Assessment of Existing General Purpose Data Bases for Highway Safety Analysis

KING K. MAK, JOHN G. VINER, AND LINDSAY I. GRIFFIN III

Safety is a continuing concern for highway officials at all levels of government. If the safety impacts of existing and proposed programs and policies for the construction and maintenance of highway systems are to be properly assessed, it is imperative that these officials be provided with the necessary supporting data. A recently completed FHWA-sponsored study critically reviewed a number of large national data bases for applicability and utility to various areas of highway safety that are of prime concern to the FHWA. Conceptual alternatives that would improve or enhance the capability and utility of these data bases from the standpoint of highway safety analyses were developed and evaluated for feasibility and practicality, and appropriate recommendations were made. The study results are to be considered as part of an effort to improve the capability and use of existing data bases and to offer a basis for improvements in ongoing and future data collection efforts so that the information needs of highway safety analyses may be better served.

Highway officials at the federal, state, and local levels are faced with the continuing task of assessing the potential safety impacts of proposed programs, policies, and alternatives in the construction and maintenance of the highway systems. To ensure that their decisions are appropriate and as cost-effective as possible, these officials must be provided with the supporting data needed to conduct the appropriate safety analyses. This work ranges from the identification of problems, causal factors, and countermeasures to the evaluation of the effectiveness, as well as the unintended effects of the countermeasures.

Various data bases have been created and are maintained at the federal, state, and local levels for a variety of reasons. Some of these data bases are intended for record-keeping purposes, with no consideration given to analysis requirements. Others are designed for a specific purpose and are of little use otherwise. A research study sponsored by the FHWA critically reviewed the existing general purpose data bases for their ability to meet the information needs of highway safety analyses (*1*). In this paper, selected major findings and recommendations from this study are presented.

HIGHWAY SAFETY ANALYSIS

The term highway safety, as used in this study, refers to those traffic safety areas that, at the national level, are of prime concern to the FHWA. In consideration of this definition, the emphasis of the study was on large data bases that are national in scope and intended for general purposes. The various components of highway safety analysis that are considered in this study are shown in Figure 1.

Studies of highway safety data bases can be categorized as either analysis or implementation. Analysis refers to the use of the data bases to address problems and questions from the standpoint of research and development, evaluation, and analysis. Implementation, on the other hand, is related to the development of warranting criteria and to project selection based on the warrants. The assessment of the data bases in this study was only from the analysis standpoint and excluded implementation.

As shown in Figure 1, highway safety analysis can be characterized by four factors:

- Type of analysis (problem identification, cross-sectional evaluation, or longitudinal evaluation);
- Unit of analysis (location or accident);
- Purpose of analysis; and
- Specificity (highway, accident, or both).

Type of Analysis

Types of analysis commonly used in highway safety studies are as follows:

- *Problem identification.* The determination of where and why the accidents are occurring;
- *Cross-sectional evaluation.* The study of the effects on or relationships to accidents of various factors, using information during a given time frame; and
- *Longitudinal evaluation.* The study of the effect of a given treatment (e.g., a countermeasure or a modification to a highway or to an environmental factor) on accidents during different time frames.

Unit of Analysis

This may be defined as location or accident. A location is defined as a roadway section, a point on the roadway, or a physical feature of the roadway, such as a bridge. An accident refers to an actual accident or a vehicle involved in an accident. The unit of analysis corresponds to the dependent variable used in the analysis. A location-based analysis is related to accident frequency or rate, whereas an accident-based analysis is related to accident severity.

K. K. Mak and L. I. Griffin III, Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843-3135. J. G. Viner, FHWA, U.S. Department of Transportation, Turner-Fairbank Highway Research Center, 6300 Georgetown Pike, Room T204, McLean, Va. 22101.

