Evaluation Framework for the Creation and Analysis of Integrated Spatially-Referenced Driver-Crash Databases

Final Report
April 2009

IOWA STATE UNIVERSITY
Institute for Transportation

Sponsored by
University Transportation Centers Program,
U.S. Department of Transportation
(MTC Project 2007-13)

About the MTC

The mission of the University Transportation Centers (UTC) program is to advance U.S. technology and expertise in the many disciplines comprising transportation through the mechanisms of education, research, and technology transfer at university-based centers of excellence. The Midwest Transportation Consortium (MTC) is a Tier 1 University Transportation Center that includes Iowa State University, the University of Iowa, and the University of Northern Iowa. Iowa State University, through its Institute for Transportation (InTrans), is the MTC's lead institution.

Disclaimer Notice

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The opinions, findings and conclusions expressed in this publication are those of the authors and not necessarily those of the sponsors.

The sponsors assume no liability for the contents or use of the information contained in this document. This report does not constitute a standard, specification, or regulation.

The sponsors do not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Non-discrimination Statement

Iowa State University does not discriminate on the basis of race, color, age, religion, national origin, sexual orientation, gender identity, sex, marital status, disability, or status as a U.S. veteran. Inquiries can be directed to the Director of Equal Opportunity and Diversity, (515) 294-7612.

Technical Report Documentation Page

1. Report No. MTC Project 2007-13	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Evaluation Framework for the Creation and Analysis of Integrated Spatially-Referenced Driver-Crash Databases		5. Report Date April 2009 6. Performing Organization Code	
7. Author(s) Tim Strauss and Lucas Geadelmann		8. Performing Organization Report No.	
9. Performing Organization Name and Address Midwest Transportation Consortium 2711 South Loop Drive, Suite 4700 Ames, IA 50010-8664		10. Work Unit No. (TRAIS) 11. Contract or Grant No.	
12. Sponsoring Organization Name and Address Midwest Transportation Consortium 2711 South Loop Drive, Suite 4700 Ames, IA 50010-8664		13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code	
15. Supplementary Notes Visit www.intrans.iastate.edu for color PDF files of this and other research projects.			

16. Abstract

This project examines the potential of, and constraints on, the integration of transportation safety databases in a spatially-referenced geographic information systems (GIS) environment. The project focuses specifically on the analysis of crash records and driver records. The objective of such database integration is to facilitate the improved analysis of issues related to transportation safety. The objective of such database integration is to facilitate improved analyses of issues related to transportation safety.

This report begins with a review of existing databases. These include traffic records databases and other databases that may enhance safety analyses. The report then outlines several existing applications, which are typically in a non-spatial framework, of integrated databases related to transportation safety. The report then focuses on existing and potential applications that make use of spatially-referenced information within these databases.

Despite the promise of spatially-referenced data integration and analysis, there are several concerns to address. Possible constraints on spatially-referenced database integration and analysis involve technical issues related to the integration of spatial data files and the data's representation in GIS; methodological issues concerning the use of spatial data; administrative concerns regarding data collection, management, and linkage; and legal and ethical concerns related to the use of confidential information. This report attempts to identify strategies to address these concerns and concludes with recommendations for future activities related to the integration of spatially-referenced data for transportation safety.

17. Key Words	18. Distribution Statement		
data integration—GIS—spatial analysis—transportation safety information system	No restrictions.		
19. Security Classification (of this report)	**20. Security Classification (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified.	Unclassified.	55	NA

EVALUATION FRAMEWORK FOR THE CREATION AND ANALYSIS OF INTEGRATED SPATIALLY-REFERENCED DRIVER-CRASH DATABASES

Final Report
March 2009

Principal Investigator
Tim Strauss
Associate Professor
Department of Geography, University of Northern Iowa

Research Assistant
Lucas Geadelmann

Authors
Tim Strauss and Lucas Geadelmann

Sponsored by
the Midwest Transportation Consortium
a U.S. DOT Tier 1 University Transportation Center
(MTC Project 2007-13)

A report from
Midwest Transportation Consortium
Iowa State University
2711 South Loop Drive, Suite 4700
Ames, IA 50010-8664
Phone: 515-294-8103
Fax: 515-294-0467
www.intrans.iastate.edu/mtc

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The authors would like to thank the Midwest Transportation Consortium for sponsoring this research. The authors also wish to thank several individuals for their insights and information, including Robert Thompson, Governor's Traffic Safety Bureau; Scott Falb, Iowa Department of Transportation; Richard Pain, Transportation Research Board; David Harkey, University of North Carolina Highway Safety Research Center; Donna Johnson and Suning Cao, Iowa Department of Public Health; Anita Gordon, University of Northern Iowa Institutional Review Board; Ben Metz, SMART Public Safety Software; Kevin Blanshan, Iowa Northland Regional Council of Governments; and Mohammad Elahi, City of Waterloo, Iowa.

EXECUTIVE SUMMARY

This project examines the potential of, and constraints on, the integration of transportation safety databases in a spatially-referenced geographic information systems (GIS) environment. The project focuses specifically on the analysis of crash records and driver records. The objective of such database integration is to facilitate the improved analysis of issues related to transportation safety. Linking spatially-referenced GIS safety data may yield better information on crash types and frequencies and on the locations associated with drivers and crashes, and the aim of such a system is to better focus countermeasures.

This report begins with a review of existing databases related to transportation safety. These include traffic records databases as well as other databases that may enhance analyses of safety. The report also illustrates that data linkage can be, and has been, used to address safety. For instance, the analysis of problem drivers, those responsible for a disproportionate number of crashes, can be improved in several ways by linking driver records with crash records. One instance may include first identifying the degree to which a small percentage of drivers may be involved in a large percentage of crashes and then analyzing the characteristics of such drivers. With the integration of spatially-referenced information, the locations of these drivers' crashes, along with spatial variations regarding the prevalence of the problem-driver issue, can be evaluated. In addition, the home addresses of drivers can be located, and the relationship between home residences and crash locations can be assessed. Driver citation histories and the relationship between these histories and subsequent crash types and locations can also be analyzed. This report outlines several potential applications of integrated spatially-referenced databases related to transportation safety.

Despite the potential of spatially-referenced data integration and analysis, there are several concerns to address. Possible constraints on spatially-referenced database integration and analysis involve technical issues related to the integration of spatial data files and the data's representation in GIS; methodological issues concerning the use of spatial data; administrative concerns regarding data collection, management, and linkage; and legal and ethical concerns related to the use of confidential information. This report attempts to identify strategies to address these concerns and concludes with recommendations for future activities related to the integration of spatially-referenced data for transportation safety.

INTRODUCTION: PROBLEM DESCRIPTION

Crashes result from a variety of causes and have a range of outcomes. To devise appropriate countermeasures related to engineering, enforcement, public policy, education, and other strategies, it is important to understand these causes and outcomes. Information related to crashes and their characteristics is collected in a variety of ways, for a variety of purposes, in several databases that are often separately developed and managed. This may result in transportation safety analyses that are not as comprehensive or informative as they could be. Key relationships and causal factors spanning across databases may go undiscovered. Furthermore, these databases contain attributes that are, or could be, spatially referenced, such as the locations of crashes, drivers, vehicles, medical facilities, and roadways. Spatially-referenced data are useful for displaying spatial patterns, integrating data from different sources, and generating new research questions and the means to address such questions to a degree not possible without such data. Other data typically developed for other purposes such as demographic and land use data can also be used to explore causal relationships.

Project Overview

This project examines the potential usefulness of, and possible constraints on, the integration of databases related to transportation safety in a spatially-referenced geographic information systems (GIS) environment. Linking spatially-referenced transportation safety databases may yield better information on crash types, frequencies, causes, and outcomes, with the ultimate aim of designing more focused countermeasures. For instance, such data linkage could be used to address the issue of problem drivers (those responsible for a disproportionate number of crashes). Linking driver records with crash records has the potential to improve analysis in several ways. For instance, the degree to which a small percentage of drivers may be involved in a large percentage of crashes can be identified, and the characteristics of such drivers can be analyzed. In addition, with spatially-referenced data, the locations of these drivers' crashes and spatial variations in the importance of the problem driver issue can be addressed. Moreover, the home addresses of drivers involved in crashes can be located, and the relationship between home residences, crash location patterns, and types of crashes can be evaluated. In addition, driver citation histories and their relationship to subsequent crash types and locations can be analyzed. Several such existing and potential applications of integrated spatially-referenced safety databases are discussed in this report.

Organization of the Report

Components of Transportation Safety Databases

This report first outlines the components of databases typically included in transportation safety information systems. Attention will be given to the attribute data contained in these databases; data that are, or can be, spatially referenced; and data that can be used to link these databases to each other. Other databases that can be used to support analyses of transportation safety are also considered.

Existing Uses of Linked Databases

The report then reviews existing transportation safety literature that makes use of linked databases. Much of the surveyed literature uses integrated crash and driver records to analyze the characteristics of drivers as these characteristics relate to the drivers' tendencies to be involved in crashes. Research efforts that make use of other forms of data integration, such as crash-injury linkage and crash-citation linkage, are also discussed. This review first considers efforts done largely in a non-spatial context. It then outlines key aspects of spatial data analysis and research that explicitly uses spatially-referenced data, and highlights examples of such work.

Potential Concerns about Spatially-Referenced Data

Despite the potential of spatially-referenced transportation safety data integration and analysis, there are several concerns to address. Possible constraints to spatially-referenced database integration and analysis involve technical issues related to the integration of spatial data files and these files' representation in GIS; methodological issues concerning the use of spatial data; administrative concerns regarding data collection, management, and linkage; and legal and ethical concerns related to the use of confidential information. Although such issues are not unique to spatially-referenced data, the use of such data creates unique concerns. This project attempts to identify strategies to address these concerns.

Related Research Effort

A research effort similar to this one in its motivation, if not its application area, is a project completed several years ago on the linking of driver license data to employment data in order to generate individual-level origin (home address) and destination (work address) data to support travel demand models and other transportation planning applications, such as welfare-to-work analysis and bypass analysis (Souleyrette et al. 1998). The two databases used for the project were already being maintained for their own purposes; linking them together created new and practical uses at comparatively little cost.

Project Objectives

The current study addresses the potential applications of integrating crash-related databases, emphasizing the use of spatially-referenced data elements. Given the number and variety of databases involved, the opportunities and complications related to spatially-referenced data, and the technical and administrative constraints to be addressed, this study focuses on exploring possible applications to evaluate the databases' potential. The overall intent of the current project, therefore, is to identify uses of integrated spatially-referenced databases for transportation safety and to address potential problems and constraints. The report concludes with recommendations for future activities related to transportation safety data integration.

OVERVIEW: TRANSPORTATION SAFETY INFORMATION SYSTEM COMPONENTS

Reliable, accurate, and comprehensive information is needed to address issues related to transportation safety. The National Safety Council has stressed the importance of "highway safety information systems" to provide information that is "critical to the development of policies, and programs that maintain the safety and the operation of the nation's roadway transportation network" (National Safety Council 1997, p. 1). Motor vehicle crashes have a variety of characteristics, causes, and outcomes, and information related to them is collected in several forms. The National Highway Traffic Safety Administration (NHTSA 2006, p. 10) lists the following six main databases related to motor vehicle crashes:

- Crash information
- Roadway information
- Driver information
- Vehicle information
- Citation/adjudication information
- Statewide injury surveillance information

Integrated data are needed to address safety issues, particularly to analyze crash causes and countermeasures. Another NHTSA report, on improving traffic safety data, highlights the following point:

> The effectiveness of informed decision making at the national, State and local levels, involving sound research, programs and policies, is directly dependent on data availability and quality. Without accurate and comprehensive data, it is not possible to determine causation or to develop countermeasures that will prevent crashes or mitigate the injury consequences of the crashes that do occur. (NHTSA 2004, p. 6)

DeLucia and Scopatz (2005) also point to the need for comprehensive transportation safety data to support sound decision making:

> An increasing emphasis on traffic records is not without justification. It has become apparent over time that appropriate, accurate, and timely information describing various aspects of the transportation system (including its crash experience) are needed to improve traffic safety and mobility…. To manage its safety programs effectively, each state needs to analyze an increasingly wide variety of information about the design characteristics of its road system, the behavior of traffic on that system, and the crash experiences of its users. (p. 5)

The authors further point out that the complex nature of crash causation requires linked data across several areas. "The underlying realization that environment, vehicles, and human factors all play a role in crash frequency and severity points directly to a need for data systems that can

link these information sources" (DeLucia and Scopatz 2005, p. 25). Conceptually, although safety-related databases are often collected, managed, and maintained by separate entities, they can be seen as part of one larger database. For instance, according to NHTSA (2006), a traffic records system (TRS)

> has been defined as a virtual set of independent real systems (e.g., driver conviction records, crash records, roadway data, etc.), which collectively form the information base for the management of the highway and traffic safety activities of a state and its local subdivisions. A more modern concept of a TRS encourages states to take a global approach and work toward compiling data into a unified, accessible resource that meets the needs for safety information. Sharing and integrating data makes such a system possible, without necessarily duplicating costly and time-consuming tasks such as data entry. Achieving integrated access to data without bringing all the data into a single database is a goal of the TRS. (p. 2)

Similarly, DuLucia and Scoptaz (2005) recommend the development of a "knowledge base to serve the highway and traffic safety community" (p. 3). This proposal is reinforced by Council and Harkey (2006), who, in their scanning study of traffic safety information systems, have recommended that all state departments of transportation (DOTs) develop data warehouses that would include both safety data and critical non-safety data. There is also increasing recognition of the importance of spatially-referenced data. Council and Harkey noted the need to develop common location referencing systems to facilitate data integration and analysis. They also highlighted the need to ensure that

> (1) all data in the warehouse are compatible with spatial referencing (e.g. GIS) because all data users are moving toward such systems, and that (2) spatial data expertise is included in the knowledge base. (p. 33)

The authors also recommended that agencies "Move as rapidly as possible to a geospatial reference (e.g., GIS) system for all types of safety data" (Council and Harkey 2006, p. 35).

