May 2005

Highway Economic Requirements System: Safety Model Assessment

Prepared For:	Federal Highway Administration Office of Legislation and Strategic Planning 400 Seventh Street, S.W. Washington, DC 20590

Prepared By:	John A. Volpe National Transportation Systems Center Office of System and Economic Assessment 55 Broadway Cambridge, MA 02142

Acronym List

Acronym	Full Name
AADT	Average Annual Daily Traffic
AASHTO	American Association of State Highway Transportation Officials
AMF	Accident Modification Factors
CODES	Crash Outcome Data Evaluation System
CRF	Crash Reduction Factors
FHWA	Federal Highway Administration
GAO	General Accounting Office
GES	General Estimating System
HERS	Highway Economics Requirements System
HPMS	Highway Performance Monitoring System
HSIS	Highway Safety Information System
HSM	Highway Safety Manual
HSRC	Highway Safety Research Center
IHSDM	Interactive Highway Safety Design Model
MVMT	Million Vehicle Miles Traveled
NHTSA	National Highway Traffic Safety Administration
SPF	Safety Performance Functions
TFHRC	Turner-Fairbank Highway Research Center

Contents

1. Basic Issues in Predicting Crash Costs

There are many reasons to be concerned with estimating the frequency and social costs of highway accidents, but most reasons are motivated by a desire to minimize these costs to the extent feasible. Competition (choices) for scarce resources is a practical necessity, and society seeks to apply those resources where they will do the most good. With highway crashes, the problem is that results can only be predicted with a probability; thus, for accident reduction, the problem is to generate sound information without ever knowing for sure if it is correct.

Causes of Highway Accidents

The HERS model applies crash prediction equations in the context of deciding which kinds of highway improvements are justified to which sections of highway. Thus it is concerned with the effects of geometric attributes on expected highway accidents. This does not imply that driver behavior or vehicle characteristics are irrelevant, only that geometric attributes must be a factor.

The roll of geometric attributes is of interest to others besides those choosing among alternative highway investments. Engineers designing highways, communities wanting to reduce the hazard they encounter, and policymakers directing research funding can draw upon knowledge of the contribution of geometric attributes to accidents. Public programs concerned with reducing driver error, and manufacturers trying to build safer vehicles both need information beyond geometric factors, but that does not reduce the importance of being able to diagnose the effects—independent as well as interactive—of geometric properties.

Relationship Between Highway Attributes and Crashes

In the HERS model, only the attributes of the highway sections are used in the determination of the expected safety costs; however, the highway attributes are only one category of factors that can combine to produce circumstances that lead to a motor vehicle crash. Crash causes are generally divided into three categories, in the following order: driver factors, roadway factors, and vehicle factors.[1]

Driver factors involve the actions taken by or the condition of the driver of the motor vehicle, including speeding, violating traffic laws, driving under the influence of alcohol or drugs, inattention, decision errors, and age. Roadway factors that contribute to, or are associated with, crashes include roadway design attributes (i.e. number of lanes, lane width, median width, shoulder width, presence of curves/grades/intersection), roadside hazards (i.e., poles, trees, animals, or embankments adjacent to the road), and roadway conditions (i.e., weather conditions, lighting conditions). Vehicle factors include any vehicle-related failures that may exist in the automobile or design of the vehicle.

1. GAO (2003), Sabey and Staughton, Treat

Distribution Among Causes

An obvious question raised in the previous section is, what percentage of crashes result from each of these categories? To answer this question, we turn to the body of research on the subject. Unfortunately, the body of research specifically focused on the overlapping impacts of vehicle, driver, and highway causal factors is largely composed of two studies completed in the 1970's. Recent research has mainly focused on analyzing the effect of a specific factor(s) (i.e., speeding, alcohol, access control, etc.) in crashes. While there are new studies underway which focus on collecting and analyzing crash causation data (Large Truck Crash Causation Study, One Hundred-Car Naturalistic Driving Study, Drive Atlanta Study), at this time, there are no results available.[2]

According to the 2003 GAO report on traffic crash causation:

> "One of the most significant studies to date on the factors that contribute to motor vehicle crashes was the Tri-Level Study of the Causes of Traffic Accidents, conducted in the 1970s by the Indiana University at Bloomington Institute for Research in Public Safety. According to NHTSA officials, the Tri-Level study has been the only study in the past 30 years to collect large amounts of on-scene crash causation data. To provide researchers with insight into the factors that contribute to traffic crashes, collision data were collected on three levels, each providing an increasing level of detail, including 13,568 police reported crashes; 2,258 crashes investigated by on-scene technicians; and 420 crashes investigated in depth by a multidisciplinary team. The study assessed causal factors as either definite, probable, or possible. The study found that crashes were caused by human (or driver-based) factors, environmental (roadway or weather-related) factors, or vehicle-related factors."

As shown in Figure 1, driver factors are the primary cause of the largest percentage of motor vehicle crashes, followed by roadway and then vehicle factors.

Figure 2 provides a different view of the data from the Tri-Level study.[3] It clearly defines the percentage of crashes due solely to roadway, driver, and vehicle-related factors as well as the percentage of crashes resulting from a combination of these factors. It should be noted that roadway factors are associated with approximately 34% of all crashes.

Another similar study was also performed by Sabey & Staughton in Great Britain in the 1970's, and the results (also shown in Figure 2) are very similar to those of the Tri-Level study.

The most pertinent result from each of these studies is the role that roadway factors play in motor vehicle crashes. In most cases where a crash occurs, a roadway design feature is not the single, definite cause of the crash. Instead, it is generally the behavior of the driver that leads to a crash; however, that does not mean roadway attributes

---

2. GAO (2003)
3. Treat

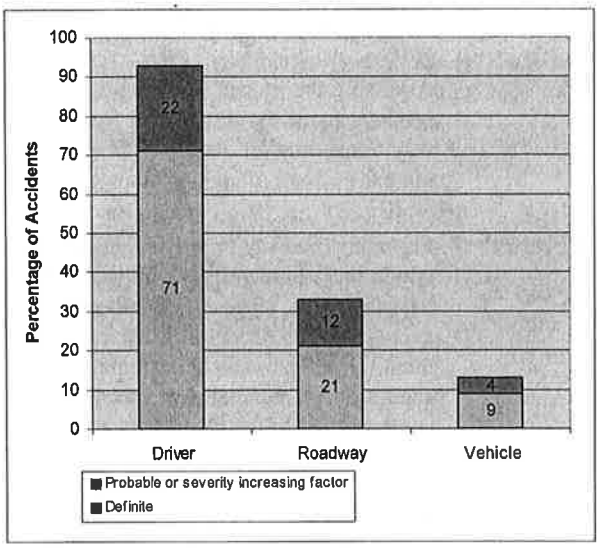

Figure 1. Crash Causes Found by Tri-Level Study

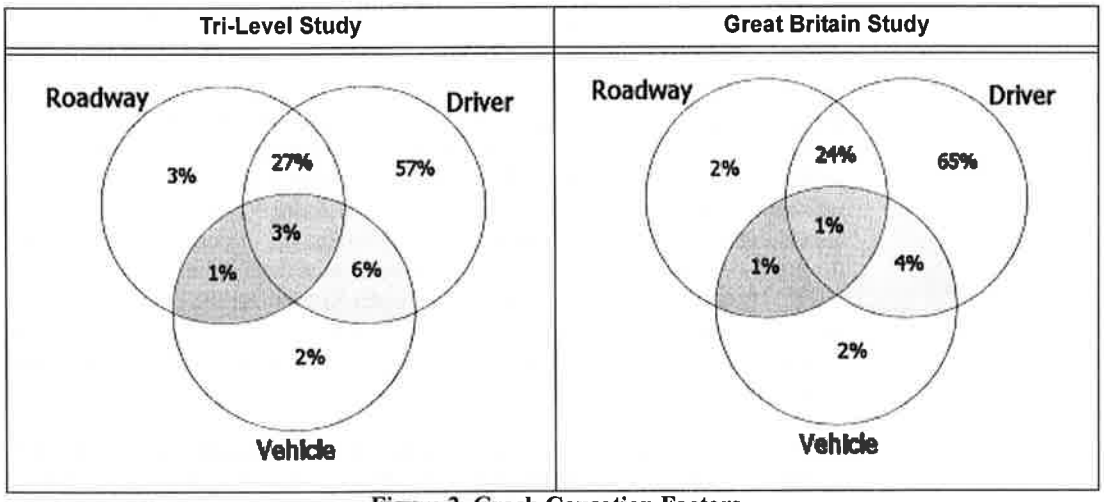

Figure 2. Crash Causation Factors

do not play a role in crashes. In fact, in approximately 27-34% of the crashes, it was a combination of roadway factors and driver factors that lead to the crash.

Crash Modeling Strategies

Assessing the impact of changes in roadway geometric characteristics is a common problem faced by virtually all federal, state, and local agencies responsible for highway transportation. At this time there are two main approaches that are used to estimate the effects of making changes to roadway sections on crashes.

Effectiveness Rates

For many roadway improvements and countermeasures, research has been conducted which seeks to identify the impact of the improvement or countermeasure on crashes. In many cases, ranges of effectiveness rates have been determined which quantify the average change in crashes that can be expected to accompany a particular improvement. These rates can be extremely specific in that they may depend on what the configuration of the roadway was before the improvement as well as what the final configuration of the roadway is after the improvement.

Given an effectiveness rate, determining the safety cost savings from an improvement is a fairly straightforward task. First, an average number of crashes needs to be determined for the roadway section being modified. Second, the effectiveness rate is multiplied by this average to determine the number of crashes prevented by the improvement. Once the number of crashes is estimated, crash severity averages are used in order to predict the number of fatalities and injuries resulting from those crashes. With the expected number of fatalities and injuries, average costs can be applied which yield the total costs of crashes occurring on a particular section.

Count Models

Count models directly estimate the number of crashes that are expected on a particular section of the roadway based on the geometric and traffic characteristics of that section. These models are typically developed using large amounts of crash and roadway inventory data. The most common functional forms for these regression models are either Poisson or Negative Binomial depending on the dispersion of the data about the mean (i.e. sample mean = sample variance). Once the number of crashes are predicted, crash severity averages are used in order to estimate the number of fatalities and injuries resulting from those crashes. With the expected number of fatalities and injuries, average costs can be applied which yield the total costs of crashes occurring on a particular section.

One of the issues with the count model approach is that typically the crash prediction and severity calculation are broken up into separate steps. Typically the crash prediction is performed by the regression model, and the severity calculation uses national averages to apportion the crashes into the categories of fatal, injury, and property damage only. From this point, more national averages are used to estimate the actual number of fatalities and injuries. While the national fatalities averages are good estimators, national injury averages are no longer published by FHWA in their annual report covering highway usage statistics. Additionally some researchers question the division between crash prediction and severity. While fatality crashes are far too rare to estimate by themselves, some work has been done to estimate property damage only crashes separately from fatality and injury crashes.

Interaction Effects

A second issue with this approach is in the actual development of the regression models. The process generally involves identifying a set of variables that are to be evaluated for their statistical significance in explaining the variation in the dependent variable. While this process is suitable, what is missing is more up front analysis of the independent variables to be evaluated.

Overall, there has been very little consideration given to the interaction between independent geometry variables. More specifically, cross-product terms are virtually nonexistent in most crash prediction models. This is quite surprising since virtually all research into the causes of crashes generally indicates that multiple factors are associated with the occurrence of a crash. Furthermore, this recommendation was made in a separate report by the FHWA in the early 1990's.[4]

Vehicle Mix

Another example of the lack of up front analysis involves the use of aggregate variables when disaggregate data are available. For example, every crash prediction model has some exposure variable, typically AADT. While exposure is a necessary variable in any model, not enough consideration is given to more disaggregate exposure variables, such as commercial vehicle AADT and passenger vehicle AADT. This approach is also supported by what is known regarding the impact of vehicle mix on crash rates.

Relationships Among Frequency, Severity, and Cost

"Frequency" is the rate at which crashes occur, generally in terms of number per 100-thousand vehicle miles of travel; "severity" is the level of damage with respect to fatalities, injuries, and property damage per incident, while "cost" is the value of the resources used to correct or compensate for the damage.

These results (severity and cost) do not occur with a fixed proportion per crash. For example, an increase in volume for a given capacity forces vehicles into closer proximity than at lower volumes, and more crashes occur; higher volume also decreases speed, however, resulting in lower severity and fewer fatalities.

The simplification of modeling crashes times fatalities per crash, then, is clearly only an approximation if it is assumed that the two rates are independent. One strategy is to model frequency and severity separately, but using the most appropriate variables (such as speed) for each model. Some variables may appear in both. An alternative approach is to model frequency and severity simultaneously.

---

4. Cirillo

2. HERS Crash Estimation Models

The Highway Economic Requirements System (HERS) is an engineering/economics model designed to estimate investment requirements for the nation's highways. To estimate future investment requirements, HERS uses an extensive set of data on a sample of highways throughout the nation (Highway Performance Monitoring System (HPMS) Sample data) to conduct project-level benefit-cost analyses of alternative improvements. The model evaluates potential improvements on each sample highway section by comparing their construction costs with the benefits accruing to highway users and agencies (i.e. reductions in travel times, vehicle operating costs, safety, etc.) to determine whether an improvement is warranted.[5]

How HERS Predicts Crash Cost Benefits

To estimate the highway user benefits associated with a particular highway improvement, HERS makes extensive use of statistical prediction models. These models calculate benefits by using highway geometric design (i.e. number of lanes, lane width, median width, presence of curves/grades/intersection) and traffic attributes of the highway section as input to the statistical models, with the output being safety, travel time, and operating costs. As improvement alternatives are "implemented" in the model, the design attributes of a highway section (i.e. widening a road, adding a lane, etc.), the highway user costs change as well.[6]

HERS uses a three-step process to calculate the total safety costs for a particular improvement alternative. The three steps are discussed in further detail in the subsequent sections. Prior to discussing the models, it is necessary to review the two methods of classifying highway sections that are used by these models.

Facility Type and Functional Class

The most common method of classifying highway sections is to group them according to the type of service or function they provide. This method assigns each highway section to one of the following general categories, which are known as *functional classes*.

- Principal Arterials - carry long-distance traffic to/from significant traffic generators
- Minor Arterials - carry shorter distance traffic to/from lesser traffic generators
- Collectors (Major & Minor) - carry traffic to/from residential or rural areas to higher functional classes

5. Camus (2000)
6. GAO (2000)

- Locals - carry traffic to/from adjacent properties and to higher functional classes

The functional class attribute indicates whether the highway section is located in a rural or urban area as well. Figure 3 shows the hierarchy of functional class values.

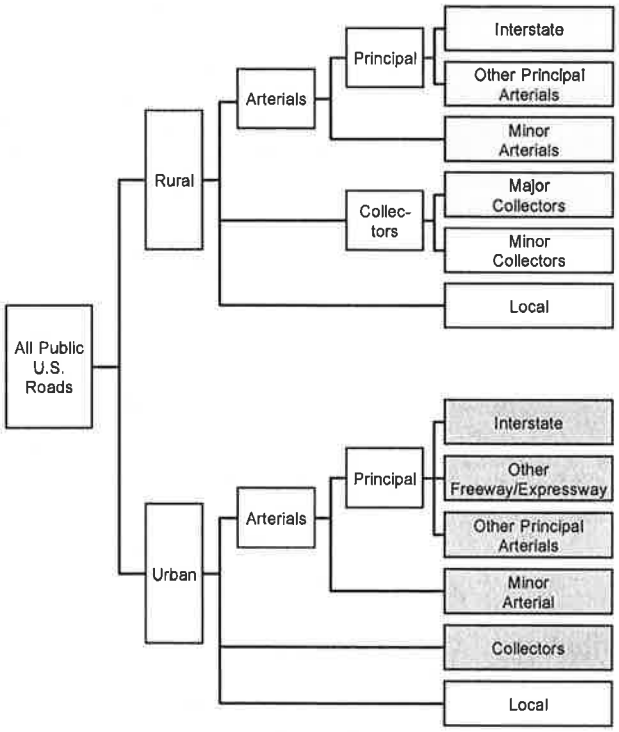

Figure 3. Functional Class Hierarchy

The second method HERS uses to classify highway sections is based on the design attributes of the highway. This method assigns each highway section to one of the following general categories, which are known as *facility types.*

- Freeways - includes all divided sections with full access control and two or more lanes per direction
- Multilane sections - includes all sections with two or more lanes per direction that do not meet the criteria for a freeway
- Two-lane sections - includes all sections with two or fewer lanes per direction

HERS Crash Estimation Models

Six different methods are used by HERS for calculating the expected number crashes per 100 million vehicle miles traveled (100 MVMT). These different methods are based on the facility type and whether the section is in a rural or urban area. A brief overview of each model is provided below.

- Rural Two-Lane Roads - This method has four composite models (four legged signalized intersections, four- and three-legged intersections with stop control on the minor approach, and non-intersections) that calculate the expected crash rate by decomposing the highway section into sub-sections based on their proximity to an intersection. The output from each of the component models is combined to create the expected crash rate for the complete section. This model, which incorporates over 15 different geometric attributes, was developed for the FHWA using negative binomial regression analyses of crash data from four states.
- Rural Multilane Roads - This model was also developed using negative binomial regression, and it incorporates 9 geometric attributes.
- Rural Freeways - This exponential model estimates the crash rate using AADT and lane width as the only input variables.
- Urban Freeways - This fifth-order polynomial model also uses only AADT and lane width to estimate the crash rate.
- Urban Multilane Surface Streets - This exponential model uses AADT and the number of signals per mile to estimate the crash rate
- Urban Two-Lane Streets - This model is log-linear in AADT as was developed by ordinary least squares. This is the least developed of all the HERS crash estimation models.

Severity Distribution

Once the expected crash rate is computed by the crash estimation models, this rate can be converted to the expected number of crashes by multiplying the number of vehicle miles traveled. At this point it is necessary to estimate the expected number of fatalities and injuries for the section of highway being analyzed. HERS uses fatality and injury rates (fatalities per crash and injuries per crash), which are different for each functional class, to estimate the number of fatalities and injuries for a given highway section. These rates were developed by using national level crash, fatality, and injury data.