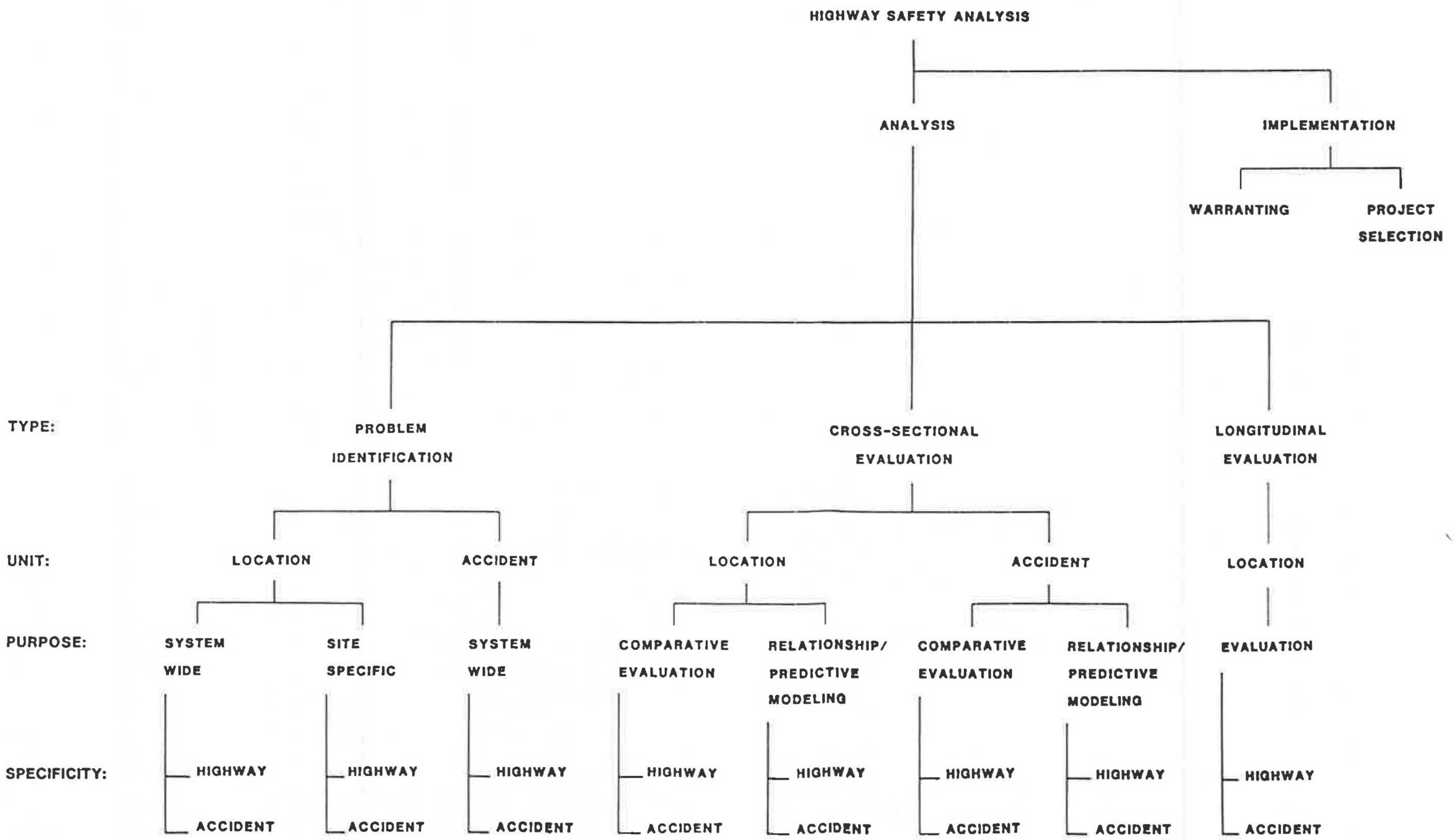

FIGURE 1 Characterization scheme for highway safety analysis.

Purpose of Analysis

This is simply the objective of the analysis, that is, what the analysis is intended to accomplish. The purpose varies according to the type of analysis being conducted.

Specificity of Analysis

This factor reflects the level of detail needed in the analysis. It includes division into subsets or constraint of the data (e.g., fatal accidents only, or two-lane rural highways only), and the inclusion of variables (e.g., highway type, accident type, etc.). The specificity for any analysis purpose can be defined by highway-related variables (e.g., highway type, curve or tangent, number of lanes, etc.), by accident-related variables (e.g., weather and surface condition, accident type, vehicle type, injury severity, etc.), or by both types.

For analysis of problem identification, the unit of analysis can be either location or accident. The purpose can be systemwide (program level) or site specific (project level). Note that accident-based analysis is not applicable at the site-specific level. By definition, a site-specific analysis refers to a location or locations and not to an accident or accidents.

In the problem-identification type of analysis, the datum sought, be it accident frequency, rate, or severity, is always related to the accident experience for a given set of conditions. Also implicit in this question is the comparison with a certain baseline to determine, first, whether a safety problem exists for the given set of conditions under study and, next, the extent of the problem.

For cross-sectional evaluation, the unit of analysis is again location or accident. The purpose of the analysis can be grouped into two general categories:

- *Comparative evaluation.* To compare the safety performance or effects on accidents between two or more different sets of conditions; and
- *Relationship or predictive modeling.* To determine or predict the effect of certain conditions or parameters on the frequency, rate, or severity of accidents.

For longitudinal evaluation, the objective is to assess the safety effects of a given treatment before and after its implementation. From the FHWA's point of view, the unit of analysis must be location, and the purpose is always evaluation.

DATA BASES SELECTED FOR STUDY

A number of candidate data bases were identified at the federal, state, and local levels from completed studies and from ongoing data-collection efforts. The candidate data bases were then categorized according to the following criteria:

- Application (primary or secondary);
- Data base purpose (general purpose or special purpose);
- Unit of analysis (location or accident); and
- Level of detail (police level, enhanced police level, or in depth).

Application

A primary data base is one that can be used directly for safety analysis. In comparison, a secondary data base can only be used indirectly for safety analysis, as a supplement to a primary data base. Only primary data bases were included in the study.

Data Base Purpose

For a primary data base, the purpose is termed either general or special. A general-purpose data base is created for general use and not for any specific application. There are very few general-purpose data bases in existence. Most data bases are special purpose in nature, that is, they were created with a specific purpose in mind. A general-purpose data base is useful for a wide variety of applications but often lacks the specificity desired for study of particular topics or questions. It must contain a large number of data elements to be general in nature, and it may not have enough depth for specific questions.