The key to such an approach to linking databases is to make the data available, from whatever files and in whatever combinations, to users who are addressing transportation safety issues. NHTSA (1998) provides a diagram of such a "distributed model" of data management, which shows the basic elements of each database, as well as key aspects of linkages between them (Figure 1).

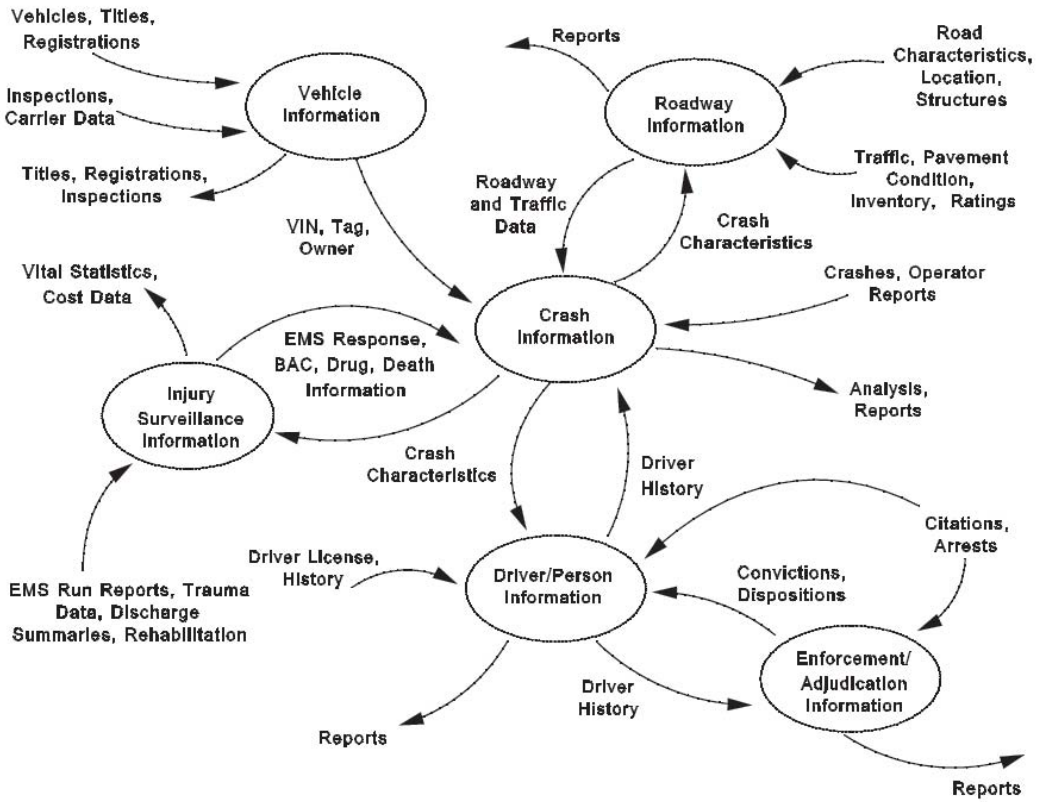

Figure 1. Model of distributed data processing in a traffic records system (NHTSA 1998)

Each of the six components of a traffic records system, listed above, is briefly discussed in the next several sections. Much of this discussion is based on the *Traffic Records Program Assessment Advisory*, which provides guidance to state agencies and "serves as a description of an ideal system—one that supports high-quality decisions that lead to cost-effective improvements in highway and traffic safety" (NHTSA 2006, p. 1). For each component, there are several important types of data elements. Most of the fields, or variables, in the database contain attribute data, i.e., characteristics of the crashes, drivers, vehicles, etc., depending on the database. In addition, there are several "key fields" or common identifiers that can be used to link to records from other databases. These key fields include such data elements as driver license number, vehicle identification number, and crash identification number, which may appear in more than one database. These can be used for "attribute joins," or the linkage of records in one database to corresponding records in another database based on common values for these key fields.

Finally, there are fields containing spatially-referenced data, or data that could be spatially referenced. These fields can include geo-coded data, i.e., data with coordinates indicating the spatial location of objects that can be mapped in a geographic information system. Examples include the point locations, expressed as X and Y coordinates, such as longitude and latitude, of crashes and the coordinates of road segments (which can be expressed as a series of ordered coordinates). Other fields that can be spatially referenced include data elements such as the addresses of drivers' residences, addresses of hospitals, addresses of vehicle owners, etc. In non-spatial databases, such information is not normally given coordinates, or geo-coded, for display,

integration, and analysis in a GIS. However, such information can be generated through procedures like address matching, which combines the text address with a "matchable" road layer in GIS that contains information on street name, address range, and area (zip code, city, county, state, etc.) for each road segment. The result is a file containing a set of coordinates that correspond to the location of each record in the non-spatial database, which can be used to map the locations of the records in GIS. Other spatially-referenced data include city, county, state, and zip code. These data provide information on the area in which the record (e.g., driver residence) is located, but not the exact point. However, this information can also be mapped in GIS, either for individual records or aggregated across groups of records located in the same area.

Just as attributes can be used to link records based on common values for key fields, spatial data can be used in "spatial joins" to link records in different databases based on location. For instance, data elements for the nearest roadway segment can be linked to each crash point in a database to incorporate roadway data (e.g., traffic volume, shoulder width) into analyses of crashes. Conversely, data from crashes can be linked to the nearest roadway segment. This can facilitate analyses of crash characteristics along roadways:

> Linkage with location-based information such as roadway inventory databases and traffic volume databases at the state level can help identify the kinds of roadway features that experience problems, allowing states to better address these needs through their various maintenance and capital improvement programs (NHTSA 2006, p. 7).

Similarly, a spatial join can be used to attach to a hospital, represented as a point, data regarding the crashes that are closer to that hospital than to any other hospital, e.g., to evaluate safety-related demand for hospital services. In addition, "point-in-polygon" spatial joins can be used to link to crash points information associated with the areas (e.g., zip code, traffic analysis zone) in which each point is located.

Crash Data

Crash databases include information on the time and location of crashes, the characteristics of the drivers involved, vehicles involved, injuries, and the circumstances of the crash (NHTSA 2006). Information is typically collected by law enforcement or by the crash participants and sent to a central location, e.g., an entity within the state DOT. Crash characteristics include information on the crashes themselves, as well as the vehicles and persons involved (e.g., drivers, passengers, pedestrians), and roadways on which the incidents took place. These characteristics include data on vehicle types (e.g., passenger car, van, motorcycle), collision types (e.g., head-on, rear-end, sideswipe), vehicle actions (e.g., turning right or left, passing, changing lanes), harmful events (e.g., rollover, collision with vehicle or other non-fixed object, collision with structure or other fixed object), roadway type (e.g., bridge, intersection type), traffic controls (e.g., traffic signals, stop signs), light conditions (e.g., day, dusk, dark, dawn), driver conditions (e.g., fatigue, alcohol), surface conditions (e.g., dry, wet, snow), weather conditions (e.g., clear, rain, wind), obscured vision (e.g., by parked vehicles, trees), and contributing circumstances (e.g., failure to yield right of way, distracted driving, excessive

speed, running through signals). In addition, information on injuries and fatalities is recorded (e.g., severity, seat belt use, air bag deployment, position in vehicle).

The Model Minimum Uniform Crash Criteria (MMUCC) guidelines list certain crash, vehicle, and person data elements as being collected directly at the scene of a crash, while others are derived from the collected data (such as age from date of birth or day of week from crash date), with still other data elements being incorporated through links with other data (NHTSA 2003). For instance, crash data collected at the scene contain only limited information on the roadway involved and typically do not include data on traffic volume, geometrics, and other important data elements. In addition to roadway information, data from other databases can be linked to crash data to yield a more comprehensive picture. For instance, information on injury severity in crash records is often based on preliminary assessments by law enforcement personnel in the field. Linkage to other records, e.g., from injury surveillance system data, can generate more complete information. Links to driver histories can produce information regarding variations in the tendencies of drivers to be involved in crashes or certain types of crashes, and links to vehicle records can facilitate analyses of crashes experienced by specific types of vehicles.

As discussed above, there are two main methods to link databases, attribute joins using common data elements (i.e., key fields) and spatial joins based on the locations of objects (e.g., crash locations and roadway locations) in each of their respective databases. There are several possible ways to link crash data to other databases. The crash itself is commonly given a unique crash identification number. The driver license number (and driver name and address) can be used in an attribute join to link the crash database with driver records, and the vehicle identification number (VIN) can be used to link with vehicle data. Depending on how the databases are structured, incident numbers and personal information (e.g., name, driver license number) may be used to link crash data to enforcement data, and crash data may be linked through probabilistic linkage with injury surveillance data by personal identifiers (if available and allowed), date, time, location, and emergency medical services (EMS) run report number (NHTSA 2006; NHTSA 1996). The location of the crash can be used in a spatial join to link with roadway data using a GIS, provided that appropriate and compatible spatial reference systems are used for both databases.

Crash data elements that are, or could be, spatially referenced include the locations of the crashes. These locations can be recorded in a number of ways, depending on the crash form and data collection procedures, as well as on the location referencing system used. Location referencing systems include simple textual descriptions of the location, street addresses, mile posts, reference points, and point coordinates. Depending on the location referencing systems used, this information can facilitate spatial joins. Other spatially-referenced data in crash reports can include the driver license jurisdiction, the addresses of the drivers, the addresses of the vehicle owners, and the addresses of the persons injured. These data can be used to analyze spatial relationships related to crash patterns.

Driver Data

Driver data include driver-specific information on name, address, date of birth, driver license number, license type and status, driver restrictions (e.g., vision), traffic violations, and crashes

(NHTSA 2006; NHTSA 1996). Key fields for attribute joins to crash data include driver license number and other personal identifiers. Citation data can be linked using citation number and case number, as well as personal data and location. Linkage to injury data can be facilitated by personal identifiers (if allowed) as well as crash date, time, and location (NHTSA 2006; NHTSA 1996). Much of the usefulness of integrating driver data with other data concerns the issue of problem drivers. As a NHTSA (1996) report notes,

> When driver information from the crash data are combined with medical cost and conviction information, this information is useful to assess the societal costs caused by repeat offenders. Linkage of the crash and driver licensing data files provides access to the SSN to facilitate linkage to insurance claims data, such as Medicaid. (p. 8)

In addition, spatially-referenced information includes the address of the driver's place of residence. This information can be geo-coded and displayed in GIS for spatial analysis:

> Driver and vehicle owner addresses are useful for geographic analyses in conjunction with crash and roadway data components. Linkage in these cases should be based on conversions of addresses to location codes and/or geographic coordinates in order to match the location coding method used in the roadway data component and in the GIS (NHTSA 2006, p. 19).

This information can be analyzed together with the locations of crashes, medical facilities, and other crash-related data. The crashes of specific subsets of drivers, e.g., those with extensive crash histories or those involved in certain types of crashes, can also be mapped and analyzed for non-random, causal spatial relationships.

Vehicle Data

Vehicle data include information on ownership and other characteristics of vehicles, such as vehicle make and model and year of manufacture. Additional information is available for commercial vehicles, including carrier information and inspection data (NHTSA 2006; NHTSA 1996). Crash data can be linked to vehicle data by VIN and personal identifiers. Citation data can be linked using personal data and location, and injury data can be linked using personal identifiers (if allowed) and crash date, time, and location (NHTSA 2006). In addition, information in the VIN can be used to access data on restraint system, vehicle weight, and other vehicle characteristics. Thus these data can, by extension, be used with other databases for more comprehensive crash analysis as the analysis relates to vehicle types and other factors (NHTSA 1996). Furthermore, "[l]inked crash, vehicle registration, census, and injury data generate information that relate specific types and characteristics of the vehicle to urban and rural crash patterns and their specific medical and financial consequences" (NHTSA 1996, p. 8). Data that can be spatially referenced include jurisdiction and the place of residence of the owner. As noted above, these can be geo-coded and displayed in GIS.

Citation/Enforcement/Adjudication Data

Citation/enforcement/adjudication data include information on citations, convictions, and sentencing. The data should allow the tracking of a citation through the legal system, from the writing of the citation to its disposition in a court of law. Data elements include type of violation, date, time, enforcement agency, court jurisdiction, and final disposition (NHTSA 2006). In addition, "[s]imilar information for warnings and other motor vehicle incidents that would reflect enforcement activity are also useful for highway safety purposes and should be available at the local level (NHTSA 2006, p. 24). Linkage to other databases can be accomplished using driver license number, personal data, VIN, citation/incident number, and location. If locations are geo-coded, spatial joins can be used to link with roadway data and relate this information to other components of a transportation safety information system, such as crashes.

Roadway Data

Roadway data can include such information as structures, classification, geometrics, pavements, traffic volumes, roadside features, and other characteristics. Such information is critical for the analysis of crashes, and, as noted above, only limited roadway data are collected at the scene of a crash. As Council and Harkey (2006) note, "The accuracy and completeness of both types of data [crash and inventory/traffic volume], as well as the integration of the two, are critical to operating a successful safety program" (p. 5).

Road segments, intersections, and interchanges are commonly given unique identifiers, which can facilitate attribute joins to other databases, such as crash databases. However, spatial joins, using location, often provide a more useful and flexible method of linking roadway data with other files:

> A location reference system should be used to link the various components of roadway information as well as other TRS information sources, especially crash information, for analytical purposes. Compatible location coding methodologies should apply to all roadways, whether state or locally maintained. When using a GIS, translations should be automatic between legacy location codes and geographic coordinates. This process should be well established and documented. Compatible levels of resolution for location coding for crashes and various roadway characteristics should support meaningful analysis of these data. (NHTSA 2006, p. 17)

Several linear referencing systems are available to locate features along roadways, just as there are several methods to locate crashes. Roadway data can also, by extension, be linked to medical cost data associated with injuries occurring along road segments to support decision making regarding road improvement and maintenance (NHTSA 1996).