Unit Costs by Severity

Finally, the expected number of fatalities and injuries for the highway section in question are converted to costs. Again HERS uses fixed unit costs (cost/fatality, cost/injury, cost/property damage only) to convert the number of fatalities and injuries to an overall safety cost for the particular highway section.

Impacts of Highway Improvements on Crash Rates

While it is important to have safety models that accurately capture the relationships between roadway attributes and crashes, it is also necessary to understand how roadway attributes are affected by the various improvement projects modeled by HERS. Table 1 shows which section attributes can be modified as a part of an improvement project. An important point to take away from this, is that while a variable may be a factor in crashes and appear in a crash prediction model, that does not imply that the variable is ever modified in a HERS improvement project. If significant relationships between crashes and particular attributes currently not modeled in HERS are discovered, it may be necessary to modify other component models of HERS in order to take advantage of this discovery.

Table 1. Section attributes potentially affected by a safety improvement

Section Attribute	Possible Changes
Number of Lanes	increase or no change
Lane Width	meet design standard or no change
Shoulder Type	existing or minimum tolerable condition, or no change
Right Shoulder Width	meet design standard or no change
Pavement Condition	recalculate
Pavement Thickness	recalculate
SN or D	increase or no change
Surface Type	meet design standard
Peak Capacity	recalculate or no change
Median Width	meet design standard or feasible
Median Type	unprotected, none, or no change
Access Control	full or partial
Grades	meet design standard
Curves	meet design standard
Passing Sight Distance	improve to average or no change
Weighted Design Speed	recalculate
Widening Feasibility	lower code or no change

Review of the HERS Crash Frequency Models

A review was undertaken for each of the six HERS crash prediction models. To develop these models, original crash prediction model literature was reviewed and pertinent models were incorporated into HERS. Table 2 provides a summary of the review of the HERS crash prediction models. The columns in the table describe the original research used to develop the model, the data used to develop the model and the characteristics of the model.

Table 2: Summary of Accident Predictions Models used in HERS

Facility Type	Basis	Data	Model Characteristics
Rural Two Lane Roads	1998 work by Vogt and Bared, "Accident Models for Two-Lane Rural Roads: Segments and Intersections"	Minnesota (1985 - 1990) and Washington (1993 - 1995) data from the Highway Safety Information System was used. This includes 1,300 segments and 700 intersections.	An extended negative binomial model was developed for: • Non-Intersections A negative binomial model was developed for: • Signalized Intersections • Non-Signalized 4-Legged Intersections • Non-Signalized 3-Legged Intersections
Rural Multi-Lane Roads	1998 work by Wang, Hughes, and Stewart, "Safety Effects of Cross-Section Design of Rural Four-Lane Highways"	Minnesota (1990) data was used in the model development. There were 622 segments covering 431 miles.	A single model was developed using the method of Poisson regression.
Rural Freeways	This model was derived from 1992 work by Persuad, "Roadway Safety - A review of the Ontario Experience and Relevant Work Elsewhere."	1987 data provided by the Ministry of Transportation of Ontario.	Persuad's original model was an exponential model with AADT as the only explanatory variable. The explanatory variable lane width was incorporated by the HERS team resulting in the following model: $CRASH = 17.64 \cdot AADT^{0.155} \cdot e^{(0.0082(12-LW))}$
Urban Freeways	This model was derived from 1996 work by Richard Margiotta, "Incorporating Traffic Crash and Incident Information into the Highway Performance Monitoring System Analytical Process."	Aggregate HPMS data was used to develop the original model.	Margiotta's original model was fifth order polynomial with AADT as the independent variable. The explanatory variable lane width, was incorporated (as with rural freeways) by the HERS team resulting in the following model: $CRASH = 1.54 - 1.203 \cdot ACR + 0.258 \cdot ACR^2$ $-0.0000524 \cdot ACR^5 \cdot e^{(0.0082 \cdot (12-LW))}$
Urban Multi-Lane Surface Streets	This model was derived from 1996 work by Richard Margiotta, "Incorporating Traffic Crash and Incident Information into the Highway Performance Monitoring System Analytical Process". This model, however, used data from the 1994 work by Bowman and Vecellio, "Effect of Urban and Suburban Median Types on Both Vehicular and Pedestrian Safety."	The accident data came from on-site collection (by video tape) of fifteen arterials in Atlanta, Phoenix, Los Angles, and Pasadena. This data was collected prior to 1994 and included 46.2 miles of urban arterials.	The model is a multiplicative model with AADT and number of signals per mile as the independent variables. The model takes the following form with different variables for *a, b,* and *c* depending on the type of section. $CRASH = a \cdot AADT^b \cdot NSIGPM^c$

Table 2: Summary of Accident Predictions Models used in HERS (Continued)

Facility Type	Basis	Data	Model Characteristics
Urban Two Lane Streets	Model was developed by HERS team.	A table of AADT and Crashes per 100 Million VMT was used to develop the model. The table was populated by HPMS data and HERS crash rates.	The model is an ordinary least squares regression with AADT as the independent variable. $CRASH = 19.6 \cdot \ln(AADT) + 7.93 \cdot (\ln(AADT))^2$

Although the crash frequency prediction models used in HERS were developed ten or more years ago, not much research has occurred since then that would warrant replacing the existing equations with improved versions. Nonetheless, the HERS equations are weak in several respects in light of current ideas on crash modeling:

- Geometric properties are missing from many equations that probably should include geometric attributes as instrumental variables, notably the 2-lane 2-way urban streets model that has no geometric attributes at all.
- Data used to fit some of the equations are thin and perhaps unrepresentative; models may have been fitted to data from a single state, without testing the model against other data.
- Changes in crash rates caused by an improvement on a section sometimes are the result of a change in facility type (e.g., adding lanes or changing access control), leading to a different crash estimation equation. There has been no coordination among the equations, however, to ensure that the differences in the resulting crash rates are a reflection of real safety improvements rather than artifacts of the equations.
- For some of the models, the methodology and theory used to design and fit the equations is below current standards for generating crash prediction equations, such that some equations could be improved (at least in the statistical sense) by refitting the equations to the same data.

3. Highway and Crash Data Sources

In contrast to the limited amount of empirical research on highway crash models that has been reported recently, the amount and quality of data available for crash modeling has been steadily improving. A finite number of such data sets exist at the national and state level, and these are described below.

Highway Attributes

Comprehensive data on highway section attributes do not include crashes, and comprehensive crash data do not include all sections or all section attributes. Thus it may become necessary to link databases, either for research or for application, or at least to transfer results from one database to another.

Highway Performance Monitoring System

Once every two years, the FHWA is required (by Congress) to create Conditions and Performance reports, which describe the current status and future needs of the road systems in the United States. These reports provide Congress with the information necessary to appropriate funds to individual states for highway maintenance and development. Originally, these reports were generated through extremely labor intensive special studies, which gathered data from each state, analyzed the highway systems, and then created the reports. In 1978 the FHWA streamlined the process with the creation of the Highway Performance Monitoring System, which replaced the biennial process of gathering highway data. The HPMS standardized the data collected by each State about the highway systems (i.e. conditions, performance, use, geometry, etc.), and it stored the data for all States in a central repository. This system also requires each State to report data annually so that the system is kept up-to-date.[7]

The HPMS data are organized into two different groups, the Universe Data and the Sample Data. The Universe data contain basic highway information (e.g., AADT, Functional Class, Number of Lanes, Pavement Roughness, etc.), which states are required to report for every highway section in the United States. The Sample data are composed of a statistically chosen sample of 10% of roadways from all functional systems with the exception of local roads (urban and rural) and rural minor collectors. For each highway section in the Sample, an additional set of highway information (48 additional attributes) is collected. These additional data include geometric attributes such as access control, median/shoulder/lane width, curves, grades, and traffic attributes such as speed limit, capacity, K-factor, and percent trucks. Table 3 provides a comparison of the centerline miles and number of sections in the HPMS Universe and Sample aggregated by functional class.

The HPMS Sample data are used as the base data for HERS, which provides input to the Congressional Conditions and Performance reports. The HPMS data are also used

[7] FHWA

Table 3. Comparison of Sample and Universe Data

Functional Class	Sample Miles	Universe Miles	Sample Sections	Universe Sections
Rural Interstate	17,005	32,078	7,333	20,216
Rural Other Principal Arterial	26,076	97,087	10,366	83,201
Rural Minor Arterial	15,269	135,664	5,759	108,923
Rural Major Collector	18,736	424,667	7,340	264,170
Rural Minor Collector	0	267,793	0	5,755
Rural Local	0	2,079,000	0	10,635
Urban Interstate	8,494	14,691	9,159	23,221
Urban Other Freeway and Expressway	4,857	9,930	5,391	16,409
Urban Other Principal Arterials	12,833	57,256	22,283	144,283
Urban Minor Arterial	13,822	94,769	24,340	212,489
Urban Collector	11,687	98,323	21,272	21,5811
Urban Local		678,589		6,918
Totals	**128,779**	**3,989,847**	**4,118,626**	**1,112,031**

throughout the transportation planning community for research and planning purposes.

While the Sample contains enough geometric data to input into the HERS crash models, these data are not used in the assessment of the models. The major problem is that neither the Universe nor the Sample HPMS data record the number of crashes occurring on a highway section, which is the key piece of data required to assess the existing crash models and even develop new models. It is for this reason that the HPMS data cannot be used for estimating crash rates, although HPMS data might be used for extracting geometric properties not included in accident data. Consideration should be given to adding historical accident data to the HPMS.

Accident Data

Highway Safety Information System

The Highway Safety Information System (HSIS), operated by the University of North Carolina Highway Safety Research Center (HSRC) and LENDIS Corporation, under contract with FHWA, is a multistate database that contains crash, roadway inventory, and traffic volume data for a select group of States. The participating States - California, Illinois, Maine, Michigan, Minnesota, North Carolina, Utah and Washington were selected based on the quality of their data, the range of data available, and their ability to merge data from the various files.[8]

8. HSIS Web Site

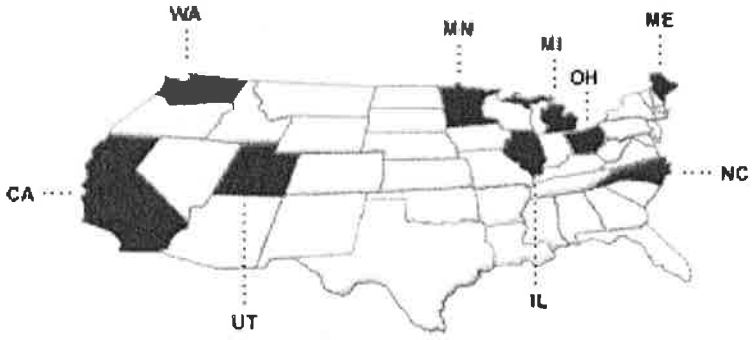

Figure 4. HSIS Participating States†

†. Highway Safety Information System Web Site

Due to contractual obligations with the States, the data in their entirety cannot be distributed; however, subsets of the data can be made available upon request. Therefore, after reviewing the attributes available for each of the participating States, the data in Table 4 were requested.

Table 4. Requested HSIS Data

State	Years
Ohio	1997 - 1999
Minnesota	1996 - 1997
California	1996 - 2000
North Carolina	1996 - 1999
Washington	1996, 1999 - 2002
Michigan	1996 - 1997

Data from the other HSIS States were not requested because of the lack of attribute data to support the evaluation of the existing crash estimation model (e.g. no curve data implies the two-lane rural model cannot be evaluated). The subset of attribute data requested was chosen to allow the existing models to be applied to the data, and also to allow some exploratory analysis of the correlation between various geometric attributes and crashes.

Fatality Analysis Reporting System

The Fatality Analysis Reporting System (FARS) was conceived, designed, and developed by the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA) in 1975 with the following goals:

- to provide an overall measure of highway safety,
- to help identify traffic safety problems, to suggest solutions, and

- to help provide an objective basis to evaluate the effectiveness of motor vehicle safety standards and highway safety programs.

FARS contains data derived from a census of fatal traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. To be included in FARS, a crash must involve a motor vehicle traveling on a roadway customarily open to the public and result in the death of a person (occupant of a vehicle or a non-motorist) within 30 days of the crash.[9]

Currently, the FARS data are not incorporated into this analysis for two major reasons. First, as its name implies, FARS only records crash data where at least one fatality occurs. All injuries and fatalities associated with a fatal crash are recorded in FARS; however, these data are only sufficient to build models predicting fatal crashes since injury crashes are not recorded in this data source or any other known data source. The second issue limiting the usefulness is the lack of roadway geometric attributes. Only a very limited number of geometric attributes are recorded in the FARS data. To address this issue, in 2000, FARS incorporated Geographic Information System (GIS) technology into the data collection system. While this allows the crash data to be linked to other GIS-based data sources (e.g., NHPN/HPMS and state highway inventory database), there is still a significant amount of work required to acquire these data sources and to link the crash data with the roadway inventory data. For these reasons, FARS will not be used to develop statistical crash prediction models. At this point, the only potential use for this data is updating the expected number of fatalities by functional class statistics.

General Estimating System

Developed in 1998 by NHTSA, the National Automotive Sampling System General Estimates (GES) provides annual national level estimates of motor vehicle crashes and the factors that contribute to those crashes. These estimates are developed from a random sample of about 50,000 police accident reports collected from 400 police jurisdictions in 60 areas. The areas and police jurisdictions are chosen so that they properly reflect geography, roadway mileage, population, and traffic density, and so that the police accident reports can be used to estimate national results. The national level estimates as well as the sample are available for analysis from the National Center for Statistics and Analysis (NCSA).

Crash Outcome Data Evaluation System

Originally conceptualized by NHTSA to report to Congress the benefits of safety belts and motorcycle helmets, a Crash Outcome Data Evaluation System is a comprehensive system linking police reported crash data with hospital recorded injury data. Currently, NHTSA has at least partially funded developed of these systems in thirty states.

Police reports alone do not provide enough information regarding the types and severity of injuries sustained as a result of a motor vehicle crash. By linking police reports with additional data sources (shown in Figure 5), this system provides a wealth of additional outcome data such as:

9. FARS Overview Web Site

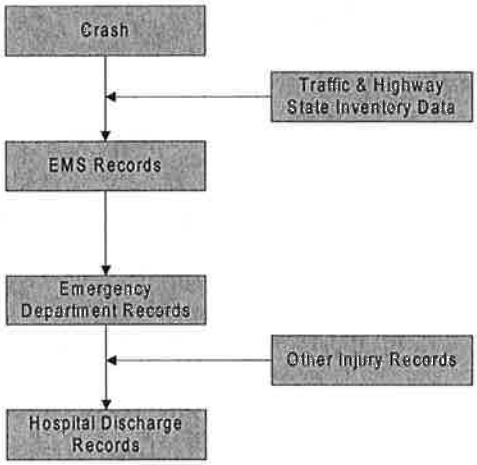

Figure 5. Possible Data Sources for CODES

- specific type of injury - head, neck, back, lower extremities, etc.
- severity of injury - requires hospitalization, intensive care, etc.
- cost of injury - hospital cost of treating injury
- medical system response - EMS response time, transfer time, hospitalization time, etc.

At this point, just over half of the states have developed CODES, and there is no standardized data model or national level data that can be used for analysis.

Summary of Data Sources

A limited number of data sources can be tapped for highway crash analysis, but the HSIS data developed from state accident sources has developed to the point that rich data sets are available from a handful of states that offer the potential for major improvements in empirical crash estimation models. To the extent that new models reveal relationships among accident characteristics, additional geometric attributes and historical crash data might be added to the HPMS sample sections.

4. Recent Research on Geometric Effects

The primary goal of this review effort is to assess the HERS crash prediction models and develop recommendations for improving these models. In order to assess the existing models it is necessary to understand how the HERS models differ from those being developed through current research efforts. It is possible that recent research has developed crash prediction models that are more accurate than those currently utilized by HERS. This field of research has a significant body of previous work as well as a number of recent efforts. This section first describes some of the major safety projects currently under development as well as some previous efforts to develop crash prediction models for specific entities.

Previous Research

AASHTO Tools

The American Association of State Highway Transportation Officials has developed two different tools for state officials in the area of crash prediction models.

User Benefit Analysis for Highways. This manual provides users with guidance for estimating the benefits that accrue to roadway users as the result of roadway improvement projects. One of the sections is devoted to estimating the safety benefits that may result from a highway improvement. This section provides a brief overview of crash prediction methodology and other resources available to a transportation official, including Highway Safety Manual, Interactive Highway Safety Design Model, SafetyAnalyst, and others. One of the resources discussed in depth is the Roadside Safety Analysis Program (RSAP) that is a companion analytic tool to their Roadside Design Guide.

Roadside Design Guide. First published in 1999, this guide provides users with a synthesis of current information and operating practices related to safety treatments that minimize the likelihood of fatality or serious injury when a driver runs off the road. Developed under NCHRP project 22-9, RSAP allows users to compare the cost-effectiveness of implementing multiple alternative roadside safety improvements. This program estimates accident costs based on roadway and roadside design features.

To estimate the safety impact of roadside improvement projects, RSAP first estimates the number of occurrences of a vehicle departing from the roadway (called encroachments). The second step in the model is to estimate the number of crashes, which are occurrences of a vehicle striking another vehicle or object. The attributes of the roadway (design speed, curves, grades, etc.) are the major inputs to the encroachment and accident models. The accident model also uses the number of encroachments as an input variable. Once the number of accidents are determined, the severity of the accidents are determined through averages and units costs per fatality and injury.

Safety Effectiveness of Highway Design Features

Completed in 1992 this compendium, which was prepared for the FHWA, reports the most probable safety effects of improvements to key highway design features, including:

- Volume I - Access Control
- Volume II - Alignment
- Volume III - Cross Sections
- Volume IV - Intersections
- Volume V - Interchanges
- Volume VI - Pedestrians and Bicyclists.

This compendium was developed as a result of the FHWA implementing one of the 23 recommendations contained in Transportation Research Board Special Report 214, "Designing Safer Roads - Practices for Resurfacing, Restoration, and Rehabilitation."[10] These reports are comparable in structure and type of information that will be contained in the Highway Safety Manual (see "Highway Safety Manual" on page 22) chapter on Knowledge, although the HSM will have more recent results.

