numbers or rates significantly higher than the group average.

1.2.2. Fixed versus Floating Segments

It is important to look at the highway network as a total system rather than merely as a combination of independent segments. In many cases, the presence of several high accident spots on a highway section may be due to more than just an isolated roadway deficiency. A roadway safety problem that extends for several miles may exist. Such a problem requires the consideration of

improvements on a broader scale than would be considered for an individual high accident spot location.

The accident rate at a location is commonly assumed to follow a Poisson process [Deacon, et al., 1975; Higle and Hecht, 1989], and hence accident rates are also unpredictable (random variables). The randomness in the accident rate and location poses a difficult problem to the highway agencies who try to identify specific locations with high accident potential.

Some accidents may not occur randomly but may have an underlying singular cause, such as an inadequate road design [Zeeger, 1982]. Agencies, therefore, also need to identify locations of particularly small radii which are extremely hazardous. In practice, such locations fall into the category of spots. Spots are very useful for accident analysts for the following reasons:

(1) Underlying causes are easier to identify for spots because, accident spots are often caused by singular identifiable factors.

(2) Subsequent corrective measures are easier and more economical.

Study in this section will be confined to the identification and analysis of spots. The traditional methods used to identify accident spots are fixed points and floating segments. These are defined as follows:

A. Fixed Point Schemes: Accidents occurring within specified control section log mile limits (0.1 increments) may constitute a

spot. Another scheme adopted by some states is to calculate the number of accidents occurring within a 0.1 mile radius of a single mile post. If the number of accidents is greater than 2, then the area around the mile post is termed a spot.

B. Floating segment methods: There are several distinct disadvantages in using fixed beginning and ending control log miles for spot identification [Zeeger, 1982]. For example, very often accidents occur at the boundary between two sections of a road. In such cases, fixed point methods attribute some of the accidents to one section and some to the other. This method results in poor spot identification. So, a floating segment is used.

To illustrate, if a roadway has a length of 10 miles and the required spot is one of 1 mile along this section, then the floating segment scheme proceeds as follows:

(1) Scan 0.0 to 1.0 log mile for the number of accidents (say N1).
(2) Now, scan 0.1 mile to 1.1 log mile for the number of accidents (say N2).
(3) If N2 is greater than N1, then choose N2; otherwise choose N1.
(4) Continue steps 1 to 3, incrementing as needed till the end of the section is reached.

This method is widely used in accident analysis systems and is better than fixed points for analysis. However, the proposed cluster analysis scheme is highly time efficient. It achieves the

same results of floating segment methods and also provides a good base for further statistical analysis of accidents.

1.2.3. Length of a Spot or a Section

An important consideration in accurately identifying high accident locations is the selection of appropriate segment lengths for which accident data are to be accumulated. Segments are generally classified as either spots or sections. A spot is a short segment (usually defined as 0.3 mile or less) of highway used to identify problem "point" locations, such as short bridges, curves, intersections, and railroad crossings. A section is usually defined as a highway segment longer than 0.3 mile and is used to identify problems due to inadequate cross section, geometries, pavement surface, a series of driveways, etc. Most state agencies define 0.1-mile segments, 0.3-mile segments, or intersections (within a certain number of feet) as spots (Deacon et al., 1975].

Louisiana uses a fixed point variable section length partly because adjacent locations are separated by geographic road elements, such as narrow bridge, etc. However, as mentioned in Section 1.2.4, since the cluster analysis was used as a method to identify the accident location, the cluster length and the radius of the cluster are more important than the section length and spot length. Cluster length and radius are discussed in the next section.

For reference, the guidelines recommended by Zeeger (1982) are also listed below:

(a) A spot or a section should have consistent characteristics in terms of geometries, traffic volumes, and class of highway. Selection of such a spot can be best accomplished by using a traffic volume and roadway file to supplement the accident file.

(b) The spot or segment length should be no smaller than the minimum increment for reporting an accident location. For example, if accidents are reported to the nearest 0.1 mile, then the spot length should be no smaller than 0.1 mile.

(c) The length should be selected to account for the suspected degree of error in reporting accident locations. If the degree of accuracy is low, then longer segment lengths will be needed to minimize the error. If a state estimates accident locations to be accurate only within about 0.4 mile, a minimum segment length of about 0.8 to 1.0 mile should be used. Using a 0.1 mile segment in this case would likely pick up a large number of incorrect locations and would not identify the truly hazardous locations.

(d) The spot length should be at least as large as the area of influence of a highway hazard. An accident scene may extend for several hundred feet, and a dangerous curve may often contribute to accidents that occur several hundred yards apart. Thus, spot lengths of 0.2 or 0.3 mile often provide more appropriate results than spot lengths of 0.1 mile or less, particu-

larly in rural areas where the area of influence of a hazardous spot (e.g., a horizontal curve or narrow bridge) may often extend for about 0.3 mile.

(e) The length of a spot has a direct impact on the reliability of the identification of high accident locations. The errors in identifying hazardous locations caused by the random nature of accident occurrences can be minimized by the use of longer spots. Too short segment lengths can also give erroneous results when accident rate or accident severity is the measure of safety hazard. Accident rates (in accidents per million vehicle-miles) become unstable and of questionable value for highway segments of short length (i.e., less than 0.3 mile) and/or with low traffic volumes (i.e., less than 500 vehicles per day), even when several years of accident and volume data are used.

(f) It is recommended that two or more lengths be used by an agency to identify locations for further analysis. One short spot length (0.2-0.3 mile) and one longer section length (1-2 miles or a variable-length section) should be suitable for most agencies.

1.2.4. Cluster Length and Analysis

Cluster analysis is a nonlinear optimization algorithm which minimizes the number of clusters (spots). The distance between two clusters (center of the clusters) of accidents is always greater than the specified radius of the required spot, i.e. 0.05 mile

radius for a 0.1 mile length spot. The cluster analysis procedure used is a non-hierarchical clustering procedure; that is, no accident may belong to two clusters at the same time. The procedure FASTCLUS in SAS provides an extremely fast clustering procedure based on Hartigan's leader algorithm and Macqueen's k-means algorithm (SAS, 1985). This method proceeds as follows:

Step 1: Generate random seeds of accident locations from the accident file and assume these as cluster centers.

Step 2: Select accidents within a 0.05 mile radius (0.1 mile diameter) of the random seeds selected in step 1 as members of the cluster.

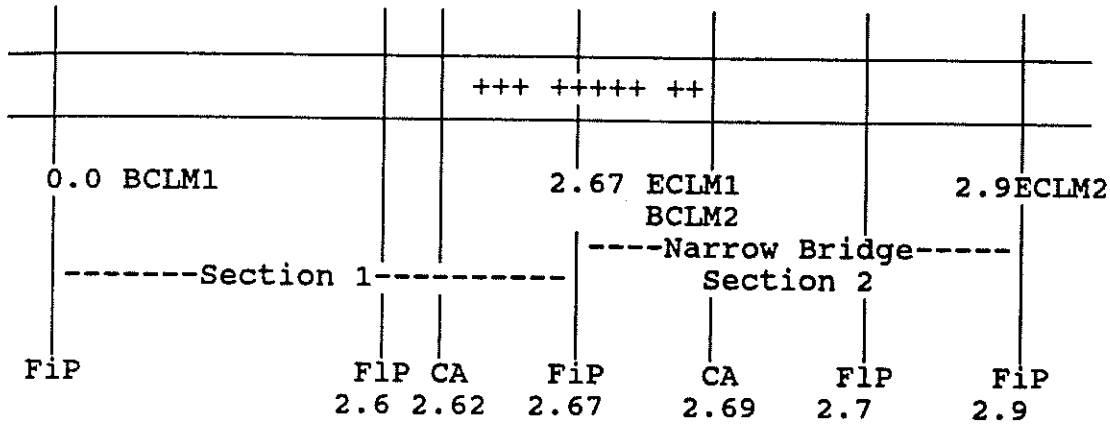
Step 3: Compute the new center of the cluster so formed in Step 2 as the current mean of the cluster.

Step 4: Repeat Steps 2 and 3 until the seed distance between iterations does not differ by 0.001 mile.

To illustrate the use of cluster analysis, assume two sections, 1 and 2, as shown in Figure 1. Section 1 starts at 0.00 log mile and ends at 2.67 log mile. A narrow bridge starts section 2 from log mile 2.67 and ends at 2.9. Let there be a cluster of accidents at the boundary, as indicated by the '+' marks.

If one were to go by fixed segments, some of the accidents would be attributed to section 1 and the rest to section 2, thereby tending to reduce the hazardous nature of the site as such. If the segment were allowed to float in steps of 0.01 mile, then the hazardous spot would lie between 2.62 and 2.72. However,

cluster analysis results would give us a spot starting from 2.62 to 2.69 as the wet hazardous one.



Legend		Log miles indicated by method
FiP	- Fixed Point Method	0.0 to 2.67 and 2.67 to 2.9
FlP	- Floating Point Method	2.6 to 2.7
CA	- Cluster Analysis	2.62 to 2.69
BCLM	- Beginning of Control Log Mile	
ECLM	- End of Control Log Mile	

Figure 1. Advantages of cluster analysis over floating point

The reduction of the radius of the spot using cluster analysis instead of floating segment is not by itself, wholly advantageous. This is because there is always a probability of error in identifying a location [Deacon, 1975] within 0.01 mile. In addition, the cause of the accidents involved in a spot may be features or factors which are at a distance from the actual cluster of accidents. For this reason, an allowance for error (say a distance of 0.02 mile) is given to each cluster after identification.

Deacon et al. (1975) recommended that a spot (cluster) length be within 0.1 and 0.3 mile. So, for analysis, this length of 0.1 mile is assumed to be reasonable. Lesser lengths are not accurate enough because of the probability of error in locating an accident.

The wet accident data for a single year are first sorted in the required order of highway class, highway number and control. Thus, a single highway belonging to a control and having several sections is analyzed separately for clusters.

The length of the cluster used is 0.1 mile, and the variable used for clustering is the computed accident point as indicated by log miles. The convergence criterion is a distance of 0.001 mile between two cluster seeds.

After clustering, the clusters so formed are analyzed using the accident analysis methods discussed later. The cluster analysis procedure has therefore several distinct advantages:
(1) It is easier to incorporate.

(2) It is faster (no querying) and more accurate (mathematically sound).

(3) It also provides a scheme of hierarchy in spot identification, i.e., the distance from the adjacent cluster, number of clusters within a 5 mile radius, etc. are known.

(4) Traditional accident analysis schemes may be run using these clusters as spots.

(5) The tendency of clusters to repeat themselves over the years may lead to a hazardous road feature such as blind spots and other obstructions to the driver.

1.3. DEVELOPMENT OF THE WET WEATHER HIGHWAY HAZARDOUS LOCATION IDENTIFICATION METHOD

In this section, we start with an overview of the current hazardous location identification methods being used in other states. A detailed description of the recommended Bayesian method is also given. Following this, a comparison of some of the existing methods with the Bayesian method is presented to delineate the strengths of the recommended Bayesian method.

1.3.1. Current literature which can be used to identify the hazardous locations on a Highway

A summary of the methods used in each state for identifying hazardous locations is given in Table 2 [Zeeger, 1982]. Most states utilize several different methods for identification

Table 2 METHODS FOR IDENTIFYING HAZARDOUS LOCATIONS

STATE	INTERSTATE							STATE							LOCAL						
	A	E	L	R	S	Y	Z	A	E	L	R	S	Y	Z	A	E	L	R	S	Y	Z
ALABAMA							X	X			X	X			X			X	X		
ALASKA							-	X			X	X			X			X	X		
ARIZONA	X			X	X			X			X	X			X				X		
ARKANSAS	X			X	X	X		X			X	X	X							X	
CALIFORNIA	X			X				X			X				X			X			
COLORADO	X			X	X			X			X	X			X				X		
CONNECTICUT	X			X		X		X			X		X		X					X	
DELAWARE	X	X			X			X	X			X									-
FLORIDA			X								X										1/
GEORGIA	X		X	X	X			X		X	X	X			X						
HAWAII	X	X		X				X	X		X				X	X		X			
IDAHO	X		X	X				X		X	X				X		X	X		X	
ILLINOIS				X							X				X				X		
INDIANA	X						X	X						X	X					X	
IOWA	X	X		X	X			X	X		X	X			X	X		X	X		
KANSAS	X			X				X			X									X	
KENTUCKY	X		X	X			-	X		X	X				X						
LOUISIANA	X			X	X			X			X	X									*
MAINE	X			X				X			X				X			X			
MARYLAND	X	X	X	X	X			X	X	X	X	X			X						
MASSACHUSETTS	X		X	X	X			X		X	X	X			X		X	X	X		
MICHIGAN	X	X	X	X	X			X	X	X	X	X	X		X	X		X	X		
MINNESOTA	X			X	X			X			X	X								X	
MISSISSIPPI	X			X				X			X				X			X			
MISSOURI	X		X	X				X		X	X									X	
MONTANA	X			X	X			X			X	X			X		X	X	X		X
NEBRASKA	X	X		X	X			X	X		X	X			X	X		X	X		
NEVADA	X			X	X			X			X	X			X			X	X		
NEW HAMPSHIRE	X				X	X		X				X	X		X				X	X	
NEW JERSEY	X	X	X	X	X	X		X	X	X	X	X	X		X	X		X	X		
NEW MEXICO	X			X				X			X									X	
NEW YORK	X	X		X	X			X	X		X	X			X	X		X	X		
NORTH CAROLINA	X			X	X			X			X	X			X			X	X		
NORTH DAKOTA	X		X	X	X			X		X	X	X			X			X		X	
OHIO	X			X				X			X				X			X			
OKLAHOMA				X	X	X					X	X	X		X				X		
OREGON	X	X		X	X			X	X	X	X	X			X	X		X	X		
PENNSYLVANIA	X			X	X			X			X	X			X				X		
RHODE ISLAND	X	X	X		X			X	X	X		X			X	X	X		X		
SOUTH CAROLINA	X			X	X			X			X	X						X	X		
SOUTH DAKOTA	X				X			X				X			U	N	K	N	O	W	N
TENNESSEE	X	X	X	X	X			X	X	X	X	X			X	X	X	X	X		X
TEXAS							*	X		X	X				X					X	
UTAH	X			X	X			X			X	X			X			X	X		
VERMONT							*							*							*
VIRGINIA	X	X		X	X			X	X		X	X			V	A	R	I	E	S	
WASHINGTON	X			X	X			X			X	X			X			X			
WEST VIRGINIA	X			X	X			X			X	X							X		
WISCONSIN	X	X		X	X		X	X	X		X	X		X	X	X		X	X		
WYOMING	X			X	X			X			X	X			X			X	X		
DIST. OF COL.	X	X		X				X	X		X										-
PUERTO RICO							-					X							X		
SAMOA							-	X		X	X	X									-
GUAM							*							*							*
VIRGIN ISLANDS							*							*							*

ot applicable.
ot reported.
omplete coverage not planned.

us Location Identification Criteria Codes

lumber of accidents.
conomic loss/accident cost.
specific number of locations (e.g., top 100).
ccident rate, including rate-quality control.
ccident severity.
ther.

Y Codes

Arkansas: Requests from district engineers, maintenance personnel, and law enforcement agencies.
Connecticut: Modified rate-number quality control method.
Idaho: Input from local jurisdictions.
Kansas: Pin maps, 402 safety studies.
Michigan: Accident patterns.
Minnesota: Local authorities criteria, priorities, and funding systems
New Hampshire: Input from public and maintenance division offices.
New Jersey: For overlay projects—percent of wet-weather accidents, skid number, on-site investigation.
Oklahoma. Field reviews.
Texas: Safety improvement index.

purposes. Nearly one-fourth of the states are currently developing a methodology for location identification for one or more types of highways. In addition, virtually all states consider for improvement all locations where a fatal accident has occurred.

Most of the available methods were briefly discussed in the introduction. A more detailed description is given here.

1.3.1.1. Accident Frequency Method

Accident frequency method is used to search the accident file for concentrations of accidents within a fixed or variable segment length. Usually, one or more segment lengths (0.01-mile, 0.3-mile, 0.5-mile, 1-mile, 3-mile, etc.) are used to "float" through the accident file in which accidents are ordered by location (e.g., by county, route number, and milepoint, or by sequential reference points with the distance and direction from each reference point), and selections that meet or exceed a predefined accident criterion are identified. Such floating segments generally advance in 0.1-mile increments through the file. When a roadway segment that meets the user-specified frequency criteria is identified, the location is printed out along with the corresponding accident information.

Several different segment lengths and/or years of accident data (usually 1-5 yrs.) are often used for file searches. Also, the accident criteria for selecting highway segments usually vary according to area type (urban, rural) or other classification variables (number of lanes, functional class, etc.). The computer

program is generally written to rank the identified highway segments in descending order by accident frequency. Many agencies utilize the frequency method as a preliminary accident file search and then apply another method (rate-quality control, severity, rate, etc.) to rank locations for further analysis.

1.3.1.2. Accident Rate Method

According to Zeeger (1982), this is an improvement over the accident frequency method as it considers the differing vehicle exposures (i.e. differing ADT). The accident rate for spots and intersections is given by :

$$Rsp = (A)\ (1{,}000{,}000) \ /\ (365)\ (T)\ (V) \qquad \ldots\ldots(2)$$

where,

Rsp = accident rate for the spot/ intersection (in accidents per million vehicle miles entering the spot/intersection),

A = number of accidents for a given analysis period,

T = time of analysis period (in years or fraction of years), and

V = average annual daily traffic (ADT) during study period.

Similarly, the formula for accident rate for sections is given by

$$Rse = (A)\ (1{,}000{,}000) \ /\ (365)\ (T)\ (V)\ (L) \qquad \ldots\ldots(3)$$

where,

Rse = accident rate for highway section (in accidents per million vehicle miles), and

L = length of section (in miles).

This method is still commonly used in most states including Louisiana as a measure of accident potential. Spots or sections are ranked in order of descending accident rate. When the computer accident file is used in implementing the accident rate method, interfacing with the traffic volume file is required. The highway reference method is the controlling variable in the program. The search through the accident file may be started in a similar manner as that described for the accident frequency method. However, in a pure accident rate calculation, every highway segment identified with one or more accidents will be located. The traffic volume file must be formatted by a compatible location reference method in the same order as the accident file.

Several problems can arise in merging the data from the volume file for purposes of rate calculation. The most significant problem involves the overlap of the intersection itself. It is desirable for the volume file to be arranged to include the cross-street volumes along the major street. Without such cross-street volumes, the computed rate for intersections from the merged files will not include the cross-street volume, and the rate will be erroneously computed, which could cause large errors in the true accident rate value, particularly when accident rates for short highway sections (less than 1 mile) are computed. In California, intersection records contain cross-street volumes, and high-accident intersections are analyzed separately from highway segments and are not duplicated in the segment analysis.

When conducting a computer search for long sections, problems can also occur, even if cross-street volumes are available in the traffic volume file. For example, suppose that a 2-mile section is used to float through the accident file and accident rates are computed by interfacing with the traffic volume file. There should be some mechanism to account for the intersections with the section. Most accident files do not allow for easy recognition of the locations of intersections along a route. An exception to this is when a reference method of links and nodes is used and accidents and volumes are tied to those nodes and links. Then, intersection accidents and corresponding volumes may be interfaced. In a similar manner, accident rates for the links can be computed by retaining the full link distances between nodes.

1.3.1.3. Frequency Rate Method

The frequency rate method [Renshaw and Carter, 1981] is used for identifying locations based on both accident numbers and rates. Usually, this method is applied by selecting a sample of locations that meet the accident frequency criterion and then ranking the selected locations by accident rate. However, some agencies identify locations by rate and then rank them by frequency. As explained earlier, in some agencies, a location must meet or exceed both a minimum number of accidents and a minimum accident rate, to be considered for further analysis.

The frequency rate method may also be applied by developing a plot of accident frequency categories (0-2, 3-5, 6-10, etc.) and

rate categories. This results in a two-dimensional accident data matrix, in which any highway location may be placed in a single matrix cell representing a given level of accident frequency and rate. The matrix cells in the upper right corner represent the most hazardous locations, which will be given top priority for further analysis. The matrix cells in the lower left corner denote the locations with the lowest priority. The frequency and rate categories on the x and y axes can be changed to best suit the type of highway, the time period of the accidents analyzed, and other user needs. To use the frequency rate method in this manner, the user must define the combinations of frequency and rate corresponding to priority 1, priority 2, etc.

1.3.1.4. Rate Quality Control Method

This method, proposed by Deacon et al. (1975), is very popular in most of the states in the United States. This method is based on the assumption that the accidents follow a Poisson distribution. Thus, the following formulae hold:

$$P(n) = e^{-\alpha m} (\alpha m)^n / n! \qquad \ldots\ldots(4)$$

where,

P(n) = probability that n accidents will occur at a given location during a given time period,

e = base of natural logarithms,

α = expected accident rate in accidents per million vehicle miles, and

m = number of vehicle miles in millions

The critical rate (i.e., the upper control limit above which the accident location is termed hazardous) is given by the following approximation [Morin, 1967]:

$$CR = \alpha + k\sqrt{(\alpha/m)} + 1/(2m) \quad \ldots\ldots(5)$$

where,

CR = critical rate for a particular road location (traffic accidents per million vehicle miles),

k = a probability factor determined by the level of significance needed for CR, and

The probability factor k is commonly chosen to be 1.645 at a significance level of 0.05. α is usually replaced by the state-wide average for the highway class to which the location belongs. The locations are analyzed within separate highway classes, i.e. rural 2- lane, urban 4-lane divided, etc.

The rate quality control method not only entails the calculation of the accident rate at each location, but also a statistical test to determine if that rate is significantly higher than accident rates for other locations with similar characteristics. The statistical test is based on the commonly accepted assumption that accidents follow a Poisson distribution. A probability level for judging the statistical significance is selected to ensure that an accident rate is sufficiently large so that it cannot be reasonably attributed to random occurrences. Selecting higher confidence levels results in fewer locations being identified as having critically high accident rates. The critical accident rate is computed for each location and compared to the

actual accident rate. If the actual accident rate exceeds the critical rate, then the location may be considered for improvement. A series of critical rate curves was developed in Kentucky for urban intersections of arterial and collector streets. In practice, some states select locations for further analysis if the accident rate is 2 or 3 times higher than the critical rate.