Roadway data can also be used for explicitly spatial analyses. Network distances and routes between residence location and crash location, or between EMS location and crash location, can

be calculated. Information from travel demand models, such as route assignments generated from origin-destination pairs, could also be integrated into crash analyses.

Injury Surveillance System Data/Medical Data

Medical data include information from a variety of sources related to injuries resulting from motor vehicle crashes. This information includes data from emergency medical services, hospital emergency rooms, hospital stays, discharge from hospitals, long-term care, outpatient services, and death certificates. These data contain more complete information than is recorded at the crash scene, including details related to injuries, medical outcomes, and costs. This information is important for the analysis of transportation safety. As NHTSA (2006) notes,

> This system should allow the documentation of information that tracks magnitude, severity, and types of injuries sustained by persons in motor vehicle–related crashes. Although traffic crashes cause only a portion of the injuries within any population, they often represent one of the more significant causes of injuries in terms of frequency and cost to the community. The [injury surveillance system] should support integration of the injury data with police-reported traffic crashes and make this information available for analysis to support research, public policy, and decision making. (p. 27)

However, linkage to other data sources can be problematic because of a relative lack of common identifiers across databases. Other methods, such as probabilistic matching, may be used in these circumstances (NHTSA 1996).

Spatially-referenced medical data can include the locations of medical facilities of varying levels, including emergency medical services, hospitals, trauma centers, and long-term care facilities. This information could be used to examine relationships between the locations of crashes and the locations of medical facilities. For instance, the distances, times, and routes from emergency medical service facilities to a set of crashes along roadways may facilitate analyses of response times and evaluations of EMS service coverage areas. In addition, this information could be used to assess the impact of response times on medical outcomes. Such analyses would require the use of roadway data (e.g., travel times, distances) in a GIS environment.

Other Data

Other types of data, often not seen as part of transportation safety information systems, can be integrated with crash-related data to support crash analysis. Census data can be used in a GIS environment to analyze the role of spatial variations in demographic factors (e.g., age, gender, socioeconomic status) on crash distributions, as well as to provide information on urban vs. rural contrasts and population density, and to provide a means to normalize and compare crash data across different jurisdictions within a state or across states (NHTSA 1996). In a GIS, such data require the use of boundary files for blocks, block groups, census tracts, zip codes, counties, etc. to support integration, mapping, and analysis.

In addition, insurance data contain information on individual drivers, their places of residence, and claims history. Information from intelligent transportation systems (e.g., from video monitoring or traffic data collection systems) can also be used, although there may be concerns about privacy. Behavioral, attitudinal, and perception data can also be generated, e.g., through surveys, and linked to empirical crash data. Other GIS resources, in addition to census data, can also be used to support crash analyses of various types, including data on land use, vegetation, and topography.

EXISTING AND POTENTIAL APPLICATIONS OF INTEGRATED SPATIALLY-REFERENCED TRANSPORTATION SAFETY DATA

Overview

This section outlines existing research related to the integration of transportation safety databases. The first several subsections focus on key research directions, typically not explicitly involving spatial analysis, using data linkage with crash-related databases, typically via attribute joins. The following subsection focuses on research that explicitly analyzes spatially-referenced data. This subsection outlines types of spatial analysis, explores how existing non-spatial research could be expanded to incorporate spatially-referenced data, and discusses research that uses spatial joins and explicitly analyzes spatially-referenced integrated safety data. This summary of existing research is not meant to be an exhaustive review of the research literature in each area. Instead, it is intended to illustrate the scope of applications that involve integration of transportation safety information system components for the analysis of safety. Research has been conducted in several countries and states, using different databases and different definitions. There has been a particular focus on the problem driver issue, mainly involving driver and crash databases, which provided much of the motivation for this study. Another important research direction is the linkage of crash and injury data. These and other applications are outlined below.

Crash and Driver Data Integration

Perhaps the most common type of data integration in transportation safety information systems involves the combination of driver characteristics and crash characteristics. The motivation behind much of this research is to address the issue of "problem drivers," i.e., drivers with higher tendencies to be involved in crashes. As Cohen and Preston (1968) posed the question, "How can we account for the fact, assuming it is a fact, that some groups of drivers become involved in more than their 'fair share' of road accidents in terms of their numbers or exposure to hazard in any period of time?" (p. 73). Researchers have looked for relationships to help predict drivers' likelihoods of being involved in a crash, particularly in a crash in which they were at fault, as a function of several variables, including involvement in past crashes, receiving citations for such traffic violations as speeding or running signals, risk-taking behavior, and various "psychological antecedents" or personality characteristics.

It is not a new idea that some drivers may be predisposed to be involved in crashes. The concept of "accident proneness" goes back at least to the 1920s (Farmer and Chambers 1926, cited in Cohan and Preston 1968). Tillman and Hobbs (1949) conducted one of the first studies to look at the role of personality in influencing involvement in crashes. The authors introduced the concept that people "drive as they live" (Willett 1964), although this concept has been criticized as an empty truism (Cohen and Preston 1968) and the research stemming from it has been criticized both for statistical and methodological reasons (Evans 2004). Even if there are indeed subsets of drivers who present higher and lower risks, the probability of being involved in a crash during any time period is so low, and the randomness factor is so high, that it becomes difficult to detect clear differences in empirical studies (Evans 2004). However, as Evans (2004) notes,

> dismissing the notion of accident proneness does not mean that individual drivers, or groups of drivers, cannot be reliably identified by other methods as posing greater than average driving risks…. For example, it can be predicted with confidence that an individual driver convicted of many traffic-law violations will have higher future crash risks if permitted to continue to drive, and it can be predicted with near certainty that a group of 20-year-old male drivers will have higher than average crash risks. (pp. 13-14)

In practice, state agencies implement programs to address problem drivers using systems that assign points to individuals based on moving violations and crashes, and these programs apply countermeasures, including license revocation, based on these points.

Efforts to identify and analyze high-risk individuals have focused on previous crash history, citations, risk-taking behavior, and personality factors. Several of these efforts have linked crash data to other databases to analyze the first two items, crash history and citations. In their summary of recent work, Chandraratna et al. (2005) state, "The literature shows that not only previous crash involvements but also accumulations of citations are good predictors of drivers' future crash risk…. many researchers have repeatedly highlighted the importance of past crash and citation records in predicting drivers' potential crash involvement" (p. 1). In their own study, they linked 3,201,620 driver records from the Kentucky Driver License database with records from the Kentucky crash database to study young novice drivers, using data for 1997–2002. "Young novice drivers" were defined as those less than 25 years of age with two years of driving experience. Each driver's two-year "before" experience as a driving novice, in terms of crashes and citations, was then compared to the crash experience in the two subsequent years. The crash database was used because the driver license database did not include detailed crash history data. For multi-driver crashes, the authors created separate records because the unit of analysis was the licensed driver. The authors found that previous crash involvement of these young drivers was positively related to subsequent crash involvement. Drivers who had an at-fault crash were about 150% more likely to be involved in another crash than those who did not, while drivers who had a "not-at-fault crash" were 125% more likely. Speeding violations were also a good predictor, with each speeding conviction increasing by 21% the chance of the driver being in a crash. Increased risk was also associated with male drivers and younger drivers. The authors suggest that the results and methodology can facilitate the identification of high-risk drivers (Chandraratna et al. 2005).

In a later article, Chandraratna et al. (2006) used two versions of the Kentucky Driver License database to obtain driver license number, age, sex, and citations. This was linked, via the driver license number, to the Kentucky crash database to incorporate data on crash characteristics and human factors for a seven-year period. The focus of the study was high-risk drivers, in particular those involved in multiple at-fault crashes, and a crash prediction model was developed to classify at-fault drivers. Drivers with at least one at-fault crash were more likely to be at fault in a subsequent crash, as were those with license suspensions and traffic school referrals. The authors suggest that the model could be used by motor vehicle licensing agencies when renewing driver licenses.

Some research has used linkages to insurance industry data to analyze crash histories. Chen et al. (1995), for instance, conducted a study of traffic infraction history to "assess the relative impact on future crash-involvement risk of a number of different infractions and also of accident history" (p. 9). The study examined nearly two million driver records of British Columbia drivers, linked to their crash histories from the Insurance Corporation of British Columbia Claim Database, over a five-year period. The researchers used logistic regression to predict driver involvement in at-fault crashes during a two-year "after" period, given the drivers' previous history of crashes and traffic violation convictions during a three-year "before" period. The authors found that previous at-fault crashes were a better predictor of subsequent at-fault crashes than were previous traffic violations. Moreover, incidents such as failure to yield right-of-way and disobeying traffic signals were those most closely correlated to later crashes.

In a related study, Cooper (1997) linked driver data over a five-year period with crash-related insurance claim data in British Columbia in order to analyze the relationship between driver violations and crash involvement. Driver fault and crash costs were derived from the insurance claim data. Driver records were matched with police accident files to integrate data on the details of each crash in which each driver had been involved. The study found that non-casualty, low-cost crashes were associated with previous right-of-way and traffic control violations more so than speeding. The number of speeding violations was linked to the number of crashes; the more serious "excessive speed" violations were linked more to subsequent crash severity than were the less serious "exceeding speed limit" violations. "The implication," as Cooper (1997) notes, "is that efforts to reduce speed-related and severe crashes should focus on the excessive speeders—those at the high end of the speed distribution" (p. 94).

Gebers and Peck (2003) point to studies that show that models using traffic convictions often do a better job of predicting crashes than do models using previous crash involvement, largely because of the greater frequency and reliability of traffic convictions. In their own study, the authors used a 1% sample of drivers from the California driver license master file, about 250,000 records, to predict both citation involvement and crash involvement. As explanatory factors, the authors used both variables at the individual level (age, gender, crash history, citation history, etc.) as well as "territorial," i.e., geographic, variables based on the zip code of the driver residence. The territorial variables, used largely to control for geographic variations in crash and citation history, included census data (on race/ethnicity, public assistance, unemployment, age, and income) and mean zip-code level citation and crash values from the driver license file. The results indicated that models to predict citations were more accurate than models to predict crashes, but that citation history was no more successful than crash history in this study in predicting future crashes.

In a study of alcohol use and transportation safety, Rosman et al. (2001) linked driver records in all reported crashes with corresponding records of all arrests for driving while intoxicated over a nine-year period in Western Australia. Data for crash-involved drivers came from the Western Australia Road Injury Database, which has data on all reported crashes. Data on drunk driving arrests came from the Integrated Numerical Offender Identification System (INOIS) database maintained at the University of Western Australia. Probabilistic linkage was used to join the files, using driver license number, date, time, and INOIS number. Comparing drunk-driving crashes with "routine enforcement" drunk-driving arrests, the authors found that both younger

and older drivers were more likely to be involved in crashes, while the Aboriginal population was more likely to be involved in routine enforcement arrests.

In another analysis conducted in Australia, Crettenden and Drummond (1994) studied younger driver crashes using the Victorian Mass Crash Database and the Victorian Driver License Database. The resulting file contained data on individuals age 18–40 involved in casualty crashes over a seven-year period. Data elements included a variety of both crash and driver characteristics, as well as the license number, which was used for matching. The authors found that only a small percentage of drivers were involved in multiple crashes and that this percentage decreased as driver age increased. The authors remarked that involvement in multiple crashes may be a function of exposure rather than risk-taking behavior or similar psychological or personality factors. They concluded that countermeasures should focus more on the "younger driver problem" than on "problem younger drivers."

Some researchers have focused on risk-taking and personality characteristics. Much of this research involves the use of questionnaires, interviews, and observations. These are then related to crash involvement, either through data linkages with driver files or through self-reporting of crash experiences. For instance, Elliott et al. (2000) analyzed young drivers in Michigan using driver license records. Self-administered questionnaires were given to school-aged individuals, from 5th grade through 12th grade, over roughly a ten-year period. Data on these participants were then linked to driver license files from the Michigan Department of State. Information in the file included sex, age, driving offenses, and the number and characteristics of crashes in which the participants were involved. The researchers found that drivers having a serious driving offense in a given year, controlling for other variables, had about double the chance of having a serious offense during the following year, and having an at-fault crash in a given year increased the chances by about 50% of having another such crash in the following year. The authors suggest that "a 'hard core' of high-risk individuals may be identified as young drivers age into their 20s" (Elliott et al. 2000, p. 241).

In a related effort, Shope et al. (2003) followed a set of students in Michigan using a series of self-administered questionnaires over a period of about eight years. These surveys elicited information on such measures as alcohol use, friends' support for drinking, and susceptibility to peer pressure. Upon completion of the survey period, the students' responses were linked to driver license data using their names and birth dates, and information on driving offenses and crashes was used to analyze relationships between driving behavior and the questionnaire responses. The authors found strong relationships, especially for predicting involvement in serious offenses, alcohol-related offenses, and alcohol-related crashes.

Several researchers attempting to identify traits of individuals that experience more crashes than average focus on psychological characteristics. Marottoli et al. (1994) conducted interviews of 283 older (age 72 and over) drivers, and information was generated on the demographic, health, psychosocial, activity, and physical performance characteristics of the interviewees. A follow-up survey was taken a year later to collect information on driving experiences. Specifically, participants reported how often they drove as well as any "adverse events", i.e., if they had been stopped or cited by police, if they were involved in an accident, and whether they had been injured or hospitalized because of an accident. In all, 13% of the participants reported an adverse

event, most of which were crashes. The researchers found several factors associated with adverse events, including chronic health conditions, antidepressants, cognitive impairment, and physical inactivity. However, the authors note that in many cases these statistical relationships were weak. The authors also mention that the self-reporting of crashes in the interview may result in differences from those obtained from official data. Underestimates may result if interviewees forget the incident or are concerned about possible consequences. On the other hand, the method may identify crashes, especially minor incidents, that are not reported, or cases in which a person is stopped by police and given a warning rather than a ticket. Overall, however, more accurate data may be obtained through the integration of driver, citation, and health data.