Current Research

Interactive Highway Safety Design Model

The Interactive Highway Safety Design Model (IHSDM) is being developed by the Turner-Fairbanks Highway Research Center (TFHRC), which is home to the Federal Highway Administration's (FHWA's) Office of Research, Development, and Technology. IHSDM is a suite of decision-support modules (Crash Prediction, Design Consistency, Intersection Review, Policy Review, and Traffic Analysis) for evaluating safety and operational effects of geometric design decisions in the highway design process. It compares existing or proposed highway designs against relevant design policy standards and estimates expected safety and operational performance of the design. The current version of IHSDM models only rural two-lane highways; however, by 2007 the application will be expanded to include rural multilane highways and urban/suburban arterials.[11]

Crash Prediction Module. The crash prediction module in the IHSDM performs a similar function to that of the HERS crash prediction models. Like the HERS models, the IHSDM crash algorithm estimates the baseline expected crash rate for a highway section based on its geometric design and traffic attributes. In fact, they both use the exact same statistical model developed by Vogt and Bared in 1998.[12] The generalized IHSDM algorithm, however, augments the statistical base models with a number of additional inputs that are intended to adapt the base estimates according to local safety conditions. The additional steps in the algorithm, which can be applied to any

10. Cirillo, Zeeger, Twomey, Kuciemba
11. ISHDM Web Site
12. Vogt and Bared (1998)

type of crash prediction model, are shown in Figure 6 and are discussed in the subsequent paragraphs.

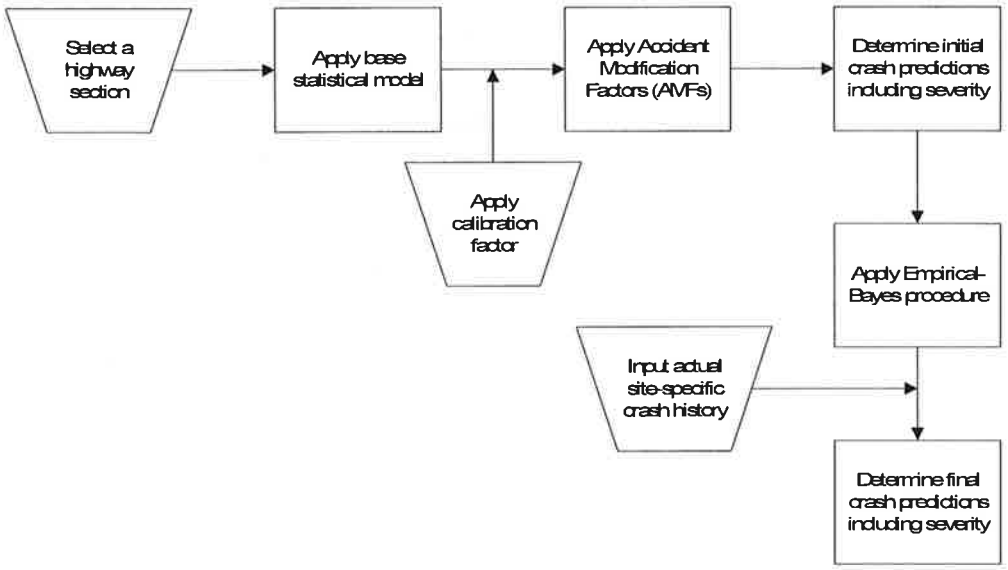

Figure 6. IHSDM Generalized Crash Prediction Algorithm

Since states differ markedly in climate, animal population, driver populations, crash reporting threshold, and crash reporting practices, and these variations may result in some states experiencing substantially more reported traffic crashes on rural two-lane highways than others. Once the base statistical models are applied to the highway section data, the results can be calibrated (increased or decreased by a multiplicative factor) by State or local agencies.

The accident modification factors (shown in Table 5) adjust the calibrated base model estimates for individual geometric design element dimensions and for traffic control features. The factors are the result of an expert panel review of related research findings and consensus on the best available estimates of quantitative safety effects of each design and traffic control feature.

The final step in the algorithm applies an empirical bayes procedure for weighted averaging of the algorithm estimate with project-specific crash history data. The weights used for the predicted and actual crashes are calculated from the overdispersion parameter of the base statistical model used to estimated the predicted number of crashes.

Table 5. Accident Modification Factors

Roadway Segments	At Grade Intersections
Lane Width	Skew angle
Shoulder Width and Type	Traffic control
Grades	Exclusive left-turn lanes
Driveway Density	Exclusive right-turn lanes
Two-way left-turn lanes	Intersection sight distance
Passing lanes/short four-lane sections	
Roadside design	
Horizontal Curves • length • radius • presence or absence of spiral transitions • superelevation	

SafetyAnalyst

SafetyAnalyst is also a decision-support application, being developed through a cooperative effort of the FHWA and thirteen state highway agencies. Unlike the IHSDM, which is applied at the project level, this application is intended to support a system-wide program of site-specific highway safety improvements.[13] SafetyAnalyst is composed of four modules:

- Network screening - This module will identify highway sites (spot locations as well as highway sections of varying length) that exhibit higher-than-expected crash frequencies, high crash severities, and high proportions of specific crash types.
- Diagnosis and countermeasure selection - This module will use collision diagrams and crash statistics at a particular site to identify specific safety issues and a set of countermeasures that could mitigate those issues.
- Economic appraisal and priority ranking - This module will use default and user provided cost data along with crash prediction estimates and countermeasure specific accident modification factors to estimate the benefits of a countermeasure for a specific site. This data will be used in the priority ranking algorithm for prioritizing safety improvements at multiple sites throughout the network.
- Evaluation of implemented improvements - This module will use the Empirical Bayes statistical approach on actual crash and traffic volume data to assess the actual impact of implemented improvements.

Safety Performance Functions. The network screening, economic appraisal and evaluation modules in SafetyAnalyst will use safety performance functions (SPF; also called crash estimation models) to estimate the expected number of crashes at a spe-

[13] SafetyAnalyst Web Site

cific site.[14] This estimate, which is adjusted for recent crash history with the Empirical Bayes approach, will be used in the following applications:

- The network screening module will compare the observed and expected crash frequencies to identify sites with higher-than-expected crash frequencies.
- The economic appraisal module will apply accident modification factors (i.e. effectiveness rates) for alternative countermeasures to the expected crash frequencies in order to estimate the safety benefits.
- The evaluation module will compare the observed crash frequency after the improvement to the observed crash frequency before the improvement in order to assess the actual impact of the improvement.

While the user of SafetyAnalyst will have the option of providing safety performance functions, a standard set of functions are being developed for this application. An interim version of the SafetyAnalyst application is being released at the end of 2004 in order to collect feedback for the final version release in 2006. The interim will contain a complete set of interim safety performance functions; however, these functions were developed to predict crash frequency using AADT as the only explicit explanatory variable. Different crash prediction models were developed for numerous categories of highway sites, as shown in Table 6.

Neither the categories nor the SPFs are intended to be the final versions, and SPFs with additional explanatory variables will be developed and incorporated into the final version of SafetyAnalyst.

Highway Safety Manual

The Highway Safety Manual is a Transportation Research Board initiative to provide the best factual information and tools in a useful and widely accessible form and to facilitate roadway design and operational decisions based upon explicit consideration of their safety consequences.[15] This manual would greatly strengthen the role of safety in road planning, design, maintenance, construction, and operations decision making. The HSM is organized into five parts:

- Introduction and Fundamentals - Outlines the purpose and uses of the HSM in addition to discuss the fundamental concepts in safety analysis (e.g. crash counts, safety performance functions, crash modification factors, etc.)
- Knowledge - This section outlines the known relationships between safety and highway attributes, including:
  - specific highway design elements (e.g. shoulders, curbs, medians, alignment, and guardrails),
  - operational elements (e.g. speed, rumble strips, signs, lighting, weather, etc.),
  - intersections and interchanges, and

14. Harwood (2004)
15. Hughes

Table 6. Interim SafetyAnalyst SPF Categories

Roadway Sections	Intersections	Ramps
Rural two-lane	Rural three-leg intersections with minor-road STOP control	Rural diamond off-ramps
Rural multilane divided	Rural three-leg intersections with signal control	Rural diamond on-ramps
Rural multilane undivided	Rural four-leg intersections with minor-road STOP control	Rural parclo loop off-ramps
Rural freeway - 4 lanes	Rural four-leg intersections with all-way STOP control	Rural parclo loop on-ramps
Rural freeway - 6+ lanes	Rural four-leg intersections with signal control	Rural free-flow loop off-ramps
Rural freeway within an interchange - 4 lanes	Urban three-leg intersections with minor-road STOP control	Rural free-flow loop on-ramps
Rural freeway within an interchange - 6+ lanes	Urban three-leg intersections with signal control	Rural direct or semidirect connection ramps
Urban two-lane arterials	Urban four-leg intersections with minor-road STOP control	Urban diamond off-ramps
Urban multilane divided	Urban four-leg intersections with all-way STOP control	Urban diamond on-ramps
Urban multilane undivided	Urban four-leg intersections with signal control	Urban parclo loop off-ramps
Urban one-way arterials		Urban parclo loop on-ramps
Urban freeway - 4 lanes		Urban free-flow loop off-ramps
Urban freeway - 6 lanes		Urban free-flow loop on-ramps
Urban freeway - 8+ lanes		Urban direct or semidirect connection ramps
Urban freeway within an interchange - 4 lanes		
Urban freeway within an interchange - 6 lanes		
Urban freeway within an interchange - 8+ lanes		

  - special facilities (i.e. grade crossings, work zones, bridges, tunnels, etc.).

- Predictive Methods - This section develops crash prediction models for the following types of roadways:
  - Rural, Two-lane Roads,
  - Rural, Multilane Highways, and
  - Urban and Suburban Arterial Highways.
- Safety Management of a Roadway System - This section discusses approaches for prioritizing and selecting improvement projects. This follows the same methodology as the SafetyAnalyst application.
- Safety Evaluation - This section discusses how to measure the actual effectiveness of an implemented improvement.

The first edition of the HSM is scheduled to be completed in 2007; however, a draft chapter on rural, two-lane roads is currently available. The content of the draft chapter has not been fully approved by the project sponsor and is subject to change before the final version is released.

A more recent series of guidebooks to assist state and local agencies in reducing injuries and fatalities in target areas are currently being develop under NCHRP Project 17-18(3). Each guidebook corresponds to one of the 22 key emphasis areas (shown in Table 7) that are outlined in AASHTO's Strategic Highway Safety Plan. Each guidebook contains a general discussion of the problem as well as strategies and countermeasures to address the problem.

Table 7. Elements of AASHTO Strategic Highway Safety Plan

Instituting Graduated Licensing for Younger Drivers	Making Truck Travel Safer
Ensuring Drivers are Fully Licensed and Competent	Increasing Safety Enhancements in Vehicles
Sustaining Proficiency in Older Drivers	Reducing Vehicle-Train Crashes
Curbing Aggressive Driving	Keeping Vehicles on the Roadway
Reducing Impaired Driving	Minimizing the Consequences of Leaving the Road
Keeping Drivers Alert	Improving Design and Operation of Highway Intersections
Increasing Driver Safety Awareness	Reducing Head-on and Across Median Crashes
Increasing Seat Belt Usage and Improving Airbag Awareness	Designing Safer Workzones
Making Walking and Street Crossing Safer	Enhancing Emergency Medical Capabilities to Increase Survivability
Ensuring Safer Bicycle Travel	Improving Information and Strategic Support Systems
Improving Motorcycle Safety and Increasing Motorcycle Awareness	Creating More Effective Processes and Safety Management Systems

Future Research

This section covers some of the in-progress work that will provide results for some of the major safety projects discussed in the previous section.[16]

Crash Reduction Factors

NCHRP Project 17-25: "Crash Reduction Factors for Traffic Engineering and ITS Improvements"

16. Transportation Research Board Web Site

The objective of this project is to develop reliable crash reduction factors (CRFs) for traffic engineering, operations, and ITS improvements. Crash reduction factors (also known as accident reduction factors or accident modification factors) provide a computationally simple and quick way of estimating crash reductions. Many states have a set of crash reduction factors that are used for estimating the safety impacts of various types of engineering improvements, encompassing the areas of signing, alignment, channelization, and other traffic engineering treatments. Typically, these factors are computed using before-and-after comparisons, although later research has suggested the use of cross-sectional comparisons. The estimated completion date of this effort is July 31, 2005, and the researching agency is the University of North Carolina - Chapel Hill.

Urban Arterials

NCHRP Project 17-26: "Methodology to Predict the Safety Performance of Urban and Suburban Arterials"

The objective of this project is to develop a methodology that predicts the safety performance of non-limited-access urban and suburban arterials and to prepare a chapter on urban and suburban arterials for inclusion in the Highway Safety Manual. This project will analyze the various elements (e.g., lane width, shoulder width, use of curbs) considered in planning, design, and operation of non-limited-access urban and suburban arterials. The estimated completion date of this effort is October 31, 2005, and the researching agency is the Midwest Research Institute.

Rural Multilane Highways

NCHRP Project 17-29: "Methodology to Predict the Safety Performance of Rural Multilane Highways"

The objectives of this research are to develop a methodology to predict the safety performance of rural multilane highways and to prepare a chapter on rural multilane highways for inclusion in the Highway Safety Manual. The methodology will apply to both highway segments and at-grade intersections but does not include full access-control highways. The estimated completion date of this effort is June 30, 2006, and the researching agency is the Texas A&M Research Foundation.

Conclusions

It should be noted that this section does not include all work relating to highway safety, as this is a large area of research; however, this section is intended to communicate the major direction of the work in this field of research. Based on the information presented here, highway safety research is focused in two major areas: development of improved regression models, and development of improved data regarding countermeasures and their effectiveness.

While there are efforts underway to develop crash prediction models developed for specific geographic areas, facility types, or functional classes, the general direction of the field is toward a more comprehensive process surrounding the estimation of crashes and the effectiveness of any countermeasures. These more comprehensive

processes build on base statistical models by incorporating adjustments for recent crash history, state to local level model calibration, and general crash modification factors.

A more comprehensive process for predicting crashes definitely improves the predictive power of the models at the local level; however, it really does not add a lot of value for the HERS model. One reason is the lack of national level data required for the additional steps in the crash prediction process. For instance, the HPMS does not require states to submit the actual number of crashes on roadway segments; therefore, empirical-bayes steps cannot be implemented. Furthermore, some steps in the process are not really intended for use at the national level, such as the calibration of the model to local conditions.

Little in the current body of research can be directly integrated in the HERS models. The crash prediction models developed by SafetyAnalyst are functions of only a single variable, AADT, and the research on crash prediction models is a year or more away from completion. In addition, very little effort is focused on the urban two-lane and urban multi-lane road facility types. This is unfortunate given the fact that the HERS crash prediction functions for these facility types are more in need of updating than the other facility types which are receiving more attention. It is for these reasons, that it was deemed necessary to acquire and analyze state inventory and crash data for the purpose of upgrading the urban two-lane crash prediction function currently utilized in the HERS model.

In order to ensure that the needs of the HERS model for crash cost estimation are met, it is essential that the HERS team participate actively in the development of suitable models.

5. Urban Two-Lane Streets

HERS predicts annual urban two-lane crash rates as a polynomial function of that section's daily traffic, a model form which appears overly simplistic in light of the poor quality of the predictions and the purpose of HERS as a safety cost/benefit model for roadway improvements. Yet, a review of published urban two-lane crash prediction studies suggests that crash prediction models for urban two-lane crash streets have not received a great deal of research. (As one example, a technical memorandum for FHWA's SafetyAnalyst[17] application describes its urban two-lane crash model, also defined strictly in terms of vehicle AADT, as flawed and a necessary research subject.) The impression is that improvements to the quality and usefulness of the HERS urban two-lane street model would also advance the general body of knowledge.

The data explorations and estimation model described by this section suggest that impressive gains in both roadway section-level crash prediction accuracy and insight can be realized, not by novel statistical methods, but by considering the *combined effects* of the roadway's geometric features and car or truck traffic (HSIS data permits distinct analyses of the two). The first volume in the Safety Effectiveness of Highway Design Features series (Cirillo 1992) paraphrases a series of studies: *"Of importance in the [Cribbins, et al] work was the consistent finding that combinations of geometric and traffic characteristics had a more significant impact on accidents than any single variable and Cribbins et al recommend against further research into the effects of single variables."*[18]

Osculations between traffic and roadway geometry are generally absent from published crash models. Prior studies generally presume roadway geometry provides no information about daily traffic and vice versa. Capturing interaction effects does not require novel or complex statistical methods. The model proposed in this section is an example of the Generalized Linear Model (GLM) technique, the method adopted by Vogt and Bared (1998) on rural two-lane roads. Subject to further validation, this approach could be applied to the entire suite of HERS crash models.

The Current HERS Crash Model

Crash Equation

HERS estimates a section's crash rate (annual crashes per 100 million vehicle miles traveled-denoted by *CRASH*) as a quadratic function, fit using ordinary least squares regression with *ln*(AADT) as the lone predictor[19]:

[17] Harwood (2004)

[18] Cirillo

[19] Camus

$$CRASH = 0.8743 \cdot (-19.6 \cdot \ln AADT + 7.93 \cdot \ln(AADT)^2) \quad [1]$$

Literature review uncovered very little in the way of data exploration or other studies that motivated this model. The HERS documentation only alludes to the data, four mean value data points from noisy data, used to fit the model and provides no reference documenting the development of this model.

Accuracy

To establish current performance, the HERS equation was applied to three years (1997-1999) of HSIS Ohio urban two-lane data. (Recall, the Ohio HPMS Sample data does not include crashes.) The HSIS data comprises Ohio's entire urban two-lane section inventory, including annual AADT. Consequently, equation [1] accuracy is assessed over nearly 15,000 data points, representing more than 18,000 accidents over three years. Figure 7, which plots annual crash rate estimates via equation [1] (curve) against observed rates (scatterplot) illustrates the poor fit of the current model. The scatterplot offers weak evidence of higher crash rates as AADT increases, a pattern presumed by equation [1][20]. The U-shape apparent in the scatterplot is actually an artifact of the section length term in the rate calculation, and will be discussed in greater detail in "Effect of Section Length" on page 32.