On the other hand, a special-purpose data base contains all the required data elements for the specific question or topic under study but little else. Sometimes it may be possible for a special-purpose data base to be used for addressing other questions or topics that are similar in nature to the one for which the data base was created, but such applications are usually limited.

Unit of Analysis

The unit of analysis corresponds with the unit used for the data record. Each data record in a location-based data base contains information on a location, which may be a roadway section, a point on the roadway, or a physical feature of the roadway, such as a bridge. In comparison, each data record in an accident-based data base contains information on an accident.

Level of Detail

Police-level accident data are obtained directly from police accident reports. These reports are readily available and are maintained on a continuous basis. However, there are well-documented problems associated with police-level accident data, such as the lack of detail, inaccuracy and inconsistency in definitions and nomenclature, and differences in reporting thresholds.

Enhanced police-level accident data, as implied by their name, are still based on police accident reports but are enhanced by the addition of supplemental information and better quality control. The Fatal Accident Reporting System (FARS) is an example of an enhanced police-level accident data base in which police accident reports on fatal accidents are supplemented with additional data elements, such as vehicle, driver, and medical data.

In-depth accident data are collected by trained accident investigators in much greater detail than are the police-level or enhanced police-level accident data, and they are tightly controlled to ensure accuracy and consistency. The costs of collecting in-depth accident data are very high and thus limit the sample size of the data base. Also, because police accident reports are used as the starting point for sampling the accidents to be investigated in depth, the problem of reporting criteria remains (e.g., unreported accidents, different reporting thresholds within and among the states, etc.).

Figure 2 illustrates how the four parameters are combined into the categorization scheme. Note that there are only seven

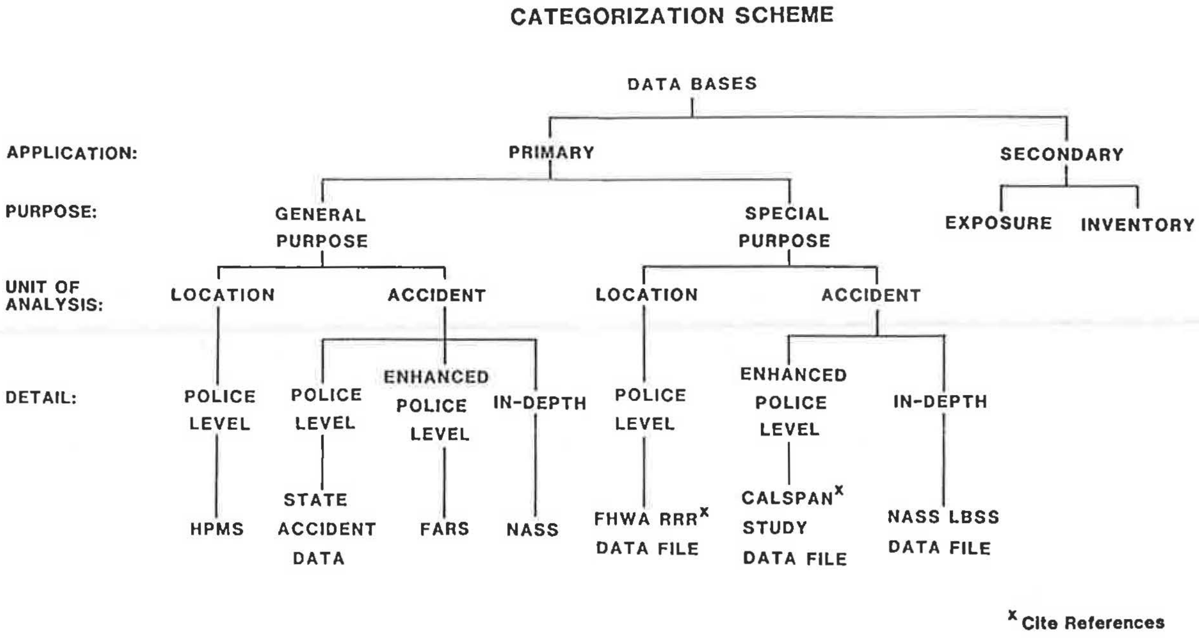

FIGURE 2 Categorization scheme for data bases.

combinations with existing primary data bases. This situation does not mean that combinations not included in the categorization scheme are inappropriate but is instead a reflection of what is currently available.

Only three general-purpose primary data bases currently exist:

- Highway Performance Monitoring System (HPMS): location-based, police-level accident data;
- National Accident Sampling System (NASS) Continuous Sampling Subsystem (CSS): accident-based, in-depth accident data; and
- Fatal Accident Reporting System (FARS): accident-based, enhanced police-level accident data.

State accident data files, particularly those integrated with road log, traffic, and other roadway and roadside data, are adaptable for use in highway safety analysis at the state level, even though they are not part of a national accident data base in the strict sense of the word. With some modifications and integration, it is conceivable that a national data base could be created from state records. State accident data files are therefore included under the category of general-purpose police-level accident data bases.

It should be noted that although the unit of analysis is shown as accident, state accident data files (and to a limited degree, FARS) can also be used for the location-based type of analysis. This is because both of these data bases are censuses of accident data, and state accident data contain location identification variables.