In a study that linked survey data to official driver data, Iversen and Rundmo (2002) distributed a survey to Norwegian drivers, randomly selected from the driver license register, to examine the relationship between personality, risky driving, and involvement in crashes. In all, 2,605 Norwegian drivers participated. The authors found that personality characteristics (e.g., sensation seeking, driver anger, and "normlessness" or resistance to rules) were related to risky driving (e.g., speeding, ignoring traffic rules), and that risky driving was strongly related to involvement in crashes. The study suggests that

> some of the personality variables could be used to predict an individual's tendency to commit risky driving.... Identification of the groups representing specific personality traits associated with risky driving and accident involvement more than others can help develop more adjusted traffic safety interventions. The results suggest that different intervention strategies may be needed, adjusted for different subgoups of drivers. (Iversen and Rundmo 2002, pp. 1258–9)

In a study that focused on risk-taking behavior, Burns and Wilde (1995) combined in-vehicle observations of taxi drivers in Canada with a questionnaire that was later administered. In addition, participants gave their permission for the authors to view their driving records from the Ontario Ministry of Transportation, which were then compared to their observed habits and their questionnaire results. The study found that drivers who rated high in sensation seeking, as determined by the questionnaire, also took more risks as drivers. A relationship was also found between sensation seeking personalities and driving violations; however, no relationship was found between collision history and either sensation seeking or risk-taking behavior. The authors note, however, that police-reported collision statistics may not be complete, as self-reported collisions tended to exceed those in the driving records. The authors conclude the following:

> Sensation seeking personality could be used to predict an individual's tendency to be a repeat traffic offender because of its significant relationship with observed fast and careless driving, and convictions for both speeding and traffic violations. However, this information may not contribute to the goal of increasing traffic safety because there may not be a relationship between [sensation seeking] motivated speeding and collision risk. (Burns and Wilde 1995, p. 277)

Finally, in a psychological study conducted at a much broader, cross-national scale, Lajunen (2001) notes that the link between personality and traffic crashes is far from clear. Focusing on a standard typology of personality from the discipline of psychology, he states that some previous

research has found, for instance, a relationship between extraversion and a person's likelihood of crashing, while other research finds no relationship. Likewise, some studies link higher levels of neurotic characteristics with risky driving behavior, while other studies find the opposite relationship. To further analyze this issue, he used data from the World Road Statistics for a five-year period, specifically traffic fatality rates per 100,000 vehicles. This was then compared to national-level data, from previous research, on "extraversion," "neuroticism," and "psychoticism." Lajunen then compared a set of 34 countries, acknowledging the limitations of macro-scale, cross-national analysis. The results statistically linked extraversion to traffic crashes and fatalities, but the relationship with "neuroticism" was more complex (with lower and higher levels of neuroticism being linked to fatal crashes, but a moderate level of neuroticism, perhaps a "healthy" level of worrying, being linked to safer behavior). The author also noted the potential importance of "safety culture."

Crash and Injury Data Integration

Several efforts have linked crash data with injury data from hospital records to support analyses of injury types, outcomes, and costs that cannot be completed using crash data alone. The Crash Outcome Data Evaluation System (CODES) is a longstanding NHTSA effort to support state-level linkage of crash data with medical outcome data (Johnson and Walker 1996; NHTSA 2004). This project has generated much research related to the health impacts of motor vehicle crashes. Sauter et al. (2005), for instance, used CODES data in Wisconsin to show the importance of helmets in reducing head injuries and deaths in motorcycle crashes. Karlson et al. (1998) examined serious injuries to lower extremities resulting from motor vehicle crashes, finding that female drivers, especially those over 60 years old, faced a higher risk of such injuries, as did persons in head-on collisions and those traveling on roads with higher speed limits. Smith et al. (2004) analyzed fatigue, intoxication, and seat belt use and their effect on hospitalization and death from crashes. The authors found improvements in increasing seat belt use and decreasing intoxication, but no improvement in reducing fatigue-related crashes.

Similar work has been done internationally. One such example is Rosman's (2001) report on the probabilistic linking process used to match hospital discharge data, containing information on injury type, severity, and treatment, to police crash reports for a ten-year period in Western Australia. Initial analysis using the resulting database indicated that half the variation in the driver injury severity of single-vehicle crashes was linked to seat belt use, controlling for crash, driver, and vehicle variables, but seat belt use in head-on collisions had much less effect. Rosman noted that the project could be developed by linking to driver license data in order to analyze speeding and drunk driving. More recent work has focused on the relationship between age and injury severity (Meuleners et al. 2006).

In a similar effort, Lyons et al. (2008) "investigate[d] the degree to which understanding of trends in serious road traffic injuries is aided by the use of multiple datasets" (p. 1406) in Great Britain. Analysis has typically relied on police reports collected into a single dataset, called STATS19; however, not all road injuries are reported to the police. To address this issue, the authors matched hospital admissions data, trauma data, and emergency department data to analyze the degree of correspondence regarding numbers and trends. In one case, the authors found that about half of the slightly injured casualties in the police record were not found in the

corresponding hospital record, and that there were several slight and serious casualties not known to the police. Overall, the authors found that several trends for pedestrian casualties (downward) and motorcycle casualties (upward) were consistent across databases; however, the STATS19 serious casualty data showed a strong declining trend for vehicle occupant casualties not seen in the other databases. The authors also discussed the relative strengths and weaknesses of the various databases and noted issues related to the matching process, such as a lack of key fields to match records.

Taiwan has constructed a Comprehensive Crash Database that includes crash, hospital, insurance, vital registration (death certificate), and traffic violation data. This database was used to adjust fatality statistics to facilitate comparisons with other countries (necessary because of varying reporting standards regarding the maximum length of time between a crash and a fatality associated with the crash, 24 hours in Taiwan and 30 days in several other countries). Crash data were linked to vital registration data, and patterns related to motorcycle vs. non-motorcycle crashes were discussed (Lai et al. 2006).

Other Applications

Although most efforts have focused on drivers or injuries, several data integration efforts have been completed in other areas. Some of these efforts have linked crash files to commercial vehicle or workplace data. For instance, Staplin and Gish (2005) analyzed several files in the Motor Carrier Management Information System to evaluate the relationship between commercial drivers' crash histories and the degree to which the drivers change jobs. The Crash File included information on fatal, injury, or tow-away crashes involving trucks or buses, about 100,000 crashes per year. The Census File contained information on 600,000 interstate carriers and shippers. The Inspection File contained data (driver, carrier, and violation) related to safety inspections; this file can also be used to generate data on drivers' work histories across carriers. The authors found that drivers' likelihoods of being involved in a crash increase with their rate of changing jobs. The authors note, however, that the results are subject to a variety of driver characteristics (e.g., age, education) not available in the dataset, as well as to variations in cargo types and geography.

In a similar study, Boufous and Williamson (2006) merged crash data and workers' compensation data to study work-related traffic crashes in New South Wales, Australia. The authors noted that it was "the first study to link compensation and crash records in order to examine the nature and circumstances of work-related traffic accidents" (p. 18), an important issue because work-related and commuting crashes make up about 25% of all work-related deaths in the U.S. and nearly 50% in Australia. The traffic analysis database contained information on all reported crashes in New South Wales, including crash date, time, and location and the age, sex, and alcohol level of the persons involved. Nearly 84,000 records of injuries and fatalities were selected from the database for a five-year period. From the workers' compensation statistics, over 61,000 records were selected that related to compensation claims stemming from crashes. These records included information on the industry, occupation, age, and sex of the employee, as well as information on the injury leading to the claim. The two databases were linked using probabilistic linkage. The results were used to analyze the

importance of driver behavior, commuting vs. work-related activity, driver fatigue, heavy trucks vs. other vehicles, and alcohol.

In a traffic engineering study, Tindale and Hsu (2005) reviewed traffic volume data, physical conditions, and traffic signal operations data, in addition to crash data, to analyze the effect of traffic signal coordination and "platoons" on crash incidence. The authors found, through a review of crashes with causes given as "disregarded traffic signal," evidence that signalization timing and one-way streets may affect the tendency of drivers at the end of a platoon to speed up and perhaps run red lights.

Another type of data linkage is seen in a study by Marques et al. (2003). The authors related data from alcohol ignition interlock devices, used by over 2,000 driving-under-the-influence offenders in Alberta, Canada, to data from questionnaires (for demographics and self-reported drinking) and driver records. The authors found that failed interlock tests were the best predictor of recidivism (compared to questionnaire data or driver records) during the period after removal of the device.

The studies cited above illustrate the usefulness of linking crash-related databases to study driver characteristics as they relate to crash experience, medical outcomes of crashes, and a variety of other safety-related issues. Most of these studies, however, did not explicitly incorporate or analyze spatially-referenced data, either because such data were unavailable or because the use of such data was out of the research scope of the studies. The next section examines the potential use of spatially-referenced data and reviews several studies in which such data were used and analyzed.

Integration and Analysis of Spatially-Referenced Data

Overview

Spatially-referenced data in crash-related databases can take several forms. Locations can be recorded by jurisdiction/area, by address, by geographic coordinates, or by other location referencing systems. Crash data, for instance, typically include jurisdiction/area (e.g., city, county) information, and often more specific information on crash location, with respect to some location referencing system. Driver data typically include jurisdiction information on the residence of the driver, as well as address information that can be converted to point coordinates in a GIS through address matching. The locations of citations are typically specified by jurisdiction and increasingly through a point location referencing system. Locations of medical facilities are often already available as GIS databases, or such databases can be readily developed. Vehicle data typically have information on the jurisdiction of the owner's address, as well as address information that can be converted to point coordinates. Finally, other data that may be useful in crash analysis, such as census demographic data, topography, land cover, and land use, are also available in GIS.

In addition, spatially-referenced data can be analyzed in several ways (O'Sullivan and Unwin 2002; Kim et al. 2001). The most common form of spatial analysis is the evaluation of spatial

distributions. Spatial distributions may be generated that require only the use of one database, such as a mapping out of crashes or crash types, drivers (once geo-coded), and so on, to assess the degree of clustering or dispersion or to identify other spatial patterns. However, spatial distributions can also be generated that require the use of two or more databases. Examples include a map that uses crash and driver data to show crashes involving the residents of a specific county, or a map that uses crash and injury data to show crashes resulting in a specific type of injury or certain injury cost. Linkage across databases may be particularly useful for generating a subset of records in one database, relating these records to corresponding records in another database, and mapping the subset of records. For instance, drivers involved in multiple crashes can be identified and selected from driver data, and their crashes can then be mapped and analyzed for spatial patterns. Similarly, the home addresses of drivers involved in multiple crashes can also be mapped (either as point locations or by aggregating these locations by area, e.g., census tract, city, zip code, county).

Beyond examining the spatial distribution of one set of objects (e.g., crashes), spatial relationships among sets of objects can be investigated. Such relationships can be considered in terms of spatial association, proximity, and adjacency. Spatial association or correlation refers to relationships between different objects or characteristics in space, for instance, the relationship between pedestrian crashes and traffic volume (Kim et al. 2001), the relationship between distracted driver crashes and signage, or the relationship between citations and crashes. Proximity relationships are those within a given distance threshold. An analysis of proximity, for instance, might evaluate the percentage of child pedestrian crashes within a given distance of schools (Kim et al. 2001). The concept of adjacency more narrowly defines distance relationships in binary terms. Objects are either adjacent or they are not (O'Sullivan and Unwin 2002). For instance, a road segment may be considered adjacent to connecting road segments or to commercial developments located along the segment. The adjacency of crashes to each other, or to other types of objects, may be similarly defined.

Spatially-referenced objects across crash-related databases (e.g., crashes, drivers, vehicles, citations, roadways, injuries) can be linked to explore the spatial relationships among the objects. A common example is the linking of crash data with roadway characteristics to examine the influence of environmental factors. However, other examples can be considered, including the mapping of both the locations of crashes and the locations of residences of individuals involved in these crashes. Linkage of spatially-referenced data can also be used to create subsets of drivers or crashes to explore more specific relationships, such as crash and residence locations involving drivers involved in multiple crashes (e.g., to evaluate the travel patterns of potential "problem drivers"), crashes at night (e.g., to evaluate whether familiarity with the roadway affects crash incidence in nighttime conditions), crashes in wintry or other poor driving conditions (e.g., to study who is more likely to have accidents or to be traveling in such conditions), and crashes involving alcohol (e.g., to study travel and crash patterns of drunk drivers with respect to home locations). Relationships can also be examined between injury crashes of a given type or severity and the locations of medical facilities of a given level, or between the locations of traffic violations and crashes in a given metropolitan area. Beyond the information contained solely in crash-related databases narrowly defined, spatial correspondence between crashes and other information in GIS, such as land use, land cover, and topography, can also be evaluated.

The role of distance relationships is a key component in the analysis of spatial association. These relationships are typically examined with respect to two or more spatial distributions (e.g., driver residence vs. driver crash, crashes vs. hospitals) with objects that can be linked on a one-to-one, one-to-many, or many-to-one basis. There are several ways to operationalize and measure distance. The quickest and simplest is by using straight-line distances. For instance, distances between crashes with different injury severities and the nearest medical facility can be calculated using X, Y coordinates and the Pythagorean theorem. If road network data and routing algorithms are available in GIS, network distances can be calculated. Further, if road characteristics are available (e.g., impedance values, speed limits, time to traverse each segment), distance measures can be converted to other units, such as time. These times can be compared to, or calibrated with, EMS records of time to crash site. Such analyses of distance relationships are being used in other areas of research, such as crime analysis, e.g., the "Journey to Crime" module in CrimeStat (Levine 2007), and there is interest in such relationships in a crash analysis context, as seen below. The identification of segments used in network shortest distance (or time) routes can potentially be linked back to the road network and employed in studies of network use.

Finally, the spatial association and analysis of spatially-referenced data commonly involves the concept of neighborhood, which relates to the region surrounding and associated with a given set of spatial objects (O'Sullivan and Unwin 2002). Crashes can be associated with the jurisdictions (e.g., cities, counties) or other areas (e.g., census tracts, zip codes) in which the crashes are located. These regions can be used to compute area-specific values (e.g., age, income, travel characteristics) to generate area-wide crash rates, or these regions can be used to control for geographic variability in exposure or other factors influencing crash incidence, for broader-scale analyses. In addition, a centroid, i.e., a centrally located point, can be used to represent each region in GIS. This can be useful for distance calculations and other spatial statistics and display methodologies.