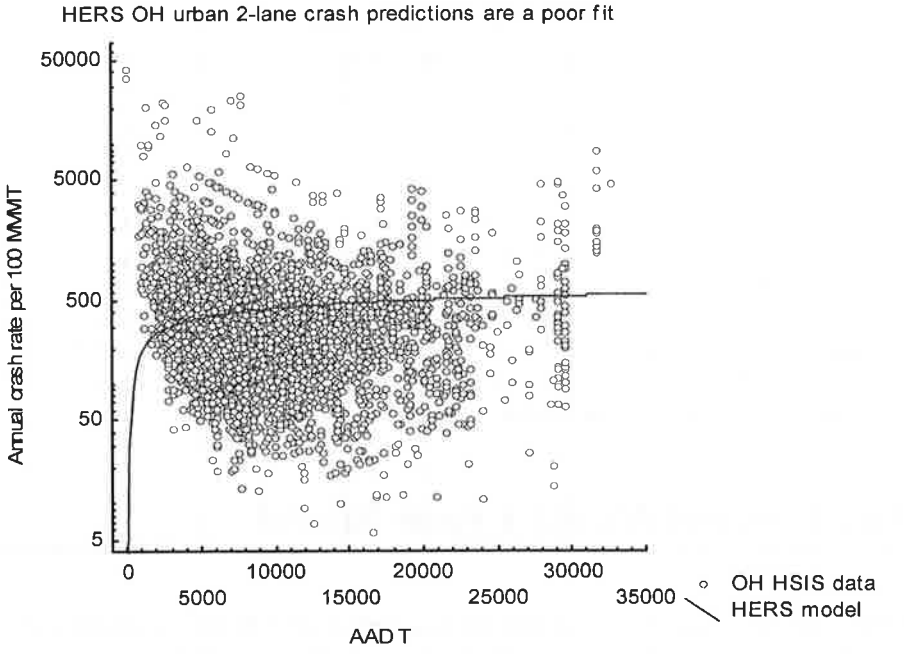

Figure 7. OH Two-Lane Road Actual Crash Data

[20] This same exercise was repeated with Minnesota HSIS data, and the resulting scatterplot is very similar. It is omitted to avoid confusion; that graph would be the only reference to Minnesota data in this report, as analyses to date have concentrated on Ohio data.

The poor fit translates into gross overestimates of Ohio urban two-lane crashes rates (Table 8).

Table 8. HERS Urban two-lane prediction errors

Predicted crashes	Actual crashes	Predicted fatalities	Actual fatalities	Cost error ($M)
75,078	18,225	186	97	$1,189

The HERS model predicts four times the number of urban two-lane crashes as actually occurred. As a result, HERS urban two-lane safety cost estimates are grossly overstated; the crash rates predicted by equation [1] overstate fatality, injury and property damage costs by $1.18 *billion* over the three years. To arrive at this figure, the current HERS three-stage method for estimating safety costs was followed. First, each section's estimated annual crash rate was converted into its annual crash count equivalent. From these, annual fatalities and injuries were estimated. Finally, economic costs per fatality ($3 million), per injury ($35,750) and property damage per accident ($6,900) were applied. The expected number of fatalities and injuries per crash were estimated from the following conversion factors: 0.00247 (fatalities/crash) and 0.34485 (injuries/crash), respectively. Data exploration considered several aspects of roadway geometry. The empirical evidence suggests that model terms reflecting particular features in combination might more accurately capture traffic and crash patterns than the existing model. Before describing that evidence, the next section discusses how Ohio HSIS data was analyzed and aggregated in order to provide better insights into the relationship between crashes and roadway geometric attributes.

Behavior

Two graphs of equation [1] are depicted in Figure 8. The left-hand graph depicts equation [1], a parabola in *ln*(AADT), while the right-hand graphs that same function in terms of untransformed AADT.

Figure 8 highlights some obvious flaws:

- Both equation [1] terms translate near-zero daily traffic flows into impossibly high annual crashes rates. The first term's negative coefficient translates *smaller* values of AADT < 1 into *more* predicted crashes. (For $AADT < 1$, $ln(AADT) < 0$ and trends towards $-\infty$ as AADT approaches zero vehicles.) The second, quadratic term, combined with the positive coefficient *exponentially* inflates predicted crash rates as AADT tends towards zero (*ln*(AADT) tends towards $-\infty$)!
- The predicted annual crash rate is *negative* for certain volumes less than 20 vehicle daily (to be precise, $1 < AADT < \exp(19.6 \div 7.93) \approx 11.8$ vehicles per day).
- Annual crash rates never approach a limiting value - or even decline - as AADT approaches or exceeds the section's designed capacity. The first deriv-

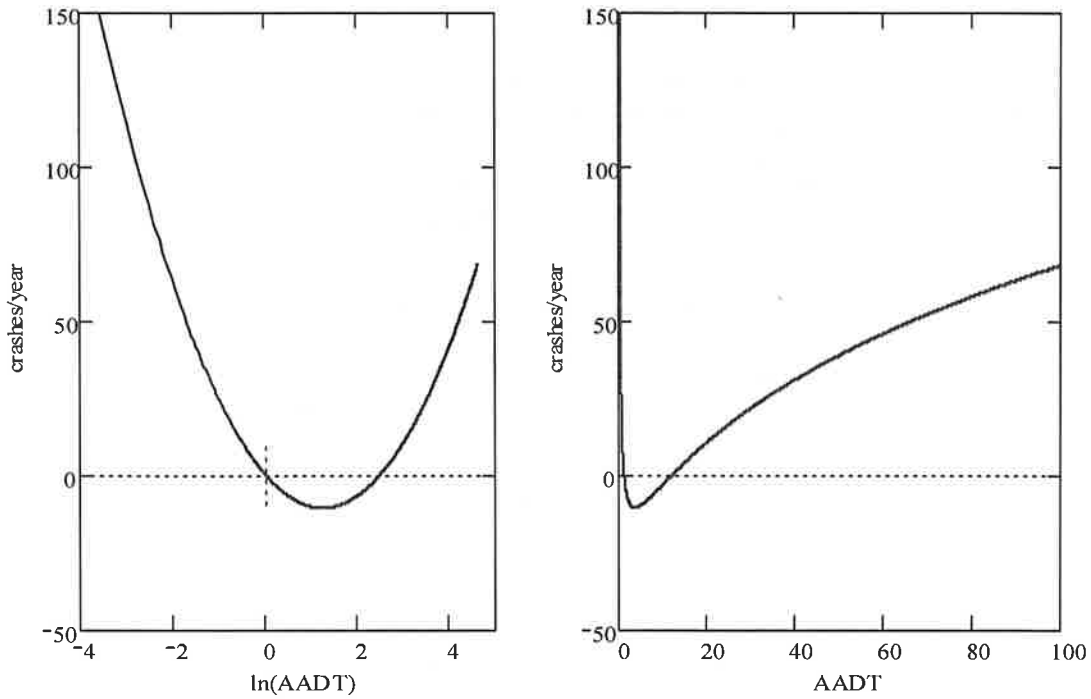

Figure 8. The existing crash equation for urban two-lane streets.

ative of equation [1] is linear in *ln*(AADT) and so each unit increase in *ln*(AADT) always translates into a fixed increase in the annual crash rate. A model unconstrained in section capacity (*v/c*) is contrary to findings relative to urban freeway crash rates versus daily traffic, published in the AASHTO 'Redbook' (suggesting those annual crash rates plateau when section *v/c* > 1.2).[21] According to equation [1], each unit increase in *ln*(AADT) (equivalently, each increase in AADT by a factor of 2.72) equates to an $0.8743 \times 2 \times 7.93 = 13.87$ -fold increase in the annual crash rate!

Preparing and Cleansing HSIS Data

In any given year, a HSIS state supplies detailed geometry of roadway section as well as specifics about crashes (exceeding a property damage threshold). A section's geometry is presented among several data files; typically, there is a data file for the state's entire roadway section inventory, another for the location and geometry of intersections and still other describing curves and grades. Files detailing drivers and vehicles involved in each crash are also provided. As a HERS data assessment, this

[21] AASHTO (2003). Page 5-27, Figure 5-5.

study focuses on the effects of roadway geometry on annual crashes ceteris paribus ("all else being equal"). Consequently, vehicle and driver specifics are not considered.

Location of Crashes and Geometric Attributes

Each HSIS roadway section is identified by its beginning and ending mileposts. Every intersection, curve, grade and crash is identified by specific mileposts. To assess the effects of roadway geometry on crashes, a necessary first step is to properly position each intersection, curve, grade and crash within the correct roadway section. To simplify the process additional lookup tables were constructed which cross-reference each roadway section with its associated intersections, curves, grades and any reported crashes in the given year. Whenever a location (milepost) is strictly between a section's endpoints, the assignment is straightforward. If an intersection, curve, grade or crash milepost coincides precisely at the end milepost of one section and the beginning milepost of the very next, by convention that feature or crash is always assigned to the latter section.

Computation of Non-Inventory Attributes

Each HSIS state has some latitude in how it submits data. For example, Ohio does not explicitly report lane and shoulder widths. Rather, the state reports roadway width defined as the total section width (in feet) excluding medians and surface width defined as the total width (in feet) of travel lanes, excluding shoulders. For urban two lane road sections[22], a raw value for lane width, computed by [2], is rounded up to the nearest foot.

$$LaneWidth = \frac{SurfaceWidth}{2} \quad [2]$$

Total shoulder surface is calculated in [3].

$$ShoulderSurface = RoadwayWidth - SurfaceWidth \quad [3]$$

Since Ohio does not report shoulder width explicitly, shoulders are assumed symmetric in each direction, so that shoulder width is computed according to [4] and rounded up to the nearest foot.

$$ShoulderWidth = \frac{ShoulderSurface}{2} \quad [4]$$

Rounding both lane and shoulder widths up might technically yield 2 additional feet; any rounding errors, however, were ignored because, after a case-by-case review of each computed {lane width, shoulder width} combination, lane widths along sections exceeding 12 feet were adjusted downward according to the assignment shown in Table 9.

[22] A small number of Ohio's urban two-lane inventory possesses medians and these were removed from further consideration.

Table 9. Lane Width Adjustment

Raw Lane Width (feet)	Adjusted Lane Width (feet)
13	12
14 or 15	10
16	11
17 and above	12

Effect of Section Length

Across a section, the annual crash rate (crashes per 100 MVMT) is computed the standard way:

$$Rate = \frac{Crashes}{AADT \cdot SectionLength} \times \frac{10^8}{365} \quad [5]$$

A scatterplot of crash rates (y-axis) versus AADT (x-axis), shown in Figure 9 suggests a U-shaped relationship (i.e. annual crash rates are particularly high at very low and very high traffic levels).

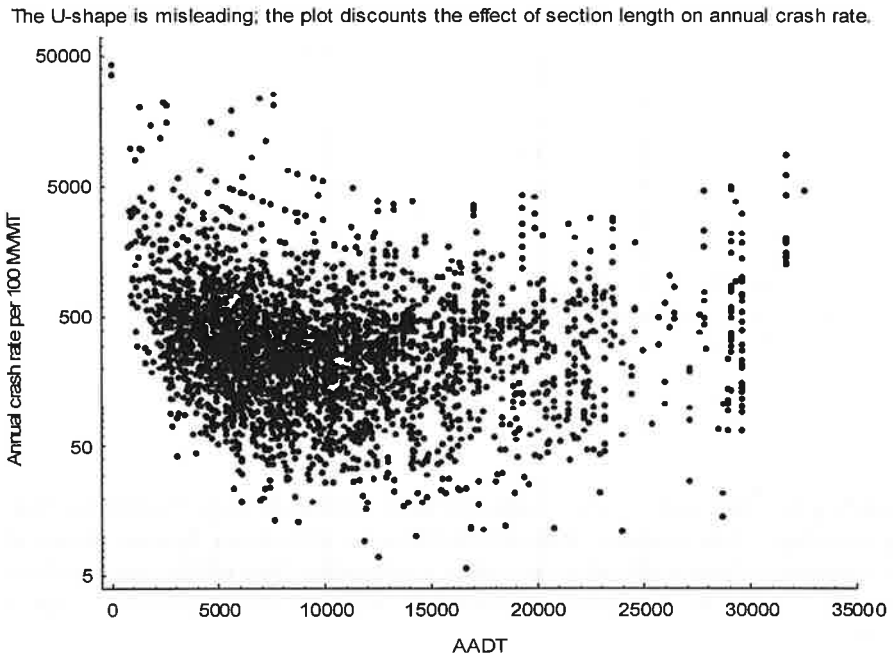

Figure 9. Relationship Between Crash Rate and AADT

Yet, to conclude that an urban two-lane prediction model must capture the U-shape via predictor variables defined such that 'crash rates are highest when traffic is either very light or very heavy' would be misguided. Equation [5] will always assign a

higher crash rate to the shorter of two sections with a single crash and identical AADT. This effect is significantly magnified when section length is under 0.1 mile, as many Ohio sections are. Given two hypothetical sections, alike in every geometrical respect except that one is 0.01 mile long and the other is 0.1 mile long, the former's annual crash rate will be 10 times the latter's. For example, most data to the rightmost side of Figure 9 in the vertical lines, correspond to different sections of the same Ohio road; thus, they share the same AADT (31,600 vehicles) and many geometric features. Yet the computed rates per span 1,250 crashes/year to 8,670 crashes/year because each section is only 0.01 mile to 0.09 mile long.

Ohio DOT may inventory such short sections to isolate specific past improvements between mileposts. When the sections are alike in other geometrical respects, this artifact of section length is undesirable; a priori, section length is not a geometrical feature that influences crash risk. To remove this artifact, consecutive roadway sections no more than 0.25 miles long are combined if each shares traffic levels (both truck and total AADT), lane width, shoulder width and each possesses intersections or is devoid of them.

Less than 1.5% of all Ohio urban two-lane sections can be joined end-to-end. About 1/3 of the resulting segments account for less than 1.5% of all urban two-lane crashes. These percentages are small, but sufficient to remove the effect of section length. In Figure 10, the left-hand graph depicts unaltered Ohio data; crash rates can reach enormous levels due to section length alone. The right-hand graph illustrates how crash rates of the newly-combined sections (circles) are entirely consistent with crash rates for Ohio sections never even candidates for combination (diamonds), namely those exceeding 0.25 miles in length.

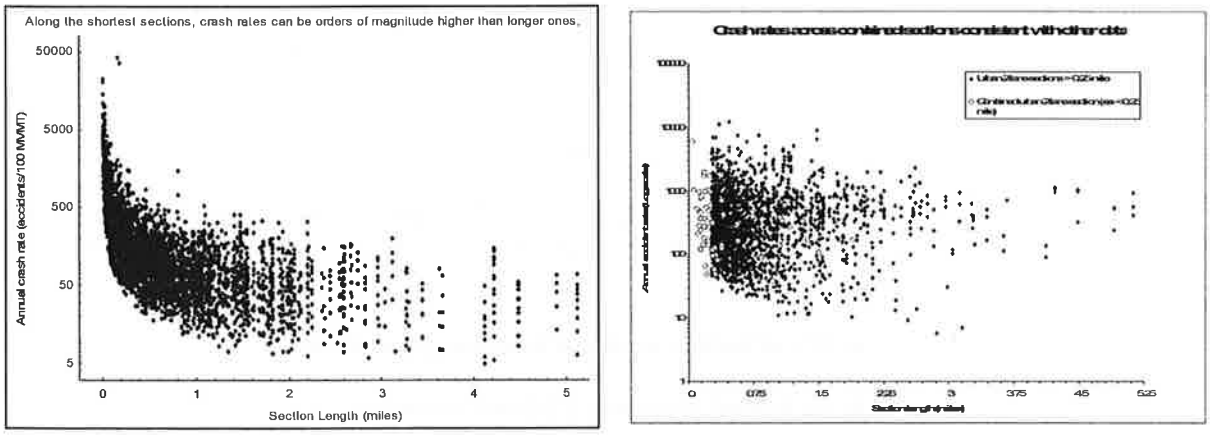

Figure 10. Effects of Combining Short Sections

Since this approach is successful at eliminating the artifact of segment length, it will be adopted in future analyses for other states and possibly other roadway functional classes. Many studies in the literature (e.g., Vogt and Bared 1998) merely discard segments less than 0.1 mile long.

Exploratory Data Analysis

Scatterplots of annual crash rates versus segment length highlight the noise present in the source data (Figure 9 and Figure 10). Estimation methods seek errors which, on average, equal zero and also possess the minimum possible variance. Noise in the source data complicates modeling, possibly manifesting itself as excessively large error variances.

Before fitting models, an extensive exploration of HSIS traffic, geometry and accidents for Ohio urban two-lane segment (spanning 1997-1999) was conducted to validate the hypothesis that HERS model [1] would generate more accurate estimates possessing tighter error variances if *associations* between variables were modeled. Association terms in regression models capture correlations between the independent variables and, as such, are not evidence of cause-and-effect relationships (e.g., lane width and daily car traffic may trend in the same direction, but one does not cause the other).

Defining categories often clarifies patterns and trends in multivariate data. Two two graphical methods are relied upon most heavily in the data exploration which follows: the boxplot and the histogram.

Section geometry and daily traffic

Intersections and AADT. Five bands of total daily traffic (vehicle AADT) are defined which partition Ohio's urban two-lane inventory:

(1) $AADT \leq 1,000$ vehicles,

(2) $1,000 < AADT \leq 5,000$ vehicles,

(3) $5,000 < AADT \leq 10,000$ vehicles,

(4) $10,000 < AADT \leq 20,000$ vehicles

(5) AADT > 20,000 vehicles.

WIthin each category, a segment is further classified by the number of intersections located across it. The histogram of urban two-lane sections by intersection count reveals that:

a) 95% of sections possess at most five intersections,

b) 3% beyond that possess 6-10 intersections,

c) still another 1.3% possess 11-20 intersections which leaves,

d) only 0.4% possessing 21 or more intersections.

Consequently, segments possessing more than five intersections are grouped according to b)-d). Each of the nine categories is labeled by the weighted average number of

intersections possessed by segments in that group. On average, group b) segments possess 7.45 intersections; group c) segments possess 13.84 intersections and group c) segments possess 28.52 intersections. Segments possessing 0-5 intersections are treated as distinct categories. The other six categories are, trivially, the common intersection count shared by all segments.

Boxplot of annual crash counts versus these categorical variables - AADT and intersections - reveal how increases in either variable, particularly in combination, are associated with more crashes (Figure 11).