Three special-purpose primary data bases were also included in the evaluation for illustrative purposes:

- FHWA RRR data file: location-based, police-level accident data;
- Calspan study data file: accident-based, enhanced police-level accident data; and
- NASS Longitudinal Barrier Special Study (LBSS) data file: accident-based, in-depth accident data.

Discussions on these special-purpose data bases were presented in the final report for the study but are excluded from this article.

EVALUATION OF SELECTED DATA BASES

The state accident data bases were evaluated on the assumption that the accident data files are linked or integrated with the roadway inventory files. It is believed that such capability exists for most of the states. If the accident and roadway inventory files are computerized and have a common location reference system between the files, it is a relatively simple task to link the files together. Also, the FARS data base is considered only for fatal accidents and fatalities.

Each selected data base was evaluated on its capability to address the 16 specific types of analyses illustrated in Figure 1 for data base applicability and utility. The applicability of a data base refers to the types of highway safety analysis for which the data base can be used. The only criterion used in the assessment is whether the data base can be used for a particular analysis. No consideration is given to the utility of the data base, which is evaluated separately.

Utility refers to how good the data base is at satisfying the information needs of the applicable analyses. This quality was assessed in terms of breadth of representation, sample size, level of detail, accuracy and consistency, and flexibility.

Breadth of representation refers to the geographical distribution and sampling scheme for the data, that is, where and how the data are collected. The geographical distribution can be national, state, or local. The sampling scheme can be a census, statistical sample, or sample of convenience. A census includes the entire population of interest, for example, all fatal accidents in the nation or all reported accidents on state-maintained highways. A statistical sample allows the sampled data to be projected back to the population of interest for representative estimates. Such a sampling scheme could be based on location or accident. A sample of convenience does not provide data that are representative of the population but depends on other controlling factors for analysis.

Sample size of a data base determines the extent and detail of analysis possible with the data base, that is, the categorization of data or classification into subsets. Sample size also affects the precision of any estimates made from the data and the power of statistical tests. For an ongoing data collection effort, either the sample size available on an annual basis or that over the entire program is considered in the evaluation.

Level of detail refers to the amount of information available from a data base. The evaluation criteria include first, the total number of data elements available per record and the number of usable data elements on highway and accident characteristics, and second, the specificity of the data elements, that is, the number of levels available per data element.

Accuracy and consistency refers to the level of quality control in the data-collection effort. In other words, are the collected data accurate, or are there major sources of inaccuracy that could threaten data validity? Also, is the quality of the collected data checked to make sure that the data are accurate and consistent?

Flexibility refers to the ease of use and integration with other data files. The ease of use is evaluated from the user's standpoint for availability, completeness, and understandability of documentation for the data base. Another consideration is the extent of data processing required to create an analysis file from the data base, including case selection and data recording or reformatting. To create a new data file for analysis, it is often necessary to integrate or merge other data files into the data base for variables that are not available from the data base itself. These evaluation criteria include the availability of identification variables for merging with other data bases and the extent of data processing required.

Assessment of the applicability of the 7 selected data bases to the 16 components of highway safety analysis is illustrated in Figure 3. Note that state accident data bases, once they have been integrated with roadway inventory data, are the only data bases that are applicable to all components of highway safety analysis. The three national general-purpose data bases and the special-purpose data bases are limited in their applications.

HPMS provides extensive data on the physical and operational characteristics of the sample panels. The data are summaries because they apply to the entire sample panel, which can be several miles long and vary in length between sample panels. Specificity is thus limited to summary variables. Accident data associated with the HPMS panels are quite limited and provided in summary form only: the data consist solely of counts of total, fatal, and nonfatal injury accidents. Applicability of the data base is thus limited to location-based analyses on systemwide problem identification, cross-sectional evaluation, and possibly longitudinal evaluation. The data base is not applicable for many analyses that are accident based or site specific.

State accident data bases are applicable to all components of highway safety analysis once they have been integrated with the roadway inventory data. As mentioned previously, it is believed that most states have the capability to merge accident and roadway inventory data files, even though these files may not currently be integrated. Without roadway inventory data, the state accident data bases will be limited to only accident-based analyses.

FARS is primarily applicable to accident-based systemwide problem identification. The applicability of the data base is severely limited because only fatal accidents are included, so there is no basis of comparison in terms of severity. Also, there are no built-in exposure data to calculate fatal accident or fatality rates. Information from other sources will be necessary if the FARS data are to be used for any type of safety analysis other than problem identification.

NASS CSS data are not applicable for location-based analyses or for accident-based, site-specific problem identification because the data base has no location information. In addition, the data base does not have any exposure information for calculation of accident rates.

Assessment of the utility of the seven selected data bases can be seen in Table 1. It should be emphasized that the evaluation is based on available documentation and related publications gathered and reviewed in the study and not on actual processing and application of the data bases. Some of the evaluation criteria are rated subjectively on a simple three-point scale of poor, fair, and good.