Spatially-Referenced Driver-Crash Data Linkage

Many of the driver-crash studies outlined above could have been extended with the availability of spatially-referenced data. Crash experience has been associated with both traffic violations and previous crash experience. It makes sense to understand more fully these relationships, including their spatial aspects. For instance, research on "problem drivers" and "accident proneness" may benefit from the analysis of spatial distributions to identify any patterns that might assist law enforcement, education, and other traffic safety activities. Cooper (1997), for example, found a link between excessive speed violations and individuals involved in multiple crashes. However, information was not available to link back to the locations at which these individuals crashed or received citations, which might be useful information from a law enforcement perspective. These locations could be compared to those of higher-end speeding violations in general or crashes attributed to excessive speed. In addition, the locations of these drivers' residences might be useful for education programs (especially if younger drivers are involved). There would, of course, be privacy concerns (discussed later in this report).

Similarly, Chandraratna et al. (2005) focused on young novice drivers. A follow-up study using integrated spatially-referenced data could consider where the crashes of high-risk younger and/or

novice drivers are located with respect to schools, place of residence, the route between school and home, or other locations (e.g., land uses associated with shopping or commercial establishments, perhaps as a proxy for leisure or work activities), as well as roadway characteristics. The study by Boufous and Williamson (2006) cited above, which linked crash records to workers' compensation data, may have interesting spatial aspects that involve the locations and times of work-related crashes, the home addresses of the drivers involved, and the work addresses from the workers' compensation database. Staplin and Gish (2005) noted in their study that their results regarding commercial drivers' job changing rates and crash histories were subject to variations in geography. The study by Rosman et al. (2001) found that younger and older drivers were more likely to be involved in drunk-driving crashes than routine enforcement arrests, while the Aboriginal population was more likely to be involved in routine enforcement arrests. The integration of crash locations and citation locations might yield preliminary conclusions regarding whether this result was due to variations in driving behavior or spatial patterns of enforcement. Although most studies of driver-crash relationships have not explicitly used and analyzed spatially-referenced data, one of the studies cited above, Gebers and Peck (2003) did use the zip code of the driver residence and a set of geographic variables to control for spatial variations in crashes and citations. Also, at a much larger scale, Lajunen's (2001) study, discussed above, linked crash data and psychological data at the cross-national level.

Spatially-Referenced Crash-Injury Data Linkage

In contrast to driver-crash studies, the potential of analyzing spatially-referenced crash-injury data was seen early in the CODES effort, partially in response to the location information typically found in CODES files, and partially because of the usefulness of spatial data in identifying problems and weighing countermeasures (Kim et al. 2001). The integration of spatially-referenced data facilitates a wide range of analyses:

> Questions are often posed in terms of point locations (which intersections produce the highest number of fatal crashes); or segment and roadway queries (which roads have the highest incidence of bicycle or pedestrian crashes); or zonal or areal tabulations and analyses (how do cities or towns or census tracts or block groups compare in terms of the frequencies of various types of crashes). With linked data, however, the questions and topics of inquiry can be expanded to include more detailed information about ambulance transports, medical treatments, hospital and insurance costs, and other elements contained within the CODES linked databases. Which highways produce the most EMS runs? Which segments of roadway have the highest costs in terms of utility pole crashes? Which produces more hospitalizations among elderly drivers involved in crashes at a particular location - broadside crashes or head on collisions? GIS is an enabling technology that allows safety researchers to fully explore relationships between crash and injury variables contained within the CODES databases. (Kim et al. 2001, p. 3)

Each of the databases has specific information on distinct aspects of crash incidents. The linkage allows these aspects to be brought together and analyzed in a common GIS environment. As Kim et al. (2001) remark, "By geo-coding the crash files, the CODES states are able to map not just

crash data, but also all of the data linked to the crash file" (p. 8). The authors further highlight several applications of CODES data in a GIS environment:

- "Mapping pedestrian crashes and injury outcomes
- Spatial correlation of pedestrian accidents and their medical and financial outcomes with socioeconomic characteristics of neighborhoods
- Identification of the medical and financial outcome for hazardous roadway segments with a high incidence of run-off-road crashes
- Mapping the benefits in terms of medical and financial outcome for variations in the temporal and spatial patterns of EMS utilization along key highways
- Installation of traffic calming devices where they will have the most impact on reducing injuries and health care costs
- Mapping locations with serious injuries and high health care costs in order to prioritize installation of red-light running cameras
- Identification of the injury, EMS, and hospital costs associated with crashes occurring at particular locations" (Kim et al. 2001)

Similarly, Rosman (2001) suggested the importance of spatially-referenced data in her study of crash/injury data linkage:

> Geo-coding of the crash sites and other location details will enable future analysis to include spatial characteristics of the crash environment. Whereas previous analyses concentrated on the risk of serious injury as a result of a road crash, inclusion of social and physical environment details will permit crash risks to be estimated for drivers and their vehicles. (p. 87)

Initial applications tended to focus on mapping and analyzing the spatial distribution of crashes, such as the use by Cromley et al. (1998) of the Connecticut CODES GIS. Later applications combined health delivery data to evaluate spatial relationships between crashes and medical care. For instance, Cromley and Wei (2001) used CODES data, locations of ambulance facilities, and network analysis to study EMS response times and service areas in northeastern Connecticut and to recommend new EMS locations to improve response.

Most traffic fatalities occur in rural areas, and several observers have noted that fatality rates from motor vehicle crashes are higher in rural areas as well (GAO 2004). This can be attributed in part to speeds and crash types, but there is interest in the effect of response times to serious crashes. The time it takes for an ambulance to reach a crash victim, the time it takes for the victim to reach the hospital, and the level of medical care available are key concerns in evaluating the outcomes of crashes (Brodsky 1990; Durkin et al. 2005).

In one study stemming from the CODES effort in Wisconsin, Durkin et al. (2005) used ten years of CODES data to analyze the effect on fatality risk of the distance between crash location and Level I or II trauma care. The authors found that, controlling for other factors (e.g., age, type of crash), the fatality rates of crashes were related to distance from trauma care, and they suggested that fatalities could be reduced by improving response times and access to trauma care,

especially in northern and western Wisconsin. It is important to note, however, that the distances used were from the centroid of the county in which the crash took place to the location of the nearest Level I or II trauma center. The authors note that a limitation of the study is that they did not have the point location of each crash. "The availability of such point data would enhance precision and allow estimation of actual response times from crash occurrence to receipt of trauma care" (Durkin et al. 2005, p. 30).

Instead of distance, Gonzales et al. (2009) assessed the importance of time as recorded in the CODES data. The authors linked crash with EMS/hospital data to examine the effect of EMS response time, scene time, transport time, overall time, and EMS distance on mortality in Alabama. Census data and GIS were used to divide the crashes into urban and rural areas. The authors concluded that higher prehospital times are associated with higher rural fatality rates. In a similar linked-data study conducted in Taiwan, Li et al. (2008) found that rural crash victims are more likely to die before they reach a hospital.

Although several studies have evaluated emergency response to crashes and the ways injury severity is affected by distance and time, there are opportunities to improve analyses through the use and assessment of network distances, routing algorithms, identification of key EMS corridors, service area coverage, and spatial patterns of available higher-level medical care. In addition, there may be opportunity to incorporate spatially-referenced data more directly into the data linkage and verification process

Spatially-Referenced Area-Level Studies

Other research has examined spatial variations in crash outcomes using census data to capture the influence of broader geographic variables. In this case, the linkage is to data outside the standard set of crash-related databases and corresponds to the "neighborhood" form of spatial association and analysis discussed above. The main difference for spatially-referenced area-level studies is that at least some data are aggregated into spatial units of observation (e.g., census tracts, counties, states), and crash data are either aggregated to the same spatial scale or each crash point has attached to it, through a spatial join, the area-level data for the region it lies within. This enables the inclusion of variables at a broader scale. In an earlier study, Baker et al. (1987) found a relationship between population density and per capita motor vehicle fatality rates at the county level. Similarly, Clark and Cushing (2004) conducted a state-level analysis and found that population density is inversely related to rural motor vehicle mortality rates.

In a more detailed study, Clark (2003) examined the relationship between traffic crash mortality and population density. Citing several previous studies, he noted that population density has an inverse relationship with per capita mortality rates from motor vehicle crashes for several hypothesized factors, including crash characteristics (e.g., speed), behavior (e.g., seat belt use), and distance to medical facilities. To analyze the role of such factors, Clark obtained data from the National Automotive Sampling System–General Estimates System (NASS-GES). This resource contains data on a random sample of police reports across the United States. The data include the characteristics of the people involved, safety equipment, injury severity, and the location of the collision (region and size category of the city or town). The data also include the zip codes of the residences of each driver. A zip code to county correspondence table was

acquired, as were census data for each county. In addition, Fatality Analysis Reporting System (FARS) data, which include the county of the crash, were also acquired and linked to the zip code and county files so that the locations of the crashes and the drivers could be compared. The driver residences were found to be an adequate surrogate for crash locations, and the zip code and county files were then linked to the NASS-GES data.

Using the data, Clark employed multiple logistic regression to evaluate possible factors influencing injury severity (population density of driver county of residence, region and size category of the city or town, age, sex, speed, seat belt use, alcohol use). Controlling for the other variables, "rural" (population less than 25,000) crash location was positively related to mortality after collision (adjusted odds ratio = 2.10), as were age over 50 years, speed over 50 miles per hour, and unbelted drivers. However, region, sex, and alcohol use did not have an independent effect. As the author notes, "The implication is that the distance from effective medical care in rural areas influenced the mortality of the victims of similar crashes with incapacitating injuries" (Clark 2003, p. 969). To improve the study, Clark recommends that NASS-GES data include the county of collision and/or FHWA roadway function data. In addition, he suggests, "Eventually, more precise geo-coding by latitude and longitude may be possible and will probably be necessary to show definitively how much mortality from vehicle collisions is affected by distance from medical resources (whose locations could also be precisely specified)" (p. 970). As more crashes are spatially referenced and linked to roadway and other data, such analyses can be more comprehensive.

In a broader-scale study on the importance of medical services, Noland (2003) examined the link between crash fatalities and medical treatment at the international level. Specifically, he reviewed various underlying causes of decreased motor vehicle fatalities in developed industrialized countries, including vehicle design, drunk driving enforcement, seat belt usage, and road design, but he suggested that improvements in medical treatment and technology also played a role. He combined data from the International Road and Traffic Accident Database (IRTAD) with health care data from the Organization for Economic Cooperation and Development (OECD). Using national-level data and several proxy measures for improvements in medical care, he found that such improvements are indeed related to decreases in motor vehicle fatalities over time.

Area-level studies of metropolitan crash patterns have also received increased interest, specifically the use of macrolevel collision prediction models (Hadayeghi et al. 2003; Lovegrove and Sayed 2006; Lovegrove and Sayed 2007). Some researchers suggest that microscale models and analyses of hot spots, while useful, are largely reactive, and these researchers recommend the development and use of neighborhood-level or macrolevel analyses as a proactive approach to complement the microscale approach. Lovegrove and Sayed (2007) analyzed 577 urban and rural neighborhoods in the greater Vancouver area using explanatory variables related to exposure, sociodemographics, travel characteristics, and network characteristics, and the authors highlighted the potential of the approach in identifying early warnings of problem areas. Similarly, Hadayeghi et al. (2003) developed a series of models for the Toronto area to predict total and severe crashes as a function of zonal characteristics, including socioeconomic/demographic, network/supply, and traffic demand variables. Notably, geographically weighted regression was used to evaluate spatial heterogeneity, i.e., spatial variation in parameter values, although results were mixed.

Other Studies

There is interest in examining the spatial and temporal relationship between enforcement activity and crash activity, particularly whether enforcement in a given area results in a reduction in crashes and how long any effect lasts. However, as Anderson (2003) notes, "There has been little attempt to merge road traffic incident reduction and road traffic policing within a spatial context" (p. 22). Similarly, Beenstock et al. (2001) suggest that, overall,

> there has been relatively little direct investigation of the effect on road safety of police enforcement per se. Most probably this reflects the absence of the necessary data on traffic policing. (p.73)

To address this issue, Beenstock et al. (2001) examined data on crashes and police reports for road segments in Israel. Their findings suggest "increasing returns to scale" to enforcement. That is, it was better to focus on large-scale efforts in a few areas rather than smaller-scale efforts over a large area. However, once enforcement was removed, the crash rate typically increased to its former level and any beneficial spillover effects to other road segments were generally weak. Moreover, while enforcement had an effect on total crashes, no such effect was found on fatal crashes.

With increased availability of data, several CODES efforts are integrating citation data into crash-related databases. These include efforts to compare the spatial patterns of impaired driving crashes with those of impaired driving citations as part of the Maryland CODES project (Kerns and Burch 2006), as well as similar activities being implemented or explored in Massachusetts (Rothenberg 2008), Indiana, and other states. Information on crashes and citations could then be used to assist enforcement efforts (Steil and Parrish 2008).

Other studies have been conducted that make use of spatially-referenced data on residences and crashes and the calculation of distance relationships between the two. Moellering's (1974) study of residence-to-crash distances in Michigan is an early example. More recently, Gary et al. (2003) incorporated the location of driver residences in their analysis of alcohol-related crashes in wet versus dry counties in Kentucky. The authors used the locations of crashes (county, highway, mile point) and drivers' residences (zip code centroids) to calculate the straight-line distance between the two. The authors found that, although there were more alcohol-related crashes in wet counties than dry, a higher percentage of the residents of dry counties had been involved in an alcohol-related crash. Moreover, the distance from residence to crash was higher for the residents of dry counties, both for alcohol-related crashes and those not involving alcohol. Residents of dry counties that do not border wet counties crashed farther from their homes than those whose counties border wet counties, suggesting that spatial patterns in the ability to purchase alcohol may influence travel behavior and the location of crashes. The authors note several limitations to their study regarding spatially-referenced data. Several crashes had to be deleted from the database because they lacked route numbers or mile points, and centroids of zip codes were used instead of home addresses. Also, straight-line distances were used rather than network distances because of insufficient linework for the roads in GIS.