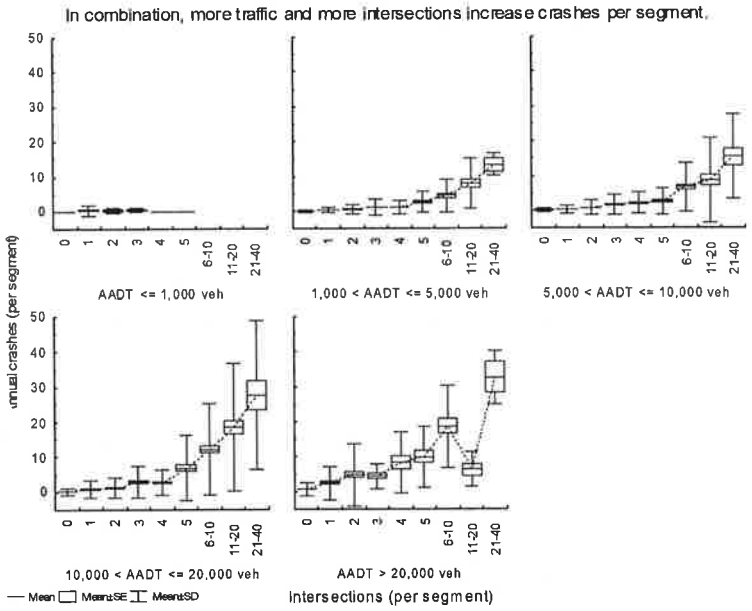

Figure 11. More intersections and higher traffic are associated with more crashes.

The combined effects of the two variables are *not identical*. Between categories changes to both variables simultaneously do not imply equal increases in annual crash counts. This is clear from Figure 12 and Figure 13 in which, at fixed levels of one variable, the boxplot means with respect to the other variable are connected.

From Figure 11 through Figure 13, two conclusions may be drawn: first, for every urban two-lane segment, higher average annual crash counts are associated with higher traffic and more intersections; second, the most successful crash prediction models would treat 'intersection count' as a *categorical*, not a *continuous* variable. A model designed to capture the evident effects across the different variable categories will produce better estimates. By contrast, modeling intersection counts only as a single continuous variable can estimate the combined effects *on average* across the categories. That 'best fit' curve may appear adequate, yet errors that appear modest will result in poor predictions especially if, as HERS model [1] does, the most appropriate model form involves the *log*-transform of traffic variables.

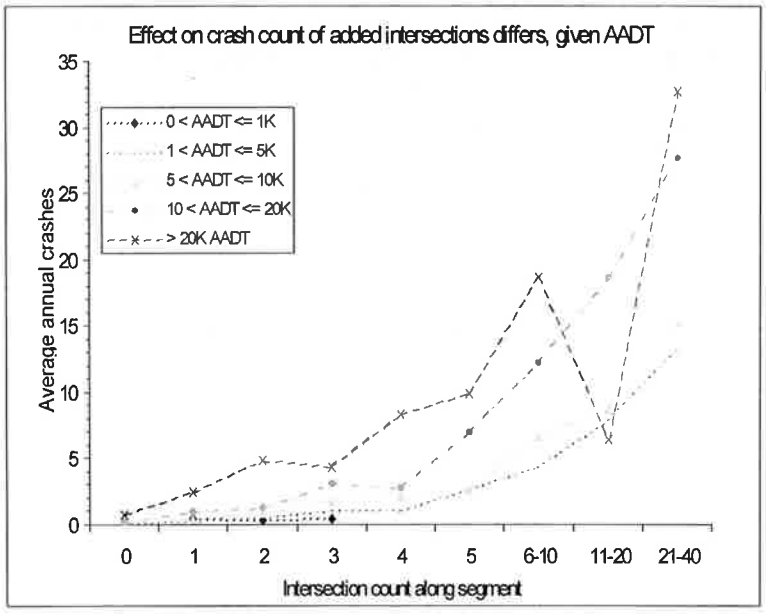

Figure 12. A categorical variable best captures impacts at fixed levels of traffic.

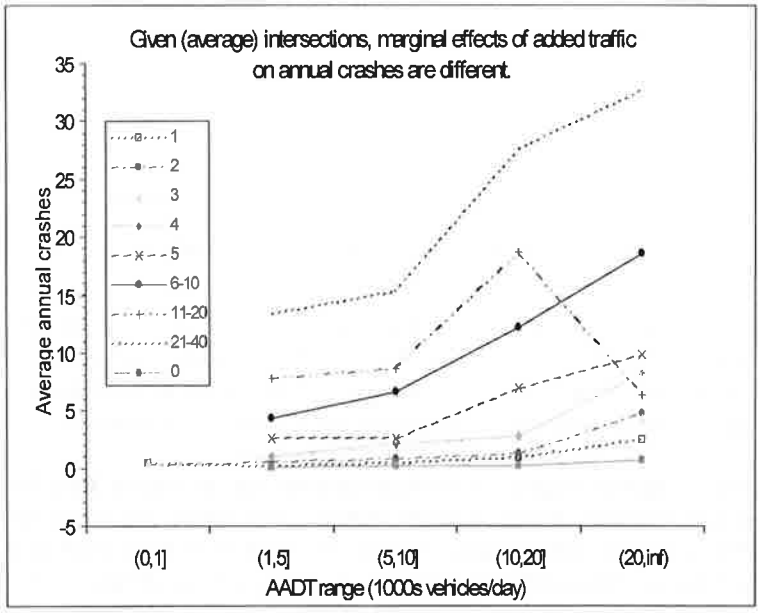

Figure 13. Given intersection label, impact from added traffic differs.

Ohio also reports daily commercial truck traffic in HSIS, that component of urban two-lane traffic is assessed the same way. First, segments are categorized across four ranges of daily truck traffic:

(1) truck $AADT \leq 500$ vehicles,

(2) truck $500 < AADT \leq 1,000$ vehicles,

(3) truck $1,000 < AADT \leq 2,000$ vehicles,

(4) truck AADT> 2,000 vehicles.

Evidently, increases in truck AADT and intersection density across a segment also translate into more annual crashes (Figure 14).

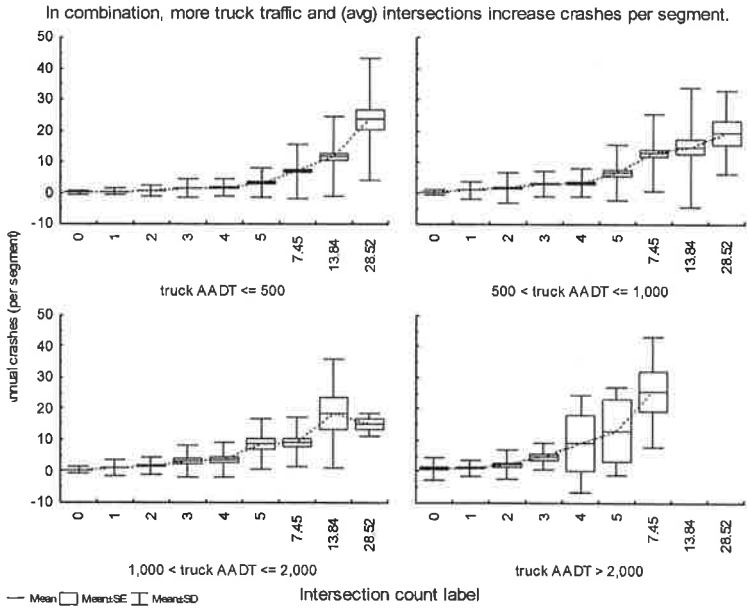

Figure 14. More intersections and truck traffic are associated with more crashes.

Perhaps not surprisingly, the behavior of combined effects of intersection counts and truck AADT (Figure 15 on page 38 and Figure 16 on page 38) are similar to those observed for total AADT. Note that Ohio crashes trend *downward* relative to increased truck AADT along segments with the most intersections (21-40). Perhaps this is suggestive of effective traffic control.

Lane Width and AADT. There are nearly 14,200 records of Ohio urban two-lane segments. Only 40 of these have reported lane widths under 10 feet, so analysis removes them from further consideration. Boxplots by lane width and AADT bands (Figure 17 on page 39) suggest that lane width would not be an effective predictor of annual crashes at any level of traffic. Only across segments travelled most heavily (AADT > 20,000) are annual crash count distributions appreciably different.[23]

[23] It is evident from the boxplot whiskers that 'zero annual crashes' is always within one standard deviation of the mean for every boxplot. The initial impression is that 'lane width' or an association between lane width and AADT would be insignificant in a regression model.

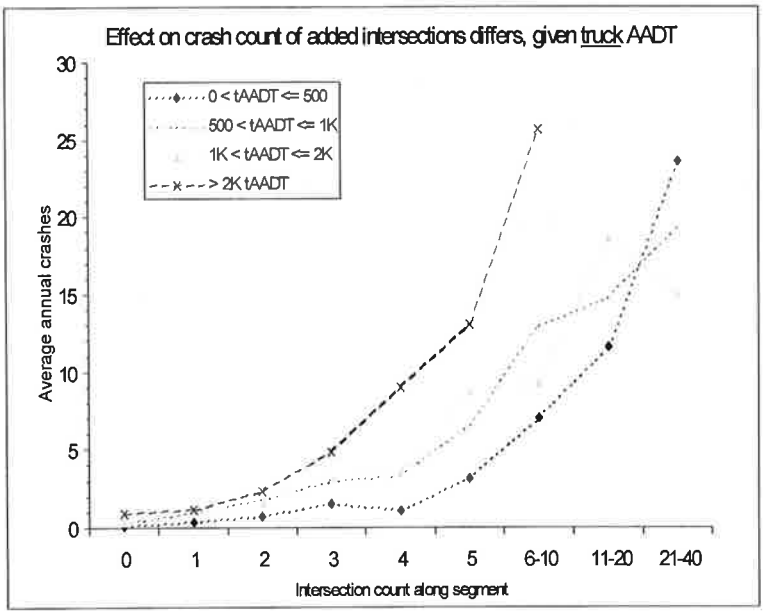

Figure 15. A categorical variable best captures effects given truck traffic.

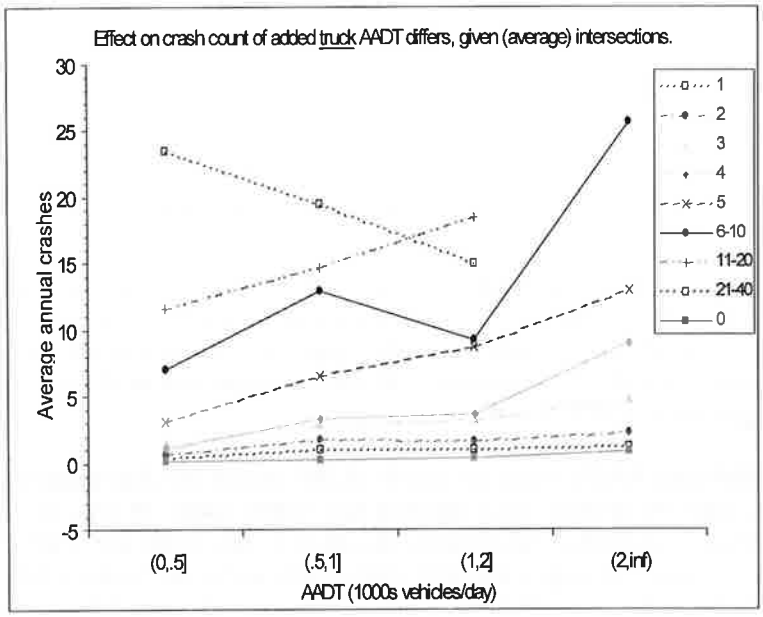

Figure 16. Given intersection label, impact from added truck traffic differs

Shoulder width and AADT. Ohio reports shoulder width to the nearest foot and a histogram of Ohio's inventory was plotted (Figure 18 on page 40). This plot moti-

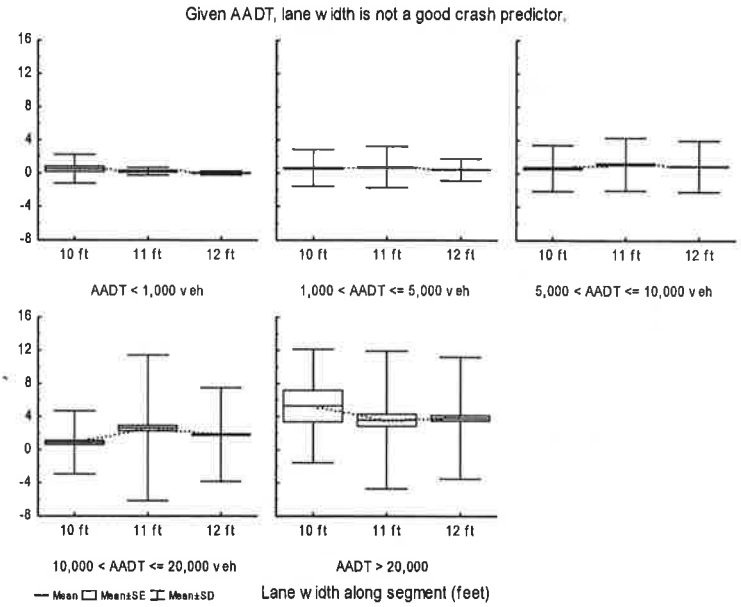

Figure 17. At any traffic level, Lane Width is not an effective crash predictor.

vated seven categories: 0 feet, 1-2 feet, 3 feet, 4 feet, 5-6 feet, 7-9 feet, more than 9 feet.

Figure 19 on page 40 suggests that, across all levels of AADT, the only significant differences are modestly lower crash tendencies across sections with 'no shoulders' or 'wide' shoulders (7 feet or more). Consequently, model fitting considered only three categories: *None* (0 feet), *Common* (1-6 feet) and *Wide* (7 feet or wider).

Access Control and AADT. The Ohio HSIS inventory file reports state and federal classifications, yet only federal codes are analyzed because the documentation Ohio submitted suggests analyses adopt that practice. A very small portion of records (82 out of 14,000) report the segment uses medians for access control. Considering this is highly unusually for urban two-lane streets nationwide (not just Ohio), these segments were eliminated from further analysis. According to the federal coding standard, three access control methods remain:

(1) (federal code '2') *access at interchange or public street; no direct private access allowed unless property retains deeded rights and then only for right turn. (Left turn may be allowed in certain circumstances.)*,

(2) (federal code '3') *no direct private access if property has another reasonable alternative access or opportunity to obtain such access; when allowed, generally for right turn*, and

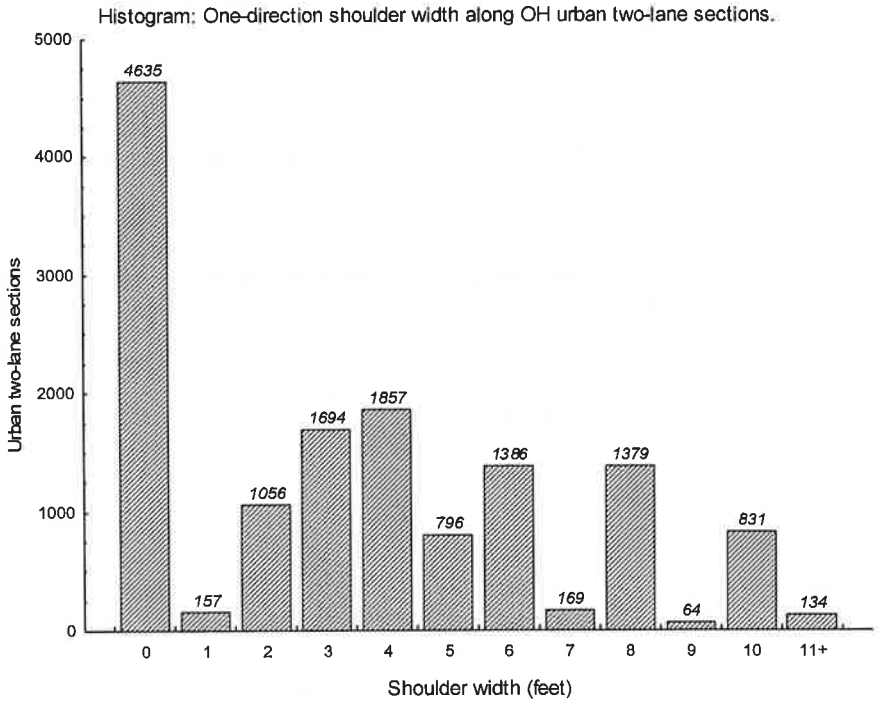

Figure 18. Histogram of Ohio urban two-lane segments by shoulder width.

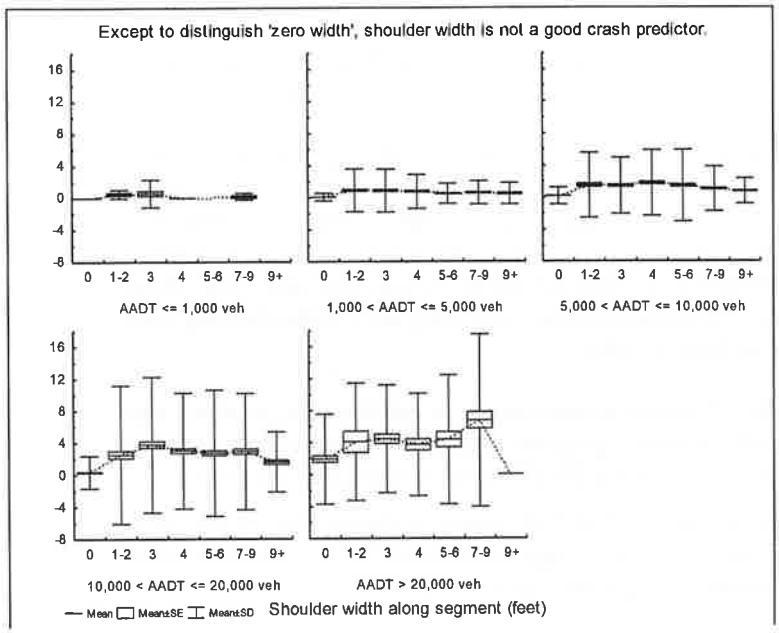

Figure 19. Crash count differences are manifested only with 'no' or 'wide' shoulder widths.