1.3.1.5. Accident Severity Method

The accident severity methods are used to identify and/or rank locations based on the number of severe accidents at each location (National Safety Council, 1976). Accident severity is defined by the National Safety Council (1976) and many states in the following categories: fatal accident, A-type injury (incapacitating) accident, B-type injury (non-incapacitating) accident, C-type injury (probable injury) accident, and PDO (property damage only) accident. One of the severity methods used for comparing highway locations is called the equivalent property damage only (EPDO) method which is given by (valid only for Kentucky):

$$EPDO = 9.5\ (F+A) + 3.5\ (B+C) + PDO \qquad \ldots\ldots(6)$$

where,

EPDO = Equivalent Property Damage Only,

F = number of fatal accidents,

A = number of A-type injury accidents,

B = number of B-type injury accidents,

C = number of C-type injury accidents, and

PDO = number of PDO accidents.

The locations are then ranked in the descending order of equivalent property damage. This method cannot be used in Louisiana due to the lack of injury data classifications such as A-type or B-type injuries, etc., in the accident data file.

With this equation, each accident is classified by the most severe injury that occurred, and an accident is counted only once in the equation. Locations are ranked by their computed EPDO number. In North Carolina, an EPDO rate is computed by taking into account both frequency of severe accidents and vehicle exposure.

Another accident severity method, called the relative severity index (RSI), is used to compute average accident costs for a particular accident type. Accident costs are based on the distribution of fatal, injury, and property-damage accidents that occur on each type of highway. RSI values can also be computed for each accident type (right-angle, rear-end, etc.) within each highway type. A third accident severity method for identifying problem locations involves the identification of locations with a minimum frequency of severe accidents (i.e., fatal plus injury accidents) in a given time period. This method can be classified as either a frequency or a severity method.

In using an accident severity method, the program is usually written to search all the severity columns and select the most severe injury to any driver or passenger in any vehicle. With a variable-length file of numerous vehicles and/or passengers, this involves searching the injury codes for each occupant. To compute

a severity rate, interaction must be made with the traffic volume file.

1.3.1.6. Hazardous Roadway Features Inventory

This method identifies locations that do not necessarily exhibit a history of high accident experience, but may deserve consideration for improvement based on a potential for high accident frequency or severity. Hazards are located by routine field inventories or by special studies conducted to locate a certain type of hazard. Dangerous roadway features are routinely identified to prevent accidents. Improvements are considered at locations having a potential for a large number of accidents due to substandard horizontal curves, improper super-elevation, Y-intersections, poor sight distance, etc.

1.3.1.7. Bayesian Methods

This method is proposed by Higle and Witkowski (1988) based upon Bayesian assumption [Morris, 1983; Berger, 1985]. The assumption of Poisson accident rates for a location as in equation (4) is carried over here:

$$P(n) = e^{-rV} (rV)^n / n! \qquad \ldots\ldots(7)$$

where,

P(n) = probability that n accidents will occur at a given location during a given time period,

e = base of natural logarithms,

r = accident rate, and

V = volume of vehicles in the time period of analysis.

Based on a gamma assumption of regional accident rates, the gamma parameters α and β are calculated for every site from the available accident data using the method of moments technique.

1.3.2. Weaknesses of the Currently Used Location Identification Methods and Need for the Bayesian Method

As indicated in Table 2, most states including Louisiana, are using one or several methods among accident number (accident frequency method), accident rate, including rate-quality control method, accident severity method, and others, such as economic loss/accident cost to identify hazardous locations. Table 3 shows the summary of advantages and disadvantages of these methods. Tables 4A-4D show the comparisons of these methods in different categories. Most of the weaknesses of each method were discussed in the previous sections (1.2 and 1.3.1). However, there are two weaknesses which would affect the accuracy of identification. They are regression-to-mean effect and counter-measure effect.

The regression-to-mean effect is the phenomenon in which locations whose accident rates were above the mean in a before_ period tended to become lower in the after_period and locations whose accident rates were below the mean in the before_period tended to increase in the after_period. This effect results from the use of total annual accidents as the criterion for selection of hazardous locations in the above-mentioned methods,unlike in the Bayesian method. This is shown to be incorrect in most cases [Hauer, 1986].

It is necessary to correct this bias for appropriate accident evaluation.

The intersection data for East Baton Rouge Parish in Louisiana for 1984-85 with 1984 as the before_period and 1985 as the after_ period were used to analyze this effect of regression-to-mean. The results are shown in Table 5. These results seem to agree well with the experimental results of Hauer (1986). See also Tables A6 and A8 in the Appendix A for additional results.

The regression-to-mean corrections may be completely avoided if the accident rate is assumed to be a random variable. This is accomplished by the Bayesian Analysis method. The second effect is the counter-measure effect which usually accompanies the regression-to-mean effect. This occurs for sites overlaid or reconstructed during the period of analysis. These locations have a lower expected accident rate when compared to others. Thus, the sample population is biased by including such overlaid sites if the current methods are used.

Table 3. COMPARISON OF METHODS OF ACCIDENT ANALYSIS

METHOD	ADVANTAGES	DISADVANTAGES
A) Accident Frequency Method	Simple and popular where section lengths are fixed.	Cannot be applied when section lengths and traffic volumes differ. Can be used only as preliminary accident file search.
B) Accident Rate Method	Considers differing vehicle exposures.	Problems likely when merging data, for eg. overlap of intersection.
C) Frequency Rate Method	The 2-dimensional accident data matrix is easy to interpret and determine from,priorities for hazardous locations. Frequency and rate categories on X & Y axes can be varied according to conditions.	
D) Rate Quality Control Method	Uses a statistical test to check if accident rate is significantly lower/higher than for locations with similar characteristics. Ensures that the rate is not attributable to randomness.	
E) Accident Severity Method	EPDO index is easy to calculate and interpret.RSI method can be used for different types of accidents such as right-angle, rear-end, etc.	Cannot be implemented in Louisiana due to lack of injury data classification such as A-type, B-type, in accident data file.

F) **Bayesian Analysis**	Corrects the bias introduced by regression-to-the-mean. This is not done in the previous methods which use total annual accidents.	
G) **Hazardous Roadway Features Inventory**	Locations that don't have a history of high accident experience, but still deserve consideration due to potentially high accident frequency or severity, can be considered.	Cannot be used in isolation and depended upon as the only method of analysis.

Table 4A COMPARISON OF THE METHODS OF ACCIDENT ANALYSIS

Analysis Methods	Characteristics for comparison: Simplicity	Applicability where section lengths & traffic volumes differ	Usefulness in isolation
Accident Frequency Method	Yes	No	No
Accident Rate Method	Yes	Yes	Yes
Frequency Rate Method	Yes	Yes	Yes
Rate Quality Control Method	Yes	Yes	Yes
Accident Severity Method	Yes	Yes	Yes
Hazardous Roadway Features Inventory	Yes	Yes	No
BAYESIAN ANALYSIS	Yes. Simple to use.	Yes. Adapts to different section lengths & traffic volume.	Yes. Can be used as the only analysis method.

Table 4B COMPARISON OF METHODS OF ACCIDENT ANALYSIS

	Characteristics for comparison		
Analysis Methods	**Adaptability to differing ADT**	**Easy to merge data**	**Easy to interpret**
Accident Frequency Method	No	Yes	Yes
Accident Rate Method	Yes	No	Yes
Frequency Rate Method	Yes	Yes	Yes
Rate Quality Control Method	Yes	Yes	Yes
Accident Severity Method	Yes	Yes	Yes
Hazardous Roadway Features Inventory	Yes	Yes	Yes
BAYESIAN ANALYSIS	Yes. The method adapts to differing ADT easily.	Yes. No problems in merging data.	Yes. Easy to interpret.

Table 4C COMPARISON OF METHODS OF ACCIDENT ANALYSIS

Characteristics for comparison

Analysis Methods	Implementation in Louisiana	Statistical check test	Correction of bias introduced by regression-to-the-mean
Accident Frequency Method	Yes	No	No
Accident Rate Method	Yes	No	No
Frequency Rate Method	Yes	No	No
Rate quality Control Method	Yes	Yes	No
Accident Severity Method	No	No	No
Hazardous Roadway Features Inventory	Yes	No	No
BAYESIAN ANALYSIS	Yes. As it is a proven statistical procedure,it can be implemented in Louisiana.	N.A Not necessary.	Yes. This is the only method which corrects the inherent bias in all other methods.

Table 4D COMPARISON OF METHODS OF ACCIDENT ANALYSIS

	Characteristics for comparison	
Analysis Methods	**Considers locations with potentially high accident frequency and severity.**	**Free from effects of randomness.**
Accident Frequency Method	N.A	Yes
Accident Rate Method	N.A	Yes
Frequency Rate Method	N.A	Yes
Rate Quality Control Method	N.A	Yes
Accident Severity Method	N.A	Yes
Hazardous Roadway Features Inventory	Yes	Yes
BAYESIAN ANALYSIS	N.A This feature can be used as a supplement.	Yes. No negative effects due to randomness.

Table 5 : Accident Count at Intersections 1984/1985*

Number of Intersections	Number of accidents per int. in 1984	Average no. of acc. per int. in 1985
136	0	1.382
581	1	0.157
136	2	0.309
59	3	0.492
26	4	0.385
20	5	1.500
18	6	1.167
6	7	1.000
4	8	4.750
3	9	3.667
1	10	0.000
1	11	0.000
1	13	0.000
1	15	8.000

*

It is evident from the table that intesections having lesser number of accidents in a year tend to have a higher accident rate in the next year and so the probability of accident is random.

Recently, several researchers [Brinkman, 1986; Hauer, 1986; Persaud, 1983, 1984; Persaud and Hauer, 1984; Higle and Witkowski, 1988; Morris, 1988] either discussed advantages of, or recommended the use of the Bayesian method over other methods. Morris stated that "Accident rate estimation is extremely uncertain because the number of accidents at any one intersection tends to be quite random and subject to the regression-to-mean phenomenon." Table 3 summarizes the characteristics of each method. Although Pendleton (1988) claimed in the discussion of the Bayesian method [Higle and Witkowski, 1988] that it is not efficient, it is clearly shown that the Bayesian method has many advantages compared to other methods. Therefore, we determine to test the effectiveness of the Bayesian method for recommendation.

1.3.3. Finding the Frequency of Wet Weather Highway Accidents

All methods discussed in Section 1.3.1. can be used for wet accident analysis by replacing the total number of accidents with the number of wet accidents, and the total traffic volume with the traffic volume when the location is wet. Further, the ADT must be corrected for wet exposure since the hazardous nature of a location also depends on the amount of time a pavement was exposed to wet weather. For example, assuming that two locations, A and B, have the same accident ranking with A being wet 2% of the time and B being wet 8% of the time, then, A is more hazardous than B because the number of accidents in B inspite of being wet are equal to A.

The method of using hourly surface observations to measure wet weather exposure proposed by Harwood et al. (1988) was used for this study. This method considers the following factors:

(1) **Minimum level of wetness that reduces pavement friction.** The study establishes that an hourly rainfall of 0.01 inch is sufficient to reduce pavement friction by approximately 75%. Therefore, all hourly rainfall data over 0.01 inch are to be taken into account for calculation.

(2) **Rainfall intensity and duration.** The model also takes into account the intensity of rainfall during an hour. Previous models assume that the entire hour is wet. In this model, if the rainfall amount for an hour is, for example, 0.05 inches, then, the mean duration of rainfall for that hour is 45.9 minutes.

(3) **Runoff period following rainfall.** The time required by the rain water to run off while rain falls is given by the following kinematic wave method. Runoff times after rainfall are assumed to be equivalent to runoff times at low rainfall intensities such as 0.01 inch/hr.

$$TC = \frac{0.94 \times L^{0.6} \times n^{0.6}}{i^{0.4} \times S^{0.3}} \qquad \text{.....(8)}$$

where,

TC = time of concentration or runoff time (min),

L = length of drainage path (ft), measuring from the crown of the pavement to the edge (half the pavement width),

n = Manning Coefficient, determined by the surface of the pavement and typically ranging between 0.01-0.05,

i = Rainfall intensity (in/hr), and

S = average slope of drainage path (ft/ft).

(4) **Pavement drying period following rainfall and runoff.** Laboratory tests from the same study indicate that the pavement drying period depends on: solar radiation, wind Speed, air temperature, relative humidity, pavement type.

(5) **Pavement wetness due to fog.** This is assumed to occur only when there is occurrence of fog and the dew point temperature is within 2°F of the ambient temperature.

(6) **Estimation of exposure to ice-and-snow conditions.** The minimum frozen precipitation level that makes an hour icy and snowy is 0.01 inch. The corrected ADT will then be given by the following formula:

Wet_ADT = ADT x percent_wet_time / 100 .....(9)

1.3.4. THE PROPOSED BAYESIAN METHOD

This section of the chapter focuses upon the reasons for the recommendation of the Bayesian Analysis as a method to identify hazardous locations in the state of Louisiana. It begins with a description of the regional Bayesian Analysis Method as presented by Higle and Witkowski (1988). It then elucidates the modification done to this method to suit the conditions of the Louisiana accident analysis system. It also describes the computer implementation of the proposed method and its performance as

opposed to some of the classical statistical methods mentioned earlier in the 'Background of Research' section.

1.3.4.1. Why Bayesian Analysis?

Historical data do not always reflect long term accident characteristics accurately. That is, a location with low accident rate (i.e., in the long run) may still have a high accident rate over a short period of time, and vice versa. Also, traffic analysts agree that the accident rate associated with a particular location is a random variable, i.e., it cannot be predicted with absolute certainty. This is true regardless of the identification method used. Moreover, although regional accident characteristics may provide some useful information regarding the accident rate at a particular location, each location must be evaluated separately and should only be compared with locations that have similar underlying characteristics. The vast differences in accident histories that one finds among various locations suggest that the random variables used to describe the accident rates should differ from location to location.

Some of these difficulties could be overcome by the use of Bayesian Analysis in the process of identifying hazardous locations. Bayesian Analysis provides a method by which the random variables representing the accident rates at the various locations are mathematically defined. This is achieved by combining regional accident characteristics and the location specific accident histories. Moreover, by using a Bayesian Analysis, one can identify

hazardous locations on the basis of the probability that accident rates exceed some level.

1.3.4.2 Steps of Bayesian Analysis

Several researchers [Hauer, 1986; Hauer and Persaud, 1983, 1984; Persaud and Hauer, 1984; Higle and Witkowski, 1988] suggested different Bayesian Analyses. Higle and Witkowski's method (1988) finds the accident rate at a particular location while other methods predict the number of accidents at a location. As mentioned before, a simple accident number without considering the ADT would not provide useful information about the accident potential of that site. Considering this fact, it was decided that the analysis suggested by Higle and Witkowski (1988) would be the best method to pursue in the State of Louisiana. The Bayesian Analysis methodology proposed by Higle and Witkowski identifying hazardous locations can be summarized in three steps.

Step I: The accident histories are aggregated across a number of locations (i.e., across all locations within an appropriately defined region). The result of this step is a gross estimation of the probability distribution of the accident rates across the region.

Step II : The regional distribution so obtained is used, along with accident history, at a particular location to obtain a refined estimation of the probability distribution associated with the accident rate at that particular location.

Step III : With the collection of refined distributions, the probability that any given location is hazardous is assessed.

The corresponding formula for each step is listed below.

STEP I: Aggregation of Accident Histories for Accident Rate Estimation

The following notation is used to describe the Bayesian Analysis identification process formally.

r_i = accident rate at location i (note that r_i is treated as a random variable)

N_i = number of accidents at location i during the period of time in question

V_i = number of vehicles passing through location i during the period of time in question

$f_i(r/N_i,V_i)$ = probability density function associated with the accident rate at location i, given the observations N_i and V_i

$f_R(r)$ = probability density function associated with the accident rate across the region.

Then, a gross estimation of the probability distribution of the accident rates across a region is given by,

$$f_R(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r} \qquad \text{.....(10)}$$

The most common techniques available to determine the values of α and β are the Method of Moments Estimate (MME) or the Maximum Likelihood Estimates (MLE) Techniques [Higle and Witkowski, 1988].

STEP II: Obtaining Refined Distributions

In the second step, the observed accident rate at each location is used in combination with the gross estimate of the regional probability distribution to obtain location specific probability density function $f_i(r/N_i, V_i)$

$$f_i(r/N_i, V_i) = \frac{B_i^{\alpha_i}}{\Gamma(\alpha_i)} r^{\alpha_i - 1} e^{-B_i r}$$

.....(11)

STEP III: Identification of Hazardous Locations

With this collection of probability functions, the identification is done by identifying location i as hazardous if the probability is significant that r_i exceed r, where r is an upper limit on the acceptable accident rates, i.e.,

$$P(r_i > r/N_i, V_i) > \delta$$

.....(12)

1.3.4.3. BAYESIAN ANALYSIS APPLIED TO LOUISIANA WET WEATHER ACCIDENT ANALYSIS

This section explains what modifications are made to the Bayesian Analysis developed by Higle and Witkowski (1988) to apply it to the Louisiana accident analysis.

1.3.4.3.1. Estimating the proportion wet time of asphalt and concrete pavements

The primary focus of the methodology is on measuring the most important quantity of interest in all methods of wet accident analysis, namely, the wet accident rate at a location, given by:

$$W_r = N_i/(V_i p_i) \quad \ldots\ldots(13)$$

where,

W_r = wet accident rate,

N_i = number of wet accidents,

V_i = volume of vehicles passing through the location during the period of analysis, and

p_i = proportion time wet.

The values of N_i and V_i for a location i are obtained from accident data such as the accident files of the Department of Transportation. However, finding p_i for every location is extremely difficult since it has to be estimated. The model developed by Harwood et al. (1988) predicts the number of wet hours in a year based on the hourly precipitation, wind speed, temperature and dew point data. The value of p_i may then be estimated with accuracy for a year by dividing the number of wet hours at a location by (365 x 24).

Unfortunately, the rainfall and hourly surface observation data is not available for every location in the state. There are very few weather stations in a state compared to the amount of road locations. Very often, weather data are available only from stations which are very far away. Specifically, four locations in

Louisiana, Baton Rouge, Lake Charles, New Orleans and Shreveport (First Order Stations), have such type of data for a good period of record, that is, 27 years. Other locations (about 75 in number, Second Order Stations) have only total rainfall amounts for the period in question and rainfall normals (30-year averages) for 1950-1980. Harwood et al. (1988) recommend the inverse distance formula for estimating the wet proportion time at every Second Order Station, that is given by,

$$p_{so} = \frac{\frac{N_{so}}{N_1} p_1 d_2 + \frac{N_{so}}{N_2} p_2 d_1}{d_1 + d_2} \qquad \text{.....(14)}$$

where,

p_{so} = proportion wet time at second order station,

p_1 = proportion wet time at nearest first order station,

p_2 = proportion wet time at second nearest first order station,

d_1 = distance of nearest first order station to the second order station,

d_2 = distance of the second order station from the second nearest station,

N_{so} = 30 year normal (average total annual precipitation) for the second order station,

N_1 = 30 year normal for the nearest first order station, and

N_2 = 30 year normal for the second nearest first order station.

This formula is therefore used to estimate proportion time wet for all second order stations in Louisiana. The airline distance from first to second order stations was computed using the arc distance formula, based on latitude and longitude measures, given as follows (Robinson, 1977]:

$$\cos D = (\sin a + \sin b) + (\cos a \cos b \cos P) \quad \ldots\ldots(15)$$

where,

D = arc distance between A and B,

a = latitude of A,

b = latitude of B, and

P = degrees of longitude between A and B.

Harwood et al. (1988) recommend plotting contours using this estimated data for second order stations and data from first order stations. Although contours are smoothed values, computation of proportion wet time for every location in the state is not computationally feasible from contour plots. Therefore, the finite element method for estimation of mean areal rainfall (Singh, 1989; Akin, 1971] is used to estimate mean areal proportion wet time. A triangular mesh of the weather stations (first and second order) is formed using the gridding procedure in SAS/GRAPH. A sufficient number of second order weather stations from Arkansas, Texas and Mississippi are also selected to cover a major part of the state (Figure 2). The mean areal proportion wet time is assumed to be constant throughout the triangle.

Every location in Louisiana is then assigned to a triangle of the mesh using control section maps. Locations within one triangle are assigned to that triangle. Locations passing through two or more triangles are assigned to the triangle through which the location passed the most distance. Thus, the proportion wet time for every location is estimated by assigning the mean areal wet proportion time of the triangle to that location.

1.3.4.3.2. Application of Empirical Bayes to Better Estimation of Proportion Wet Time p_i

The proportion wet time, p_i, for a pavement varies from year to year for a single location. Also, p_i is dependent on the annual amounts of rainfall every year at a place. The annual amounts of rainfall vary every year geographically and seasonally (Singh, 1989]. Also, the amount of rainfall at a place can be derived from those of nearby places. In fact, many methods for estimating missing rainfall data use this fact (Singh, 1989]. Thus, the following assumptions are valid:

p_{ij} is a random variable, where j is the index of the year. As the amount of rainfall varies from year to year, instead of assuming that the true mean of p_{ij} is a point parameter Φ_i, it can be assumed that the true means of proportion wet time, Φ_{ij}, varies every year with a common prior distribution (Berger, 1985].