One approach to integrating origin-destination and route data with crash data is found in a study

by Kam (2003), who suggested a linkage between crash data and travel data in a transport modeling approach. The author proposed using GIS to match crash records to travel survey data to develop measures of crash risk that are disaggregated by age, sex, time of day, and day of week. The study notes that crash rates that adjust for distance traveled and time spent traveling have been criticized for assuming linearity in rates with increasing distance and time, when drivers on freeways can face lower risk for each mile or minute traveled than drivers on other roads. Kam's (2003) approach suggests using spatially-referenced data as a link between the locations of crashes and the addresses of trip origins and destinations. In this approach, travel routes and crash locations would be linked in GIS. Data on the characteristics of travelers, mode, start and end times, and trip purpose would be linked to the route. Data on driver characteristics, crash characteristics, and crash times would be linked to the crash locations. Linking the files would permit researchers to analyze the relationships between them, including relationships regarding the exposure rates of sub-populations. This approach is illustrated using data from the Victorian Activity and Travel Survey (VATS) and crash records. Estimates were developed of crash rates of Melbourne residents in different age-sex groups according to time of day and day of week. The results varied significantly from the "U-shaped curve" of crash rates by age normally found using an aggregated approach.

CONSTRAINTS TO DATA INTEGRATION AND ANALYSIS

Despite the potential benefits of integrating and analyzing spatially-referenced safety data, there are several constraints to address. Such constraints can be categorized into several groups, including technical constraints, analytical constraints, administrative issues, and concerns about confidentiality.

Technical Constraints

There are several types of technical issues related to safety data integration that relate specifically to the management and analysis of spatially-referenced data. A key component of spatially-referenced crash-related databases is the referencing system used to locate crashes. As DeLucia and Scopatz (2005) remark, "The use of a precise location reference method is a critical aspect of crash data, whether analyzing the location of crash occurrences or using the location reference to link crashes to other data sources" (p. 11).

In particular, crashes must be geo-coded relative to a corresponding base roadway layer in GIS in order to join successfully to roadway data. However, all spatial data have some degree of positional inaccuracy. For instance, the TIGER line files used by the Census Bureau were originally developed using Digital Line Graphs of the U.S. Geological Survey, which were created at a scale of 1:100,000. Using National Map Accuracy Standards, this scale corresponds to an accuracy of about plus or minus 167 feet. Many GIS road databases used now have much better positional accuracy. However, the key issue is how the crashes and roads relate to each other, and this relation is only as good as the least spatially accurate layer. Even if one layer has a high level of spatial accuracy (e.g., GPS coordinates for crashes), if the other layer has a lower accuracy level, then crashes may be placed at the wrong location along the road, off the road, or on another road. DeLucia and Scopatz (2005) highlight the issue of spatial data precision in their report on crash record systems:

> The use of coordinates alone can create difficulties in trying to merge data files because of the level of precision needed to match the locations. For example, a roadway file may identify a location to a particular point, whereas a crash location code may identify a spot several meters from that roadway point…. Knowledge of the roadway and a well-defined linear referencing system allows the effective correlation between the various coordinate locations to form a meaningful picture of crash experience. (p. 11)

Thus, spatial joins to attach roadway data to crash data, or vice versa, may result in erroneous information. One benefit of directly geo-coding crash locations onto a specific road network in GIS is that the crashes may be more likely to be located at the appropriate points along the roadway. More generally, positional inaccuracies may affect other types of spatial joins, e.g., crash to jurisdiction, crash to land use zoning, or crash to rail line.

A related constraint concerns the availability of high-quality spatial and attribute roadway data. Roadway information available in GIS varies, especially for local roads (NHTSA 2004). DeLucia and Scopatz (2005) elaborate on this issue:

> Although states are increasing their use of geographic information systems (GIS) technology, they are not adequately maintaining or linking a record of the roadway characteristics associated with specific locations. Core data elements such as location control, number of lanes, lane widths, shoulder widths, median type, and median width are missing in many systems that define roadway characteristics. Items such as horizontal curve, vertical grade, intersection features, and interchange features are virtually non-existent. (p. 5)

This limitation affects several aspects of integrating and analyzing spatially-referenced databases. For instance, several items in crash-related databases that can be spatially referenced are addresses, e.g., of drivers and vehicle owners. Address matching is the process that takes address data, combined with GIS road data containing street names and street addresses for each road segment, to create a point in GIS for each address. While useful, address matching "hit rates" can suffer from incomplete or missing GIS road linework, street names, or address ranges; misspelled data; and incorrect prefixes/suffixes (e.g., N, S, E, W), street types (e.g., St., Rd., Ln., Ct.), and the like. Similar problems in address data, e.g., in driver records, can also affect hit rates.

Moreover, for shortest-path or other routing algorithms, constraints include incomplete intersection, speed, and time data; missing or incorrect connections between road segments; topological inaccuracies (e.g., regarding underpasses, overpasses, on-ramps, off-ramps); and direction of travel information (e.g., one-way vs. two-way streets). Other technical issues, not unique to spatially-referenced data, relate to incomplete, inaccurate, or missing identifiers used to link files. Issues regarding file formats and data storage must also be addressed.

Analytical Constraints

Despite the potential insight to be gained through spatial data analysis, it is important to highlight a few pitfalls of spatial data (O'Sullivan and Unwin 2002). For instance, area-wide studies of the type outlined above are susceptible to the "ecological fallacy," which involves improperly transferring results acquired at one level of aggregation to a more detailed level of aggregation. This is a common issue in geographic analysis, such as in epidemiological studies or economic geography. A state-level study that relates income, for example, to disease, ethnicity, or voting patterns cannot necessarily be applied to individuals. The issue tends to be more prevalent when there is much variation within spatial units of observation (e.g., very high and low incomes, so that a single value for each state may not be representative), and when the object of study is a small subset of the population and/or a relatively rare occurrence (as would tend to be the case in crash analysis).

A related point is that the results of spatial analysis vary by spatial scale and level of aggregation (Anderson 2003). That is, relationships found at the local level may not be applicable at the state

or national level. The combination of scale effects and aggregation effects results in the "Modifiable Areal Unit Problem," i.e., the spatial patterns that are displayed and the processes and relationships that are used to explain these patterns depend on how researchers construct their spatial units of observation, be they census tracts, traffic analysis zones, or road segments.

Two additional analytical issues include the concepts of spatial dependence and spatial heterogeneity. Spatial dependence is simply the idea that what happens in one place affects what happens in other places. In a methodological context, this concept can affect the application of statistical methods that assume the independence of units of analysis (e.g., counties, intersections). For instance, countermeasures applied in one location may affect other locations, e.g., in cases of potential crash migration. Spatial heterogeneity is the concept that relationships vary spatially. In an applied context, this concept implies that the parameters that relate a set of independent variables to a dependent variable will not be spatially constant, applied to all observations, but will instead vary from place to place. Statistical techniques are available to address these issues. For instance, geographically weighted regression can be used to evaluate spatial heterogeneity, as in the Hadayeghi et al. (2003) study discussed above.

Moreover, geographic analysis typically requires a demarcated study area and is thus affected by the issue of edge or boundary effects, interactions between locations inside and outside of the study area. These interactions will affect state-level analysis, as not all drivers and vehicles that crash within a given state are represented in the relevant databases of the state, and this issue may be especially prevalent in urban areas that border other states.

A related set of analytical constraints concerns the degree to which spatially-referenced data are complete and up to date. Driver records are not static, and spatial data elements such as addresses may be the most dynamic component of them. This fact affects all aspects of data integration and analysis. From a data management perspective, for instance, data linkage is seen as a way to maintain data quality:

> Integrated data should enable driver license and vehicle registration files to be updated with current violations, prevent the wrong driver from being licensed, or keep an unsafe vehicle from being registered…. Data linkage is an efficient strategy for expanding the data available while avoiding the expense and delay of new data collection (NHTSA 2006, p. 6).

However, drivers frequently change addresses, and some drivers may slip through the cracks:

> When driver identification data cannot be validated because of the lack of real-time linkages, drivers can obtain multiple driver licenses in various States by changing the personal identification data given to DMV personnel at the time of application. A particular concern is when anticipation of a license revocation in the home State of record causes a problem driver to cross State lines and obtain a driver license from another State. The duplicate licensure is not flagged, since no licensing system currently exists that identifies all licensed drivers in the United States (NHTSA 2004, p. 17).

From an analytical perspective, this may create particular issues for the evaluation of "problem drivers," who may have an extra incentive to provide less reliable address information, have a license outside a given state's jurisdiction, or not have a license at all. Likewise, analyses of vehicle-related factors will typically be limited to vehicles that are registered within a given state (NHTSA 2006, p. 21). As noted above, these effects may be more prevalent in urban areas that border other states.

Administrative Issues

At an administrative level, data integration efforts may be hindered by a "lack of mutual understanding by various data owners" (NHTSA 1996, p. 14). Databases are developed and managed by individuals having specific objectives (crash analysis, health care, driver records, etc.) that may not extend beyond the attributes of the databases they manage:

> The current reality in most states is that no single agency has control of all the necessary data to make up a complete traffic records system. Most components of a traffic records system serve a primary operational purpose that may be far removed from highway and traffic safety analysis. It is through the work of practitioners and the cooperation of stakeholders that anything approximating a comprehensive traffic records system can be created. (DeLucia and Scopatz 2005, p. 25)

In practice, the dispersed nature of data collection and management can conflict with the need for integrated analysis to address transportation safety issues:

> [M]ost routinely collected and accessible State traffic safety data have been initially collected and maintained for agency-specific purposes without consideration of the potential for integrating these data. There is now a rising sense of urgency to understand trends and patterns of the increasingly complex traffic safety and vehicle issues. Motor vehicle crashes are more likely to be viewed as a major public health problem, one that can be reduced by actions grounded in careful vehicle and traffic safety data analysis. As a result, there is more focus on the benefits of integration for more effective enforcement and evaluation of injury outcomes. (NHTSA 2004, p. 20)

The benefits of addressing cross-cutting questions that require data integration may go unrealized. In addition, there are issues related to communication breakdowns, in that one entity may not know the content of data files maintained by other entities or not know the availability of such files. As Council and Harkey (2006) note in their discussion of crash and roadway data, such issues can exist both within and between agencies:

> Although both types [crash and roadway data] are housed in many States in the same DOT, the distinction between the two is drawn because of significant differences in who collects, computerizes, and stores the data; differences in the primary users; and differences in current national efforts to improve the two data

> types, among other factors. Significant problems with crash data continue to exist, particularly those related to data accuracy and data completeness, and solving these problems is difficult because the primary data collectors are in multiple police agencies, each with their own priorities and policies. (p. 5)

A related administrative concern is the cost of the databases themselves and of data integration, specifically the costs of formatting, processing, saving, and sending data to meet requests (NHTSA 1996). The creation and maintenance of a "knowledge base" requires an ongoing commitment of resources (financial, training, etc.), including resources for electronic data collection (NHTSA 2004; DeLucia and Scopatz 2005). In addition, "sparse local and State resources for data analysis often limit the effective use of the traffic safety data being collected" (NHTSA 2004, p. 19).

In a practical sense, spatially-referenced data integration may also be hindered by the jurisdictional nature of data collection and management. Locations of crashes may be reported using different collection methods and location referencing systems in different states or in different cities within a state, which may hinder data aggregation across agencies. Also from a jurisdictional perspective, efforts to link crash histories of different vehicle characteristics within a given state will be limited to vehicles registered within the state unless measures are taken to integrate data across state lines (NHTSA 2004; NHTSA 2006). This problem can affect the data universe available in analyses of issues that cross state or other administrative boundaries, such as in studies of metropolitan areas split across two or more states or areas affected by through traffic. Similarly, injury data linkage may be hindered if a crash occurs in one state but the injured individual is taken to another state for treatment, unless data-sharing agreements are made with other states.

Concerns about Confidentiality

A key set of constraints to data integration and access stem from important legal requirements and ethical standards related to the management of confidential data. In particular, data integration efforts may be limited by concerns regarding the confidentiality of individuals' personal information. This issue is especially relevant in the management of driver records, injury data, and related medical records. In particular, there are concerns about access to data through which individuals can be identified. Crash data made available to researchers, for instance, typically do not include identifying information on the individuals involved in crashes. Similar constraints exist in access to other databases.

A distinction is typically made between direct identifiers (e.g., name, home address, social security number), each of which can be used on its own to identify an individual or household, and indirect identifiers (e.g., age, gender, date of birth), which cannot be used individually to identify specific persons. In addition, to preserve confidentiality, researchers should not be able to combine indirect identifiers to identify specific individuals. Requirements are more stringent for the handling of data with direct identifiers.

Custodians of databases with confidential data must comply with legal requirements to ensure that confidentiality standards are maintained and that procedures are followed to meet such standards. Several pieces of legislation directly govern the use of databases used in transportation safety information systems. The Drivers Privacy Protection Act (DPPA) regulates the use of driver data. Specifically, it declares that state agencies may not disclose personal information obtained in connection with motor vehicle records. Several permissible uses are outlined, including research activities and statistical reports, "so long as the personal information is not published, redisclosed, or used to contact individuals" (LII 2009). Such personal information is defined as information that identifies an individual, including social security number, name, driver ID, address, and telephone number. Personal information does not include information on crashes, violations, or five-digit zip code. Furthermore, "highly restricted personal information," including social security number, can only be disclosed for a much narrower set of uses, largely involving legal actions but not including research. Some states have more restrictive requirements, and the agency creating and managing databases under this framework may have additional procedures to ensure compliance with requirements.