(3) (federal code '4') *one direct access allowed per parcel; additional access may be allowed if the Department determines it meets access safety, design, and operational standards.*

Safety analyses typically consider three types of access control: a) use of medians, b) 'limited' access strictly at intersections, signals and the like or c) 'uncontrolled' access, i.e., driveways are permitted for each abutter. Control via medians (federal code '1') was ignored, and definitions of the next two categories are similar enough to warrant combining them into a single 'limited access' category. The final category is considered 'uncontrolled access'.

Access control effects on Ohio urban two-lane crashes were explored, yet the variable would apparently be an ineffective predictor; federal access control codes are absent from 80% of Ohio records. Hence, Ohio segments can be naturally categorized by whether access control is reported and (among segments which are coded) whether each permits 'uncontrolled' or 'limited' access. In this context, segments possess one of three access control labels: 'Unknown', 'Limited' and 'Uncontrolled'. As with the other geometric features, categorized boxplots were examined across ranges of vehicle AADT (Figure 20).

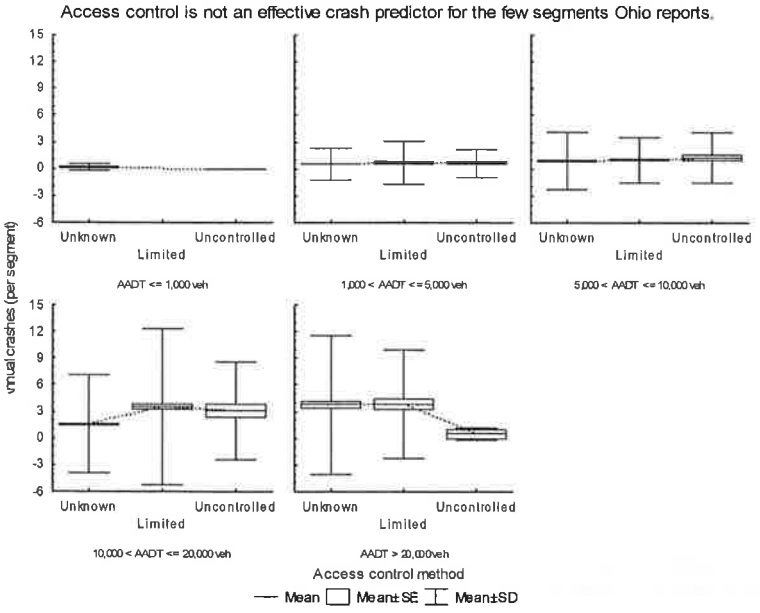

Figure 20. Because access control is largely unreported, it is a poor predictor.

If average traffic is below 10,000 vehicles/day, there is no evidence access control has any effect on crashes. Beyond 10,000 vehicles/day, crash count averages distinguish themselves, yet the respective variances (whiskers) are so wide that model fitting would conclude they are truly different. To further put the results in perspective: the

five boxplots labeled 'Unknown' comprise 11,335 Ohio records, those labeled 'Limited' comprise 2,130 records, and those labeled 'Uncontrolled' comprise a scant 311 records.[24]

Geometry associations

The previous graphs demonstrate how Ohio HSIS data empirically support hypotheses that associations between a geometric feature and daily traffic influence annual crashes. The following exploratory data analysis examines whether empirical evidence also suggests associations between pairs of geometric features, independent of daily traffic, impact annual crashes.

Lane Width in Conjunction with Shoulder Width. There is no evidence supporting hypotheses that, across Ohio urban two-lane segments, lane and shoulder width associations are useful predictors of annual crashes (Figure 21). Clearly, the boxplot means imply, on average, very few accidents are experienced with any particular combination. This, coupled with the fact that the annual crash variance is extremely wide (i.e., 'zero crashes' is always well within one standard deviation of the mean), suggest that an association term between lane and shoulder width would unlikely be statistically significant.

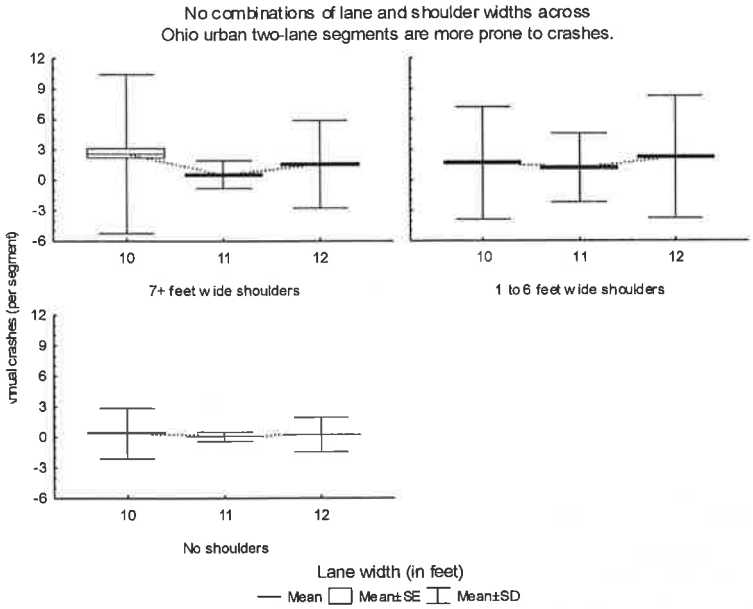

Figure 21. Lane and shoulder width associations to not influence Ohio crashes.

[24] The very last boxplot, the crash count distribution given uncontrolled access and traffic > 20,000 vehicles/day, comprises only two records.

Lane Width and Intersections. There is also no evidence supporting hypotheses that lane width and intersection count associations are useful predictors of annual crashes (Figure 22). The third row of graphs suggest that most Ohio crashes occur

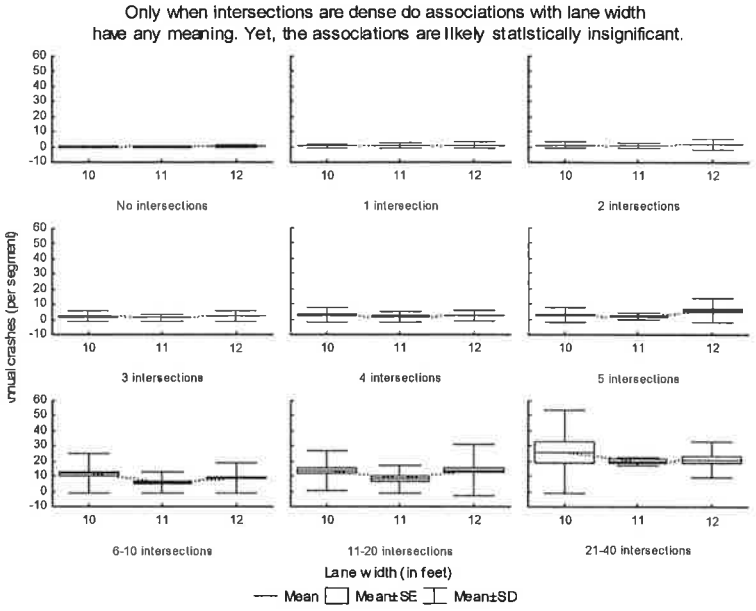

Figure 22. Associations between lane width and intersections are not helpful.

along segments dense in intersections, particularly those with more than 20 of them (perhaps corresponding to downtown streets in the larger cities). Yet, given intersection count, further partitioning by lane width does little to change the crash count distributions. Once again, boxplot means are so similar and crash variances are so wide that a lane width-intersection count association term would unlikely be statistically significant. Note the 'No intersections' graph (upper left corner). The boxplot means are virtually zero and standard deviations (whiskers) about each are negligible. Crash count variance is appreciable only when at least one intersection is present. Thus, Figure 18 illustrates another empirical fact about Ohio urban two-lane accidents (confirmed throughout the data exploration phase); virtually all of them occur near intersections.

Shoulder Width and Intersections. Segments with 'wide' shoulders and which and dense in intersections experience account for the most accidents (Figure 23). The boxplots across these categories (bottom row) are especially distinct. Direct confirmation wasn't possible from Ohio HSIS data, yet these accident-prone segments might correspond to downtown arterials which permit on-street parking. This association term was included as a candidate variable to model fitting.

It must be pointed out that the bottom row of graphs comprise the fewest segments in the Ohio data set. The boxplot triple of the lower right-hand graph (for which the

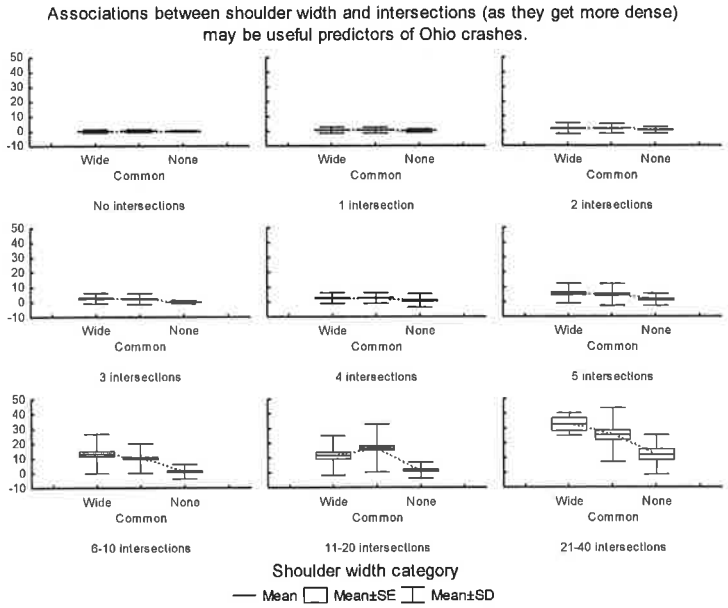

Figure 23. Subject to sample size, shoulder width-intersection association may be useful.

association appears most convincing) are composed of only 51 segments between them. By contrast, along the top row (where there is no convincing evidence of an association) each boxplot triple is composed of at least 2,200 segments and as many as 5,650 segments. This illustrates the careful consideration modeling must give to sample size considerations and the intended purpose of the regression equation. An association term may prove statistically insignificant because so few segments are involved. Yet, if those 51 segments are truly of interest (i.e., it can be confirmed that they are all indeed downtown arterial permitting parking), then the most useful model may need to include that term.

Access control and intersections. A regression term for access control and intersection count associations would offer no additional insight into annual crashes. Data exploration strengthens the case for 'intersection count' as a predictor, regardless of access control method (Figure 24). Crash count boxplots among segments with 'uncontrolled' access exhibit the least variation between category means, and the crash count variances grow progressively wider. Regardless of the number of intersections located along the segment, there is no evidence that annual crashes are associated with permitting driveways or curb cuts. Yet, the conclusion that this interaction would be ineffective is the fact that Ohio does not report access control for the great majority of segments. (Recall, only 311 records report 'uncontrolled access', while 11,335 omit the access control entirely.)

Exploratory data analysis suggests associations between daily traffic and individual features of the segment, as opposed to paired features, would be more effective predictors in a regression model estimating crash counts. To validate this, the statistical

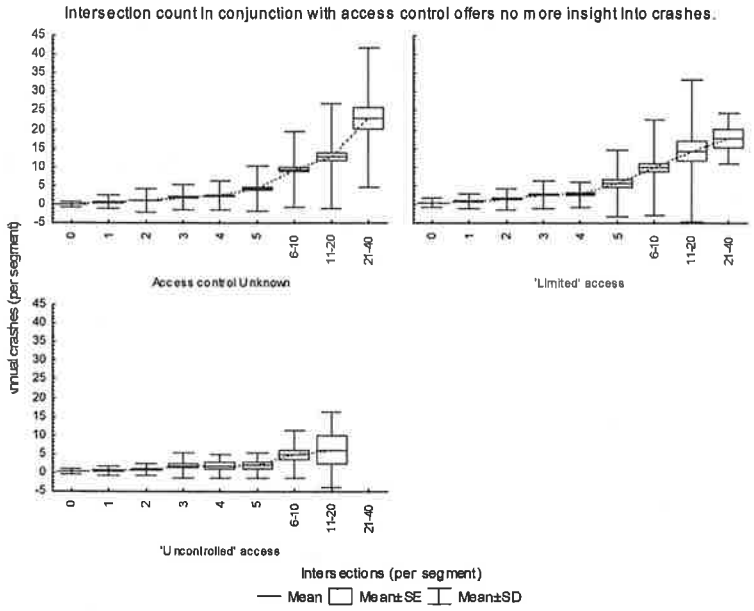

Figure 24. No evidence of intersection-access control interaction relative to crashes.

significance of association terms between pairs of geometric features were tested during model fitting.

Modeling - Variable Definitions

Segment geometry

Intersections Per Segment. A histogram of segments by intersection count suggested that reasonable labels for these categorical variables are: 0, 1, 2, 3, 4, 5, 6-10, 11-20 and 21-40 intersections. Labels for the first six categories are, trivially, the observed intersection count. The label is a weighted average for each of the last three categories. For example, the label among segments of the '6-10 intersections' category is

$$7.45 = 0.318 \times 6 + 0.269 \times 7 + 0.172 \times 8 + 0.126 \times 9 + 0.115 \times 10$$

where the integers are intersection counts and the fractions are respective proportions within the group. The nine category labels are presented in Table 10.

Table 10. The categorical variable for intersection count

Intersection Count	Label
0	0
1	1
2	2
3	3
4	4
5	5
6-10	7.45
11-20	13.84
21-40	28.52

The nine-element 0-1 vector $\boldsymbol{I}_t$ identifies the category of each segment t. For $j = 1, 2, \ldots 9$, $((\grave{I}_t)_j = 1)$ if segment t belongs to category j; otherwise, $(\grave{I}_t)_j = 0$.

Average Miles Between Intersections. Several studies in the published literature (e.g., Vogt and Bared 1998) establish the spacing of intersections along segments after meticulous reviews of state DOT photologs. This assessment relied solely on Ohio HSIS data, so a variable *approximating* intersection spacing is defined. It is straightforward to determine the length of any segment (the difference between its beginning and ending mileposts) and the number of intersections located along it (by cross-referencing segment mileposts from the Ohio inventory file with the individual mileposts from the intersection file locating the intersections). The continuous variable $M_t = L_t/(k+1)$ is defined for each segment t of length L_t miles and k intersections.

Thus M_t divides the segment into k+1 equidistant portions (Figure 25, top). Admittedly M_t will be an inaccurate average if an intersections are co-located with the

beginning or ending mileposts (Figure 25, bottom. The average distance between intersections 1,2 and 3 is really $2M_t$, not M_t.) However, this exception is extremely rare among Ohio urban two-lane segments.

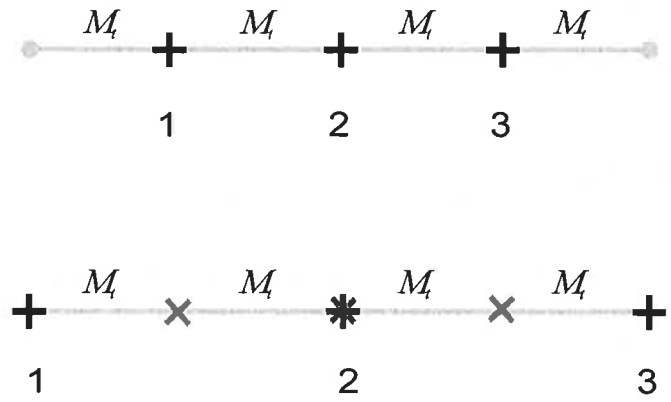

Figure 25. The variable M_t approximates intersection spacing.

Lane Width. Every segment t possessing lanes 10 feet or wider was placed in one of three categories (Table 9 on page 32). The three-element 0-1 vector $\boldsymbol{LW}_t$ identifies the lane width. If segment t belongs to category $i = 1, 2, 3, ((\overrightarrow{LW}_t)_i = 1)$; otherwise, $(\overrightarrow{LW}_t)_i = 0$.

Shoulder Width. The three-element 0-1 vector $\boldsymbol{SW}_t$ categorizes shoulder with of segment t as '0 feet', '1-6 feet' and '7 feet or more'. These three categories were patterned after the boxplots of Figure 19 on page 40. If segment t belongs to category $k = 1, 2, 3, ((\overrightarrow{SW}_t)_k = 1)$; otherwise, $(\overrightarrow{SW}_t)_k = 0$.

Access control. The three-element 0-1 vector $\boldsymbol{AC}_t$ identifies the access control method, or that it is unknown. If segment t has 'limited' access control, $(\overrightarrow{AC}_t)_2 = 1$ and the other two elements equal zero. If segment t has 'uncontrolled' access, $(\overrightarrow{AC}_t)_3 = 1$ and the other two elements equal zero. Otherwise, $(\overrightarrow{AC}_t)_1 = 1$ and the other two elements equal zero.

Average Daily Traffic

Daily Car Traffic (Car AADT). Ohio HSIS data reports both total and commercial truck AADT for each urban two-lane segment. Consequently the continuous variable C_t = *(Total AADT - Truck AADT)* / 1,000 represents daily car traffic, per 1,000 vehicles, across segment t. Scaling C_t is this manner desirable prior to model fitting. A very small number of Ohio segments were discarded when obvious data quality errors implied $C_t < 0$ across the segment.

Daily Truck Traffic (Truck AADT). The continuous variable T_t / 1,000 represents daily truck traffic, per 1,000 vehicles across segment t. A very small number of Ohio segments were discarded because $T_t = 0$. (It will become evident, $T_t = 0$ is undefined in the best-fit model.)

Only 40 Ohio segments out of nearly 14,000 in the HSIS database were discarded due to vehicle (car or truck) AADT < 0.

Dependent Variable - Annual crashes

The dependent variable is A_t, the annual accident count along Ohio urban two-lane segment t. A modeling technique appropriate to estimating counts and widely adopted by published studies was selected.

Estimating Ohio crash counts

The Negative Binomial GLM Method

The boxplots of annual Ohio crashes provides strong empirical evidence that (a) segments are heterogeneous (their geometrical features, traffic flows and accident histories are quite diverse) and that (b) assuming accident counts are Poisson distributed is invalid due to *overdispersion* in the data. If annual crash counts were Poisson distributed, by definition the crash count variance must equal the crash count average λ, the necessary parameter of the Poisson distribution. Exploratory data analysis illustrates how boxplot whiskers are commonly too wide for this assumption to be plausible.