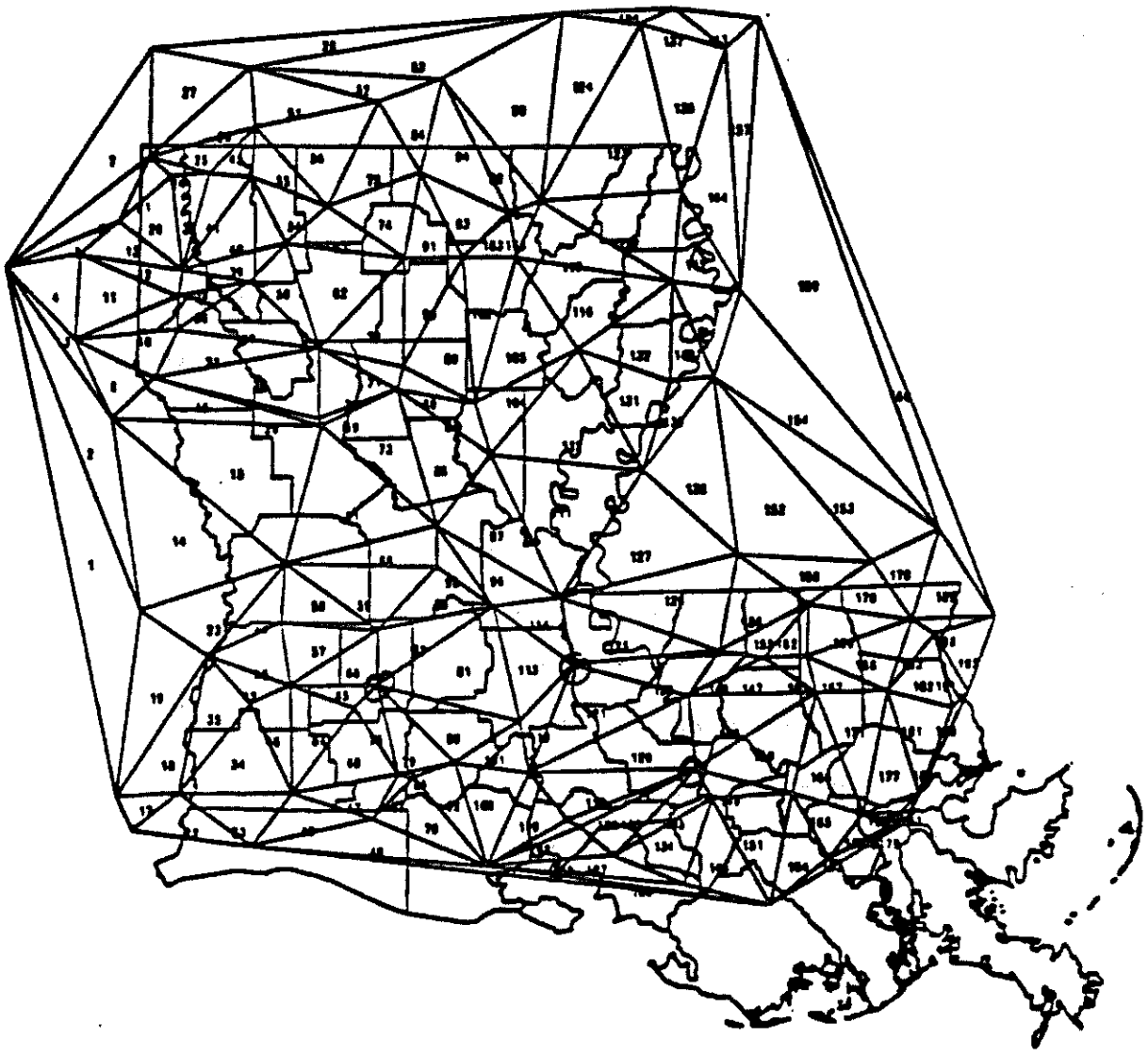


Figure 2. Triangular mesh of weather station in Louisiana

This is a Bayesian assumption of the true means Φ_{ij}. For a Bayesian estimate of the yearly Φ_{ij}, a theoretical distribution is needed to be assumed for the data. The only information available is the sample data p_{ij} which form the marginal distribution of the proportion wet time.

Quantile-Quantile plots for various years' data for some first order stations were observed for best linear fit (Figures 3-4 for asphalt pavements and Figures 5-6 for concrete pavements).

Although the proportion wet time would indicate a β distribution, owing to the central limit tendencies of the sufficiently large data values (Morris, 1983] and ease of parameter estimation, the normal distribution was tested for goodness of fit. The normal assumption seemed to provide adequate information about the data.

Consequently, the normality test provided by the Shapiro-Wilk statistic indicated a good fit to the data. In all cases, the Shapiro-Wilk statistic yielded a value greater than 0.91 and in most cases a value of about 0.94 (see Tables 6-7).

As the marginal distribution is normal, the prior distribution may then be assumed to be normal (Berger, 1985]. Therefore, p_{ij} was assumed to be observations from independent $N(\Phi_{ij}, \sigma_f^2)$ distributions. It can also be assumed that these Φ_{ij} are from a common prior distribution as the proportion wet time for all stations were within 0.0001 for every year.

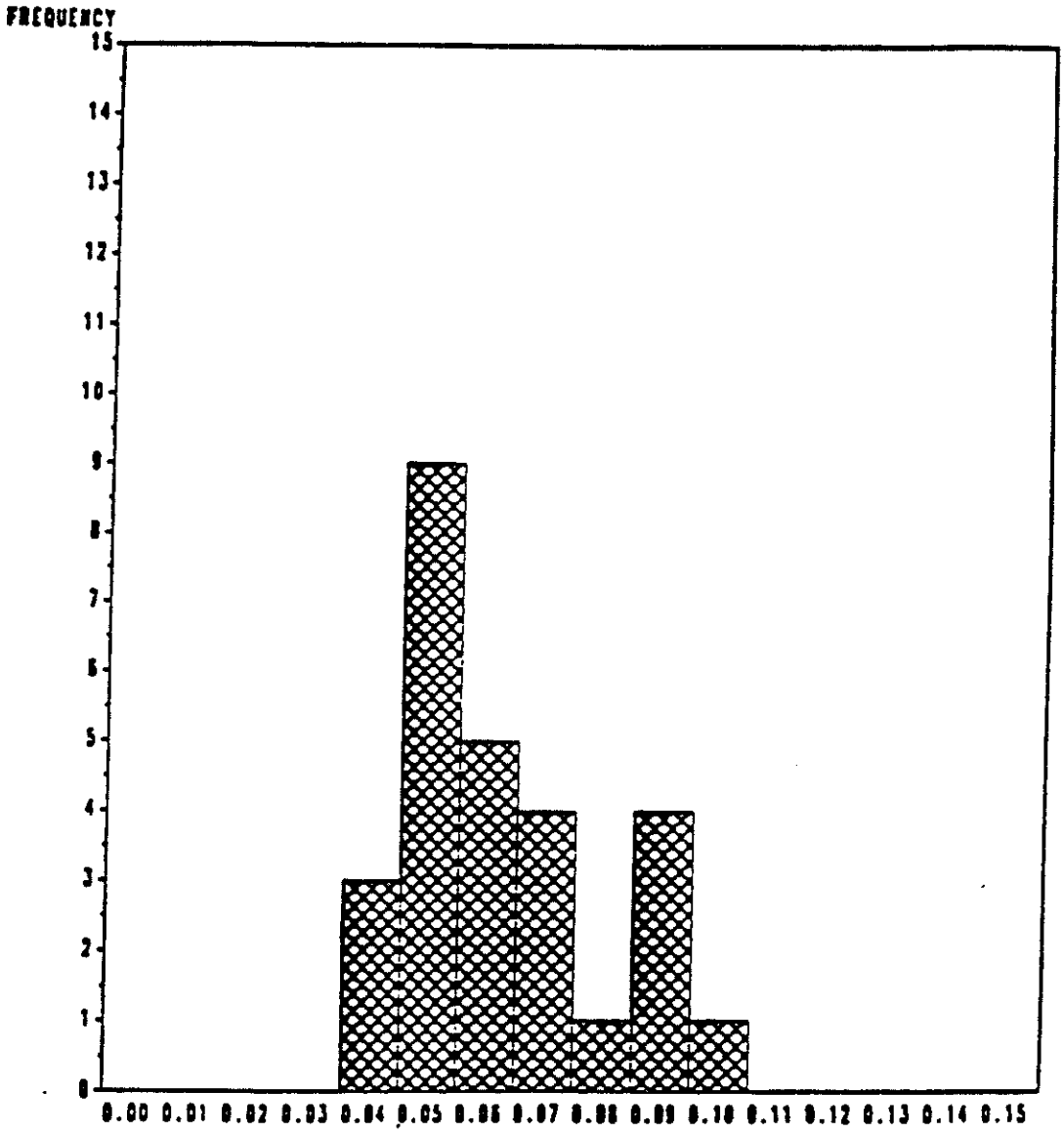


Figure 3. Histogram for distribution of proportion of wet time for a single Asphalt Pavement for 27 years.

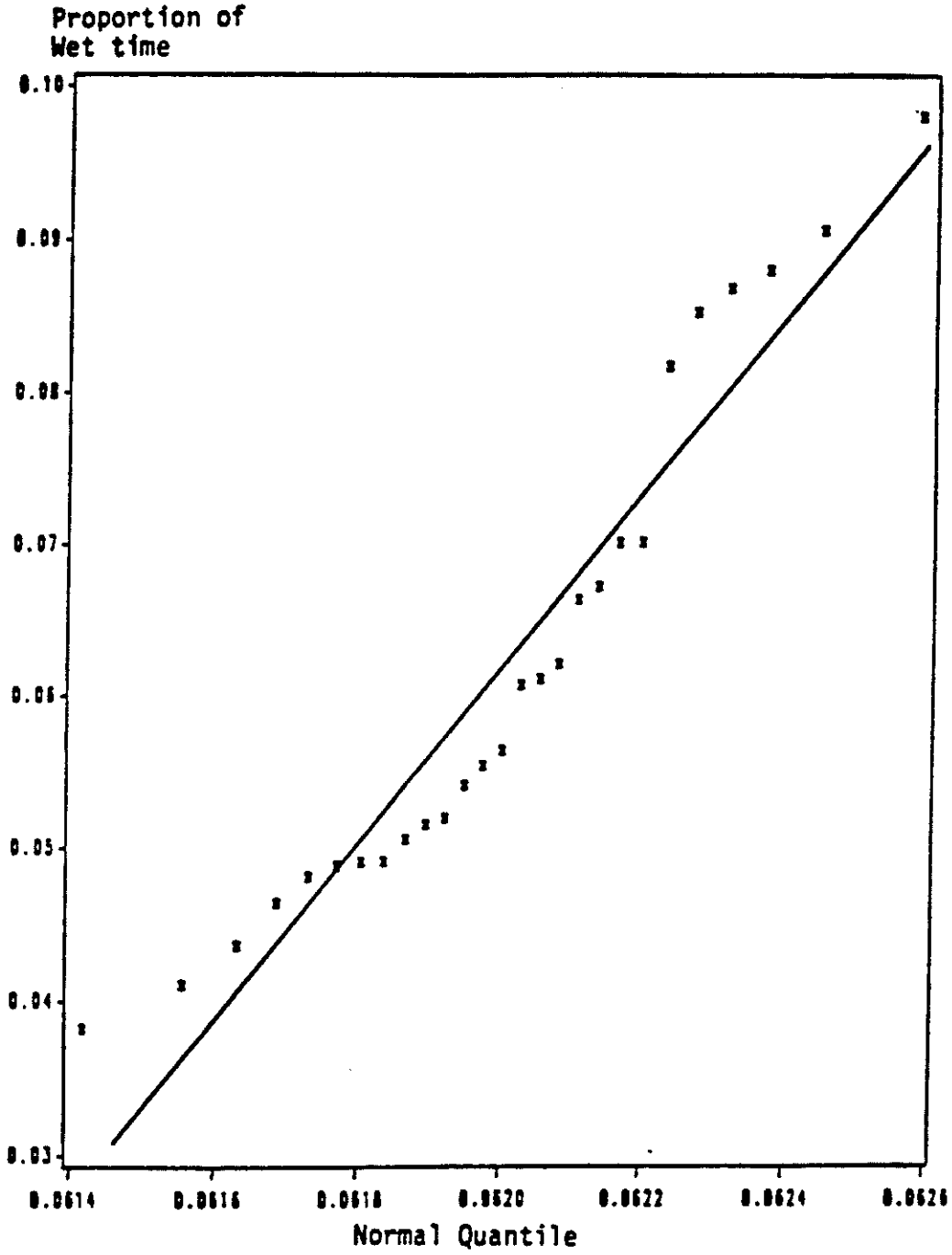


Figure 4. Quantile plot for the marginal distribution of proportion Wet Time for a single asphalt pavement for 27 years

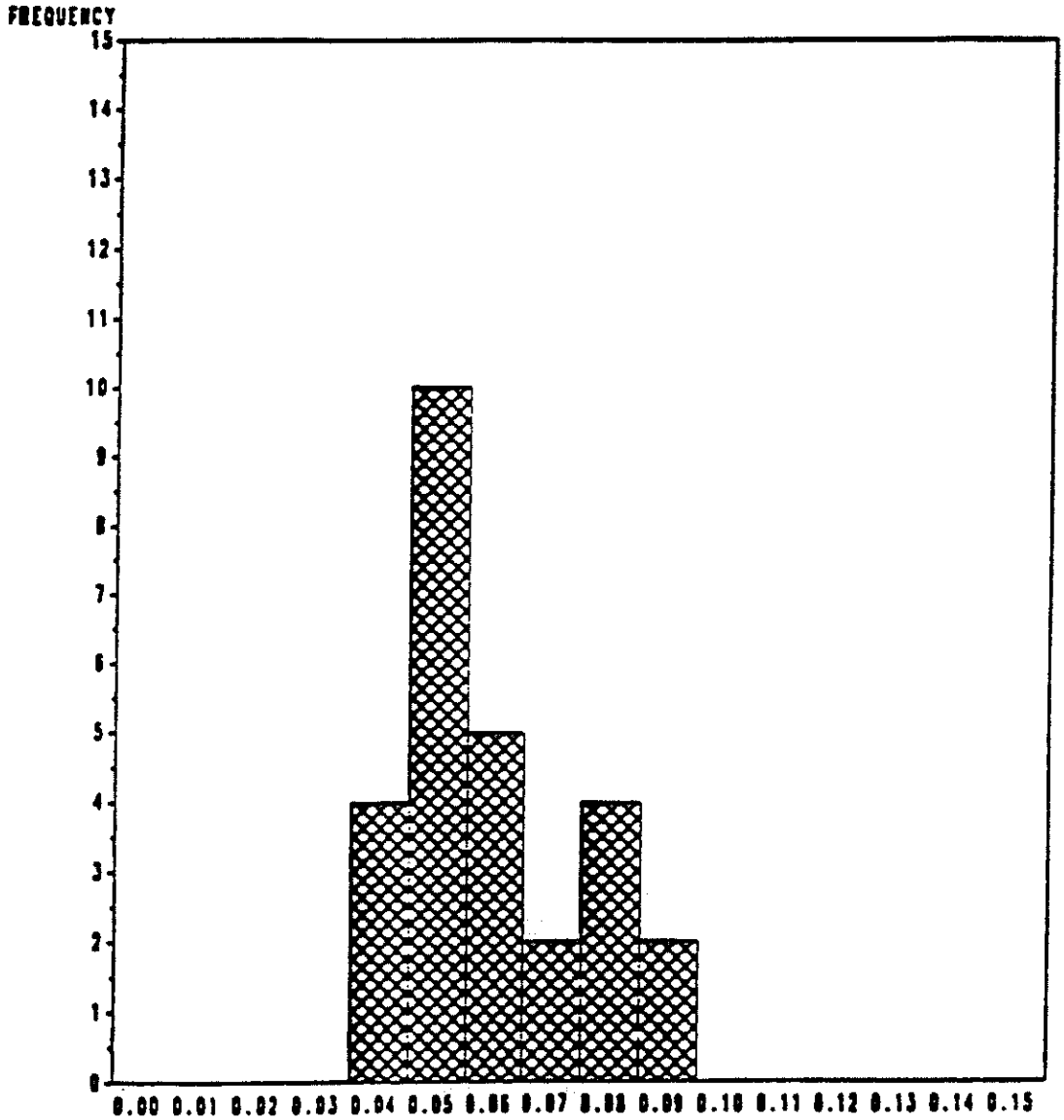


Figure 5. Histogram for distribution of proportion of wet time for a single concrete pavement for 27 years

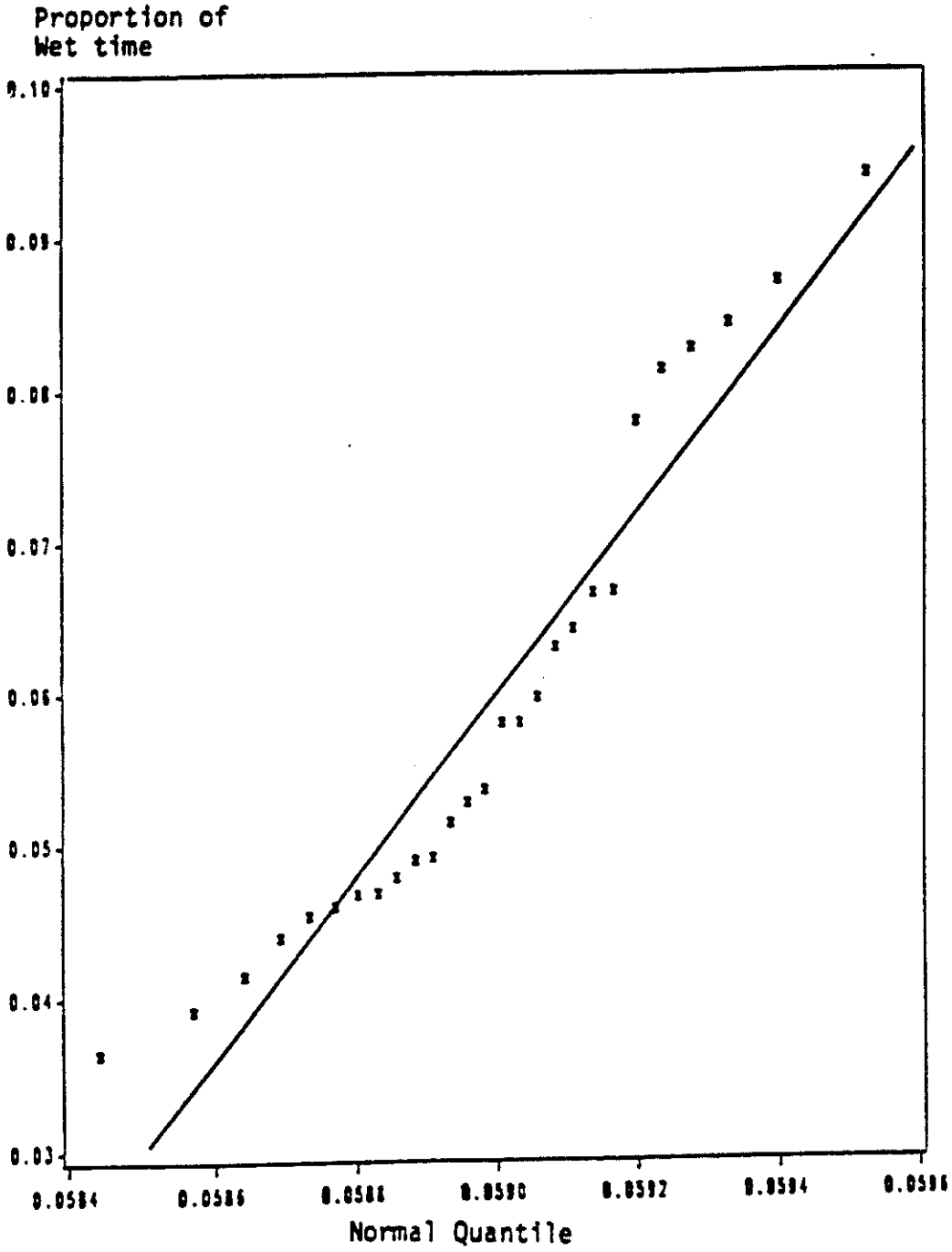


Figure 6. Quantile plot for the marginal distribution of proportion wet time for a single concrete pavement for 27 years

Table 6 Normality test results for the marginal distribution of the proportion of wet time for asphalt pavements for a single location.

Statistic	Value
Number of observations	27
Mean	0.063522
Standard Deviation	0.0076625
Maximum	0.0814753
Minimum	0.0495137
Median	0.0634223
Mode	0.0495137
t-test: mean=0	43.076
Shapiro-Wilk Statistic (W)	0.9444872
Probability < W	0.0001

Table 7 Normality test results for the marginal distribution of the proportion of wet time for concrete pavements for a single location.

Statistic	Value
Number of observations	27
Mean	0.0609525
Standard Deviation	0.00739512
Maximum	0.0786111
Minimum	0.0474657
Median	0.0608847
Mode	0.0474657
t-test: mean=0	42.828
Shapiro-Wilk Statistic (W)	0.943316
Probability < W	0.0001

The Bayesian estimate of the proportion wet time is then estimated by Parametric Empirical Bayes (PEB) (Morris, 1983]. The PEB method assumes that the Φ_{ij} are independently $N(\mu_\pi, \sigma_\pi^2)$, the hyperparameters, μ_π and σ_π^2, being unknown. The posterior mean and variance of the true means Φ_{ij} are then given by the normal posterior distribution parameters, assuming a normal prior of Φ_{ij} (Morris, 1983].

The empirical Bayes estimate of the posterior mean and variance is then given by:

$$\mu_{ij}^{EB}(p) = p_{ij} - B_e(p_{ij} - \bar{p}) \qquad \ldots\ldots(16)$$

where,

μ^{EB}_{ij}= Empirical Bayes mean estimate of proportion of wet time for a single year j for a location,

p = vector of proportion of wet time of all locations,

p_{ij} = proportion of wet time for year j for a single location,

B_e = estimate of the shrinkage parameter B,

p = mean proportion of wet time of all locations considered for analysis,

and

$$V_j^{EB} = \sigma_f^2 \left(1 - \frac{(k-1)}{k} B_e\right) + \frac{2}{(k-3)} B_e^2 \, (p_j - \bar{p})^2 \qquad \ldots\ldots(17)$$

V^{EB}_{ij}= Empirical Bayes estimate of the variance of proportion of wet time,

k = number of years of wet time data available,

σ^2_f = variance of the assumed prior distribution

$$B_e = \left(\frac{k-3}{k-1}\right)\frac{\sigma_f^2}{\sigma_f^2+\sigma_{\pi e}^2}, \qquad \sigma_{\pi e}^2 = \max\left(0, \frac{s^2}{(k-1)} - \sigma_f^2\right)$$

.....(18)

where,

$\sigma^2_{\pi e}$ = variance hyperparameter of the parameter θ_{ij}'s,

s^2 = sample variance, and

e = the estimator of the quantity and an overline, the mean of the quantities.