HPMS is a national data base with information on approximately 100,000 roadway sections from all 50 states, the District of Columbia, and Puerto Rico. The roadway sections are sampled on the basis of a statistical scheme so that national estimates can be made from the data. The large sample size should allow for very detailed analysis. The level of detail is fair to good on highway-related data elements, with information available on general roadway and traffic characteristics. However, because the highway data elements are applicable to the entire section, which may be several miles long, the specificity of the analyses must be general in nature. The level of detail on accident data elements is very poor because only summary accident data are available.

The data are submitted by the states, and the extent of quality control is somewhat limited. Thus the accuracy and consistency of the data base are judged to be only fair. On the basis of review of the documentation available for the data base, the data-processing requirements for the data base appear to be fairly complicated and are thus rated fair. The data base does have the capability, albeit indirectly, of merging with other data bases by using a location-matching process.

State accident data bases are maintained by each state and are a census of all reported accidents. The sample size varies by state and is very large for analysis purposes. If it is assumed that the accident and roadway inventory data are integrated, information on general roadway and traffic characteristics is available, although the level of detail varies from state to state. For example, the Utah accident data base is a fully integrated system and contains much more information than the Texas accident data base, which has only roadway inventory data merged with the accident data.

Police-level accident data are limited in detail and are subject to inaccuracies in areas such as location identification and definitions or interpretation of data elements. Quality control of the data generally ranges from poor to fair. Data processing for state accident data bases requires large computer facilities

Analysis Characterization Scheme				General Purpose Data Base				Special Purpose Data Base		
Type of Analysis	Unit of Analysis	Purpose	Specificity	HPMS	Integrated State Accident Data Bases	FARS	NASS CSS	FHWA RRR Data File	Calspan Study Data File	NASS LBSS
Problem Identification	Location	System Wide	Highway	●	●					
			Accident		●					
		Site Specific	Highway		●					
			Accident		●					
	Accident	System Wide	Highway		●	●	●		●	●
			Accident		●	●	●		●	●
Cross-Sectional Evaluation	Location	Comparative Evaluation	Highway	●	●					
			Accident		●					
		Relationship/ Predictive Modeling	Highway	●	●					
			Accident		●					
	Accident	Comparative Evaluation	Highway		●		●		●	●
			Accident		●		●		●	●
		Relationship/ Predictive Modeling	Highway		●		●		●	●
			Accident		●		●		●	●
Longitudinal Evaluation	Location	Evaluation	Highway	●	●			●		
			Accident		●			●		

FIGURE 3 Applicability of selected data bases to highway safety analysis.

TABLE 1 UTILITY OF SELECTED DATA BASES FOR HIGHWAY SAFETY ANALYSIS

	General-Purpose Data Bases				Special-Purpose Data Bases		
Evaluation Criteria[a]	HPMS[b]	State Accident Data	FARS	NASS CSS	FHWA RRR Data	Calspan Study Data	NASS LBSS Data
1. Breadth of representation							
a. Geographical representation	National	Individual state	National	National	11 states	6 states	National
b. Sampling scheme	Statistical	Census	Census	Statistical	Convenient	Convenient	Convenient
2. Sampling size	~100,000 sections	Varies by state	~45,000 accidents/yr	~12,000 accidents/yr	196 sites (as of 5/83)	7,972 accidents	~300 accidents/yr
3. Level of detail							
a. Highway	Fair/good	Fair/good	Poor	Fair	Fair	Fair/good	Good
b. Accident	Poor (summary data)	Fair	Fair	Good	Poor (summary data)	Fair/good	Good
4. Accuracy and consistency	Fair	Poor/fair	Fair/good	Good	Fair	Fair/good	Good
5. Flexibility							
a. Data processing	Fair	Fair	Good	Poor/fair	Good	Good	Poor/fair
b. Integration	Yes	Yes	No	No	No	No	No

[a] Evaluation criteria 3–5a were graded on a three-point scale: Poor, Fair, Good.

[b] The evaluation is based on the assumption that the state accident and roadway inventory data are integrated in the data base.

because of the massive amounts of data, but the process itself is fairly straightforward. Merging with other data bases is achieved through a matching process that is based on location or other identifiers, such as vehicle or driver license numbers.

FARS contains a census of all fatal traffic accidents in the United States, which is approximately 45,000 accidents per year. The file size is large enough for most analyses but not for analyses involving rare events (e.g., accidents involving crash cushions) or great specificity (e.g., vehicles of certain year, make, and model). The level of detail is similar to that available from police-level accident data. The key advantages of the FARS data base over state accident data bases are the improved accuracy and consistency of the data and the use of standardized coding formats for all states. The FARS data base is fairly simple to process and use. However, it is not possible to merge the FARS data base with any other data base because all identifiers are deleted from the data.

NASS CSS data contain a statistical sample of reported accidents, designed to provide national estimates of accident statistics. Accidents are selected by a method of disproportionate probabilities from 50 localities, known as primary sampling units (PSUs), within the continental United States. Note that not all of these 48 states are included in the sample, but the PSUs are scattered across the country. About 9,000 accidents were sampled each year in the Continuous Sampling Subsystem (CSS) during 1982–1984, and a larger sample size of 12,000–13,000 accidents was collected in later years. The sample size is adequate for making general national estimates but may become insufficient when greater specificity is needed.