For injury data, the Health Insurance Portability and Accountability Act (HIPAA) governs how protected health data may be used for research purposes, which is relevant for CODES programs and related efforts to analyze crash injury data. Research is defined as "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge" (Health and Human Services 2003, p. 2). To "de-identify" health data for research purposes, all information that could be used to "identify the individual or the individual's relatives, employers, or household members" must be removed. In addition to the data elements regulated by the DPPA, HIPAA also includes health-related account numbers, email addresses, and biometric identifiers, e.g., fingerprints. Furthermore, "All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes" must be removed to "de-identify" the data (Health and Human Services 2003, p. 10). An exception is given for three-digit zip codes if the resulting spatial units have at least 20,000 people. A "limited data set" used for research or related purposes may contain city, state, and zip code, without other identifiers, but such data are still considered protected health information. In limited circumstances, data containing private information may be available for researchers through a waiver, but only in situations where the research could not be conducted without access to such information, and only if procedures to protect this information from improper use and to delete such private information as soon as possible are established.

Other legal requirements may apply to other databases that could be integrated into a transportation safety information system. For instance, the Family Educational Rights and Privacy Act (FERPA) covers the use of educational records, which may also be relevant in certain circumstances, e.g., data related to younger drivers.

Academic researchers have requirements that are administered by their university's Institutional Review Board (IRB) to ensure adherence to legal requirements and ethical standards in research involving human participants. Human participants include anyone from whom information is collected through intervention or interaction, such as surveys or interviews, or about whom private information is obtained. This data collection includes the use of databases in which individuals can be identified. In general, the more sensitive the information collected and

analyzed, the more stringent the procedures must be to ensure that such information is protected. These procedures can cover who has access to the data, how access to others is precluded, how the data are stored and managed, and the disposal of the data once the project is completed.

Despite the potential value of spatially-referenced data in safety analysis, the use of such data creates unique issues with respect to confidentiality. A recent National Research Council (NRC) study (Gutmann and Stern 2007) of privacy issues related to the growth of geographic data resources concluded the following:

> The increasing use of social-spatial data has created significant uncertainties about the ability to protect the confidentiality promised to research participants. Knowledge is as yet inadequate concerning the conditions under which and the extent to which the availability of spatially explicit data increases the risk of confidentiality breaches. (p. 2)

In a transportation safety context, a key spatial component of crash databases is the address information of drivers (and vehicle owners, hospital patients, etc.), which can be located in GIS through the process of address matching. Even if textual descriptions of confidential information are deleted, the points representing confidential locations, which have coordinates generated through address matching, are considered direct identifiers as well and thus also have stringent requirements regarding their disclosure and use. Moreover, if the custodians of this confidential information do the address matching themselves and delete the addresses before sharing the spatial data, the points representing individuals' addresses in GIS can still be considered direct identifiers. Strategies to address this issue are discussed in the next section.

STRATEGIES, RECOMMENDATIONS, AND CONCLUSIONS

Specific strategies can be offered to address each of the constraints identified above. Some initial strategies are outlined below.

Strategies for Addressing Technical Issues

On the technical side, there is the need for continued work on the development of systems and databases to support the geo-coding of spatially-referenced data and spatial data linkage in GIS. As mentioned earlier, Harkey and Council (2006) note the need to ensure that warehoused data are compatible with GIS, and the authors recommend that agencies move to a geospatial reference for all types of safety data as quickly as possible. The authors further suggest that

> the move to GIS-based or other geospatial referencing systems in all safety files will lead to better data, both in terms of greatly improved crash and inventory item locations, linkage of existing safety files (e.g., crash to inventory), and linkage of existing files to new files (e.g., linkage of roadway inventory to weather data and maintenance records). We recommend that States accelerate this process to the extent possible. (p. 35)

Much of the required effort hinges on accurate spatial and attribute data for roadways and well-designed linear referencing systems. Other efforts rely on the recording of exact coordinates of crash locations, e.g., longitude and latitude, at the crash location to reduce the need for later, perhaps error-prone, data processing and to facilitate data linkage (NHTSA 2004).

Strategies for Addressing Analytical Issues

Methodological issues can be addressed mainly by an increased awareness of and sensitivity to such issues as they relate to spatial analysis. These issues (e.g., scale effects, Modifiable Areal Unit Problem, boundary effects) are simply inherent aspects of working with spatial data. Some of these issues can be addressed through newer methods, such as geographically weighted regression to incorporate spatial heterogeneity, adjustments to clustering algorithms to account for edge or boundary effects in the analysis of spatial distributions, and methods to account for spatial dependence and autocorrelation. Several such issues and the methodologies designed to address them actually provide opportunities to analyze key spatial relationships, such as spatial variation in parameter values in models of crash patterns. Similarly, methods to examine spatial dependence and autocorrelation can be used to evaluate the impacts of changes in one part of a road network (e.g., enforcement, safety improvements) on other network locations.

Strategies for Addressing Administrative Issues

To address administrative issues, there needs to be continued intra-agency and inter-agency cooperation on data collection and data sharing standards. In general terms, data integration

requires cooperation and trust among those who generate, manage, administer, and analyze data (NHTSA 1996). There has been progress in this area:

> As the capabilities of computer systems and software have grown in recent years, the ability to support large-scale integrated databases at a reasonable cost has become a reality. At the same time, states have worked to overcome institutional barriers to sharing data with authorized users both within and outside of government agencies. (DeLucia and Scopatz 2005, p. 25)

Several observers point to the importance of inter-agency entities formed to coordinate the management of traffic safety data, in particular highlighting the role of State Traffic Records Coordinating Committees (TRCCs). NHTSA (2004) outlines the role of such entities and highlights Iowa's efforts in this area. In addition to facilitating data linkage, TRCCs can also promote the use of such data by addressing obstacles:

> The U.S. DOT Highway Safety TRCC, along with the State TRCCs, should promote use of traffic safety data for public health and safety purposes. The U.S. DOT Highway Safety TRCC should ensure that training is available for State TRCCs and highway safety offices to assist public health entities in using the data to develop effective public policy. The TRCCs should support overcoming privacy and confidentiality issues at the State level that lead to unintended restricted access. Although these issues are legitimate concerns for data users, the TRCCs should take a leadership role in ensuring that State legislation/administrative policies are clarified to support access to protected health information for traffic safety outcome studies. (NHTSA 2004, p. 40)

Strategies for Addressing Confidentiality Issues

There is increasing interest in accessing and linking confidential data sources for research purposes. This can create administrative burdens for database custodians, as specific and detailed agreements, in response to individual circumstances, must be made with respect to the use of such databases in order to adhere to legal and ethical confidentiality requirements. Many issues must be addressed explicitly and in detail. Such issues may include data storage (e.g., stand-alone vs. networked computers, laptop vs. desktop computers, portable storage media, data backups), access to the data (e.g., for principal investigators, graduate students, others), security (e.g., computer password protections, the use of dedicated rooms and access to the rooms, locks, and storage cabinets), length and conditions of access to and final disposition of confidential data (e.g., date of return, confirmation of deletion, methods to ensure a "clean sweep" of data storage, disposition of databases created and altered during the analysis), and the reporting of results (e.g., level of individual or geographic detail presented). Every situation is unique, but there may be potential in further developing methods and guidelines to facilitate such agreements in order to promote appropriate and beneficial data use while ensuring adherence to confidentiality requirements and standards.

As noted above, locations on a map, such as point locations of the addresses of drivers, vehicle owners, or patients, can still be considered directly identifiable confidential information, even without the data appearing in text form in the database. There are several strategies to address this issue (Armstrong et al. 1999; Rushton 2007). The goal is to preserve confidentiality while also preserving geographic information to facilitate analysis. Perhaps the simplest method is to aggregate point observations into areas such as zip codes, census tracts, counties, and states. This includes the aggregation of individuals within such areas, so that data are at the area level rather than the individual level; such data would typically be displayed as polygons in GIS, although centroids within polygons can also be used. If data are released to researchers in this form, confidentiality may be preserved, but in the process much information is lost that would be useful for addressing key issues in a research context. Alternatively, depending on confidentiality requirements, area-level attribute data (e.g., zip codes and counties, but not address data) of individuals could be released so that these individuals, and groups of individuals residing in the same area, can be represented as centroids within polygons in GIS. Although there is some loss in spatial detail, the individual-level data are preserved.

In addition to aggregation as a means to "mask" spatial data, another strategy is to add a random component in coordinates of spatial objects, so that the X coordinate of a given point is given as X plus or minus some random value, and likewise for the Y coordinate (Rushton et al. 1996; Rushton 2007; Armstrong 1999). This approach was used in a study of infant mortality and birth defects; the researchers address-matched mothers' residences from birth certificate data, as well as addresses from birth defect and infant death records. Before mapping, the point locations were randomly displaced an average of 0.5 kilometers in each cardinal direction to ensure confidentiality. However, the statistical analyses were performed on the original coordinates, before displacement, to ensure analytical integrity (Rushton et al. 1996). The random component would have to be large enough to ensure confidentiality but small enough not to cause analytical difficulties. Other methods to mask databases with spatial information include data suppression, data swapping, and data alteration of non-spatial and/or spatial attributes (Armstrong et al. 1999; Gutmann and Stern 2007; Rushton 2007).

A challenge in masking data is to ensure that confidentiality is truly preserved. As with the suppression of demographic and economic data by government agencies to preserve confidentiality, there is the issue of researchers and others being able to partially or fully recreate the original data. Curtis et al. (2006), for instance, discuss the "re-engineering" of residential point locations and illustrate this concept by showing how a published generalized map of mortality locations from Hurricane Katrina could be re-engineered back to the original house locations.

In the recent NRC study on confidentiality issues with spatial data, it was concluded that, "[b]ecause technical strategies will not be sufficient in the foreseeable future for resolving the conflicting demands for data access, data quality, and confidentiality, institutional approaches will be required to balance those demands" (Gutmann and Stern 2007, p. 3). Several recommendations were offered in the report, including the sponsoring of research into methods to preserve confidentiality while maintaining data usefulness; training of faculty and organizations in the ethical use of spatial data; the development of expertise within Institutional Review Boards regarding how to balance access, confidentiality, and data quality; and the design of projects by researchers using spatial data that explicitly consider such trade-offs. In particular,

the authors suggest “establishing tiers of risk and access and developing data-sharing protocols that match the level of access to the risks and benefits of the planned research” (Gutmann and Stern 2007, p. 3).

In the context of transportation safety information systems, several such “tiers of risk and access” can be preliminarily outlined. An initial tier, to minimize risk by minimizing access, might be to have no data sharing with outside researchers or other entities in order to ensure confidentiality. However, the level of “benefits of the planned research” is also minimized.

Another tier, or set of tiers, might offer researchers access to data (e.g., drivers, vehicles, injuries) that are spatially aggregated at varying levels, for instance at the county level, city level, zip code level, and so on. Data custodians would prepare the summary databases for distribution to others. Summary information (e.g., on the number of drivers, number of problem drivers, number of citations, number of vehicles of given types, etc.) would be available only for areas, not individuals. This approach would allow some area-level studies of the type outlined above, perhaps with additional demographic and socioeconomic data at the same scale, and it would allow simple mapping of spatial distributions, but it would not permit detailed analysis. Also, the interpretation of statistical correlations may be problematic and subject to the ecological fallacy described above. Smaller spatial units of observation (e.g., blocks instead of counties) create more statistical variation in the database, which is useful for analytical purposes, but such units introduce greater risk in terms of confidentiality, especially when the area-level data for a given variable represent a small subset of an area’s population, depending on the specificity of attribute data (e.g., victims of injuries of a certain type).

Another tier of risk and access would be to have data available at the individual level, but with spatially-referenced data available only at a limited spatial scale. The records of the database would represent individuals (not counties, zip codes, etc., as above), but with only the county (or city, or zip code, etc.) of each individual made available. This would allow the linkage and analysis of individual-level characteristics across databases (e.g., driver vs. crash, driver vs. injury), but without the direct identifier of a street address from which point locations of residences can be generated. Distance relationships (e.g., home vs. crash locations for a subset of drivers or crashes) could be estimated using the centroids of the spatial units of observation, as was done in a few studies discussed above (Gary et al. 2003; Durkin et al. 2005). However, the authors of these studies noted that the use of centroids rather than specific addresses meant that distances could not be calculated precisely.

Still another tier of risk and access would be to make individual-level address information available to researchers, including individual-level address information. This would allow address-matching of the residences of individuals and households as points in GIS. Such information would support a full range of spatial analysis methods related to spatial distributions, proximity, adjacency, and distance at high levels of spatial precision. This tier, however, involves much higher confidentiality risk than the others and requires the preparation of and adherence to detailed data-sharing agreements, of the type noted above, regarding such issues as data access, security, data disposition, and the detail at which results can be published.

In addition to "tiers of risk and access," there are also trade-offs at each tier concerning the analytical precision available and the preparatory effort necessary to achieve that precision. Analyses of area-level data in GIS require that boundary layers (for counties, cities, etc.) in GIS are acquired and that the spatial references (e.g., zip codes, counties, and cities) in traffic safety databases are linked with these layers by an appropriate identification code in GIS. This is a fairly routine process that typically does not require much preparatory work. However, as noted above, the level of analysis possible is fairly limited.

Likewise, at another tier, individual-level databases that contain locations only as areas (not addresses) can be integrated into GIS using the same process of matching the appropriate areas in GIS with the area-level information in traffic safety databases. The locations of individuals can be displayed as polygons or by the centroids of the polygons in which the individuals are located. Again, this is a fairly routine process, but the availability of individual-level data does enable analyses (e.g., comparing individual residence location with individual crash location) that are not possible with area-level data. Thus, a significant improvement in analytical ability is possible with little additional preparatory work in GIS to facilitate the analysis.