Ordinarily, the value of σ_f^2 is known, but it needs to be estimated in this case. The value of σ_{fe}^2 may be estimated by (Berger, 1985] :

$$\sigma_{fe}^2 = \frac{1}{n(n-1)k} \sum_{m=1}^{n}\sum_{j=1}^{k} (p_j^m - \overline{p_j})^2$$

.....(19)

from other available independent samples from a $N(\theta_{ij}, \sigma^2)$ distribution, in this case, a group of triangles whose mean areal rainfall needs to be estimated independently. However, it must be noted that independence between triangles cannot be "naively" assumed since the proportion wet time was estimated from three first order stations. In order that the independence assumption be justified, m triangles, not quite nearby, were grouped together to estimate the required σ_{fe}^2.

The statistically better Bayesian estimate of the mean is then used instead of the observed p_{ij} values for proportion wet time. Thus, the wet accident rate would now be given as :

$$Wr_i = \frac{N_i}{V_i \ \mu_{ij}^{EB}}$$

.....(20)

1.3.4.3.3. Assumption of Poisson Process for Wet Accident Number

The Bayesian methods of accident analysis are modified to analyze wet accidents. The fundamental assumption is that the number of wet accidents, given the wet accident rate, follows a Poisson distribution. This assumption can be proved by the following theorem (Ross, 1985]:

Theorem 1 : If a Poisson process with rate α has two types of events I and II with probability p and (1-p), respectively, then the occurrences of events I and II each follow independent Poisson processes with rates αp and $\alpha(1-p)$, respectively.

This theorem shows that if the total accidents follow a Poisson process with rate α, then by definition, accidents which may be wet (type I event) or non-wet (type II event) also follow Poisson processes with rate αp and $\alpha(1-p)$ with p being unknown.

This fact was also verified empirically by taking five years of data for a high accident intersection in East Baton Rouge. Monthly accident totals were computed, and a Chi-square test for

goodness of fit to the Poisson distribution was acceptable for a significance level of 0.05 ($X^2 = 2.687 < X^2_{0.05}$).

Therefore, the Poisson distribution is assumed for the wet accident rate at the location. Since the Poisson distribution was assumed, the regional wet accident rate was also assumed to follow a gamma prior distribution as in literature [Glauz et al., 1985; Hauer and Persaud, 1984; Higle and Witkowski, 1988; Morin, 1967; Norden et al., 1956].

1.3.4.3.4. Identification of Hazardous Location in Bayesian Analysis for wet accident analysis

Since information about the accident rate at each location and regional accident rate is provided, the identification of the hazardous location can be performed. Higle and Witkowski(1988) introduced two different methods, B1 and B2, which can be used to identify the hazardous location using Bayesian methods. Since we assumed the gamma prior distribution in the previous section, we modified methods B1 and B2 so that they can be used with $N_i/(V_i * \mu_{ij}^{EB})$, instead of N_i/V_i. The B1 and B2 methods are shown below:

Method B1: A location i is said to be hazardous if the probability of the actual accident rate ri, exceeds the average accident rate across the region, is greater than a confidence level δ. Common values of δ are 0.99 and 0.95.

Thus, if

$$P\{ ri>x|Ni,Vi \} > \delta \qquad \ldots\ldots(21)$$

where,

x = the observed average rate across the whole region,

ri = the true accident rate at location i,

$$x=\Sigma\left(\frac{N_i}{V_i}\right)$$

Ni = the number of accidents in a given time period at location i, and

Vi = the traffic volume in that time period at location i, then the location is said to be hazardous. The above probability is calculated by using the gamma cumulative distribution formula.

Method B2: A location i is said to be hazardous if the probability of the actual accident rate (ri) exceeds the observed regional accident rate xR, is greater than the confidence level δ.

Thus, if

$$P\{ ri>xR|Ni,Vi \} > \delta \qquad \text{.....(22)}$$

where,

$$X_R=\frac{\Sigma N_i}{\Sigma V_i}$$

employing the same notations as before, then the site is hazardous. This probability, again, is calculated using the cumulative distribution function of the gamma distribution. These two Bayesian methods show much promise but are computationally intensive.

1.3.5. Computer Implementation of the Identified Methods

The methods that are relevant to the present problem of identifying wet hazardous accident locations were compared with the Louisiana data. The methods selected were the accident rate method, the rate quality control method and two Bayesian methods. Both methods B1 and B2 [Higle and Witkowski, 1988] were tested for the Bayesian method. Although Zeeger (1982) recommended the accident frequency method as an initial method and the accident severity at least as a supplemental method, since the accident frequency method does not consider the traffic volume and Louisiana does not have data for the accident severity, these methods were not included for comparison with other methods. The procedure for the accident rate method can be used to implement the accident frequency method. It differs in only that wet accident per mile criterion is used for ranking the wet hazardous locations instead of the wet accident per million vehicle miles criterion (see lines 33 and 37 of the pseudo-code of the accident rate method in section 1.3.5.3.1). For the purpose of this study, comparison and recommendation of the best of the above four methods is based on Louisiana's wet accident data.

1.3.5.1. Methodology for Comparison

In this research, the selection of the method for identifying hazardous accident locations was based upon simulation and minimization of missing hazardous location (false negative in the

techniques of Higle and Hecht (1989) which is briefly explained below.

This technique is based on a simulation of available wet accident data. This method assumes that the wet accident rate in a time period represents the "true rate". A random Poisson number with mean "true rate," is then generated for every site to represent the possible wet accident data available to the researcher. In addition, a random normal number with mean μ_{ij}^{EB} and variance V_{ij}^{EB} is generated for every accident location.

The advantage in this simulation is that we know the "true" wet accident rate at every site. Therefore, we know whether a site is truly hazardous (H), or not (NH). The simulated data is then used to calculate wet hazardous locations, which are either flagged (F) by each one of the four methods listed above or not (NF). 30 simulation runs across all highway classes are performed at different confidence intervals (probability levels of 0.90, 0.95 and 0.99) for the available wet accident data.

In summary, the methodology consists of:

(1) Estimating the proportion wet time for asphalt and concrete pavements using approximations.

(2) Developing an Empirical Bayes estimate for the mean proportion wet time of a pavement from past and present data.

(3) Computing different measures of hazardous nature, as required by each of the four methods compared.

(4) Comparing the methods by means of simulation, and finally,

(5) Recommending a method or methods based on simulation.

1.3.5.2. Data Used in the Research

The data used in the research are as follows:

(a) *Accident data (for Louisiana) from 1984-1988*. This contains information about the accident site, exact location, weather conditions, pavement type, etc.

(b) *Projects data (for Louisiana) from 1982-1988*. This contains details of the construction and counter measure efforts taken at every location.

(c) *Hourly surface observations and precipitation data (for Louisiana) for the years 1962-1988*. This contains rainfall, temperature, wind speed, dew point and fog data on an hourly basis.

The above accident and project data are available in the form of magnetic tapes from the Louisiana Department of Transportation and Development, Baton Rouge, Louisiana and the weather data from the National Climatic Data Center, Asheville, North Carolina.

1.3.5.3. Programming Existing Methods

Four methods of accident analysis were used for predicting the trigger level. Programs are written in SAS. They are as follows:

1.3.5.3.1. Accident rate method

The accident rate method is written as a single SAS program name. The input, output and pseudo-code of the program are given below:

Input: Master accident file for single year (typically). Pairs of years are also to be used for other types of analysis.

Output: Top 200 wet accident locations for the state of Louisiana and top 20 wet accident locations for each parish and district as determined by the accident rate method.

The program also provides results for sections and intersections, and each highway class separately. The following definitions of an accident hold:

An intersection accident is defined as an accident that occurs within a 0.1 mile radius of a junction. A section accident is an accident that occurs on longer and variable length road segments. A spot is defined as a part of the section with an accident frequency of 2 or more, with a 0.1 mile radius.

Corrected Accident rate method:

For the case of wet accidents, the wet pavement accident rate is divided by the wet exposure time as given by the following formula:

$$Rw = Aw/Ew \qquad \ldots\ldots(23)$$

where,

Rw = wet pavement accident rate (in million vehicle miles),

Aw = number of wet pavement accidents, and

Ew = wet pavement exposure (vehicle miles).

Ew is to be calculated from hourly surface observation data and the WET_TIME exposure model (Harwood et al., 1988) for each location. A program is written in SAS for calculating the total wet exposure time of the pavement. In case of non-availability of weather data, we may use secondary estimates of Ew such as

E = (Ed + Ew) or Ewa = (Ed * wet_time), ...(24)

where,

Ed = dry pavement exposure,

E = total exposure, and

wet_time = % time nearest weather station was wet.

Using the above criterion, the accident rate pseudo-code would use wet_ADT instead of ADT as the denominator in the calculation.

The following is the algorithm written in pseudo-code:

```
1  Procedure accident_rate;
2  Input: master accident file for 1 year;
3  Delete incorrect records;
4  Classify a location as intersection or section;
5  Eliminate highway types - local and arterial roads;
6  Provide district numbers to parishes;
7  begin sort;
8     Sort accident locations by
9     location type - section and intersection
10    highway class - interstate, etc.
11    district number
12    parish number
```

```
13    highway type - urban divided, etc.
14    highway number - I010, etc.
15    Control, section numbers
16    Beginning of Control log mile.
17    end sort;
18  loop:
19     provide frequency of accidents:
20     location type - how much in intersections, etc.
21     surface condition - dry, wet, snowy, muddy, etc.
22     weather conditions - sunny, rainy, etc.
23     2 - way cross tabulations of the above.
24     print frequency table;
25     if wet accident then go to 29 ;
      end loop;
/* analysis for wet accidents */
26   eliminate accidents with dry surfaces;
27    for other missing data check whether the weather is raining
      and road condition is flooded;
28   go to loop;
29   calculate total accidents in each section;
/* accidents per million vehicle miles calculation */
30   for each unique location do;
31   wet million vehicle miles = length * 365 * wet_ADT/1000000;
/* Use wet_ADT instead of ADT in wet accident analysis */
32      wet accidents per wet million vehicle miles =
total wet accidents / wet million vehicle miles
```

33 /* used for accident frequency method */

wet accidents per mile = total wet accidents / length of location

34 end;

35 calculate state average for 32 and 33 quantities;

36 merge with accident file;

37 if wet accidents per wet million vehicle miles > 2* state wide average then location is hazardous;

/* use wet accident per mile criterion for accident frequency method */

38 sort hazardous locations by decreasing order of wet accidents per wet million vehicle miles;

39 print top 200 locations by highway class and highway type according to line 38;

40 do the same lines from 35 to 39 for each district and each parish;

41 end procedure;

1.3.5.3.2. Rate Quality Control Method

In this case, the critical rate CR for each highway class and type is be calculated as follows:

$$CR = \alpha + k\sqrt{(\alpha/m)} + 1/(2m) \quad \text{.....(25)}$$

where,

CR = critical rate for a particular road location (traffic accidents per million vehicle miles),

α = expected wet accident rate in accidents per million vehicle miles,

m = number of wet vehicle miles in millions,

k = a probability factor determined by the level of significance needed for CR (here k= 1.645), and

L = Length of freeway section, for spots and intersections, this value is 1.

Accident locations are then to be sorted by descending order of CR, first state-wise, district-wise and finally parish-wise. Thus, we see that the above three methods have more or less the same pseudo-code except for a few changes.

1.3.5.3.3. Bayesian Methods

The Bayesian methods B1 and B2 developed jointly in a single program, using the calculations of Higle and Witkowski (1988). A new Bayesian methodology is derived using the following steps:

(1) Calculation of distances from first order to second order stations by longitude, latitude data. Airline distances between stations were calculated based on the latitude longitude data provided. We approximate using the data of 60 minutes 5 seconds longitudinal variation per mile and 52 minutes and 3 seconds latitude variation per mile. These were then used to derive the closest and second closest first order stations to every second order station.

(2) Choice of second order stations. In order to approximate the normal distribution for Bayesian analysis, we need at least

30 years of annual precipitation records. So, all second order stations with such records,about 80 of them, will be used to construct the isoexposure contour map.

(3) Approximate prediction of exposure. The approximate prediction of exposure for second order stations is possible only if we have both distances and annual precipitation records. The annual precipitation records were taken from Climatological data annual summaries from years 1956-1988. The exposure prediction is then done by the inverse-distance weighted average formula discussed in Harwood et al.(1984).

A listing of this program is listed in the appendix B.

1.3.5.3.4. Criteria for flagging hazardous location by each method

The criteria for flagging an intersection by each method are as follows (Higle and Hecht, 1989]:

C1: Location i is flagged as hazardous if

$$\overline{\alpha_i} > \overline{x} + k_\delta s \qquad \ldots\ldots (26)$$

where α_i is the true wet accident rate at the location, x is the sample mean, s is the sample variance and k_δ is the z-value for the associated DELTA value.

C2: Location i is flagged as hazardous if

$$\overline{\alpha_i} > x_R + k_\delta \left(\frac{x_R}{V_i \, \mu_{ij}^{EB}} \right) + \frac{1}{2 V_i \, \mu_{ij}^{EB}} \qquad (33)$$

where x_R is the regional accident rate and V_i is the volume of vehicles at the location and μ_{ij}^{EB} is the Bayesian mean of proportion wet time.

B1: Location is flagged as hazardous if

$$P(\overline{\alpha_i} > \overline{x}) > \delta, \qquad \ldots\ldots\ldots (27)$$

and

B2: Location is flagged as hazardous if

$$P(\overline{\alpha_i} > x_r) > \delta \qquad \ldots\ldots\ldots (28)$$

where the probabilities are computed as discussed earlier and δ is the confidence level (0.90, 0.95 or 0.99).

1.3.5.4. Criteria for Selecting the Analysis Methods

In 1982, when the Bayesian method for the hazardous location identification did not exist, Zeeger introduced the following guidelines for selecting methods for identifying high accident locations regardless of the specific methods:

(a) Although the accident frequency method alone does not consider traffic exposure and accident severity, it is useful in initially identifying a group of locations for further analysis and ranking. If the rate quality control method is used in conjunction with the frequency method, it is unnecessary to compute accident rates and critical rates for every location in the state having at least one accident. A sample could be

selected from locations that exceed a fixed number of accidents per year before the rate quality control method is applied.

(b) It is desirable to consider accident severity at least as a supplemental method when identifying locations. Some high-speed locations consistently exhibit numerous injury and fatal accidents without necessarily meeting an established accident frequency or rate criterion. For example, a spot location with five fatal and severe-injury accidents should certainly justify a higher priority for further analysis than another location with six property-damage-only accidents. Although the severity of a particular accident is subject to many factors unrelated to the accident (use of seat belts, age and health of occupants, size of vehicles, etc.), a consistent history of severe accidents at locations should be the basis for further review.

(c) The criteria (or level of confidence) for identifying and ranking locations should be based largely on the number of locations that can actually be handled by an agency. For example, if an agency can only realistically analyze and review 100 locations per year, it is unnecessary to identify and rank the top 1,000 locations. In this situation, it may be useful to set the criteria so that 150 to 200 locations are identified for further analysis. The number of identified locations can be established primarily by raising or lowering the "cut-off" accident criteria (frequency, rate, etc.), or by modifying the level of confidence. A change in the segment length (e.g., 0.3

mile, 1 mile, 3 miles) can also affect the number of locations that will be identified for further analysis.

(d) It is desirable to consider various types of accident identification methods. In general, a single identification method will allow only for the selection of a sample of locations worthy of further consideration. Consideration of several valid indications (frequency, rate, statistical reliability, accident severity, roadway features, etc.) will help to improve the reliability of the identification process. The data requirements for each method must be considered before the method is selected.

It is important to look at the highway network as a total system rather than merely as a combination of independent segments. In many cases, the presence of several high accident spots on a highway section may be due to more than just an isolated roadway deficiency. A roadway safety problem that extends for several miles may exist. Such a problem requires the consideration of improvements on a broader scale than would be considered for an individual high accident spot location. The use of accident data files in combination with other data files is valuable in producing a list of sites that warrant further study for safety improvements.

As mentioned in Section 1.3.5.1, the simulation techniques used by Higle and Hecht (1989) were used to perform 30 simulation runs across all highway classes at different confidence intervals (probability level of 0.9, 0.95, and 0.99) for the available wet

accident data. The methods were then compared on the basis of false negatives and false positives [Hauer and Persaud, 1984]. A "false negative" occurs when a method fails to identify a truly hazardous location, that is, (H ∩ NF). A "false positive" occurs when a method identifies a location that is not hazardous (NH ∩ F). Four important statistics for false positives and negatives were collected for the 30 iterations. They are:

(1) Average number of misidentifications over the 30 iterations.

(2) Average error per misidentification.

(3) Maximum number of misidentifications over the 30 iterations.

(4) Number of iterations in which there were no misidentifications.

It is clear that the "false negative errors" are more serious than the "false positive" ones to any safety analyst. The method that performs poorly with respect to false negatives is therefore not good. False negative and false positive fractions were also collected for all the location types and highway classes analyzed.

Threshold values (such as 2 accidents per million vehicle miles and above, etc.) to identify hazardous locations were implemented and tested before further study.

Simulation results were used to select the best method(s) as follows:

(1) Method(s) with the most "false negative" errors were eliminated from further consideration.

(2) If methods exhibited similar "false negative" results, methods with smaller "false positive" errors were recommended.

1.3.5.5. Time Period Used for Selecting the Analysis Methods

Deacon et al. (1975) recommend that the accident analysis time period be either 1 or 2 years. Their study showed that the reliability of the analysis scheme increased with increase of the time period of analysis. For the purpose of the present study we choose 2 years as the time period of analysis for the following reasons:

(1) Maintenance action is not started until about one to one and a half year of the year under analysis.

(2) The coding and conversion of the available data typically takes about 6 months.

(3) For statistically significant results, sufficient number of data samples are required, i.e., in the order of 20-30. This is possible only if the accident analysis time is one year as data is available for 25-40 years.

1.3.6. Summary

So far, the comparison of the above four methods is based on Louisiana's wet accident data. To compare the four methods beyond this domain, the methodology of Higle and Hecht (1989) is used. This technique is based on a simulation of available wet accident data. This method assumes that the wet accident rate in a time period represents the "true rate". A random Poisson number with mean "true rate" is then generated for every site to represent the possible wet accident data available to the researcher. In

addition, a random normal number with mean μ_{ij}^{EB} and variance V_{ij}^{EB} is generated for every accident location.

Although the analysis was run for all the location types, only the results for intersections are presented in this section. Intersections in the urban interstate class and the rural two lane class are presented here. Four methods are then compared: B1 and B2 - the two Bayesian criterions and, C1 - Classical Accident Rate and C2 - Rate Quality Control Method. These are compared at three levels of probability, that is, 0.90, 0.95 and 0.99. The results are presented in Tables 8A-8D of Section 5.1 in Chapter 5.

1.4. DEVELOPMENT OF THE SKID RESISTANCE DATA MANAGEMENT SYSTEM

1.4.1. Database Design Process

The objective of the skid resistance data management system (SRDMS) is to facilitate information processing and storage for the accident analysis process, skid resistance testing, and other related tasks. The system will enable enhancement of the safety of the highway system to the extent of policy making and repavement. Development of a new data management system for skid resistance should encompass a process to accommodate changes in data requirements and provide an assurance that our proposed system as developed is an accurate and complete reflection of the user's requirements. The users of the data management system include LDOTD, LTRC, and Highway Needs, Priorities and Programs Engineer. Considering that the user has good comprehension of the task to be

performed by the data management system, we propose to employ a user-centered process as the system development strategy.

There are five basic phases in the user-centered database design process [Chong et al., 1992]:

(1) Requirements Analysis: to specify the functional and operational requirements for the database system to be designed;

(2) Conceptual Framework: to map the requirements into a well defined "blueprint," which shows the basic interrelationships between different segments of the organization or information entities;

(3) Logical Model: to show precise specifications of the database model in the "blueprint";

(4) System Implementation: to develop a database structure and application programs according the database model so far created; and

(5) Testing and Revision: to make sure that the system works as expected.

In constructing a database, corporate managers, database designers, and programmers have several Computer-Aided Software Engineering (CASE) tools to expedite the process. A microcomputer-based CASE tool (Chen 1989) using the entity-relationship (ER) approach was used to design the database. The ER approach was first introduced by P.P. Chen in 1976 and is now widely used by industries [Batini, 1989]. It begins with an assessment of database requirements based on the view of the entire organization

(termed "enterprise schema") before translating it to mechanical elements (user schema). The process attempts to identify the organization's functionally-organized parts (entities) and the interactions (relationships) among them by means of a graphical representation called the ER diagram. An entity is identified in the ER diagram as a rectangle, while a relationship is a diamond. Entities and relationships are interconnected with lines, and cardinalities are shown by numbers or variables (Figure 7).

The scope of the SRDMS includes from the initial point of merging coded accident reports with traffic and location data files at LDOTD, through the analysis of wet weather accident data and the transmittal of the list of abnormal test locations to LTRC, to the submission of the summary of skid resistance test results to the Highway Needs, Priorities and Program Engineer. A sequence of investigation, into the skid resistance database development has been performed and reported in the following sections.