The level of detail available on accident data elements is extensive but fairly limited for highway-related data elements. The data are quality controlled exhaustively for good accuracy and consistency. The data base is somewhat difficult to learn and understand for users who are not familiar with the NASS program because of the complexity of the program and the voluminous documentation.

There is also a problem with incompatibility between data from the early years (1979–1981) for some of the data elements, a problem that requires extensive recoding and reformatting to merge between years. This problem of incompatibility between years has been resolved for the years 1982–1984, although a major revision was implemented in 1985. It is not possible to merge the NASS data base with any other data base because all identifiers have been eliminated from the data.

CONCEPTUAL ALTERNATIVES

Numerous conceptual alternatives to improve and enhance the capabilities and utilities of the existing general purpose data bases were considered. Some were rejected out of hand for being infeasible or impractical and others for not being cost effective. Brief discussions on some of the alternatives that were rejected will be presented in this section.

The ideal alternative is to have a single data collection system that would satisfy all the information needs for highway safety analysis. By modifying and integrating the four existing general-purpose data bases, a single data base could, in theory, be created for this purpose. The individual components would serve different functions but would complement each other, so that data needs for various safety analyses might be satisfied by using one or more of the individual components. This alternative would require major, fundamental changes to the various data collection systems, both technically and administratively. In theory, this might be a feasible and even desirable concept, but it is obviously not a practical alternative.

The FARS and NASS data bases are accident based and lack the capability of merging with other data bases. It is not possible to modify these two data bases to a location-based system without totally redesigning the data collection systems. Any improvements or enhancements short of major redesign and restructuring of the data collection systems would not extend the applicability of these two data bases beyond that of an accident-based analysis. Therefore, no alternative was considered for the FARS or NASS data bases. It follows that any alternative to enhance and improve existing data bases to meet the needs of highway agencies would evolve from the HPMS and state accident data bases.

The HPMS data base is location based, and the panels or highway sections are sampled on a statistical basis to provide national representativeness. The data base has a large enough sample size to handle most of the highway safety analyses of interest and the capability of being merged with other data bases, albeit indirectly. However, the HPMS data base also has some major shortcomings:

- Only summary accident data are provided; and
- The roadway and operational data elements apply to the entire sample panel, which varies in length, resulting in a lack of specificity for some of the data elements.

The lack of detailed information on accident data could be remedied by merging the state accident data files with the HPMS data base through a location-matching process. With an appropriate system design, the merged data base could provide for both location- and accident-based analyses.

One alternative is to keep the current HPMS data base unchanged. Accident records would be matched with the HPMS panels through a location-matching process on an individual state basis. No modification would be made on the accident records. This alternative would be easy to implement with a minimum of effort, provided that the states already have the capability to merge their accident files with the HPMS files. Unfortunately, this is not currently the case, which means that a merging process would have to be developed for those states that do not have this capability.

If it is assumed that such a merging capability is developed, the users would have to extract the required data from the merged data base to create an analysis file suitable for use with the intended analysis. The burden of converting the merged data base into a usable analysis file would be borne by the users, a factor that would probably discourage the use of this data base. The level of detail would be limited to that available from the HPMS data base and police-level accident data, and anyone attempting an analysis that required greater detail or specificity would have to resort to special data files created for that purpose.

Another alternative is to create a safety analysis subsystem within the HPMS data base. Appropriate modifications would be made to the HPMS and state accident data collection systems to create a data base that would be suitable for safety analysis. This would involve changes to the HPMS data elements and the state accident data records. The users would be provided with a single safety analysis data base for analytical purposes, with a minimum of data manipulation required. The level of detail is again limited to that available from the HPMS data base and police-level accident data, so special purpose data files would have to be created for any analysis that required more details or greater specificity.

The major obstacle to setting up a single accident data base is the wide variation among the states in their accident report forms and reporting thresholds. The accident report forms vary among the states in terms of available data elements, format, definitions, and coding levels. The effort that would be required to merge all the state accident data bases into a single standardized format is monumental. In some past and ongoing research studies, several state accident data bases were merged into a single data base for the purpose of analysis. The process was very tedious and time consuming. Moreover, the level of detail on the data elements for the merged data base is usually reduced to the lowest level available from the individual states.

The use of varying reporting thresholds among the states is even more problematic. The reporting thresholds used in most states are based on certain levels of property damage (e.g., $250 or above), towed vehicles, or injury to one or more of the involved occupants. However, the reporting thresholds vary among the states; a given accident might be reported in some states but not in others. It is clear that such differences in reporting thresholds would introduce unknown biases into the data base.

The ideal solution would be to have the states standardize their accident record systems by using a standardized accident reporting form and the same reporting threshold, thus eliminating the problem totally. A lot of effort has been devoted to this goal through the years, with only limited success, and there is no reason to believe that the situation will improve in the foreseeable future.

The problem with the lack of uniformity and compatibility in the accident reporting form could be minimized by using an approach similar to that of the FARS system. A subsample of accidents (e.g., a total of 250,000 accidents) occurring within the HPMS panels could be selected on the basis of a statistically representative scheme. These sample accidents would be recoded onto a standardized form and entered into a single data base. Supplemental information on accident, roadway, traffic, and other data elements could be added to the data base. More rigorous quality control checks could be instituted to improve the accuracy and consistency of the data. In short, the data quality would be improved to that of the enhanced police level.