Finally, the availability of individual-level address data makes possible a full range of analyses. However, as outlined above, more detailed spatial analysis also requires detailed and accurate GIS databases and likely some initial preparation or modification of these databases. Current road files are needed that reflect recent changes in transportation infrastructure. For address-matching to generate points from addresses, street names and address ranges linked to the road segments also must be current. For network analysis methods such as routing, the road network should have accurate data on direction of travel, permitted and unpermitted turns, intersection delays, speed limits, topology (overpasses/underpasses vs. at-grade intersections), and so on. Once the databases are prepared, the address matching process is largely automated but can involve manually matching some addresses and additional database edits. Network analysis can also require preliminary analyses to verify the validity of the network and the routing process.

Given the possible required preparatory work, much initial effort may be necessary to undertake a single study or topic. To make the best use of information resources, data-sharing efforts, and GIS and other database management activities, a range of research questions could be identified and developed in order to spread the necessary investments in time and cost across a variety of projects addressing transportation safety issues. Certain states might be better situated than others to undertake such research. Iowa, for instance, is known for the quality of its spatially-referenced crash database with a strong link to its GIS road data; its leadership in incident data collection technology; its history of inter-agency cooperation, including its nationally recognized TRCC; its work in linear referencing systems; and its CODES program. Such an effort to integrate and analyze spatially-referenced transportation safety databases would require the cooperation of several entities and discussions regarding potential applications or questions to address, expected benefits, data resources, confidentiality concerns, data-sharing, and analytical procedures.

REFERENCES

Anderson, T. 2003. *Review of Current Practices in Recording Road Traffic Incident Data: With Specific Reference to Spatial Analysis and Road Policing Policy*. London: Center for Advanced Spatial Analysis, University College London.

Armstrong, M.P., G. Rushton, and D.L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18: 497-525.

Baker, S.P., R.A. Whitfield, and B. O'Neill. 1987. Geographic variations in mortality from motor vehicle crashes. *The New England Journal of Medicine* 316: 1384-87.

Beenstock, M., D. Gafni, and E. Goldin. 2001. The effect of traffic policing on road safety in Israel. *Accident Analysis and Prevention* 33: 73-80.

Boufous, S. and A. Williamson. 2006. Work-related traffic crashes: A record linkage study. *Accident Analysis and Prevention* 38: 14-21.

Brodsky, H. 1990. *Geographic Perspectives on Improving Emergency Notification in Road Accidents*. AAA Foundation for Traffic Safety.

Burns, P.C., and Gerald J.S. Wilde. 1995. Risk taking in male taxi drivers: Relationships among personality, observational data and driver records. *Personality and Individual Differences* 18: 267-78.

Chandraratna, S., N. Stamatiadis, and A. Stromberg. 2005. Potential crash involvement of young novice drivers with previous crash and citation records. *Transportation Research Record: Journal of the Transportation Research Board* 1937: 1-6.

Chandraratna, S., N. Stamatiadis, and A. Stromberg. 2006. Crash involvement of drivers with multiple crashes. *Accident Analysis and Prevention* 38: 532-41.

Chen, W., P. Cooper, and M. Pinili 1995 Driver accident risk in relation to the penalty point system in British Columbia. *Journal of Safety Research* 26: 9-18.

Clark, D.E. 2003. Effect of population density on mortality after motor vehicle collisions. *Accident Analysis and Prevention* 35: 965-71.

Clark, D.E. and B.M. Cushing. 2004. Rural and urban traffic fatalities, vehicle miles, and population density. *Accident Analysis and Prevention* 36: 967-72.

Cohen, J. and B. Preston. 1968. *Causes and Prevention of Road Accidents*. London: Faber and Faber.

Cooper, P.J. 1997. The Relationship Between Speeding Behaviour (as Measured by Violation Convictions) and Crash Involvement. *Journal of Safety Research* 28: 83-95.

Council, F.M. and D.L. Harkey. 2006. *Traffic Safety Information Systems International Scan: Strategy Implementation White Paper*. FHWA-HRT-06-099. McLean, VA: U.S. Department of Transportation, Federal Highway Administration.

Crettenden, A. and A.E. Drummond. 1994. *The Young Problem Driver versus the Young Driver Problem – A Review and Crash Data Analysis*. Federal Office of Road Safety, Australia.

Cromley, E.K., M. Kapp, and B.R. Pope.1998. Analyzing motor vehicle injuries with the Connecticut Crash Outcome Data Evaluation System. *GIS Proceedings of the 1998 Geographic Information Systems in Public Health Conference*. Agency for Toxic Substances and Disease Registry.

Cromley, E.K. and X. Wei. 2001. Locating facilities for EMS response to motor vehicle collisions. *International Health GIS Conference Proceedings*. ESRI, November 13.

Curtis, A.J., J.W. Mills, and M. Leitner. 2006. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics* 5: 44.

DeLucia, B.H. and R.A. Scopatz. 2005. *Crash Records Systems: A Synthesis of Highway Practice.* NCHRP Synthesis 350. Washington, DC: Transportation Research Board, National Research Council.

Durkin, M., J. McElroy, H. Guan, W. Bigelow, and T. Brazelton. 2005. Geographic analysis of traffic injury in Wisconsin: Impact on case fatality of distance to Level I/II trauma care. *Wisconsin Medical Journal* 104: 26-31.

Elliott, M.R., P.F. Waller, T.E. Raghunathan, J.T. Shope, and R.J.A. Little. 2000. Persistence of violation and crash behavior over time. *Journal of Safety Research* 31: 229–42.

Evans, L. 2004. *Traffic Safety*. Bloomfield Hills, MI: Science Serving Society.

Farmer, E., and E. Chambers. 1926. A psychological study of individual differences in accident rates. *Report of the Industrial Fatigue Research Board*, 38, London.

Gary, S., L. Schulte, L. Aultman-Hall, M. McCourt, and N. Stamatiadis. 2003. Consideration of driver home county prohibition and alcohol-related vehicle crashes. *Accident Analysis and Prevention* 35: 641-48.

Gebers, M.A. and R.C. Peck. 2003. Using traffic conviction correlates to identify high accident-risk drivers. *Accident Analysis and Prevention* 35: 903-12.

General Accounting Office. 2004. *Highway Safety: Federal and State Efforts to Address Rural Road Safety Challenges*. GAO-04-663. Washington, D.C.: United States General Accounting Office.

Gonzalez, R.P., G.R. Cummings, H. A. Phelan, M.S. Mulekar, and C.B. Rodning. 2009. Does increased emergency medical services prehospital time affect patient mortality in rural motor vehicle crashes? A statewide analysis. *The American Journal of Surgery* 197: 30-34.

Gutmann, M.P. and P.C. Stern (eds.). 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington, D.C.: National Research Council.

Hadayeghi, A., A.S. Shalaby, and B.N. Persaud. 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. *Transportation Research Record: Journal of the Transportation Research Board* 1840: 87-95.

Health and Human Services. 2003. *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule*. NIH Publication 03-5388. U.S. Department of Health and Human Services.

Iversen, H. and T. Rundmo. 2002. Personality, risky driving and accident involvement among Norwegian drivers. *Personality and Individual Differences* 33: 1251-63.

Johnson, S.W. and J. Walker. 1996. *The Crash Outcome Data Evaluation System (CODES).* Report DOT HS 808 338. National Highway Traffic Safety Administration.

Kam, B.H.. 2003. A disaggregate approach to crash rate analysis. *Accident Analysis and Prevention* 35: 693-709.

Karlson, T.A, W. Bigelow, and P. Beutel. 1998. *Serious Lower Extremity Injuries from Motor Vehicle Crashes, Wisconsin 1991–1994*. National Highway Traffic Safety Administration.

Kerns, T., and C. Burch. 2006. Impaired driving - combining crash and citation data. Paper presented at the 32nd Annual Traffic Records Forum, Palm Desert, California.

Kim, K., T. Kerns, T. Hettinger, and M. Pease. 2001. *Geographic Information Systems Using CODES Linked Data (Crash Outcome Data Evaluation System).* Report DOT HS 809 201. National Highway Traffic Safety Administration.

Lai, C.-H., W.-S. Huang, K.-K. Chang, M.-C. Jeng, and J.-L. Doong. 2006. Using data linkage to generate 30-day crash-fatality adjustment factors for Taiwan. *Accident Analysis and Prevention* 38: 696-702.

Lajunen, T. 2001. Personality and accident liability: Are extraversion, neuroticism, and psychoticism related to traffic and occupational fatalities? *Personality and Individual Differences* 31: 1365-73.

Levine, N. 2007. *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations* (v3.1). Ned Levine Associates, Houston, TX, and the National Institute of Justice, Washington, DC.

Li, M.-D., J.-L. Doong, K.-K. Chang, T.-H. Lu, and M.-C. Jeng. 2008. Differences in urban and rural accident characteristics and medical service utilization for traffic fatalities in less-motorized societies. *Journal of Safety Research* 39: 623-30.

LII. 2009. *U.S. Code Collection.* Legal Information Institute, Cornell University Law School. http://www4.law.cornell.edu/uscode/18/2721.html

Lovegrove, G.R., and T. Sayed. 2006. Using macrolevel collision prediction models in road safety planning applications. *Transportation Research Record: Journal of the Transportation Research Board* 1950: 73-82.

Lovegrove, G.R., and T. Sayed. 2007. Macrolevel collision prediction models to enhance traditional reactive road safety improvement programs. *Transportation Research Record: Journal of the Transportation Research Board* 2019: 65-73.

Lyons, R.A., H.Ward, H. Brunt, S. Macey, R. Thoreau, O.G. Bodger, and M. Woodford. 2008. Using multiple datasets to understand trends in serious road traffic casualties. *Accident Analysis and Prevention* 40: 1406-10.

Marottoli, Richard A., Leo M. Cooney, D. Raye Wagner, John Doucette, and Mary E. Tinetti. 1994. Predictors of automobile crashes and moving violations among elderly drivers. *Annals of Internal Medicine* 121: 842-46.

Marques, P.R., A.S. Tippetts, and R.B. Voas. 2003. Comparative and joint prediction of DUI recidivism from alcohol ignition interlock and driver records. *Journal of Studies on Alcohol* 64: 83-92.

Meuleners, L.B, A. Harding, A.H. Lee, and M. Legge. 2006. Fragility and crash over-representation among older drivers in Western Australia. *Accident Analysis and Prevention* 38: 1006-10.

Moellering, H. 1974. *The Journey to Death: A Spatial Analysis of Fatal Traffic Crashes in Michigan*. Michigan Geographical Publication No. 13. Ann Arbor, Michigan: Department of Geography, University of Michigan.

National Highway Traffic Safety Administration. 1996. *So You Want to Link Your State Data*. Report DOT HS 808 426. National Association of Governors' Highway Safety Representatives, National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. 1998. *Traffic Records Advisory*. Washington, DC: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. 2003. *Model Minimum Uniform Crash Criteria (MMUCC) Guideline*. Second Edition. Washington, DC: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. 2004. *Initiatives to Address Improving Traffic Safety Data*. Washington, DC: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. 2006. *Traffic Records Program Assessment Advisory*. Washington, DC: National Highway Traffic Safety Administration.

National Safety Council. 1997. *A National Agenda for the Improvement of Highway Safety Information Systems*. Itasca, IL: National Safety Council.

Noland, R.B. 2003. Medical treatment and traffic fatality reductions in industrialized countries. *Accident Analysis and Prevention* 35: 877-83.

O'Sullivan, D., and D.J. Unwin. 2002. *Geographic Information Analysis*. Hoboken, NJ: John Wiley and Sons.

Rosman, D.L. 2001. The Western Australian Road Injury Database (1987–1996): ten years of linked police, hospital and death records of road crashes and injuries. *Accident Analysis and Prevention* 33: 81-88.

Rosman, D.L., A.M. Ferrante, and Y. Marom. 2001. A linkage study of Western Australian drunk driving arrests and road crash records. *Accident Analysis and Prevention* 33: 211-20.

Rothenberg, H. 2008. Beyond crash data: using multiple datasets in safety analyses. Paper presented at the ITE Technical Conference and Exhibit, Miami, Florida.

Rushton, G. 2007. Privacy and confidentiality in health GIS. Paper presented at the ESRI Health GIS Conference, Scottsdale, Arizona.

Rushton, G., R. Krishnamurthy, D. Krishnamurthi, P. Lolonis, and H. Song. 1996. The spatial relationship between infant mortality and birth defect rates in a U.S. city. *Statistics in Medicine* 15: 1907-19.

Sauter, C., S. Zhu, S. Allen, S. Hargarten and P.M. Layde. 2005. Increased risk of death or disability in unhelmeted Wisconsin motorcyclists. *Wisconsin Medical Journal* 104: 39-44.

Shope, J.T., T.E. Raghunathan, and Sujata M. Patil. 2003. Examining trajectories of adolescent risk factors as predictors of subsequent high-risk driving behavior. *Journal of Adolescent Health* 32: 214-24.

Smith, R., L.J. Cook, L.M. Olson, J.C. Reading, and J.M. Dean. 2004. Trends of behavioral risk factors in motor vehicle crashes in Utah, 1992-1997. *Accident Analysis and Prevention* 36: 249-55.

Souleyrette, R., D. Plazak, T. Strauss, M. Imerman, and T. Johnson. 1998. *Improved Employment Data for Transportation Planning*. Ames, IA: Center for Transportation Research and Education.

Staplin, L., and K.W. Gish. 2005. Job change rate as a predictor for interstate truck drivers. *Accident Analysis and Prevention* 6: 1035-39.

Steil, D., and A. Parrish. 2008. Reducing traffic fatalities using GIS data as a deployment guide. Paper presented at the 2008 ESRI International User Conference, San Diego, California.

Tillmann, W.A. and G.E. Hobbs. 1949. The accident-prone automobile driver. *American Journal of Psychiatry* 106: 321-31.

Tindale, S.A. and P. Hsu. 2005. Crash data and signal coordination: A one-way pair case study. *Journal of Safety Research – Traffic Records Forum Proceedings* 36: 481-82.

Willett, T.C. 1964. *Criminal on the Road: A Study of Serious Motoring Offences and Those Who Commit Them.* London: Tavistock Publications.