1.4.2. Requirements Analysis of the SRDMS

There are four sources describing the requirements for skid resistance database: (1) the current reports generated from the DOTDACC file at LDOTD, (2) reports required at LTRC, (3) the database requirements documented in the 'request for proposal'

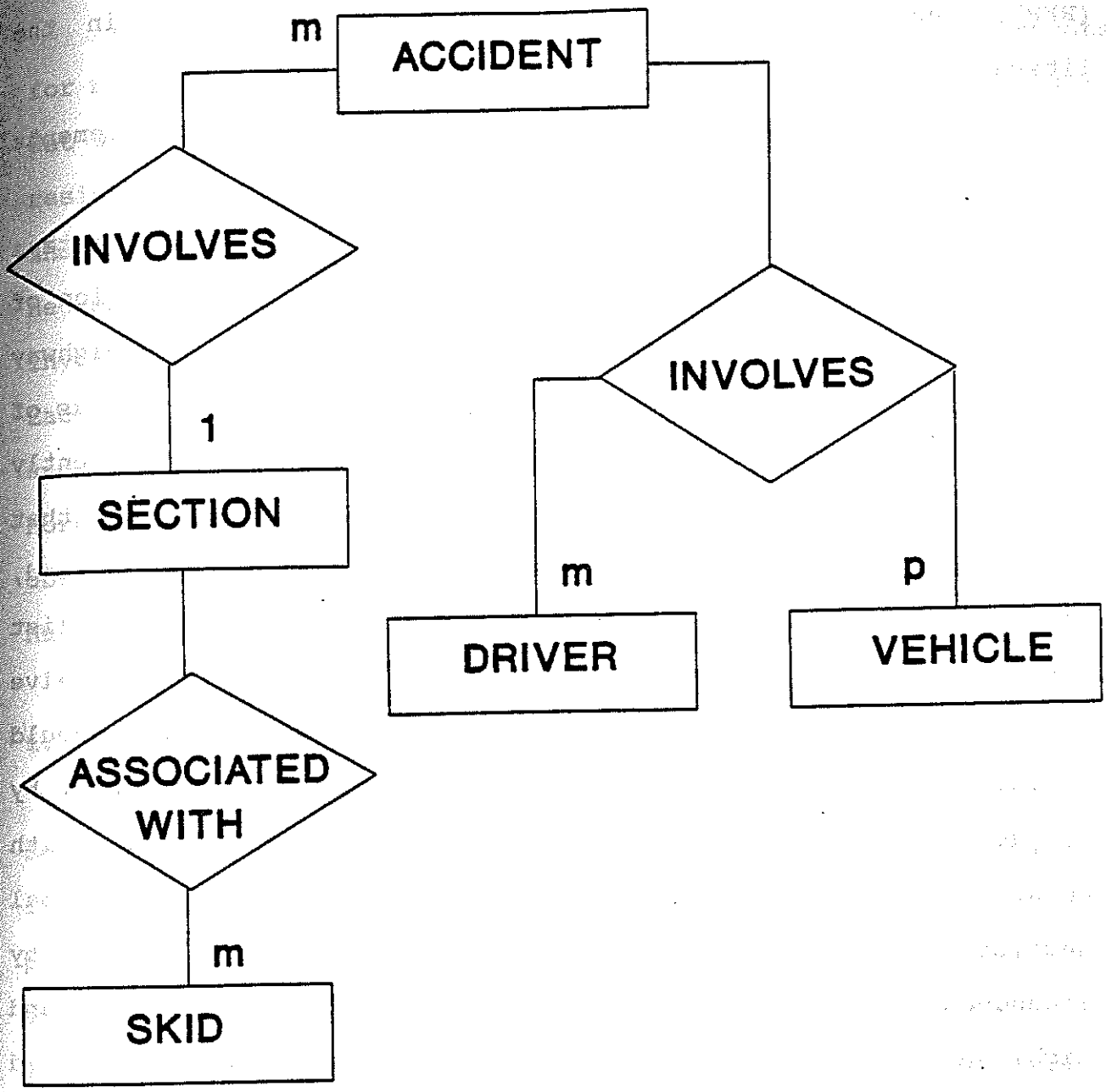


Figure 7. The overall ER diagram for the SRDMS

(RFP), and (4) the database requirements documented in the literature.

The following paragraph summarises the database requirements documented in the RFP:

The skid resistance test data is stored in the field on tape and transferred to the LTRC computer data files at completion of testing. Data is summarized by LTRC and submitted to the Highway Needs, Priorities and Programs Engineer. A historical analysis of abnormal sections or skid resistance test data is not currently performed. A historical database would be beneficial if that trends in skid resistance at a particular site can be observed and, if a particular site consistently appears on an abnormal list through the years, it can be flagged or highlighted to receive priority consideration for corrective action. The database should include the capability of maintaining historical data files by accident analysis period and location; integrating with construction and maintenance files to determine if an abnormal section has been corrected; ranking section test results by inadequacy of skid resistance and importance of facility; and highlighting locations that have an adequate level of skid resistance, but where, as indicated by a continual appearance on the abnormal list, other hazardous conditions may exist. In addition, the database system must be of flexible design such that it will readily integrate with LDOTD's proposed Pavement Management System as outlined in LTRC Research Report 195, "An Integrated Pavement Data Management and Feed System (PAMS)."

The following paragraph summarizes the database requirements for reducing skidding accidents during wet weather (Hankins, 1971: Wet weather accident data should be integrated with all the skid resistance data into an integrated database. A summary of each individual accident which occurs within a skid prone location is necessary. Greater consideration is to be given to accidents that involve only single vehicles and occur on wet pavement surfaces. Each section is identified by a construction section number (CSN) and a district number and contains its location characteristics. Moreover, a section is associated with its skid resistance data obtained by determination of the low, average, and high skid numbers.

1.4.3. Conceptual Framework of the SRDMS

An overall ER diagram based on the requirements described in Section 1.4.2 is shown in Figure 7. The purpose of the ER diagram is to communicate with the users of the system as effectively as possible. After an analysis of the user requirements and the date available, it was found that there were five entities i.e.,

(1) ACCIDENT: This entity contains information pertaining to the description of accident, site of accident and other global parameters like the average daily traffic and weather (primary source - accident file).

(2) SECTION: This entity contains information describing the various controls and sections of Louisiana state (primary source - control section file).

(3) SKID: This entity is for all the skid related information (primary source - JNSORT files).

(4) VEHICLE: This entity is for information pertaining to vehicles involved in the accident (primary source - accident file).

(5) DRIVER: This entity stores information pertaining to all the drivers involved in an accident (primary source - accident file).

1.4.4. Logical Model of the SRDMS

The ER diagram was converted to the following five relational tables:

ACCIDENT (*Accident_Number*, Date, Year, Accident_Class, Control, Section, Begin_Control_Log_Mile, End_Control_Log_Mile, Mile_Post, Parish, Computed_Location_of_Accident, Day_of_Week, Hour_of_Day, Highway_Type, Highway_Number, Number_Killed, Number_injured, Number_Involved, Surface_Condition, Road_Condition, Type_of_Accident, Hit_and_Run, Intersection, Highway_Class)

DRIVER (*Record_Number*, Driver_Number, Alcohol_Test, Age)

VEHICLE (*Record_Number*, Vehicle_Number, Vehicle_Type, Direction_Travel, Estimated_Speed, Posted_Speed)

SKID (*Control*, Section, Beginning_Control_Section_Logmile, Date, Logmile, Skid_Number, Type, Surface, Direction)

SECTION (*Control*, Section, Beginning_Control_Section_logmile,

District, Parish, Alignment, Highway_Type, Road_Condition, Intersection, Intersection_Quadrant, Highway_Class, Pavement_Width, Pavement_Type, Average_Daily_Traffic).

1.4.5. System Implementation of the SRDMS

The SRDMS is implemented in SAS. The package is chosen based on the following reasons: (1) SAS provides an interactive relational data management environment; (2) SAS provides a powerful statistical computing environment for the proposed highway accident analysis methods; (3) SAS is available at DOTD; and (4) Since the DOTD and LTRC offices are located in Baton Rouge, access to technical expertise at the LSU campus located nearby, is easily available.

Once the skid resistance database is implemented, the next step is to develop application programs using the database. The SRDMS is the integration of the database and the application programs. Since the SRDMS is intended for transportation managers to use, a user-friendly interface is necessary. One way to design this user-friendly application system is to allow the selection of the applications through a menu. The menu-driven system is discussed in Section 5.2.

1.4.6. Testing and Revision of the SRDMS

The menu-driven system has been tested by users at the LDOTD and LTRC. Several iterations have been carried out to make the entire system sufficiently robust for release for general use.

CHAPTER 2

DISCUSSION OF RESULTS

2.1. BAYESIAN ANALYSIS

As reported in section 1.3.2, the Bayesian methods encompass some very determining factors like the Regression-to-Mean effect and the Counter-Measure effects which have never been included in any of the classical analysis methods. The adaptability of Bayesian methods to some other global parameters like differing section lengths and differing ADTs is another major advantage. With regard to the method's performance in identifying hazardous locations, simulation techniques were used to establish the superiority of Bayesian Methods (as per the design mentioned in section 1.3.5.4). This section describes, in detail, the results obtained after performing 30 simulation runs (explained in section 1.3.6) on the accident data. The results of the simulation for intersections are shown in Tables 8A-8D. Table 8A shows the distribution summary for intersections in the urban interstate class.

The variable HAZARD, denoted by H and NH for hazardous and not hazardous locations respectively, indicates the true rates for the location as assumed by the simulation. The variable FLAG, denoted by F and NF for flagged and not flagged locations, indicates the performance of the method, indicated by the variable METHOD. METHOD has values B1 and B2 for the two Bayesian criterions and C1 and C2 for classical accident rate and quality control method, respectively. δ indicates the levels of probability taken into

consideration, 0.90, 0.95 and 0.99 for the flagging and hazardity of the sites.

Methods B2 and C2 showed consistently few 'false negatives' but a good number of 'false positives' for all confidence levels (δ values). C2 showed the lowest number of 'false negatives', closely followed by B2 for all confidence levels (δ values). As the number of locations increased, the number of 'false negatives,' as expected, also increased for all methods in this highway class.

For the **rural 2 lane** class (Table 8B), the number of 'false negatives' was unusually high. This can be attributed to the very high number of locations. The classical method C1 performed the best , yielding the least 'false negatives', especially at higher δ values.

The classical method C1 showed the least 'false positives' (Tables 8A-8B). The method B1 also showed a comparatively less number of 'false positives' than B2 and C2. Thus, the methods showing more 'false negatives' tended to show less 'false positives' and vice versa. This supports the conclusions drawn by Higle and Hecht (1989). However, the behavior of the methods seemed to change as the numbers analyzed increased, such as in the rural two lane intersections class. Higle and Hecht (1989) did not consider samples of more than 100 sites in any of their analyses. The 'false negative' fractions (Table 8D) for the rural 2 lane highway class is prohibitively high. Values such as 0.85 and 0.89 are inadmissible errors in the methods B1 and B2. There is no difference in the performance of B1, B2 and C2. However C1

performs better in the rural 2 lane class when the other three methods yield more 'false negatives.'

As the iterations are independent events, the number of 'false negatives' may be assumed as independent identically distributed.

Table 8A. Distribution summary for intersections in Urban Interstate highway class.

\ Haz-\ard Flag \		δ = 0.90		δ = 0.95		δ = 0.99	
		H	NH	H	NH	H	NH
B1	F	14.767	6.867	8.933	11.533	6.000	11.400
	NF	0.233	86.133	0.067	87.467	0.000	90.600
B2	F	14.600	5.267	8.800	9.967	6.000	9.900
	NF	0.400	87.733	0.200	89.033	0.000	92.100
C1	F	12.367	1.167	7.567	2.167	4.367	0.800
	NF	2.633	91.833	1.433	96.833	1.633	101.200
C2	F	14.600	5.967	8.900	10.700	6.000	11.100
	NF	0.400	87.033	0.100	88.300	0.000	90.900

Table 8B. Distribution summary for intersections in Rural Two Lane highway class.

\ Haz-\ard Flag \		δ = 0.90		δ = 0.95		δ = 0.99	
		H	NH	H	NH	H	NH
B1	F	12.300	19.500	6.033	14.767	1.033	7.400
	NF	17.700	399.500	16.967	411.233	8.967	431.600
B2	F	15.067	48.700	8.533	38.067	1.500	20.733
	NF	14.933	370.300	14.467	387.933	8.500	418.267
C1	F	15.667	10.767	11.167	8.833	5.300	5.767
	NF	14.333	408.233	11.833	417.167	4.700	433.233
C2	F	16.267	51.833	10.633	45.133	3.233	32.067
	NF	13.733	367.167	12.367	380.867	6.767	406.933

Table 8C. False negative summary statistics for intersections in Urban Interstate and Rural Two Lane classes.

\ Statistic \ Method\		Urban Interstate				Rural two Lane			
		I	II	III	IV	I	II	III	IV
0.90	B1	0.233	0.086	2.0	24.0	17.70	0.407	23.0	0
	B2	0.400	0.109	2.0	19.0	14.93	0.351	20.0	0
	C1	2.633	0.086	6.0	2.0	14.33	0.382	20.0	0
	C2	0.400	0.088	2.0	19.0	13.73	0.506	18.0	0
0.95	B1	0.067	0.046	1.0	28.0	16.97	0.402	21.0	0
	B2	0.200	0.051	1.0	24.0	14.47	0.356	19.0	0
	C1	1.433	0.058	3.0	5.0	11.83	0.421	17.0	0
	C2	0.100	0.069	1.0	27.0	12.37	0.536	17.0	0
0.99	B1	0.000	0.000	0.0	0.0	8.97	0.362	10.0	0
	B2	0.000	0.000	0.0	0.0	8.50	0.295	10.0	0
	C1	1.633	0.015	4.0	6.0	4.70	0.440	7.0	0
	C2	0.000	0.000	0.0	0.0	6.77	0.463	9.0	0

Statistic I : Average number of false negatives
Statistic II : Average error per false negative
Statistic III : Maximum number of false negatives
Statistic IV : Number of iterations with zero false negatives.

Table 8D. False negative fractions and t test results for differences in means of false negatives.

\ Stati \stic		Urban Interstate				Rural Two Lane			
Method\		FNF	tB2C2	tC1C2	tC1B2	FNF	tB2C2	tC1C2	tC1B2
0.90	B1	0.016	-	-	-	0.590	-	-	-
	B2	0.027	ns	-	sl	0.497	sh	-	ns
	C1	0.176	-	sh	sh	0.478	-	ns	ns
	C2	0.027	ns	sl	-	0.458	sl	ns	-
0.95	B1	0.007	-	-	-	0.738	-	-	-
	B2	0.022	ns	-	sl	0.629	sh	-	sh
	C1	0.159	-	sh	sh	0.515	-	ns	sl
	C2	0.011	ns	sl	-	0.538	sl	ns	-
0.99	B1	0.000	-	-	-	0.897	-	-	-
	B2	0.000	na	-	na	0.850	sh	-	sh
	C1	0.272	-	na	na	0.470	-	sl	sl
	C2	0.000	na	na	-	0.677	sl	sh	-

FNF : false negative fractions
tB2C2 : t test results for difference in number of false negatives between methods B2 and C2.
tC1C2 : t test results for difference in number of false negatives between methods C1 and C2.
tC1B2 : t test results for difference in number of false negatives between methods C1 and B2.
t test results ($\alpha = 0.05$) explanation:
ns - no significant difference
na - not applicable
'-' - not between methods indicated
sh - significantly higher
sl - significantly lower

As the number of iterations are fairly large (n=30), it may also be assumed that the number of 'false negatives' follow a normal distribution. Table 8D gives an overall picture of the

simulation analysis performed on the chosen highway classes. As shown, the most critical number for any analysis is the number of false negative (shown by the false negative fraction FNF). The t-test is therefore performed for the difference in means of false negatives of Methods B2 and C2 (indicated by t2B2C2 in Table 8D), methods C1 and C2 (tC1C2) and methods B2 and C1 (tC1B2).

The t-test results seem to indicate that in most cases, method C2 produced lesser 'false negatives' than B2 or C1. When the number of sites under analysis was high, the method C1 seemed to perform the best, with the least number of 'false negatives.' For results of other location types, see Tables 8A-8D which include detailed result tables and t test analyses. In general, the intersection results can be extended to the other location types with little difference.

As this is a simulation, some of the locations may be incorrectly flagged or may be wrongly represented as hazardous. The normal 90th percentile, 95th percentile, etc. do not truly represent the population percentiles, because the exact distribution may vary. Therefore, these simulation results should not be considered highly accurate, but these are threshold values and must therefore be sufficient indicators.

As the number of locations of analysis increases, C1 tends to perform better than B2 or C2. This can be attributed to the fact that the Bayesian technique is based on the method of moments for estimating the regional gamma parameters α and β. Instead, the empirical Bayes estimate of the mean and variance may be used (see

discussion by Morris, Higle and Witkowski, 1988 & Morris 1984. This was also tested as criterions B3 and B4 (See program listing in the appendix). The probability values are very less for all locations, so DELTA values of 0.90 and above cannot be used. In essence, this type of simulation cannot be used to compare the revised methods, B3 and B4.

In general, a large number of locations are excluded from the analysis by most highway agencies. In this case, the quality control method and the Bayesian criterion 2 perform equally well. The percentage of false positives are also quite minimal in the case of B2 and C2 if the number of locations analyzed are small (50 - 500). In cases where the number of locations are large, i.e., greater than 1000, the population should be subsetted to get optimum results.

From the above analysis of the simulation results it is evident that the performance of Bayesian Analysis (methods B1 and B2, and in particular B2) is congruous with the classical methods (C2). This result when combined with other advantages of Bayesian Analysis can be a strong motivation for the Accident Analysis Experts to use the Bayesian Methods for Accident Analysis.

2.2. THE MENU-DRIVEN INFORMATION SYSTEM

As an end product of the project, a menu-driven computer information system is implemented. The system includes an integrated skid resistance relational database, and a user-friendly application system, taking into account current LDOTD and LTRC

requirements and forthcoming enhancements. Wet weather accident data is integrated into the relational database with all the skid data including inventory, new materials, legal, special requests, etc. The user-friendly application system supports information retrieval and update on the database. The four components of the menu are: maintenance of the skid resistance database, a reporting system, the highway accident analysis system, and the archive management. The user manual for the information system is illustrated in VOLUME III.

2.2.1. Database Maintenance

The maintenance function of SRDMS has been provided to carry out general maintenance of the database. Using this facility, one can access the relational tables and perform the following functions: (1) browse the tables, (2) update the tables, and (3) obtain a Hard-Copy of any table.

To browse any of the tables, the 'browse' option has to be selected which takes the user to the table-selection menu. Upon selection of the desired table, a display of the selected table appears. Care has been taken to make sure that punching any wrong key on this display does not affect the table in any way.

The Update function consists of the following sub-functions: Delete/Edit any particular record, Insert a new record to the end of the table, and Confirm Changes function.

Upon selecting the Delete/Edit function, one again reaches the table selection menu. Once a table has been selected, one reaches

the top record of that table. Any changes made on this display are stored as new values. To access any particular record, one has to write the record number on the command line. Deletion is also performed in a similar fashion. To insert a new record, one selects the 'Insert' option which takes the user to the bottom of the selected table. A new record can be added and stored here.

Selection of the 'Confirm Changes' option reruns the analysis programs on the modified tables and updates the output file which is used for generating reports.

2.2.2. Report Generation function

This selection gives a display of the various reports generated by the DOTD for the most recent data available. In addition to that, this selection also supports some user requests and queries for LTRC. Upon selecting this option, the user is asked to make a selection, i.e., whether the person wishes to see either the DOTD reports or the LTRC requests.

(a) The DOTD Reports

This option consists of the reports generated by the DOTD for Sections, Intersections, and Spots. The format of the reports generated here is similar to the format which the DOTD is currently following, the only difference being that the two 'section' reports have been combined into one single report. Also, the database system has been incorporated with a 'cluster' report as an addition to the above-mentioned three reports. This has been a ramification of the 'Analysis method' recommended by our research team. The

user has been provided with a large amount of flexibility for viewing the above reports. One can access the hazardous locations using one of three channels Statewide, Districtwide and Parishwide. As the titles suggest, 'Statewide' report gives an overall picture of the hazardous locations over the entire state of Louisiana. Options 'Districtwide' and 'Parishwide' classify the locations districtwide (all nine districts) and parishwide (all sixty four parishes). Further, analysis has also been made available separately for each 'highway class' (i.e., 2 lane rural, 4 lane rural etc,.).

(b) The LTRC Requests

This selection takes care of some user-specific queries of LTRC. Request1 gives a skid resistance report for the most recent year. Request2 is for viewing 'inventory' type data.

Several significant contributions of the menu-driven report generation can be summarized as follows: (1) it is very easy to use; (2) it allows every manager to see all the reports generated from the DOTDACC file which evidently will provide valuable information to the manager; and (3) the computer-based reporting system allows us to generate reports which are really needed. By so doing, it reduces the number of hard copies generated and saves money.

2.2.3. Wet Weather Highway Accident Analysis

An important mission of the project is to conduct the wet weather highway accident analysis. The analysis is performed in

Section 1.3. This selection recommends the best method of wet accident analysis for intersections and clusters (spots) based on the results of the simulation experiments performed.

2.2.4. Archive Management

The archive management selection allows the user to conduct the previous three selections (i.e., maintenance of the database, report generation, and wet weather highway accident analysis) in the last year. The archiving activity is believed to be very important.

2.3. DATA QUALITY OF THE INFORMATION SYSTEM

The sources of the data stored in the skid resistance database are (1) the DPSACC file, (2) the DOTDACC file, and (3) the Control Section Data file. In order to evaluate the data quality of the information system, one needs to understand completely the current process of data transfer from the DPSACC file and the Control Section Data file to the DOTDACC file.

The DPSACC file obtained from the Department of Public Safety contains the accident information collected from the site of the accident in its entirety. The DPSACC file contains this accident information in the format of four records:

(1) Record 1 contains the accident description.

(2) Record 2 contains the vehicle description.

(3) Record 3 contains the occupant description.

(4) Record 4 contains the pedestrian description.

Before the DPSACC file is used to generate accident reports, a process of editing needs to be conducted. This entire process is presented in Figures 8 and 9. Figure 8 shows the various steps involved seen at the 'File' level. Figure 9 shows the effect of the various steps on a particular accident record.

The editing process in Figures 8 and 9 includes the following steps.

(1) The TATA 8010 program restructures the DPSACC file by inputting four records per accident into one record. The TATA 8030 program performs some consistency checks. This information is extracted using an EASYPLUS program. To take care of the DOTDACC file inconsistencies, a first temporary file and an error report are generated. Errors identified in the error report are sent to the DOTD for correction.

(2) These errors are corrected manually either by simply observing the type of error or by referring to the accident report.

(3) The corrected report is sent back to DOTD for implementation of those corrections and further checks for other inconsistencies (e.g., wrong specification of 'highway class'). The TATA 8050 program does the job and generates a second temporary file and an error list.

(4) The second temporary file is then corrected by the TATA 8060 program to generate a third temporary file. Also, the program does a vertical deletion of some supposedly irrelevant fields (e.g., primary contributing factor, secondary contributing factor etc.).