The extra human resources required could be provided through contracts with appropriate state agencies. State personnel, paid by the contracts, would be used for the coordination, recoding, collection of supplemental data, quality control, and data entry functions. The data would then be compiled at a centralized location to create the data base. The number of accidents included in the data base would be a function of funding available.

The problem with differences in reporting thresholds among the states is much more difficult to resolve. Any changes in the reporting thresholds would probably require legislation at the state level. Also, there are variations even within a state: for example, some major metropolitan areas have adopted the policy of reporting only injury and fatal accidents, leaving the reporting of property damage–only accidents to self-reporting by drivers.

Another drawback is the lack of specificity for some roadway data elements. This problem occurs because the data elements apply to entire HPMS panels, which vary in length up to several miles. This problem could possibly be alleviated by merging state roadway inventory files into the HPMS data base or by subdividing the HPMS panels into shorter sections of equal length. Either approach would require considerable effort.

Integrated state accident data bases can be used for both accident- and location-based analyses through a location-matching process. In theory, a national data base could be created by merging across the states. This data base would be applicable to all components of highway safety analysis, as discussed previously. As a census of all reported accidents in the nation, the sample size would be enormous.

The problem of lack of uniformity and compatibility among the states in their accident and inventory record systems makes the merging of state accident data files into a single data base a monumental task. Also, the data base would be too large for most applications. Extensive computer facilities would be required, and the associated data processing costs would be very high because of the large number of records. This would not be a viable alternative, and thus it is not considered any further.

RECOMMENDED ALTERNATIVE

The alternative that was found to be most cost effective and that is thus recommended is to select a small number of states (e.g., four) with integrated accident data bases and merge them into a single data base. The states would be selected on the following criteria:

- Geographical representation;
- Existing ability to integrate the accident data files with the roadway inventory, traffic, and other pertinent data files; and
- Compatibility among the states in their accident and roadway inventory record systems.

Geographical representation is the major concern with this approach, due to the possible lack of credibility because the combined data base does not provide true national representation. This concern can be partially alleviated by dividing the nation into four regions and then selecting one state from each region. Even so, the extrapolation of the analysis results to states other than those included in the data base is inferential and not statistical in nature. Caution should be exercised in interpreting the analysis results, especially if the analysis of interest is susceptible to the influence of regional characteristics, such as weather and terrain conditions, design standards, ages of the facilities, and so forth.

The ability to integrate accident record systems is a requirement of the candidate state data bases. A number of states have established or are in the process of developing integrated accident record systems; examples include Michigan, Montana, New York, North Carolina, Ohio, Texas, Utah, and Washington. This list is by no means all-inclusive, and the degree of integration varies among the states. Of course, it would be desirable to select systems with as high a degree of integration as possible.

Compatibility in terms of the data elements and reporting thresholds among the selected states is important in order to minimize the effort required to merge the state data files into a single data base. It is recognized that true compatibility is not attainable at this time. However, because the number of states involved is small, a high degree of compatibility could be achieved. Also, there is a better chance of improving the degree of compatibility by working closely with the states and possibly by staging demonstration projects.

Another consideration is the relative size of the individual state data bases. It is conceivable that the size of one state, in terms of number of highway miles and accidents, could be so great relative to the other states in the data base that the analysis results would be heavily biased toward that state. This problem could be minimized by selecting states with roughly equal proportions. On the other hand, it may be desirable to have each state in proportion to the relative size of the region in which it is located so that the data base would be more representative on a national basis.

The level of detail of the accident data is limited to police report level, while the level of detail on the roadway and traffic data elements is restricted to what is available from the state roadway inventory files. Such levels of detail are usually adequate for analysis of a general nature but not for specific applications.

It may be desirable to supplement the roadway inventory data with a limited number of data elements that are deemed essential but are unavailable from the existing inventory files in one or more of the states. However, use of this option should be kept to a minimum and should preferably be performed on an ad hoc basis. There is always a tendency to try to satisfy the information needs of as many users as possible. This effort could result in the inclusion of too many data elements that are only used infrequently, and the cost for the data collection and processing could be increased substantially without a corresponding increase in the benefits.

It is probably more cost effective to maintain in the data base on a continuous basis only those data elements that are most essential or those that are currently available from the state data files. Then, special studies could be conducted for particular applications that require more details than are available from the continuous data base.

Data quality would be subject to the same problems as in the individual state data bases because no additional quality control checks are incorporated. One way to improve the data quality to that of enhanced police-level data is by recoding the data and instituting additional quality control checks. The costs associated with this enhancement are substantial and probably not cost effective.

The privacy and accessibility of the data base is another area of concern. Tort liability has been a growing concern among state governments in recent years. It may be difficult to enlist the cooperation and assistance of the states without considerations for the privacy and accessibility of the data. It would be ideal if the data base could be protected under some form of legislation and be restricted to research and development applications only. Otherwise, safeguards against requests by attorneys or nongovernmental agencies through subpoenas or the Freedom of Information Act should be developed to maintain data privacy. The safeguards should also be extended to limit the accessibility of the data base to authorized users only. At a minimum, the data must be sanitized to remove identifiers relating to individual accidents.