Prediction models for counts generally turn to the technique known as Generalized Linear Models (GLM) because it cannot be assumed that a) prediction errors follow the Normal distribution or b) the counts themselves follow the Poisson distribution. The class of GLM most often applied is Negative Binomial GLM.[25] This study applies the technique, yet by introducing variable association terms it demonstrates an approach, to the authors' knowledge, has been overlooked in the relevant literature.

The most common model formulation, and the one adopted, is the so-called *log link*. The linear combination of independent variables is equated with the natural logarithm of the dependent variable:

$$\log\mu = \beta_0 + \sum_{j=1}^{n} \beta_j \cdot x_j . \qquad [6]$$

The SAS GENMOD Procedure. The GENMOD procedure included with SAS/STAT™ software (SAS Institute 2001) was used for all model fitting. Since Negative Binomial GLM is an example of Maximum Likelihood Estimation, models with categorical variables must be fit using one fewer degree of freedom than the number of categories defined. That is, Maximum Likelihood models which include a variable defined over J categories are fit against J-1 of them. The GENMOD procedure permits some flexibility in assigning this last *baseline category*. In the discussions which follow, the baseline of each statistically significant geometric variable in the best-fit model is the category *least prevalent* among Ohio segments. (For example, the baseline category of I_t, intersection count along segment t, is the class of segments pos-

[25] An excellent text is Alan Agresti's book *Categorical Data Analysis*, 2nd Edition. 2002. Wiley & Sons.

sessing 21-40 intersections, since these represent the smallest portion of the Ohio urban two-lane data set.)

The Best-fit equation

Fitted against nearly 14,000 Ohio urban two-lane streets, the structural form of the HERS urban two-lane crash prediction model is:

$$\log A_t = \beta_0 + \vec{\beta}_1 \cdot \vec{I}_t + \beta_2 \cdot \log T_t + \vec{\beta}_3 \cdot (M_t \times \vec{I}_t) + \vec{\beta}_4 \cdot (\log C_t \times \vec{I}_t) + \beta_5 \cdot (\log C_t \times M_t) + \beta_6 \cdot (\log T_t \times M_t) + \vec{\beta}_7 \cdot (\log C_t \times \vec{LW}_t) + \vec{\beta}_8 \cdot \vec{SW}_t \quad [7]$$

For segment t, intersection count (vector $\boldsymbol{I}_t$), truck traffic (T_t) and shoulder width (vector $\boldsymbol{SW}_t$) are statistically significant linear predictors of crashes ($\log A_t$). Statistically significant associations are between:

a) intersection spacing and intersection count $(M_t \times \vec{I}_t)$,

b) car traffic and intersection count $(\log C_t \times \vec{I}_t)$,

c) car traffic and intersection spacing $(\log C_t \times M_t)$,

d) truck traffic and intersection spacing $(\log T_t \times M_t)$,

e) car traffic and lane width $(\log C_t \times \vec{LW}_t)$.

Linear predictors and association terms involving access control variables $\boldsymbol{AC}_t$ were candidates, yet none are statistically significant. Also insignificant are the remaining 2- and 3-order interaction terms between the defined variables. That higher order interaction terms are not included is beneficial in that model [7] is both easier to interpret and, by definition, explains enough about crash count variances that higher-order terms are not helpful

Interpreting the model

Baseline categories. If segment t belongs to baseline b of a categorical variable (variables denoted as vectors in [7]), then by definition $(\vec{\beta}_t)_b = 0$ (the b^{th} element equals 0). In other words, all else being equal, a geometrical feature for each segment of the baseline has no effect on estimated crashes. Also, since car and truck AADT are in units of 1,000 vehicles/day and also transformed on a (natural) logarithmic scale, that traffic level acts as a baseline (for example, in the new scale $C_t = 1$, so that $\log C_t = 0$). When vehicle AADT is precisely 1,000/day, the respective terms of model [7] have no effect on estimated crashes. In this context, signs of the non-zero coefficients indicate whether the term implies an *increase* expected annual crashes (sign is positive) or a *decrease* expected annual crashes (sign is negative).[26]

Empirically, e^{β_0} represents predicted accidents expected of a hypothetical Ohio urban two-lane segment possessing baseline geometrical features and traffic. Model [7] adjusts estimates for every segment t relative to e^{β_0}. (For Ohio urban two-lane data,

[26] As mentioned, segments for which $\log C_t < 0$ or $\log T_t < 0$ are discarded *a priori*.

$e^{\beta_0} = e^{0.8152} = 2.260$ annual crashes.) Note that, because it is formulated as a GLM using the log link, model [7] will never predict 'zero accidents'.

Crash count distribution. By assumption of the method, annual crash counts A_t, are distributed Negative Binomial with $E(A_t) = \mu$ and $var(A_t) = \mu + K\mu^2$. The SAS/SAT GENMOD procedure estimates the parameter K, called the *dispersion parameter.* For K small, the distribution of A_t approaches a Poisson distribution with $E(A_t) = \mu$ because $\lim_{K \to 0} var(A_t) = \mu$. Hence, the GENMOD estimate of K validates the choice of negative binomial GLM for modeling. For model [7], $K = 2.502$, strong evidence of overdispersion ($var(A_t) \gg E(A_t)$), confirming Poisson GLM of A_t would be inappropriate.

Linear terms

Intersection count. The baseline category comprises Ohio segments possessing 21-40 intersections, so $(\hat{\beta}_1)_9 = 0$. The remaining coefficients $(\hat{\beta}_1)_j$ are all negative because fewer intersections imply relatively fewer predicted annual crashes (Table 11, center). All else being equal, segments of intersection category j are predicted to experience $e^{(\hat{\beta}_1)_j}$ fewer annual crashes than the baseline category (segments possessing the most intersections of all).

Table 11. Fewer accidents are generally associated with fewer intersections.

Intersection category	Coefficient	Crash reduction multiplier
0	-3.493	97.0%
1	-4.360	98.7%
2	-3.220	96.0%
3	-2.484	91.7%
4	-3.123	95.6%
5	-3.275	96.2%
6-10	-2.699	93.3%
11-20	-2.108	87.9%
21-40	0	--

These reductions are not monotonic, which is empirically consistent with Ohio HSIS data. Plots of average annual crashes across intersection categories (Figure 12 on page 36) illustrate these are not strictly increasing once traffic exceeds 10,000 AADT. Since those segments account for the most Ohio accidents, naturally the best-fit regression model mimics those variations.

The coefficient $(\hat{\beta}_1)_2 = -4.360$ for the '1 intersection' category requires further analysis. As the most negative coefficient, segments of this category have the fewest predicted crashes, all else being equal. Yet, Figure 11 on page 35 and Figure 12 on page 36 suggest that segments *without* an intersection generally account for fewer acci-

dents. It is possible that some segments possessing no intersections are outliers with abnormally high annual crashes, which would explain this discrepancy (outliers are omitted from boxplots to limit compression along the *y*-axis).

Daily truck traffic. The coefficient $\beta_2 = 0.1438$. For segment *t*, all else being equal, the current truck traffic level T_t (in 1,000 trucks/day), translates into an predicted change in annual crashes by a factor of $(T_t)^{0.1438}$. If truck AADT across segment *t* exceeds 1,000/day, all else being equal, $(T_t)^{0.1438}$ is associated with more annual crashes. Saturating the segment with truck traffic always increases predicted crashes, since $(T_t)^{0.1438}$ is unbounded from above. Ohio urban two-lane truck AADT never exceeds 5,000 vehicles/day. Hence, all else being equal, model [7] attributes no more than a $5^{0.1438} = 26.0\,\%$ increase in annual crashes attributed to truck AADT across a particular Ohio segment. By contrast, if truck AADT across segment *t* is under 1,000/day, all else being equal, $(T_t)^{0.1438}$ is associated with fewer annual crashes. Since $\lim_{T \to 0} (T_t^{0.1438})$, model [7] predicts segments with no truck traffic will experience no accidents at all (all else being equal).

Shoulder Width. The baseline category comprises Ohio segments possessing 'wide' shoulders (7 feet or wider). By definition, for any segment in this category, shoulder width has no effect on predicted crashes ($(\hat{\beta}_8)_3 = 0$). The coefficients $(\hat{\beta}_8)_s$ for the other two shoulder width categories have opposite signs (Table 12).

Table 12. Segments with no shoulders have the fewest accidents.

Shoulder width	Coefficient
None	-1.1162
Common	0.1685
Wide	--

For a segment with *no* shoulders the term predicts $e^{-1.1162} = 67.2\,\%$ fewer crashes, relative to any segment with *wide* shoulders, all else being equal. For an analogous segment with *common* shoulders (1-6 feet wide), the term predicts $e^{0.1685} = 18.4\,\%$ more crashes, relative to any segment with *wide* shoulders. These coefficients are consistent with the boxplots of Figure 19 on page 40. All boxplots labeled '0' (no shoulders) consistently have average crash counts below those among segments of the 'wide' category (right-most two boxplots in each graph). Likewise, the remaining boxplots, except at the highest levels of AADT, consistently have average counts above segments in the 'wide' category.

Variable associations

Relative to the four types of segment geometry examined: intersections, lane width, shoulder width and access control, segment *t* average daily car traffic, C_t, is statistically significant only in association terms, not as a linear predictor. The structure of

model [7] offers support of this study's fundamental hypothesis: *all else being equal, more cars do not necessarily imply more crashes*. An urban two-lane segment's daily traffic flows, *in conjunction with* the segment's geometry, is more indicative of the segment's tendency to experience accidents.

Intersection Spacing and Intersection Count. This association term is not quite equivalent to 'segment length'. By definition, the length of segment t possessing k intersections equally spaced M_t miles apart is $L_t = (k+1) \times M_t$ miles. Yet, values of this association term are $k \times M_t$. For model [7], coefficients $(\hat{\beta}_3)_j > 0$ for all categories of intersection count $j > 0$ ($(\hat{\beta}_3)_0 = 0$). This is intuitive, because $M_t > 0$ if and only if segment t possesses intersections, and virtually all Ohio accidents occur near intersections.

Model [7] predicts, due to the *combined* effects of the number (category $j > 0$) and spacing of intersections, that segment t will experience more annual crashes by a factor of $e^{(\hat{\beta}_3)_j \times M_t}$ than an urban two-lane segment without intersections (all else being equal).

For two segments in the same intersection category $j > 0$, model [7] estimates more annual crashes along the one with more separation between intersections. If intersections are spaced (on average) $d_{21} = M_2 - M_1 > 0$ miles farther apart along segment 2 than along segment 1, then the interaction term predicts segment 2 will experience more crashes by a factor of $e^{d_{21}}$. Considering the similarity between the association term and 'segment length', this is sensible; longer segments are more likely to experience more accidents.

Finally, if the two segments differ *both* in intersection count (suppose segment 1 belongs to category j and segment 2 belongs to category $j' > j$), and spacing (also suppose $d_{21} > 0$), the association term predicts segment 2 will experience

$$e^{((\hat{\beta}_3)_{j'} - (\hat{\beta}_3)_j) \times d_{21}} \quad [8]$$

as many accidents as segment 1. Table 13 presents these factors when one segment possesses more intersections (category $j' > j$) spaced slightly farther apart ($d_{21} = 0.1$ miles). For convenience diagonal elements (j,j) present coefficients $(\hat{\beta}_3)_j$. Below the diagonal, cells (j',j) are the reciprocals of cells (j,j'), and thus are not shown.

Intersection Count and Car Traffic. For model [7], coefficients $(\hat{\beta}_4)_j > 0$ for all categories of intersection count j. At fixed levels of one variable, the term attributes more accidents with increases in the other variable. If car traffic C (in 1,000 cars/day) across two segments is identical, yet segment 1 possesses j intersections and segment 2 possesses $j' > j$ intersections, this association term estimates more accidents across segment 2 than segment 1 by a factor of

Table 13. More intersections, spaced farther apart, typically imply more crashes.

Label j	Label j'								
	0	**1**	**2**	**3**	**4**	**5**	**7.45**	**13.84**	**28.52**
0	*0*	3.188	3.001	4.053	4.621	6.091	5.362	8.547	3.727
1		*11.595*	0.941	1.271	1.449	1.910	1.682	2.681	1.169
2			*10.989*	1.351	1.540	2.030	1.787	2.848	1.242
3				*13.995*	1.140	1.503	1.323	2.109	0.920
4					*15.306*	1.318	1.160	1.850	0.807
5						*18.068*	0.880	1.403	0.612
7.45							*16.793*	1.594	0.695
13.84								*21.456*	0.436
28.52									*13.157*

$$C^{((\beta_4)_{j'} - (\beta_4)_j)} . \qquad [9]$$

For two hypothetical segments, these factors are presented in Table 14. For $j' > j$, table element (j,j') presents $C^{((\beta_4)_{j'} - (\beta_4)_j)}$ for car traffic C = 2,000/day. For convenience diagonal elements (j,j) present coefficients $(\hat{\beta}_4)_j$. Below the diagonal, each element (j',j) possesses the reciprocal value of element (j,j'), and thus are not shown.

Table 14. Given traffic, more crashes not always attributed to more intersections.

Label j	Label j'								
	0	**1**	**2**	**3**	**4**	**5**	**7.45**	**13.84**	**28.52**
0	*0.448*	1.717	1.489	1.254	1.578	1.793	1.767	1.470	1.332
1		*1.227*	0.867	0.730	0.919	1.044	1.029	0.856	0.776
2			*1.022*	0.842	1.060	1.204	1.187	0.988	0.895
3				*0.774*	1.259	1.430	1.409	1.173	1.062
4					*1.106*	1.136	1.119	0.931	0.844
5						*1.290*	0.986	0.820	0.743
7.45							*1.269*	0.832	0.754
13.84								*1.004*	0.906
28.52									*0.861*

If, on the other hand, the two segments are both labeled with j intersections yet segment 2 has $C_2 > C_1$ more daily car traffic than segment 1, model [7] estimates segment 2 will experience more accidents by a factor of

$$(C_2/C_1)^{(\beta_4)_j}. \quad [10]$$

Intersection Spacing and Vehicle Traffic. The association terms $(\log C_t \times M_t)$ and $(\log T_t \times M_t)$ are statistically significant. Therefore (when [7] is transformed to estimate A_t), this association term is expressed in units of

$$\left(\frac{vehicles}{day}\right) \times \left(\frac{miles}{intersection}\right) = \left(\frac{(vehicle \cdot miles)/(day)}{intersection}\right) \quad [11]$$

which is the (average) daily VMT per intersection along the segment. In model [7], $\beta_5 = -4.043$ for cars and $\beta_6 = 1.1875$ for trucks.

Suppose across segments 1 and 2 car traffic C (in 1,000 cars/day) and truck traffic T (in units of 1,000 trucks/day) are identical. Further suppose segment 1 intersections are spaced M_1 miles apart, on average; segment 2 intersections are spaced M_2 miles apart, on average. All else being equal, this association term estimates segment 1 has

$$C^{\beta_5 \cdot (M_1 - M_2)} \quad [12]$$

as many accidents as segment 2. Likewise, for fixed T, the term estimates segment 1 will experience

$$T^{\beta_6 \cdot (M_1 - M_2)} \quad [13]$$

as many accidents as segment 2.

On the other hand, suppose the two segments possess identical intersection spacing, M, yet their car (truck) AADT levels are C_1 (T_1) and C_2 (T_2), respectively. All else being equal, with respect to car traffic this association term estimates segment 2 has

$$(C_2/C_1)^{\beta_5 \times M} \quad [14]$$

as many accidents as segment 1. With respect to truck traffic, the term estimates segment 2 has

$$(T_2/T_1)^{\beta_6 \times M} \quad [15]$$

as many accidents as segment 1 (all else being equal). Whether more or fewer crashes are attributed to segment 2 relative to segment 1 are functions of the respective and (constant) intersection spacing M.

It is useful to consider the case when $C_2 = C_1+1$. Then,

$$(C_2/C_1)^{\beta_5 \times M} = \left(1 + \frac{1}{C_1}\right)^{\beta_5 \times M} \text{ and } \lim_{C_1 \to \infty} \left(1 + \frac{1}{C_1}\right)^{\beta_5 \times M} = 1. \quad [16]$$

In terms of segment t car AADT capacity, this behavior is meaningful. As car traffic levels approach segment capacity, each additional 1,000 cars/day has a more negligible impact on estimated crashes than the previous 1,000 cars/day. The same behavior holds with respect to truck traffic, yet

$$\lim_{T_1 \to \infty} \left(1 + \frac{1}{T_1}\right)^{\beta_6 \times M} = 1 \qquad [17]$$

from *above* because the exponent has the opposite sign.)

The easiest way to depict behavior of these multipliers is to fix the *difference* in (average) intersection spacing, M_1 - M_2, and allow vehicle AADT. Figure 26 graphs the percentage changes in annual crashes associated with $C^{\beta_5 \cdot (M_1 - M_2)}$ (dotted lines) and $T^{\beta_6 \cdot (M_1 - M_2)}$ (solid lines) when M_1 - M_2 = 1.0 mile, M_1 - M_2 = 0.1 mile and M_1 - M_2 = -0.1 mile ($M_1 < M_2$).

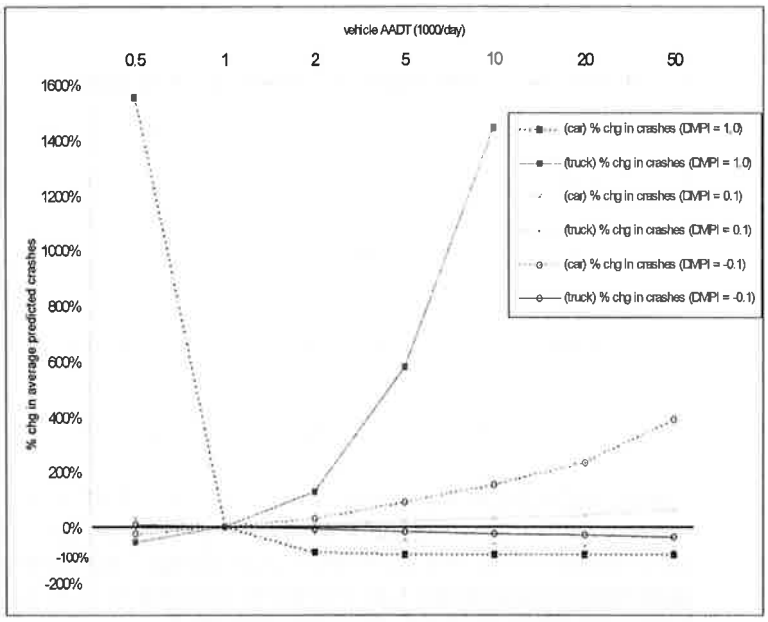

Figure 26. Interaction effects between vehicle AADT and intersection spacing.