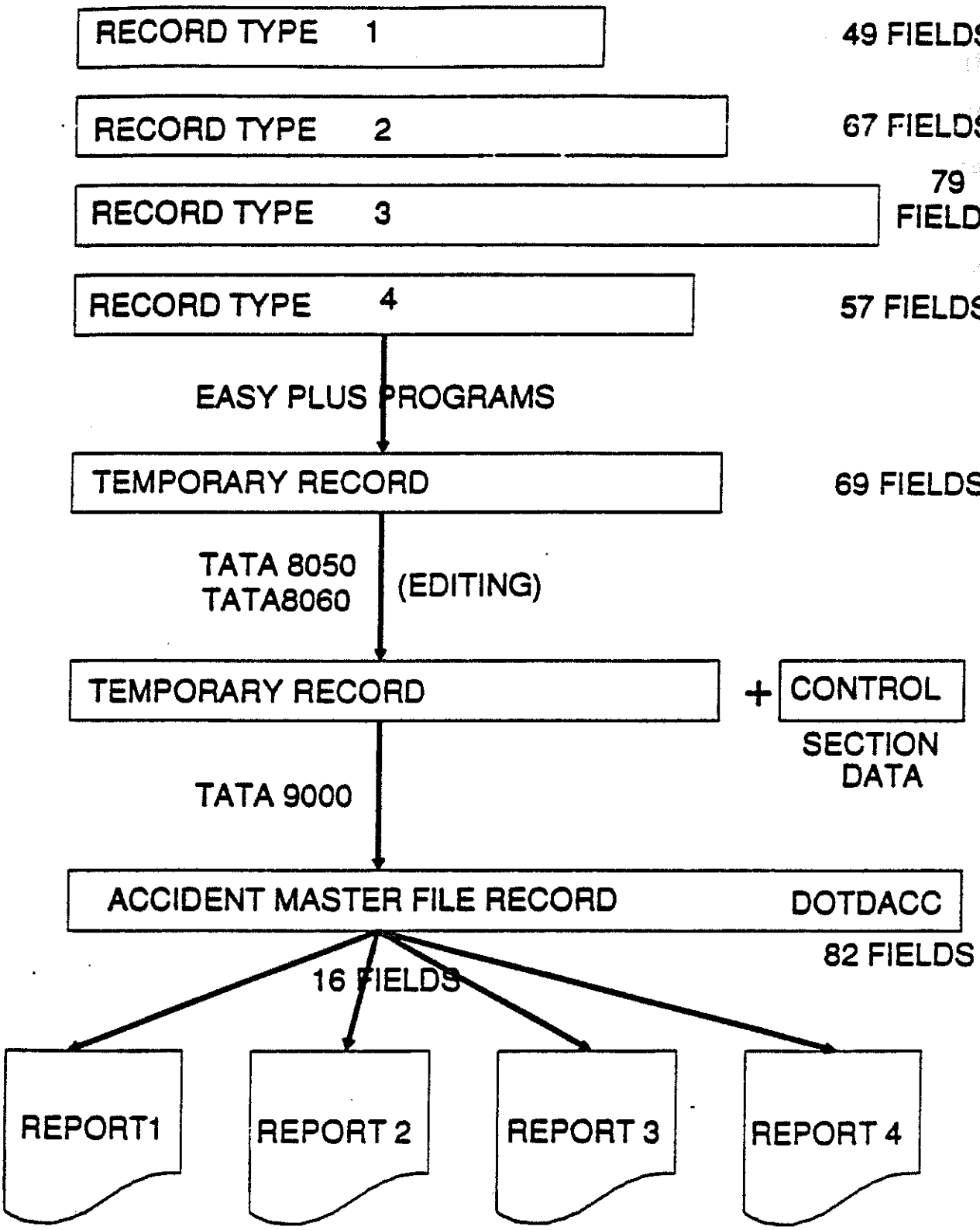


Figure 8. The record view of data transfer from the DPSACC file

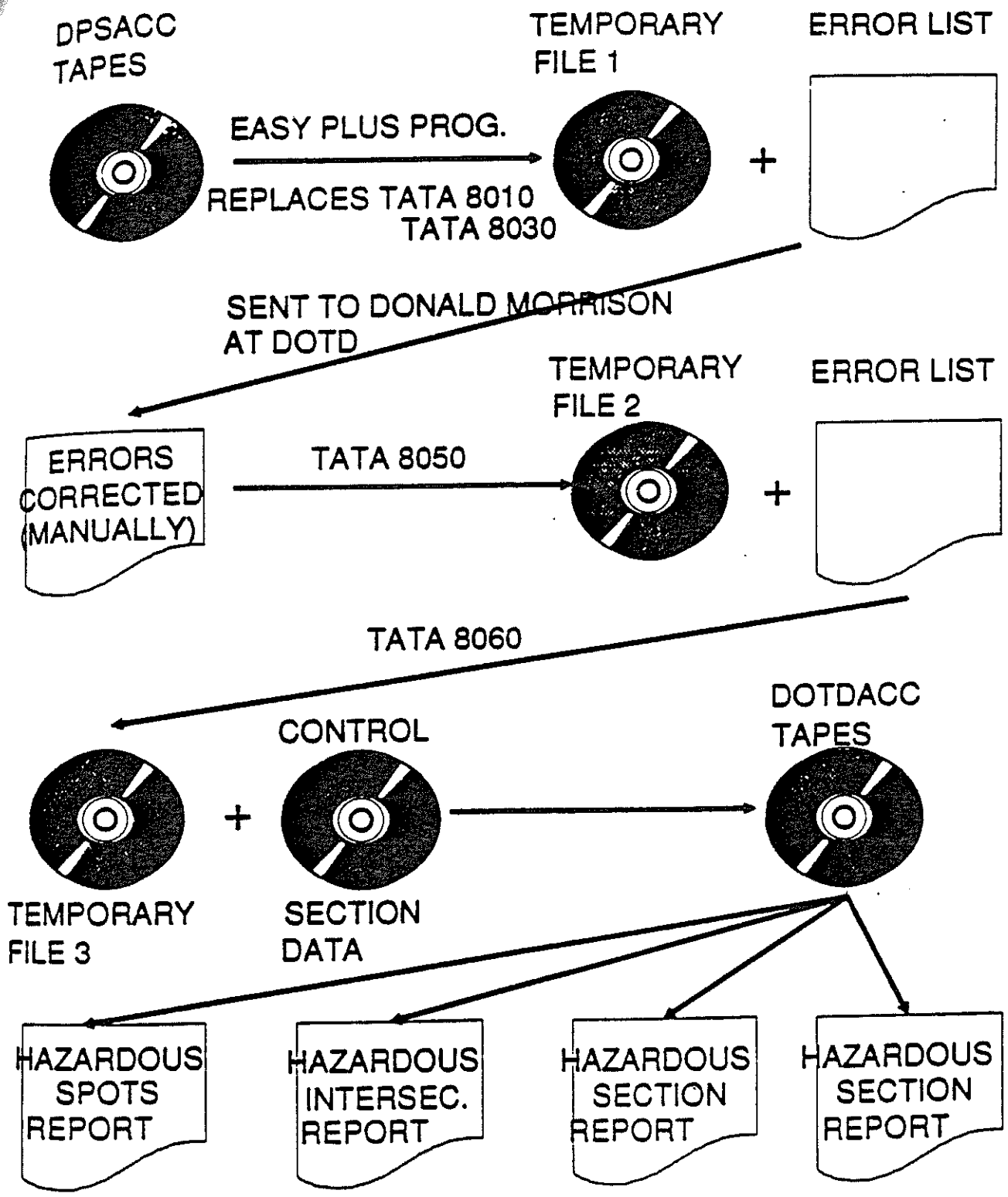


Figure 9. The file view of data transfer from the DPSACC file to the DOTDACC file

(5) The TATA 9000 program involves merging of 'surface type log/cross reference file' (i.e., the Control Section Data file) and the third temporary file. This results in the generation of final Master Accident File (or the DOTDACC file).

(6) This DOTDACC file is used and analyzed to generate four reports: Hazardous Spots report, Hazardous Intersection report, Hazardous Section report 1, and Hazardous Section report 2.

2.4. RECOMMENDATION OF A DATA RECOVERY METHODOLOGY TO ENHANCE THE QUALITY OF EXISTING DATA

The proposed methodology identifies a need for an expert system to substitute for the manual edits. An expert system will not only regenerate data from the available data, but will also make sure that there is consistency when similar situations are encountered. This cannot be guaranteed by a human expert. At the same time, the proposed methodology also perceives that having an expert system alone will not suffice. This is because of the very nature of the data. There are so many attributes being handled at the same time that it will be very difficult to come up with a sound knowledge base and an exhaustive set of rules to run the system. To take care of this shortcoming, a neural network, which has the property of identifying patterns, working in conjunction with an expert system is proposed (an integrated expert system neural network approach).

2.4.1 Objectives of Data Recovery

The objective of data recovery should be three fold:

(1) Identification of errors/inconsistencies: This step is of extreme importance to the given scenario of wet accident analysis. Out of the total data available on accidents for a year, it has been found that, on an average just about 25% percent of the accidents can be classified as wet weather accidents. Therefore, the reliability of this figure cannot be over-emphasized.

(2) Rectification of errors/inconsistencies: Using probabilistic methods, we establish relationships between various fields under consideration (wherever possible). Consider the fields 'surface condition', 'road condition', and 'weather'. A 'dry' surface condition with 'water on roadway' road condition and 'raining' weather is impossible (erroneous data). Similarly, if the surface condition is 'wet' and road condition is 'flooding,' then the weather will most likely be 'raining' (missing information).

(3) Identification of related fields: Fields which can prove to be of consequence in generating the reports can be identified. Once the top 200 hazardous locations have been identified it can become difficult for the State Government to recommend all these locations for repair because of budgetary constraints. In this situation, the manager will have to further prioritize these 200 locations in some way so as to identify the most critical spots/locations. This can be achieved by considering fields like 'road type', 'kind of location' and 'type of road' for analysis. For example, a spot somewhere near a school should be given top priority. This will help generating information which will be more valuable thereby making the decision making process risk-free.

Phase I is initiated by determining any kind of missing values present in the data. The exercise is carried out only on those fields which are being used for Wet Skid analysis. There are a total of 16 such fields presently. The printouts exhibit only 7

fields since the other fields do not have any prior basis of classification (e.g., bclm, beginning control mile, can have any nonnegative value). The data has been analyzed for years 1986, 1987, and 1988.

The objective of phase II is to generate probability matrices for the fields for which the data was found missing in the Master Accident Data File (Figure 10). As shown in the previous phase, out of the sixteen fields which are being employed for accident analysis, only three fields were found to have missing data values. These fields and the corresponding percentage of missing data values are listed below:

Surface Condition (1986-1988) - <1%

Road Condition (1986-1988) - <1%

Weather (1986-1988) - 1-14%

Since the field 'Weather' exhibits a lot of data missing, an attempt is made to recover these values first. This is done by generating a matrix of Surface Condition Vs. Road Condition Vs. Weather with only those records which had no missing values for these fields. For this case, it is also possible to establish a correlation between 'Weather' and 'Surface Condition'. Another matrix is generated exhibiting those records which did not have 'Weather' values. By picking up all the possible values of the field 'Surface Condition' one by one, corresponding values of Weather in the matrix are traced. The most likely value (defined by the value which showed maximum occurrences) of the field weather are then identified.

TABLE OF WEATH BY SURCON

WEATH SURCON

FREQUENCY / PERCENT / ROW PCT / COL PCT	A	B	D	E	C	TOTAL
A	31479	268	103	52	3	31905
	55.90	0.48	0.18	0.09	0.01	56.66
	98.66	0.84	0.32	0.16	0.01	
	74.50	1.99	26.55	26.53	12.00	
B	9901	3378	57	34	6	13376
	17.58	6.00	0.10	0.06	0.01	23.75
	74.02	25.25	0.43	0.25	0.04	
	23.43	25.12	14.69	17.35	24.00	
C	179	9328	7	9	13	9536
	0.32	16.57	0.01	0.02	0.02	16.93
	1.88	97.82	0.07	0.09	0.14	
	0.42	69.36	1.80	4.59	52.00	
E	415	345	0	8	2	770
	0.74	0.61	0.00	0.01	0.00	1.37
	53.90	44.81	0.00	1.04	0.26	
	0.98	2.57	0.00	4.08	8.00	
H	228	65	1	90	1	385
	0.40	0.12	0.00	0.16	0.00	0.68
	59.22	16.88	0.26	23.38	0.26	
	0.54	0.48	0.26	45.92	4.00	
D	7	60	220	2	0	289
	0.01	0.11	0.39	0.00	0.00	0.51
	2.42	20.76	76.12	0.69	0.00	
	0.02	0.45	56.70	1.02	0.00	
G	26	1	0	1	0	28
	0.05	0.00	0.00	0.00	0.00	0.05
	92.86	3.57	0.00	3.57	0.00	
	0.06	0.01	0.00	0.51	0.00	
F	18	3	0	0	0	21
	0.03	0.01	0.00	0.00	0.00	0.04
	85.71	14.29	0.00	0.00	0.00	
	0.04	0.02	0.00	0.00	0.00	
TOTAL	42253	13448	388	196	25	56310
	75.04	23.88	0.69	0.35	0.04	100.00

Figure 10. Probability Matrix for 'weather' vs 'surface condition'

The results obtained are as follows. Weather condition before recovery showed that approximately 15% values were missing. After recovery it was found that with a hit rate 72.6%, the data is recovered. The hit rate was calculated by using the recovery method on the data which was already available but was assumed to be missing.

In phase III, fields found to be associated with the identified field in some way or the other are as follows:

(1) Posted Speed (e.g. a location near a school or a playground will never have a very high posted speed)

(2) Surface Type (e.g. a 'business continuous' location will never have 'brick', 'dirt' or 'gravel' type of road surface.

Data Statistics: (missing data)

	1986	1987	1988
Kind of Location	15%	2.3%	7.1%
Posted Speed	4.8%	4.3%	4.7%
Surface Type	0.8%	0.7%	0.9%

Although presently this is not being used in the analysis, this field can be of utmost importance in deciding the priority in which the identified hazardous locations can be reworked. For example, a hazardous location identified near a 'school or a playground' should have a priority over a location identified in 'open country'.

2.4.2 Data Recovery - Problem Statement

All the data pertaining to dry accidents is deleted from the data file at the time of analysis since the ultimate objective of the project is to identify "wet" hazardous locations. The way this is implemented is as follows: In the data file there is one single record (with approximately 150 different attributes) completely describing the accident, the site of that accident and some other global parameters like weather, average daily traffic etc. To identify a wet accident, the analysts refer to the attribute "weather" (at the time of accident). This attribute can have the following values: A for clear weather, B for cloudy weather, C for raining, D for snowing/sleeting, E for fog, F for smoke, G for dust (a complete listing of the layout of this data file is in the Appendix). If a particular accident record has its weather attribute marked as either C or D then the accident is considered a wet accident. Any other value of weather implies a dry accident. Again, if the value of weather is missing, then the record is dropped from the analysis. This is where the problem occurs. It was found for year 1988's data that out of a total of 60,000 records, only 30% percent were wet accidents and 18-20% of the data was missing.

Problem Statement

Presence of some keypunch errors and inconsistencies in a data file lead to exclusion of a large chunk of data from analysis. The problem is to devise some methodology which could recover this data from the existing data in the best and most reliable way.

Several statistical and probabilistic models have tried in the past to overcome these kind of impediments. None of those have been very successful. One of the major drawbacks of these kind of systems is their static nature. Once the probabilities have been assigned, they stay on forever. There is no means of changing the probabilities dynamically unless they are recomputed and then reassigned.

2.4.3 Proposed Methodology for Data Recovery

An Integrated Expert System and Neural Network methodology is proposed to tackle the above problem. An expert system is used as front end for data collection and the conclusion analysis phases.

Data Collection:

The original data, that is the data which has missing as well as complete attribute values, is first inputted to the integrated system. The input is received by the expert system. Here, a small modification is made to the original Hillman's model. In the original model, the first module was responsible only for data collection. Whereas in this model, a pre-check of the data is carried out to filter out data which is contradictory. The need for this pre-check is to take care of some keypunch errors. Since a large database is being handled, there are numerous situations where there is no missing data but imprecise data.

Data Evaluation:

Once this phase is over, the filtered data is then sent to a neural network for data evaluation. The actual computing is carried

out by a neural network which exists in the background and is transparent to the user. This neural network, using a standard learning technique like *back propagation, delta learning* etc. learns the desired patterns. On the basis of this learnt data, the network then generates an output.

Conclusion Analysis:

The neural network sends its output to the same expert system for data validation. The expert system has some rules based on which it validates the data. These rules are developed based on intuition and the past history of this data. Human experts are also consulted to account for rules based on human experience. After the validation, the data is outputted to another file.

Such a system offers many advantages. In addition to those mentioned in Section 2.3, one major benefit of such a system is its dynamic nature. A neural network, by its virtue of being able to learn new patterns, can do wonders in such a situation. As the network will encounter newer and newer patterns, it will be possible to train the network so as to accommodate new inputs.

2.4.4 Detailed Design of the Data Recovery System

Data Analysis

Upon close examination of the DPS data file, it was found that there existed two other attributes, namely, Surface Condition (of the pavement at the time of accident) which was closely related to the weather attribute and the Road Condition. Surface condition has options like: A for Dry, B for Wet, C for Muddy, D for Snowy/Icy,

E for Other (any other condition). This attribute had relatively lesser number of values missing (about 1%). Similarly road conditions also has options like: A for defective shoulders, B for holes, C for deep ruts, E for loose surface material, F for construction repair, G for overhead clearance limited, F for construction, no warning, I for previous accidents, J for flooding, K for water on roadway, L for orthogonal faults on surface, M parallel faults on roadway, N for other defects, O for no defects. Again, it was found that less than 1% of records were missing from this field.

After a detailed examination of these three fields, it was concluded that there was some relationship between these three fields. It was difficult to derive a one-to-one mapping between the fields, but from intuition it was clear that these fields exhibited some correlation. For example, if the surface condition is dry (B) and the road condition is NOT water on the road (J) NOR flooding (K), then the weather cannot be raining (C) or snowing (D). Nature of this relationship cannot be determined since there are a lot of parameters involved and nothing could be stated with absolute certainty. Although there are some facts which could be stated with absolute certainty, for example, whenever the surface condition is muddy (D), the weather is never snowing (D).

Proposed Integrated System

For the above situation, wherein there is some degree of fuzziness in the relationship for certain data points and there are also some facts existing for other data points, an integrated

expert system neural network approach is tailor-made. The reason for this is the expert system in the integrated system will take care of the facts and will recover data using these facts. The neural network will take care of the fuzzy areas by first identifying and learning a pattern from the valid data. Then it will recover data using the pattern learnt. Also, it will be possible for the expert system to determine the appropriateness of using a neural network for a particular data point. Stated in simple terms once again, if the data to be recovered is fact-based, then the expert system will handle the situation and the neural network stage will be bypassed. Otherwise, the expert system will transfer the control to a neural network to recover data on the basis of the patterns learnt. The decision of selecting the right system (expert system or neural network) will be taken by the expert system.

For this study, the system is developed using the following steps.

Step 1: *An analysis of the data is carried out to extract as many facts from the data as possible.*

Step 2: *For the data from where it is not possible to gather a fact, an edit program is run to delete records which have missing values for any of the above three fields.*

Step 3: *Output from step 2 is used to train neural network(s) to identify and then later regenerate the learnt patterns. The technique used for training is the Back Propagation Algorithm (for Evaluation Phase).*

Step 4: *Using the facts of Step 1, a knowledge base is created for a Forward Chaining type Expert System (for Data Collection and Conclusion Analysis phases).*

Step 5: *Both the systems are integrated so that they can process data in conjunction.*

Now the system is ready for use. The Flow Chart of the system has been shown in Figure 10.

Validation of the output

Only the neural network output needs to be validated since the expert system output is based on facts. For validating the neural network output, the N-cross validation technique is used. In this technique, the data set, which is used for training the network, is first divided into N subsets. Then the network is trained using N-1 subsets and tested using the Nth subset. The process is repeated N times by taking a different Nth set every time.