A rigorous validation procedure is needed to double-check any analysis results derived from the data base. As an example, suppose that a predictive model is developed by using a data base that includes data from four states. The model should be applied to each of the four states individually to check whether the results are consistent across the states. It would also be desirable to apply the model to one or two states that were not included in the data base as an external check. If the results were consistent across all of the states tested, the results would probably be applicable to other states. However, if there were wide variations in the results among the states, it would be necessary to reanalyze the data to determine the cause(s) of such variations and to develop appropriate adjustment factors to account for the differences.

In other words, analysis results derived from the data base should be validated by applying the results to the individual states included in the data base and, if possible, to other states outside of the data base. This would also reduce biases that had been introduced into the results by the unequal sizes of the states in the data base. This validation process is relatively straightforward and inexpensive because it involves only the application, not the development, of the analysis results.

Proper administration of the data base is essential to its success. The data base should be administered through a single agency, be it the FHWA or a contractor to the FHWA. The data base manager has to have intimate knowledge of the individual state data bases and must constantly keep abreast of any new developments in the states. The manager should also have a good understanding of the requirements for highway safety analysis so that he or she can assist the users in proper analysis of the data base. This user interface is critical to the acceptance of the data base by the users.

This alternative is relatively inexpensive and provides a reasonably good data base for safety analyses. The startup cost consists of identifying and selecting the state integrated accident record systems for inclusion in the data base and acquiring the data bases from the selected states. Because the individual state accident data bases are already integrated, it is only necessary to develop the required software to merge the individual data files into a single data base. A startup cost of $250,000 (1988) should be sufficient for the purpose.

The annual operating cost, which includes the acquisition of the state accident data bases and update of the software for merging the files, is correspondingly low. An annual budget of $200,000 (1988) should be sufficient. This estimate does not include any reporting requirements or applications.

SUMMARY

Existing data bases were identified and grouped into seven categories according to a categorization scheme. Seven data bases, one for each category, were then selected for study. The emphasis of the evaluation is on the four general-purpose data bases that are national in scope. The three special-purpose data bases are included to illustrate the use of special studies to address specific questions that cannot be answered with the general-purpose data bases.

The various components of highway safety analysis, as used in this study, are defined and characterized. The applicability and utility of these selected data bases for use in highway safety analysis were evaluated on the basis of available documentation. The three national general-purpose data bases (HPMS, FARS, and NASS CSS) and the special-purpose data bases are rather limited in their applications. State accident data files, when integrated with roadway inventory data, are the only data bases that are applicable to all components of highway safety analysis. The utility of the data bases varies somewhat among the general-purpose data bases, but none of them had any major problems. In devising conceptual alternatives, the emphasis is thus on improving the applicability (instead of the utility) of the data bases.

Conceptual alternatives that would improve or enhance the capabilities and utility of the existing general-purpose data bases to better serve the information needs of highway safety analysis were developed. These identified alternatives were studied and analyzed for their feasibility and practicality, and appropriate recommendations were made.

The recommended alternative involves the selection of a small number of states (e.g., four) with existing integrated accident data bases and merging these data bases into a single data base. Geographical representation is approximated by dividing the nation into four regions and then selecting one state from each region. Because the individual state accident data bases are already integrated, it is only necessary to develop the software required to merge the individual files into a single data base. A startup cost of $250,000 (1988) is estimated. The annual operating cost is estimated at $200,000 (1988), including the acquisition of the state accident data bases and update of the software for merging the files.

The recommended data base would provide, at a reasonable cost, the needed information for most highway safety analyses of a general nature from the standpoint of FHWA. It should be recognized, however, that the data base also has its limitations. Some analyses are better addressed with specially designed studies, and others are simply not answerable with accident analysis.

EPILOGUE

A functional prototype of a merged multistate integrated accident data base, the recommended alternative of this study, is being pursued in a current FHWA research contract (2). The findings of this effort will be used to determine the feasibility of developing such an operating system.

The FHWA decided to delete accident data from the HPMS system, partially on the basis of the results of this study. In an unrelated event, the National Highway Traffic Safety Administration decided to restructure the NASS CSS effort during 1988. This restructured NASS CSS program is limited to data on passenger vehicle and light van crashworthiness. The applicability of the resulting data file to highway safety analyses is much less than the former NASS CSS system. These two events increased the importance of the development of an alternate source for national-level highway safety data. The system recommended in this paper is one such alternative.

ACKNOWLEDGMENTS

The work reported in this paper was sponsored by the Federal Highway Administration, U.S. Department of Transportation.

REFERENCES

1. K. K. Mak and L. I. Griffin III. *Assessment of Existing Data Bases for Highway Safety Analysis.* Report FHWA/RD-85/117, FHWA, U.S. Department of Transportation, Nov. 1985.
2. *Highway Safety Information System.* FHWA Contract DTFH61-87-C-00009, Highway Safety Research Center, University of North Carolina, Chapel Hill, 1988.

The opinions, findings, and conclusions presented in this paper are those of the authors and do not necessarily reflect the official views or policies of the Federal Highway Administration.

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