Traffic is scaled in units of 1,000 vehicles/day. So, when vehicle AADT is precisely this value, the association term has no impact on predicted crashes (in model [7] $(\log C_t \times M_t) = (\log T_t \times M_t) = \log 1 \times M_t = 0$). The (dotted) car and (solid) truck curves pivot about this point, through the *0*% threshold as M_1 - M_2 changes. If $M_1 > M_2$, model [7] predicts *fewer* accidents across segment 1 when *car* traffic exceeds 1,000/day (the dotted curves fall below the *0*% line). Furthermore, the car traffic curves asymptotically approach *-100*%, which is sensible (the association term never reduces

accidents by more than what model [7] would otherwise predict). So long as $M_1 > M_2$, saturating segment 1 with more car traffic always *decreases* predicted accident, all else being equal. Thus, the $(\log C_t \times M_t)$ interaction term is consistent with the principle (demonstrated by other studies) that accident risk does not necessarily increase as traffic approaches segment capacity. If $M_1 \gg M_2$, the association term approaches this lower bound very quickly, illustrated by the dotted 'DPMI = 1.0' curve (Figure 26).

These effects are reversed when $M_1 < M_2$ (the 'DMPI = -0.1' curve for car traffic pivots through the *0*% threshold when $C > 1000$/day). Predicted crashes along segment 1 are unbounded from above when $M_1 \ll M_2$ (Figure 26). In other words, when $M_1 < M_2$ saturating segment 1 with car traffic always *increases* predicted crashes relative to segment 2. They are also reversed for $C < 1{,}000$/day; when car traffic is light, *more* crashes are attributed to segments with intersections spaced farther apart, while *fewer* crashes are attributed to segments with intersections spaced closer together.

All effects just described are reversed relative to car traffic effect, because the coefficient on the analogous truck traffic term is of opposite sign ($\beta_6 = 1.1875$). The marginal impacts (rates of change) as truck traffic changes are also less dramatic since $|\beta_6| < |\beta_5|$.

Car traffic and intersection variables are conditionally independent. The terms $(\log C_t \times M_t)$, $(\log C_t \times \grave{I}_t)$ and $(M_t \times \grave{I}_t)$ are all statistically significant. The higher-order association term $\log C_t \times M_t \times \grave{I}_t$ was found to be statistically insignificant. Model [7] therefore concludes these three variables are *conditionally independent*; any pair of these variables is independent at fixed levels of the third. For example, in probability notation, stating car AADT (C_t) and intersection count $(\grave{I}_t)$ are conditionally independent at fixed levels of intersection spacing (M_t) means:

$$P(C_t = i | \grave{I}_t = j, M_t = k) = P(C_t = i | M_t = k). \qquad [18]$$

Conditional independence seems reasonable. Model [7] states that, controlling for segment t intersection spacing, the joint probability $P(C_t, \grave{I}_t)$ of any particular combination of car AADT and intersection count changes, yet each of the two variables is independently distributed. Likewise for the behavior of $P(M_t, \grave{I}_t)$ controlling for C_t as well as the behavior of $P(C_t, M_t)$ controlling for $\grave{I}_t$.

Lane Width and Car Traffic. The baseline category comprises Ohio segments possessing lanes 11 feet wide (the least common in the Ohio data set), so $(\grave{\beta}_7)_2 = 0$. Consequently, for segments with 11-foot lanes, there is no car AADT-lane width interaction effect on crash predictions. For the other two lane width categories, coefficients $(\grave{\beta}_7)_k > 0$ (Table 15). For any segment t with car AADT $C_t > 1$ (in 1,000 cars/day), the association term predicts *more* accidents, by a factor of $(C_t)^{0.1639}$ for seg-

Table 15. Ohio segments with 10- and 12-lanes experience more crashes.

Lane width	Coefficient
10 feet	0.1639
11 feet	--
12 feet	0.1937

ments with 10-foot lanes and a factor of $(C_l)^{0.1937}$ for segments with 12-foot lanes. Note that, for $C_t < 1$, the effect is reversed; the factors estimate *fewer* crashes relative to segments with 11-foot lanes (multipliers $(C_l)^{0.1639} < (C_l)^{0.1937} < 1.0$). At all traffic levels, this interaction term estimates more crashes along a segment with 12-foot lanes relative to one with 10-foot lanes. Rather than reflecting accident risk inherent to wider lanes, this observation probably represents a foregone conclusion; a substantial majority of Ohio segments possess 12-foot lanes.

From exploratory data analysis, boxplots across lane width categories did not suggest lane width would be an effective predictor. Additional boxplots of Ohio crashes categorized by lane with and strata of car AADT (Figure 27) suggest that the strength of this association term is relevant only at the heaviest traffic levels. (Categorizing graphs by whether intersections are present reinforces the fact that virtually all Ohio accidents occur near intersections.)

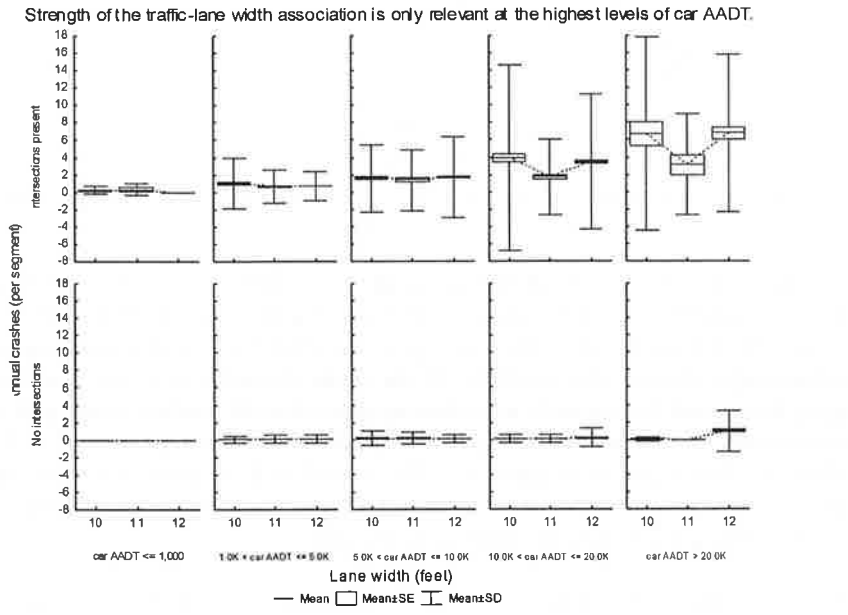

Figure 27. Traffic-lane width association is only relevant at heaviest traffic levels.

Below 10,000 cars/day, lane width is not an effective predictor. So few Ohio segments possess car AADT below 1,000 cars/day that the behavior noted of this association is essentially academic. In conclusion, despite being identified as statistically significant, the empirical evidence suggests that any combined effect of lane width and daily car traffic on annual crashes is, in fact, weak.

Model performance

Analysis of Residuals. Model [7] is a tremendous improvement over the existing HERS model [1]. A scatterplot of actual versus predicted annual crashes for every Ohio segment is displayed in Figure 28 (the hypothetical reference line $y = x$ represents perfectly accurate predictions).

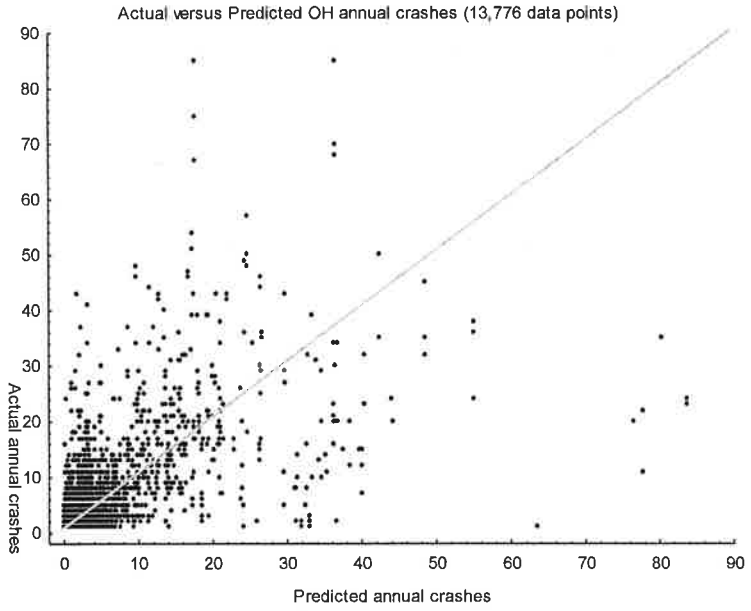

Figure 28. Scatterplot of segment-level annual crash counts versus predictions.

To visualize prediction error distributions, the entire data set of Ohio urban two-lane segments was categorized by actual crash count. The first ten categories are reserved for exactly 1 crash to 10 crashes in a year, respectively. The remaining categories define ranges of ten: 11-20, 21-30, 31-40, etc. up to 80 crashes in a year. The final category is reserved for segments experiencing more than 80 crashes in one year (comprising only two segments). Boxplots of prediction errors represents 3,417 Ohio urban two-lane segments (Figure 29). The size of each category is written directly above or below the boxplot. The remaining 10,359 segments never experienced an accident during the three years spanned by the data.

Attempts to predict 31 crashes or more are biased, have residuals with very large standard errors, or both. Yet, these segments represent a miniscule proportion of the Ohio data set. By contrast, residuals are nearly zero on average and residual standard errors are tight among segments experiencing 30 or fewer annual crashes.

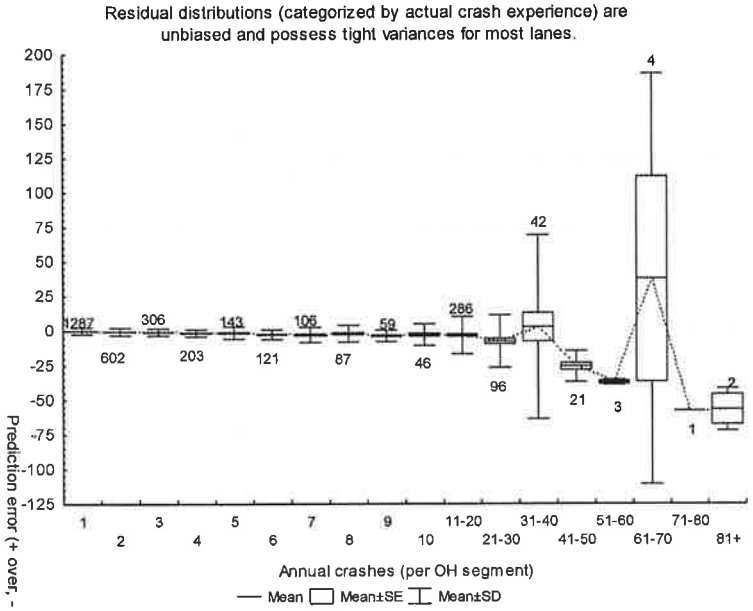

Figure 29. Regression model residuals, by annual crash count (Ohio HSIS data).

Residual boxplots associated with 30 or fewer annual crashes are reproduced in Figure 30 to illustrate a 'sanity check' conducted for these residuals. The solid line represents the residual implied if the regression model were to always return the smallest prediction possible, 0 crashes (the line $y = -x$, the implied error for segments experiencing x crashes in a year). Generally, values $-x$ are very far in the tails of the error distributions. Since the last two boxplots combine all segments with 11-20 and 21-30 accidents in a year, respectively, implied errors are the vertical bars. These are also far in the tails of the error distributions.

As mentioned, a substantial majority of segments never experience an accident. For these model [7] predicts, on average, $A_t = 0.591$ annual crashes. (Partially because model [7] includes an intercept term.) The variance associated with these prediction errors is also tight. But, because the 95% confidence interval for average A_t - (0.549, 0.633) - excludes zero, model [7] estimates for these segments are slightly biased.

Estimated safety costs. As with model [1], model [7] also tends to overpredict total Ohio crashes. Yet, judging by the estimate of total 1997-1999 Ohio crashes, relative to the observed number model [7] is incomparably superior to model [1] (Table 16, left).

Certain segments were removed from consideration during data exploration, and still others were discarded for data quality reasons by the SAS/STAT GENMOD procedure. Cross-referencing Ohio segments actually used by GENMOD with the original urban two-lane HSIS data yielded the proper set of segments for fatality and costs comparisons. Converting crash rates to counts is unnecessary since model [7] esti-

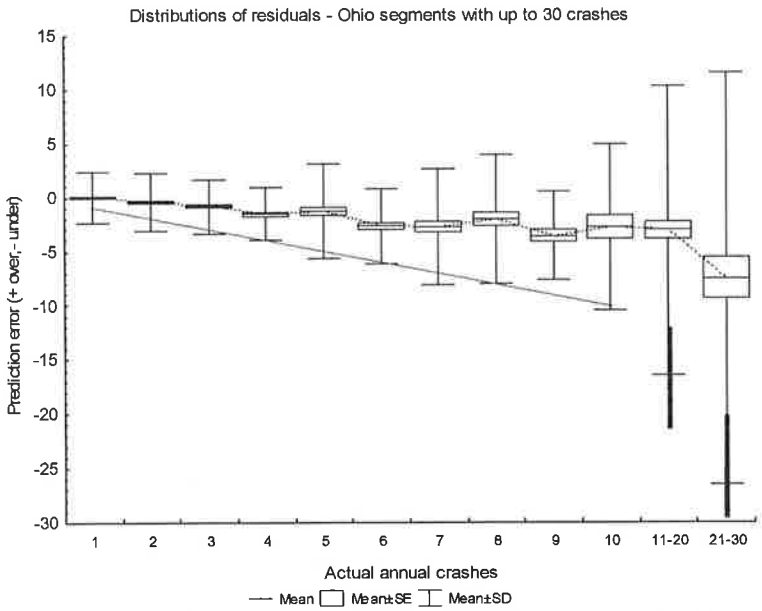

Figure 30. Up to 30 annual crashes, model residuals are well-behaved.

Table 16. Fatality and safety cost estimates (model [7])

Predicted crashes	Actual crashes	Predicted fatalities	Actual fatalities	Cost error ($M)
20,134	17,710	50	95	$(259.3)

mates those directly for every segment. The same conversion factors used in the opening discussion are applied to estimate fatalities (0.00247 fatalities/crash) and injuries (0.34485 injuries/crash). The same unit costs are applied as well: per fatality ($3 million), per injury ($35,750) and per accident property damage ($6,900).

Despite over-estimating total Ohio urban two-lane crashes, fatalities, injuries and hence economic costs are under-estimated (Table 16, right)! Alternatives to the entire HERS procedure for estimating safety costs are beyond the scope of this study. Yet, the fatality multiplier (0.00247) is half as large as the true value for Ohio urban two-lane roads (0.00536 = 95/17,710); the same holds for the injury multiplier (0.34485) relative to the Ohio estimate (0.6135 = 10,865/17,710).[27] It appears that the dramatic overestimates of safety cost by model [1] (Table 8 on page 29) are primarily due to its

[27] Ohio HSIS data reports fatalities injuries using the KABCO scale. The 10,865 injuries reported are the sum total of injuries in the A, B and C categories.

inflated crash estimates, as opposed to overstated fatality and injury multipliers. Future research could explore methods for estimating crash fatalities and injuries more directly, perhaps using HSIS data.

Conclusions

The fundamental purpose of HERS is to evaluate possible improvements to a roadway segment, in part in terms of safety benefits and costs. Yet, HERS model [1] for urban two-lane roads omits geometric features entirely, presuming crashes are strictly a function of AADT. This degree of simplification in crash prediction models is not unique to HERS, and relatively little has been published, about urban two-lane roads especially, on more sophisticated methods. Numerous studies have derived crash prediction models in terms of both AADT and roadway geometry. Appropriately, they also rely on Negative Binomial GLM for model fitting. Yet, these studies overlook or exclude variable associations even though they are not mathematical complications.

This study intended to formulate a HERS urban two-lane crash prediction model which (apart from improving performance) more appropriately serves its intended purpose. It started with the hypothesis that daily traffic flows *in conjunction with* roadway geometry would generate more accurate and useful predictions. Exploratory data analysis confirmed, despite very noisy Ohio HSIS data, that association terms between judiciously defined continuous and categorical variables are beneficial. The best-fit regression model [7] performs admirably well. Parameters are insightful and prediction errors are well-behaved. Inaccuracies in the subsequent estimates of accident severity and safety costs will be topics for future research.

Urban two-lane crash prediction model [7] incorporates traffic, roadway geometry and associations between them. Its structure would be valuable for the types of 'what-if' scenario analyses HERS is designed to support. For example, if the local roadway improvement project is to alter lane width, which in turn is expected to alter cross-segment traffic flow, HERS could explicitly quantify the expected impact on annual crashes of the new lane width in conjunction with a different level of traffic (model [7] possesses this association term). Likewise, if the local project is to introduce an intersection where there currently is none, and the segment is heavily used by commercial trucks, HERS could explicitly quantify the expected impact on annual crashes. With the current HERS model [1], similar analyses would be very difficult, if not impractical.

Further Research

HSIS data was obtained for several participating states. Rather than implement model [7] in HERS software strictly on the basis of Ohio results, it would be prudent to repeat the exploratory data analysis and model fitting exercises on other states. If project scope calls for revisiting the current HERS crash prediction model for each functional class, naturally, this remains to be accomplished. So long as the objective is to predict count data (crashes, fatalities, etc.), significant results may be realized by repeated application of the same data exploration and model fitting techniques.

The HERS three-step process of estimating safety costs

(1) Convert crash rates to crash counts.

(2) Estimate overall severity.

(3) estimate overall cost)

may be inherently inaccurate, introducing substantial uncertainties with each step. Since each participating HSIS state reports crash severity data, focus could shift towards estimating crash severity directly. Data from outside sources is a prerequisite for analyzing the accuracy of HERS safety costs estimates and proposing alternatives.