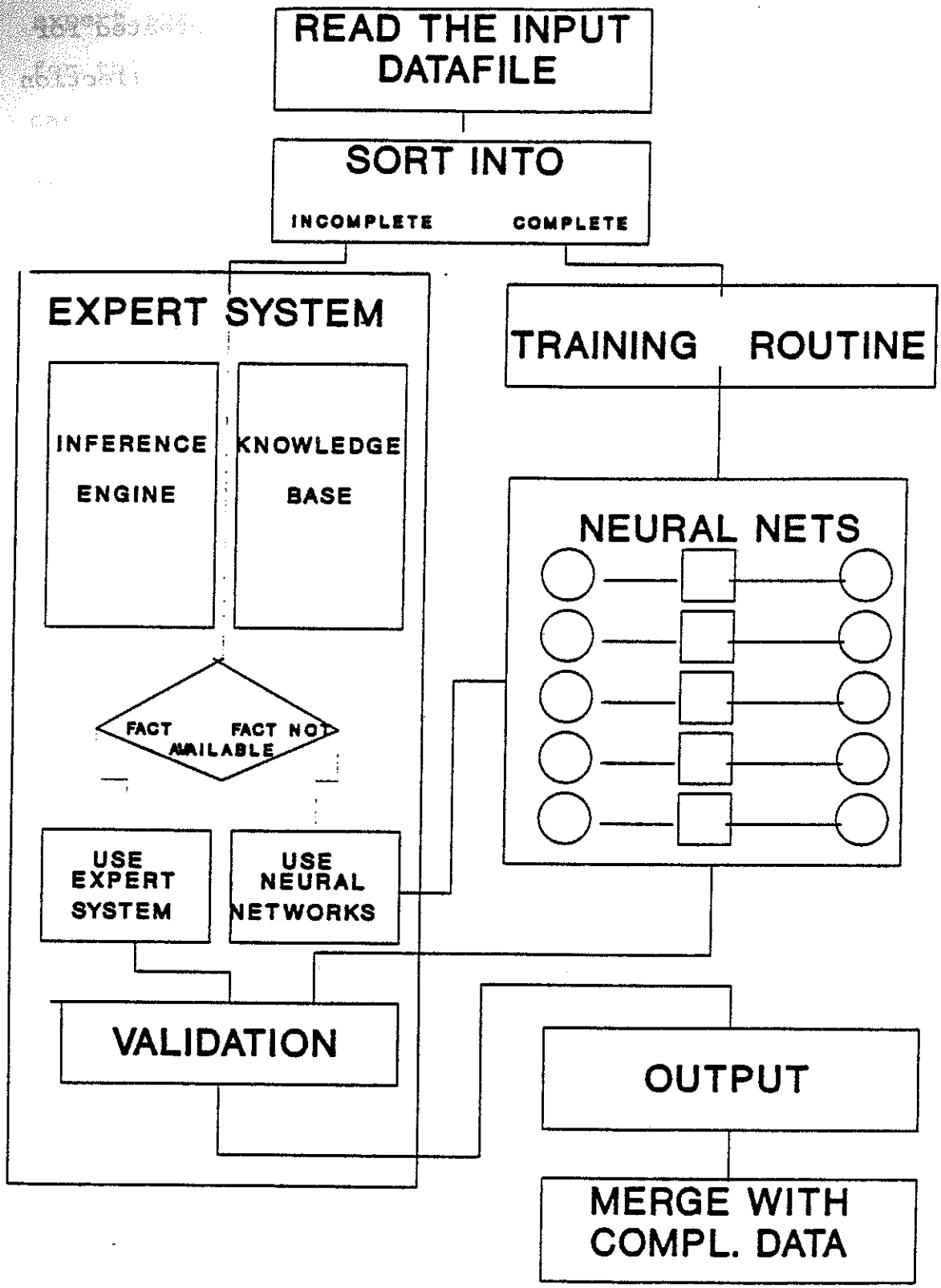


Figure 11. Integrated System for Data Recovery.

2.4.5 Program Implementation

A prototype system was developed based on the design presented in the preceding sections. This system was implemented on a DEC VAX 11/780 System in UNIX environment. The code was written in language C. This choice was primarily based on the extensive string manipulating features offered by the chosen language and the environment. A brief description of the implementation is given below.

Program NEUREX : This program is the brain of the entire system. It is basically a program written at the shell level. This program, to begin, with reads the RAW data, i.e., the data which has complete as well as incomplete data. Next, the program separates the complete data and the incomplete data into two files. This is done to facilitate the process of learning of the neural network. Once the data has been classified into complete and incomplete data, the program sorts both the files by Surface Condition. After sorting, it splits each of the above two files into five smaller files. These files have just one surface condition but a varying road condition.

Next, NEUREX invokes an encoding program called ENCODER.LEX. This program converts all the data into a '0' and '1' code since the input to the neural networks is given in the form of these bits. After the encoding has been done, NEUREX delivers all the incomplete data to the expert system. The expert system performs a preliminary check to determine the presence of any factual data. If some factual data exists,then the program does the data recovery

and gives the output. If the information received is not factual, then NEUREX redirects this data to the appropriate neural network (LEARNER.C).

Once the neural network has recovered the data, the data is sent back to the expert system for validation. The expert system validates the data by matching the data present in the knowledge base. If the recovered data (by neural network) gets an affirmative from the expert system, the result is outputted. The files used for pattern matching are PAT.1, PAT.2, PAT.3, PAT.4, PAT.5. Otherwise the expert system just gives an error message saying that the data cannot be recovered.

ENCODER.LEX :This program converts the raw data into an encoded form.

DECODER.LEX : This program converts the encoded form into a character set.

LEARNER.C :This program is a neural network program which has been based on a back-propagation learning algorithm. It takes the complete (for learning) and the incomplete data (for recovery) files as an input from the expert system. This program has one hidden layer having six neurons. The input layer has eight neurons and the output layer has four neurons. T1 and T2 are the initial weights assigned to the input/hidden and hidden/output layers.

A complete program/data files view of the system has been shown in Figure 12. The input to the program is stored in data file LTRC.data and the output gets stored in file OUTPUT.

The output of the program showed reasonable results, although it did not perform as well as the variation presented by the input data. Possible reasons for this could be the boundary constraints under which the program was developed (e.g. the number of hidden layers was one, the number of neurons in the hidden layer were six). Results will definitely be different and possibly better if these two parameters are changed. Another possible reason could be the nature of the data. The data presented one particular pattern for most of the records. This could have led to memorization by the network.

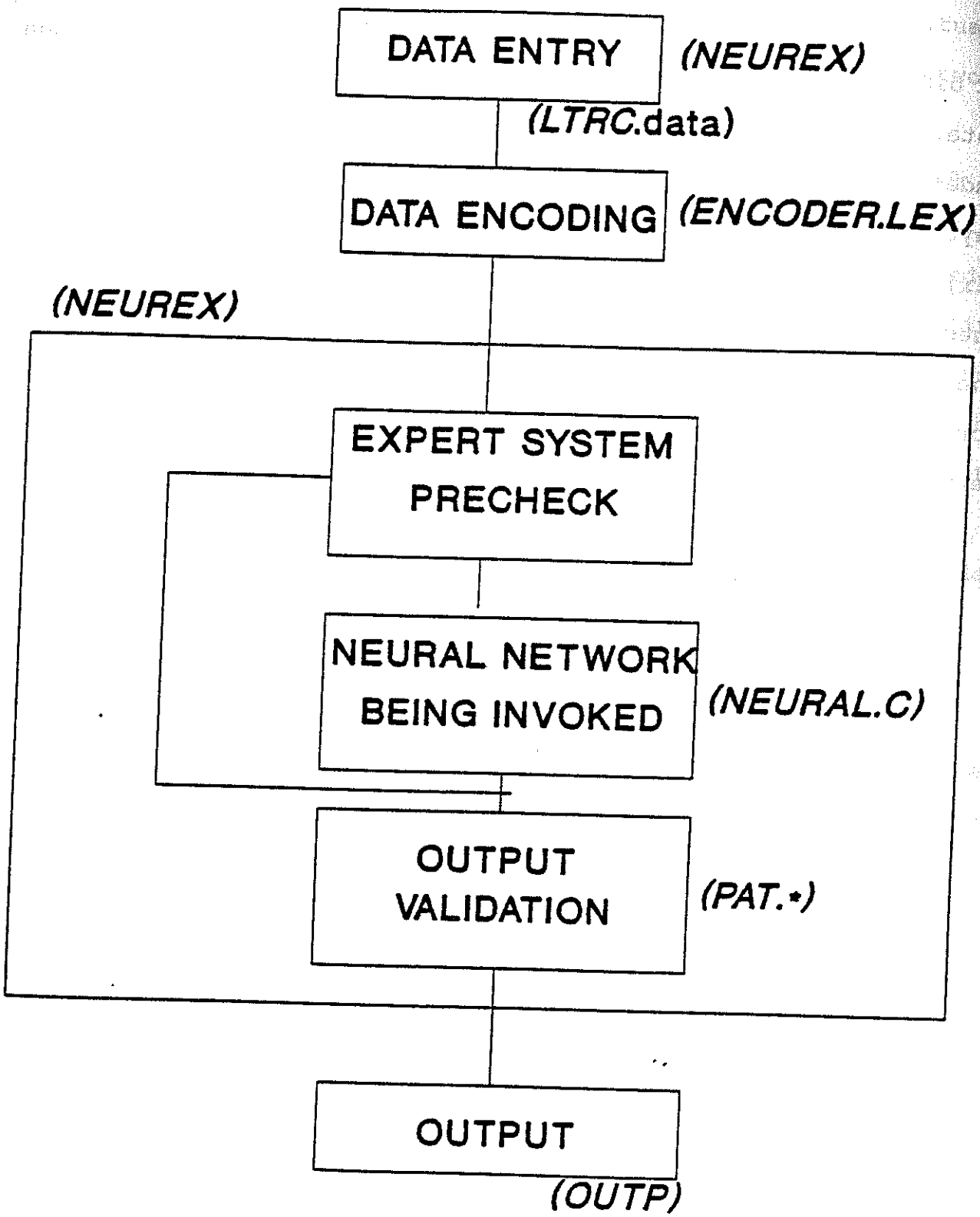


Figure 12. Program/data files view of the data recovery system

2.4.6 Illustrative Example

The prototype developed was tested by running a small data set of 5000 observations. A section of this data set is shown in the input file LTRC.data. An analysis of the output was conducted after obtaining the results (Outputs Decoded.1, Decoded.2, Decoded.3, Decoded.4, Decoded.5). It was found that the results matched fairly well with the existing pattern. One can see that for an input of 'Dry' surface condition (pattern 'A' in the first column), an output of clear weather was obtained for most of the records (pattern of 'A'in the output). Similarly, consistent results were obtained for the 'Wet' surface condition. The output exhibited 'Rainy' weather for this input. These results were later compared with the results of a frequency analysis of the complete data. Frequency analysis exhibited about 74% 'dry' surface condition matching with a 'clear' weather (consistent with intuitive results also). The expert system has also been provided with the capability of validating an existing complete data set (as mentioned in Section 3.2). That is, when the expert system encounters a record which has totally inconsistent data (e.g. 'Dry' surface condition with 'Rainy' weather), an error message is prompted. After the execution of the prototype, it was found that about one percent data did have this problem. Output in file Decoder.3 and Decoder.5 were invalidated by the expert system for having these kind of inconsistencies. A complete listing of this test-run is given in the following pages. A sample of the input file is shown in file

called LTRC.data. The output files of the program have been stored in DECODED.1,.2,.3,.4,.5.

2.4.7 Advantages of using the proposed methodology for Data Recovery

The primary goal of the prototype model was to investigate the potential for integrating neural networks and expert systems to recover data in the skid resistance data management system. Following are the conclusions reached after the study:

1) The prototype demonstrated that integration is possible, practical and productive.
2) Complex problems like the DATA RECOVERY problem can be easily solved by using an integrated approach rather than using an expert system or a neural network.
3) A neural network provides greater efficiency in capturing cases for learning where pattern recognition is of prime importance.
4) Time necessary to capture experience for training a neural network is reduced, as opposed to encoding all rules for all possible cases in a problem.

The recovery of the weather attribute is very essential and a recovery of any other attribute may not be as important. But given the same data set, the concepts of integration may be applied to recover other fields. One such example is recovering attribute "surface type" from attributes "kind of location" and attribute "posted speed". Also, by the very nature of the attributes chosen (weather, surface condition, road condition), it was almost

impossible to correlate any other attributes with the selected ones. This may not be the case with the example cited above and therefore, there could be a larger number of attributes to recover the data from, which may lead to more accurate results.

CHAPTER 3

INTUITIVE SPATIAL DATABASE

3.1 INTUITIVE SPATIAL DATABASE

A user-friendly application was developed with an intent to implement a spatial database. The purpose of this application was to provide users in the various state police troops the capability of entering and analyzing data concerning wet weather accidents around the state. The primary goal was to create an application that was easy to use, and that required a minimum amount of computer knowledge or skill to operate effectively. These goals were achieved and demonstrated on a limited scale prototype of the system.

A spatial database management system (SDMS) is one that displays and manipulates its data in a graphical, geometrical manner. A Geographical Information System, or GIS, is a large SDMS which displays geographical information. These advanced mapping applications allow users to determine relations between surface features on a computer generated representation of a map. Maps can include items such as roads, intersections, tributaries, historic landmarks, etc., to facilitate simpler location of important features.

The majority of GIS's use expensive, sophisticated mini-computers. In order to keep the cost of implementing the SDMS for this project to a minimum, it was decided that the application would be created to run on IBM-and-compatible personal computers,

hereafter referred to simply as PC's. Fast computers are necessary to manipulate effectively the large amounts of data required to display mapping features, so the PC's, which the application was developed on and for, contain the Intel 80386 and 80486 central processing unit, and an optional math co-processor to further speed computations. The data files require a large amount of space to store, so to meet these needs the computers also had fixed (or hard) disk drives. Effective display of the data and map information required color display capability for the computer, and color monitors. Higher resolution from the display resulted in better display of the data.

In order to provide a stable and broad operating system for developing the SDMS, the Microsoft Windows version 3.0 operating shell graphical user interface (GUI) was chosen. This OS runs on all PC's that meet the hardware requirements described in the previous paragraph. MS Windows also provide several capabilities, namely device independence and Dynamic Data Exchange (DDE), which can prove very useful in the development of graphical applications. Device independence allows applications developed on one computer to look and behave virtually the same on any other device, no matter how dissimilar they are. For instance, an application developed on a PC with a high resolution color display system would look the same on a system with a monochrome, except of course, for the lack of color. Dynamic data exchange is the ability of programs developed for the Windows environment to share data. By inserting data from a spreadsheet into a database the data in the

database and the spreadsheet could be changed simultaneously. All of these characteristics of the Windows GUI made it an excellent OS for the development of the SDMS.

The SDMS displayed a map of the state that the user could navigate across using a pointing device such as a mouse. Circles represented major cities, and a star represented Baton Rouge. Also shown on the overview were major waterways and interstates. A user could "zoom in" on a city by pointing at its representation (icon) and pressing a button (clicking) on the pointing device. The view would show the major throughways, interstates, and state highways of the city. The user could detail the view further by drawing a rectangle (window) around a particular region of the city and expanding it into subdivisions and surface streets. Clicking on an intersection could allow the user to retrieve or enter database information on that particular intersection. Red highlighted the problem intersections, and yellow showed the normal intersections. Another portion of the IMSE project performed the statistical analysis necessary for determining if an intersection was a problem or not. The only city available in the prototype model was Baton Rouge, and only the major thoroughfares were mapped for the project. Note that the actual creation of the maps is very labor intensive.

The implementation of the SDMS showed the improved capabilities for data entry and retrieval for people who are not computer experts that are made possible by displaying data in an

intuitive graphical form. This method could benefit many different data entry and retrieval applications for the LTRC in the future.

CHAPTER 4

CONCLUSIONS AND RECOMMENDATIONS

Three significant results of the research are presented below in the order of their importance.

First, an effective wet weather highway accident analysis procedure is developed. The procedure allows only needed locations to be identified, tested, and reported. As a result, the procedure enhances the safety of the highway system to the extent policy and funding allows. Additional benefit of the procedure is to save cost for unnecessary testing and eventual reduction of accidents which would save many lives, medical costs, and insurance costs.

Second, an SRDMS management system is implemented and tested. The system is menu-driven and user-friendly. The menu-driven system integrates the above mentioned wet weather highway accident analysis procedure and three other significant components including: (1) the maintenance of the skid resistance database, (2) the report generation at the LDOTD and the LTRC, and (3) analysis schemes, (4) archive management.

Finally, a data recovery procedure is proposed which will enhance the data quality of the menu-driven computer information system. As a result, the reports generated from the information system are more accurate and reliable.

The following are the recommendations:

(1) A better measure of the proportion wet time of pavements,

through Parametric Empirical Bayes analysis, is suggested for determining the wet accident rate for most methods of wet accident analysis.

(2) Clusters or floating point segments are recommended for spot identification over fixed point segments.

(3) The performance of Bayesian Method 2 and Quality Control Method are quite effective for medium sized population (n=50 to 500) of locations.

(4) Methods B2 and C2 are most efficient if the number of locations to be analyzed are less than 1000. In a situation where the number of locations are more than 1000, it is recommended that the population is subsetted. For example, a criterion of more than 2 total annual accidents may be used to cut off smaller values.

(5) Empirical Bayes method for estimating the regional wet accident rate parameters α and β may be used in place of the method of moments estimate.

After the project is evaluated, actual accident data can be used to run the proposed wet weather highway accident analysis procedure. The menu-driven information system can be used to generate important reports at the LDOTD and the LTRC. Finally, the proposed data recovery procedure can be implemented to enhance the data quality of the computer information system.

REFERENCES

Batini, C., ed. *Proceedings of the 8th International Conference on the Entity-Relationship Approach,* North-Holland, 1989.

Brodsky, H. and Hakkert, A.S., Risk of a road accident in rainy weather, *Accident Analysis and prevention,* 20(3), pp.161-176, 1988.

Brude, U. and Larsson, J., The use of Prediction models for eliminating effects due to regression-to-the-mean in road accident data, *Accident Analysis and Prevention,* 20(4), pp. 299-310, 1988.

Chen, P.P., The entity-relationship model: Towards a unified view of data, *ACM Transactions on Database Systems,* 1, pp.9-36, 1976.

Chen, P.P., The entity-relationship approach, *Byte,* pp. 230-232, April 1989.

Chen, Y.S., An entity-relationship approach to decision support and expert systems, *Decision Support Systems,* 4(2), pp. 225-234, 1988.

Chong, P., Chen, Y.S., and Justis, R.T., A venture capital model using CASE database, *Data Resource Management,* pp. 55-61, Winter, 1992.

Chong, P., Chen, Y.S., and Justis, R.T., A PC-based decision support and expert system for labor market analysis, *Journal of Microcomputer Applications,* 1992.

Deacon, J.A., Zeeger, C.V. and Deen, R.C., Identification of hazardous rural highway locations, *Transportation Research Record,* 543, pp. 16-33, 1975.

Everitt, A., *Cluster Analysis,* Plenum Press, 1971.

Gupta, S.C. and Kapoor, V.K., *Fundamentals of Mathematical Statistics,* Sultan Chand and Sons, 1983.

Hankins, A., A skidding accident systems model, *Transportation Research Record,* 475, pp. 245-251, 1971.

Harwood, D.W., Blackburn, R.R., Kulakowski B.T. and Kibler D.F., *Wet weather exposure measures,* Federal Highway Administration, FHWA-RD-87-105, 1988.

Hauer, E., Reflections on methods of statistical inference in research on the effect of safety countermeasures, *Accident Analysis and Prevention,* 15(4), pp. 275-285, 1983a.

Hauer, E., An application of the Likelihood/ Bayes approach to the estimation of countermeasure effectiveness, *Accident Analysis and Prevention*, 15(4), pp. 287-298, 1983b.

Hauer, E., On the estimation of expected number of accidents, *Accident Analysis and Prevention*, 18(1), pp. 1-12, 1986.

Henry, J.J., Saito, K. and Blackburn, R., *Predictor model for seasonal variations in skid resistance Vol. II : Comprehensive report*, Federal Highway Administration, FHWA/RD-83/005, 1984.

Higle, J.L. and Witkowski, J.M., Bayesian Identification of Hazardous locations, *Transportation Research Record*, 1185, pp. 24-36, 1988.

Maher, M.J. and Mountain, L.J., The identification of accident blackspots: a comparison of current methods, *Accident Analysis and Prevention*, 20(2), pp. 143-151, 1988.

Morin, D.A., Applications of Statistical Concepts to Accident Data, *Highway Research Record*, 187, pp. 72-79, 1967.

National Transportation Safety Board, *Special Study - Fatal Highway Accidents on wet pavement* - The magnitude, Location and Characteristics, NTSB-HSS-80-1, 1980.

Neter, J., Wasserman, W. and Kutner, M.H., *Applied Linear Statistical models*, Richard D. Irwin Inc., 1985.

Okamoto, H. and Koshi, M., A method to cope with the random errors of observed accident rates in regression analysis, *Accident Analysis and Prevention*, 21(4), pp. 317-332, 1989.

Page, B.G. and Butas, L.F., Evaluation of Friction requirements for California State Highways in terms of highway geometrics, *Federal Highway Administration*, FHWA/CA/TL-86/01, 1986.

Robbins, H., An Empirical Bayes approach to statistics, *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp. 157-163, 1961.

SAS User's Guide : Basics and Statistics, SAS Institute, 1985.

Wright, C.C, Abbess, C.R. and Jarrett, D.F., Estimating the regression-to-mean effect associated with road accident black spot treatment : towards a more realistic approach, *Accident Analysis and Prevention*, 20(3), pp. 199-214, 1988.

Zeeger, C.V. and Deen, R.C., Identification of Hazardous locations on City Streets, *Traffic Quarterly*, pp. 549- 570, 1977.

Zeeger, C.V., Highway Accident Analysis Systems, National Cooperative Highway research program - *Synthesis of highway practice*, Report no. 91, 1982.

APPENDIX A: ACCIDENT STATISTICS (1984-1988)

Table A1 : Frequency distribution of accident by surface condition of pavement for the year 1985

Conditions	Frequency	Percentage
Dry	61055	77.5
Rainy	17259	21.8
Snowy/ icy	52	0.1
Muddy	64	0.1
Other	232	0.3
Missing	389	0.5

Table A2 : Frequency distribution of accident by surface condition of pavement for the year 1985

Conditions	Frequency	Percentage
Dry	14831	74.2
Rainy	4611	23.1
Snowy/icy	378	1.9
Muddy	7	0.0
Other	66	0.3
Missing	107	0.5

Table A3 : Frequency distribution of accident by surface condition of pavement for the year 1986

Conditions	Frequency	Percentage
Dry	38183	76.7
Rainy	11123	22.4
Snowy/ icy	97	0.2
Muddy	42	0.1
Other	138	0.3
Missing	176	0.4

Table A4 : Frequency distribution of accident by surface condition of pavement for the year 1987

Conditions	Frequency	Percentage
Dry	36312	74.0
Rainy	12348	25.2
Snowy/ icy	108	0.2
Muddy	23	0.0
Other	148	0.3
Missing	115	0.2

Table A5 : Frequency distribution of accident by surface condition of pavement for the year 1988

Conditions	Frequency	Percentage
Dry	45012	74.7
Rainy	14318	23.7
Snowy/ icy	417	0.7
Muddy	25	0.0
Other	272	0.5
Missing	245	0.4

Table A6 : Accident Count at Intersections 1985/1986

Number of Intersections	Number of accidents per int. in 1985	Average no. of acc. per int. in 1986
585	0	1.487
153	1	0.326
50	2	0.660
16	3	0.500
7	4	1.571
9	5	2.333
6	6	1.167
1	7	2.000
1	8	6.000
2	9	5.000
1	12	5.000

Table A7 : Accident Count at Intersections 1986/1987

Number of Intersections	Number of accidents per int. in 1986	Average no. of acc. per int. in 1987
479	0	1.280
474	1	0.511
117	2	1.179
39	3	1.846
12	4	2.417
6	5	6.333
6	6	3.333
4	7	5.000
1	8	0.000
1	9	6.000
1	10	0.000
1	13	6.000

Table A8 : Accident Count at Intersections 1987/1988

Number of Intersections	Number of accidents per int. in 1987	Average no. of acc. per int. in 1988
367	0	1.289
521	1	0.470
129	2	1.101
34	3	2.059
19	4	2.000
15	5	4.333
10	6	4.000
6	7	5.667
3	8	3.667
2	9	3.000
1	10	7.000

APPENDIX B:

PSEUDO-CODE FOR WET ACCIDENT ANALYSIS

The following is the algorithm for wet accident analysis written in pseudo-code:

```
Procedure accident_rate;
  Input: master accident file for 1 year;
         Delete incorrect records;
  Classify a location as intersection or section;
  Eliminate highway types - local and arterial roads;
  Provide district numbers to parishes;
  begin sort;
   Sort accident locations by
          location type - section and intersection
         highway class - interstate etc.
         district number
         parish number
         highway type - urban divided etc.
         highway number - I010 etc.
         Control, section numbers
         Beginning of Control log mile.
  end sort;
```

```
   loop:
    provide frequency of accidents:
      location type - how much in intersections etc.
      surface condition - dry, wet, snowy, muddy etc.
      weather conditions - sunny, rainy etc.
      2 - way cross tabulations of the above.
   print frequency table;

   if wet accident then go to totacc;   end loop;
/* analysis for wet accidents */
    eliminate accidents with dry surfaces;
    for other missing data check whether the weather is raining
    and road condition is flooded;
go to loop;
totacc:    calculate total accidents in each section;
/* accidents per million vehicle miles calculation */
  for each unique location do;
  wet million vehicle miles = length * 365 * wet_ADT/ 1000000;
/* Use wet_ADT instead of ADT in wet accident analysis */
     wet accidents per wet million vehicle miles =
         total wet accidents / wet million vehicle miles
     /* used for accident frequency method */
     wet accidents per mile = total wet accidents / length of
     location
end;
 calculate  state average;
```

```
  merge with accident file;
  if wet accidents per wet million vehicle miles > 2* state wide
     average  then location is hazardous;
/* use wet accident per mile criterion for accident frequency
method */
  sort hazardous locations by decreasing order of wet accidents
per wet million vehicle miles;
  print top 200 locations by highway class and highway type;   do
the same lines from  35 to 39 for each district and each parish;
  end procedure;
```

The flow-chart for the wet accident analysis is given in Figure 13. The flow-chart for simulation method is given in Figure 14.

FLOWCHART FOR WET ACCIDENT ANALYSIS

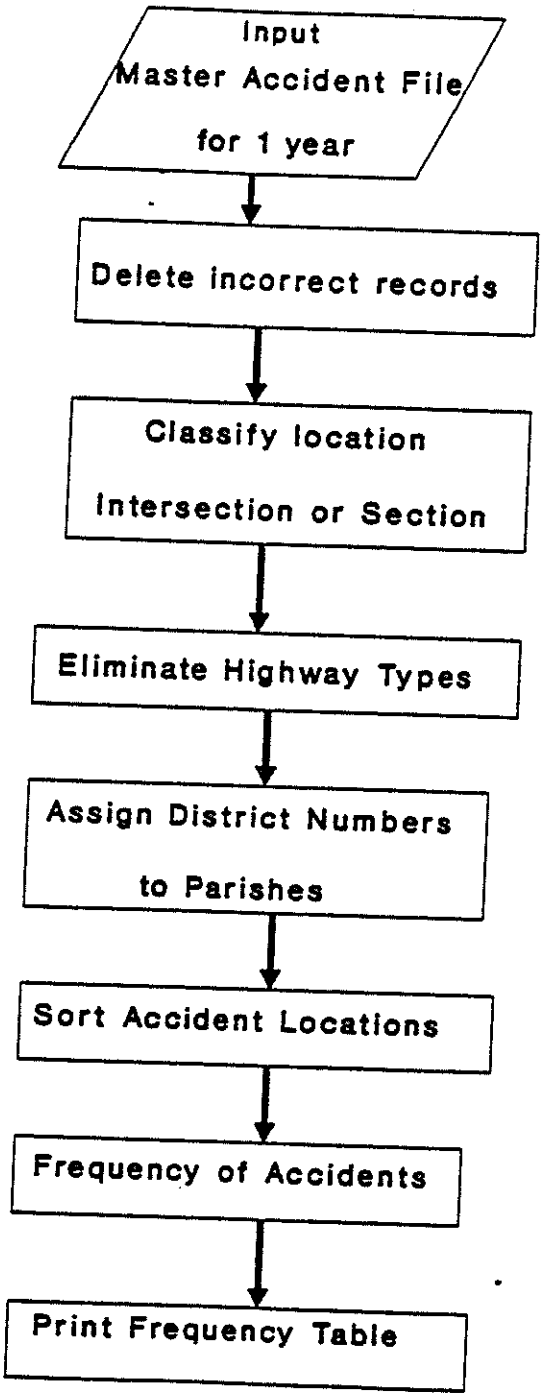


Figure 13. Flow Chart for Wet Accident Analysis

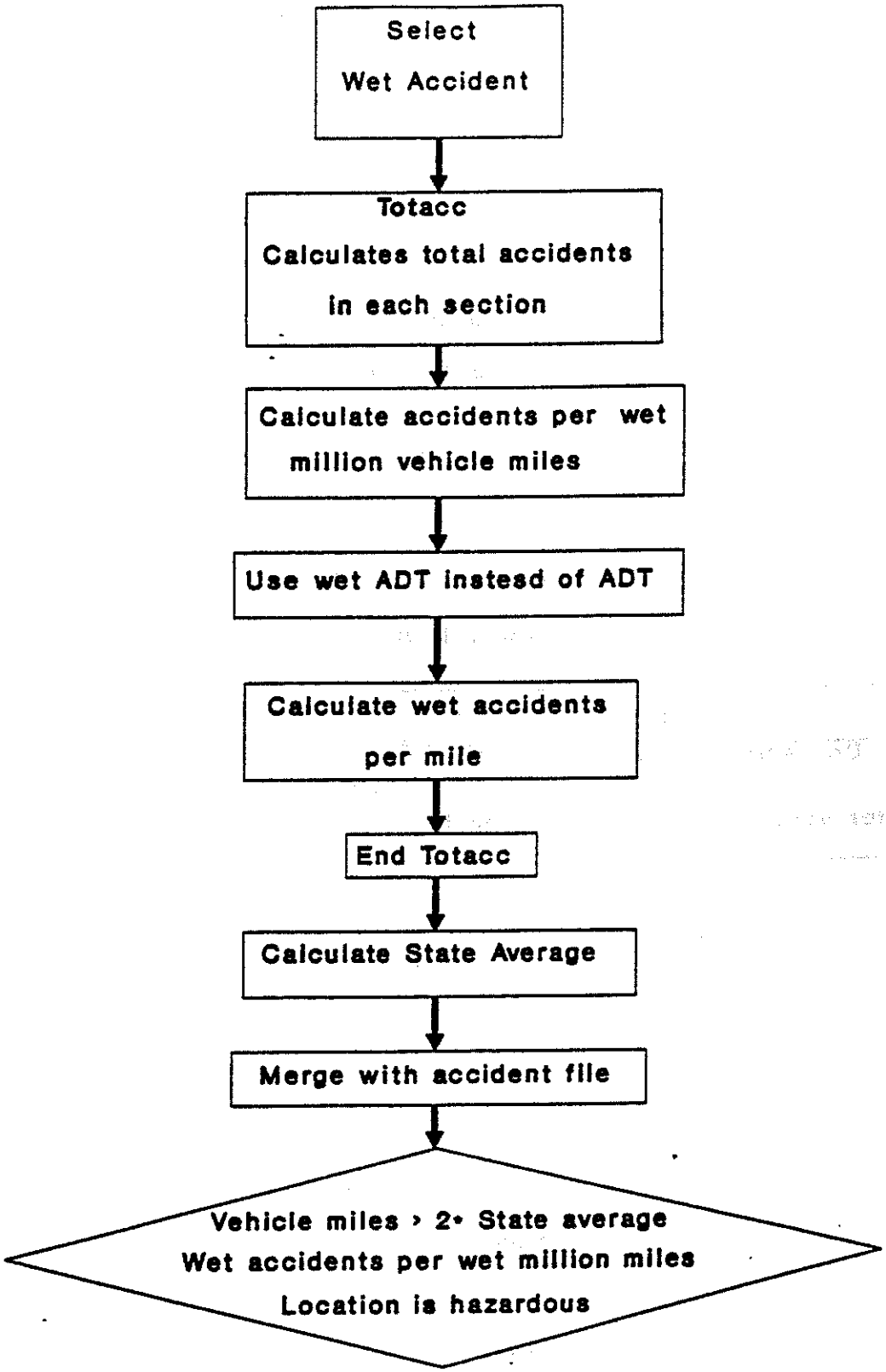


Figure 13. (Contd.)

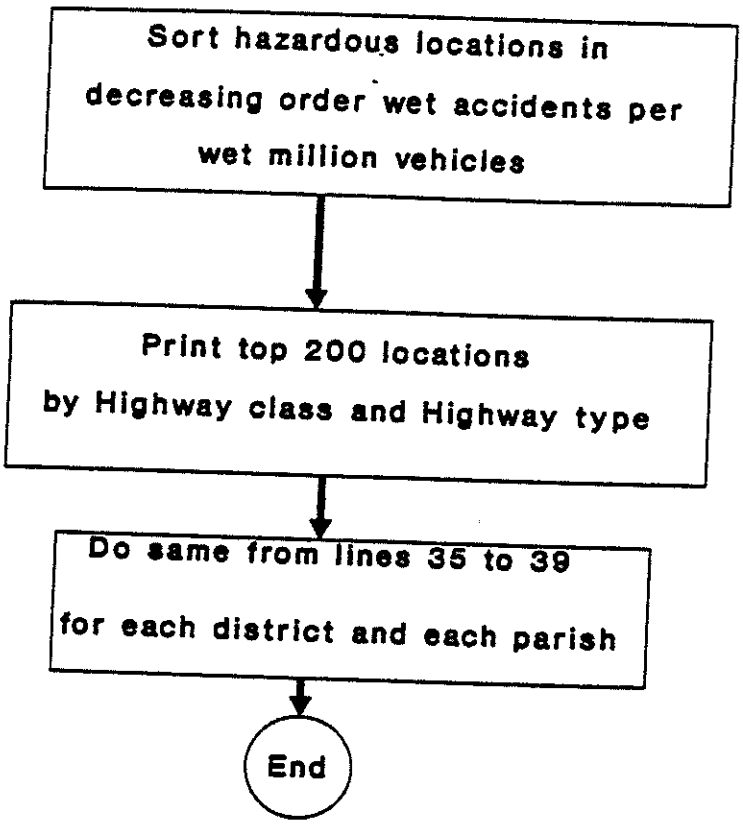


Figure 13. (Contd.)

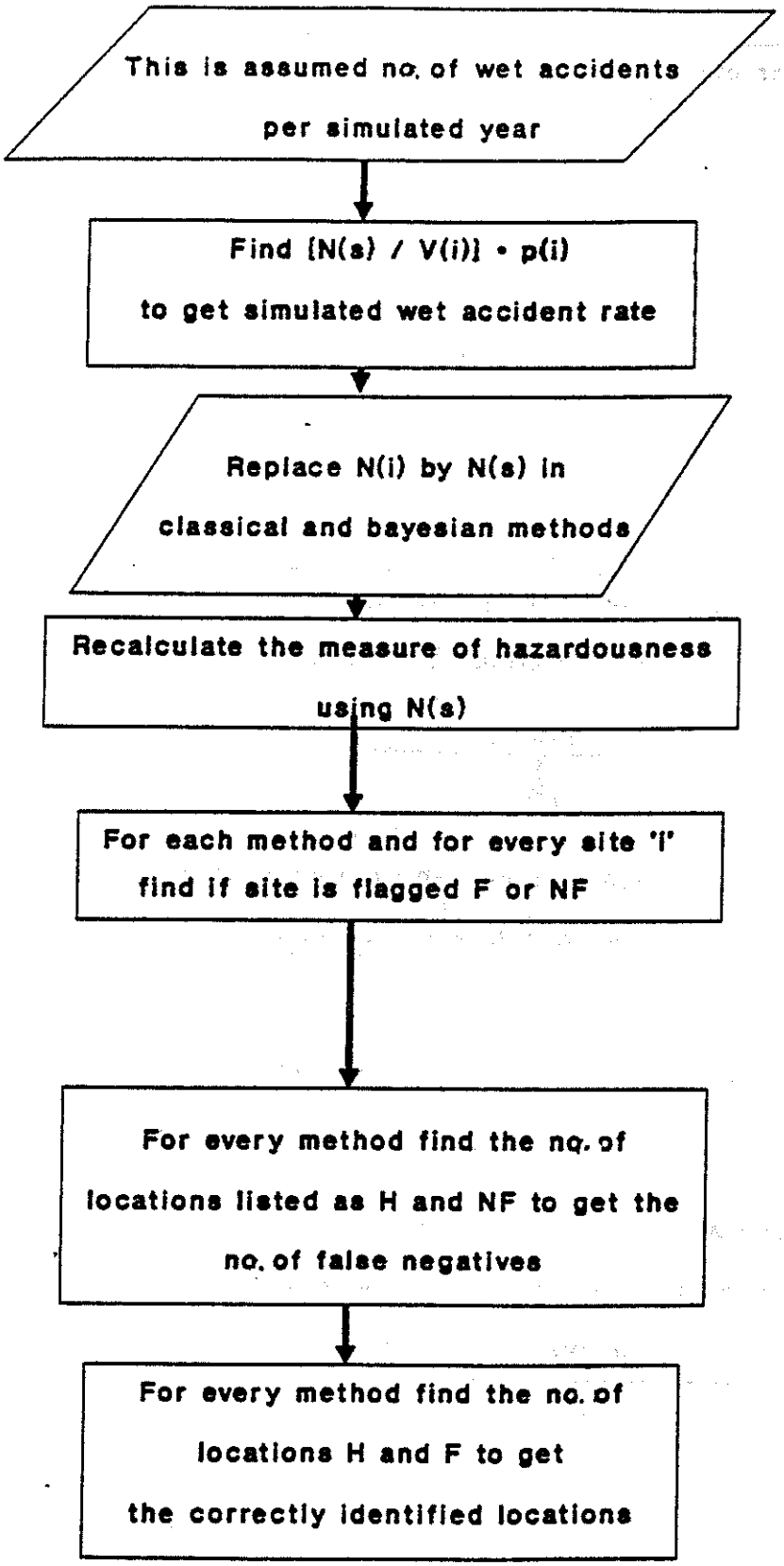


Figure 14. Flow-chart for Simulation Method.

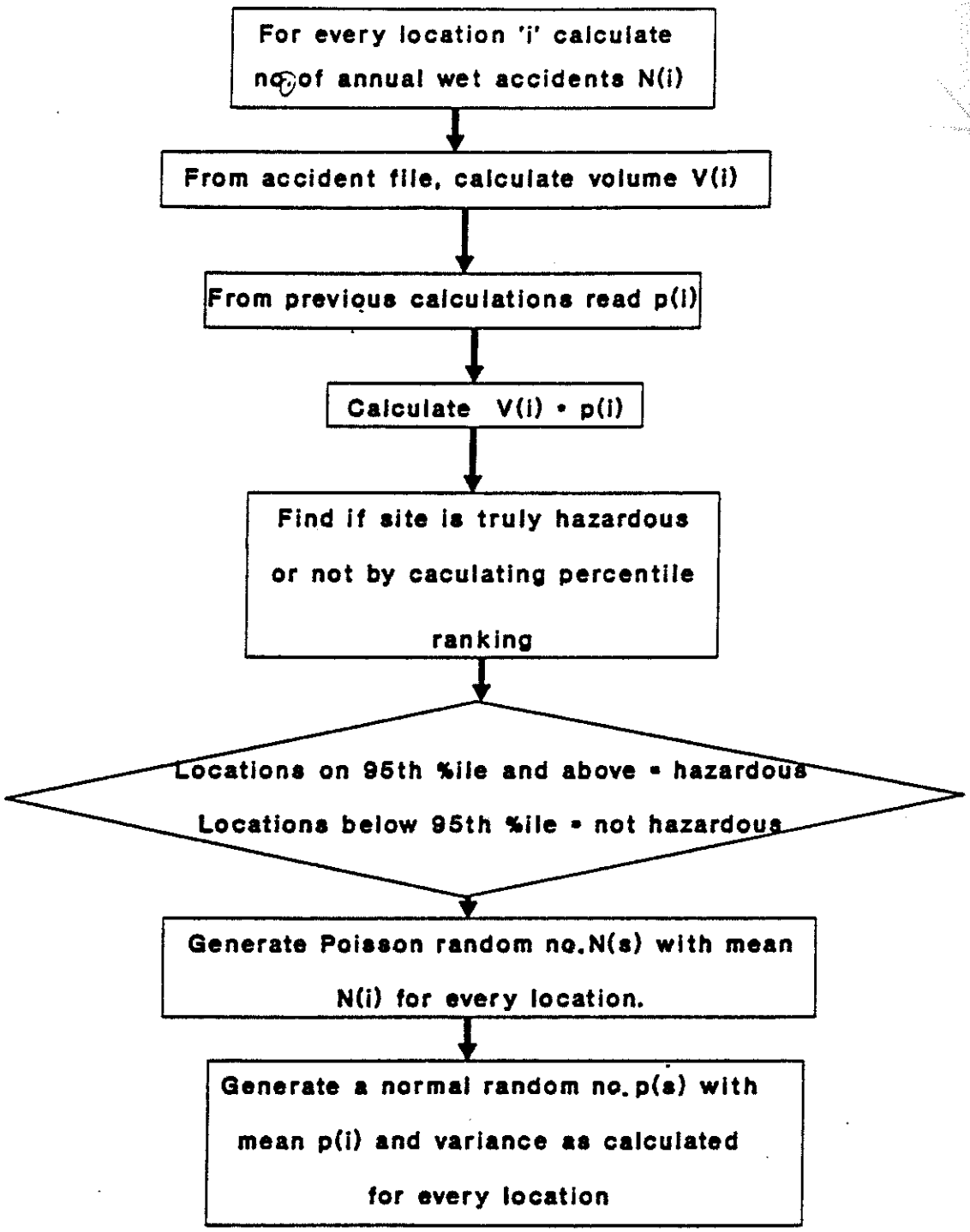


Figure 14. (Contd.)

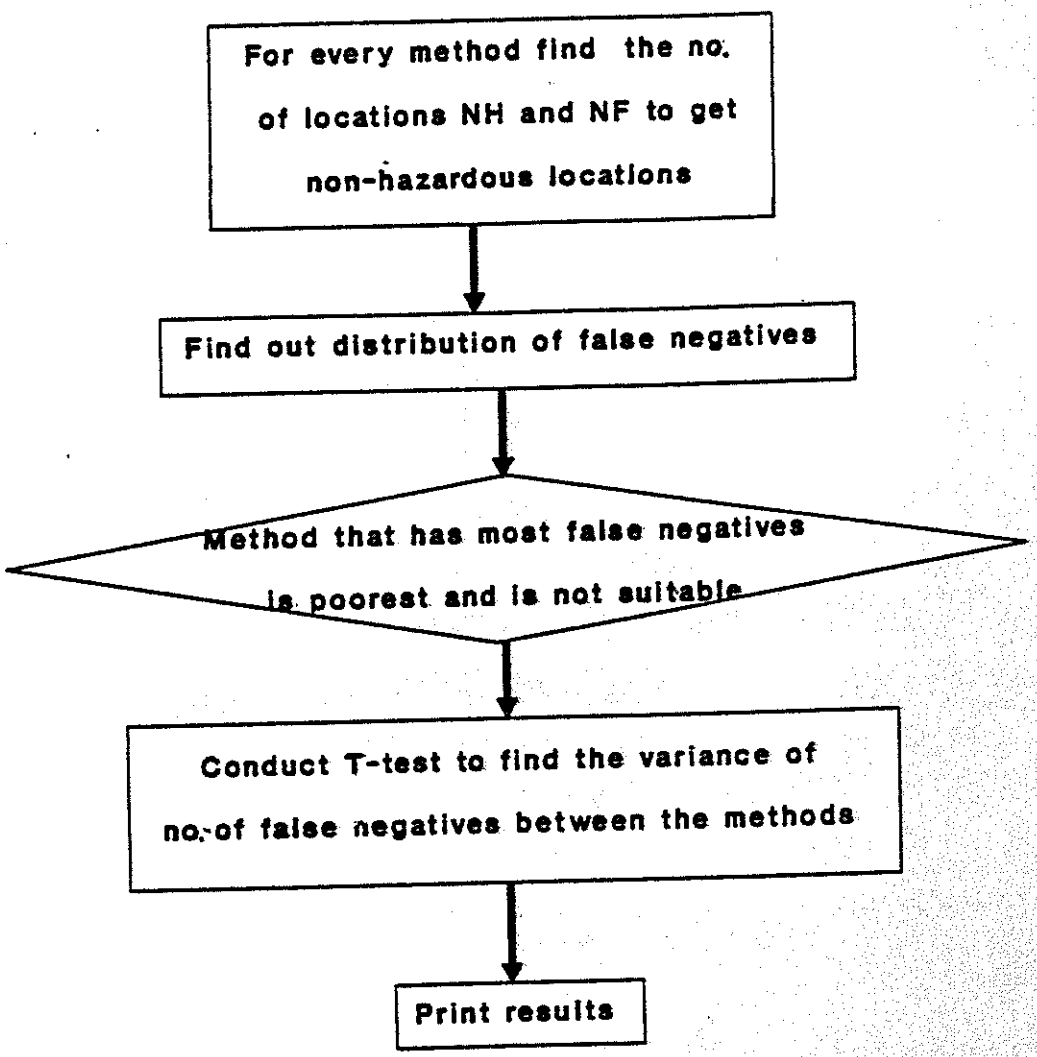


Figure 14. (Contd.)