Final Report VTRC 10-R22

Virginia Transportation Research Council

research report

Causal Factors for Intersection Crashes in Northern Virginia

http://www.virginiadot.org/vtrc/main/online_reports/pdf/10-r22.pdf

JOHN S. MILLER, Ph.D., P.E.
Associate Principal Research Scientist

NICHOLAS J. GARBER, Ph.D., P.E.
Professor of Civil Engineering
Department of Civil and Environmental Engineering
University of Virginia

SANTHOSH K. KORUKONDA
Graduate Research Assistant

Virginia Transportation Research Council, 530 Edgemont Road,
Charlottesville, VA 22903-2454, www.vtrc.net, (434) 293-1900

Standard Title Page - Report on Federally Funded Project

1. Report No.: FHWA/VTRC 10-R22	2. Government Accession No.:	3. Recipient's Catalog No.:
4. Title and Subtitle: Causal Factors for Intersection Crashes in Northern Virginia		5. Report Date: June 2010 6. Performing Organization Code:
7. Author(s): John S. Miller, Ph.D., Nicholas J. Garber, Ph.D., and Santhosh K. Korukonda		8. Performing Organization Report No.: VTRC 10-R22
9. Performing Organization and Address: Virginia Transportation Research Council, 530 Edgemont Road, Charlottesville, VA 22903		10. Work Unit No. (TRAIS): 11. Contract or Grant No.: 80574
12. Sponsoring Agencies' Name and Address: Virginia Department of Transportation, 1401 E. Broad Street, Richmond, VA 23219; Federal Highway Administration, 400 North 8th Street, Room 750, Richmond, VA 23219-4825		13. Type of Report and Period Covered: Final 14. Sponsoring Agency Code:

15. Supplementary Notes:

16. Abstract:

Intersection crashes cost the nation more than $40 billion annually, account for more than one-fifth of all highway crash fatalities nationally, and totaled almost 75,000 in the Virginia Department of Transportation's (VDOT) Northern Virginia District for the period 2001 through 2006. Although VDOT maintains several databases containing more than 170 data elements with detailed crash, driver, and roadway attributes, it was not clear to users of these databases how these data elements could be used to identify causal factors for these intersection crashes for two reasons: (1) the quality of some of the data elements was imperfect, and (2) and random variation is inherent in crashes. This study developed an approach to address these two issues.

To address the first issue, the completeness and accuracy of the 179 data elements that comprise the VDOT CRASHDATA database were assessed. For the 76 data elements for which the quality of the data was imperfect, eight rules for using these elements were developed. The rules indicate which data elements should be used in certain circumstances; which data elements are incomplete; and how to manipulate the data for certain applications.

To address the second issue, classification trees and crash estimation models (CEMs) were developed. The trees showed that specific causal factors, such as the approach alignment or surface condition, successfully indicate whether a given crash was a rear-end or angle crash. By extension, the trees suggested that intersection crashes were not purely random. Accordingly, it was feasible to develop CEMs that for 17 intersection classes predicted the number of crashes for a 1-year period for four crash types: rear-end, angle, injury, and total. The 68 CEMs showed deviance-based pseudo R-square values between 0.07 and 0.74, suggesting that the causal factors explained some, but not all, of the variation in intersection crashes. The CEMs varied by intersection class.

Two actions with regard to crash data analysis may be taken as detailed in this report. First, the eight crash data quality rules developed in this study should be considered for use on a case-by-case basis for studies requiring intersection crash data. Second, when they are collected at the crash scene, the variables that successfully classified rear-end and angle crashes may be given increased attention such that every effort is made to ensure these data elements are accurately recorded.

17 Key Words:	18. Distribution Statement:
Safety management, safety programs, transportation safety, planning, safety conscious planning, urban planning	No restrictions. This document is available to the public through NTIS, Springfield, VA 22161.

19. Security Classif. (of this report):	20. Security Classif. (of this page):	21. No. of Pages:	22. Price:
Unclassified	Unclassified	61	

FINAL REPORT

CAUSAL FACTORS FOR INTERSECTION CRASHES IN NORTHERN VIRGINIA

John S. Miller, Ph.D., P.E.
Associate Principal Research Scientist

Nicholas J. Garber, Ph.D., P.E.
Professor of Civil Engineering
Department of Civil and Environmental Engineering
University of Virginia

Santosh K. Korukonda
Graduate Research Assistant

In Cooperation with the U.S. Department of Transportation
Federal Highway Administration

Virginia Transportation Research Council
(A partnership of the Virginia Department of Transportation
and the University of Virginia since 1948)

June 2010
VTRC 10-R22

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Virginia Department of Transportation, the Commonwealth Transportation Board, or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

ABSTRACT

Intersection crashes cost the nation more than $40 billion annually, account for more than one-fifth of all highway crash fatalities nationally, and totaled almost 75,000 in the Virginia Department of Transportation's (VDOT) Northern Virginia District for the period 2001 through 2006. Although VDOT maintains several databases containing more than 170 data elements with detailed crash, driver, and roadway attributes, it was not clear to users of these databases how these data elements could be used to identify causal factors for these intersection crashes for two reasons: (1) the quality of some of the data elements was imperfect, and (2) and random variation is inherent in crashes. This study developed an approach to address these two issues.

To address the first issue, the completeness and accuracy of the 179 data elements that comprise the VDOT CRASHDATA database were assessed. For the 76 data elements for which the quality of the data was imperfect, eight rules for using these elements were developed. The rules indicate which data elements should be used in certain circumstances; which data elements are incomplete; and how to manipulate the data for certain applications.

To address the second issue, classification trees and crash estimation models (CEMs) were developed. The trees showed that specific causal factors, such as the approach alignment or surface condition, successfully indicate whether a given crash was a rear-end or angle crash. By extension, the trees suggested that intersection crashes were not purely random. Accordingly, it was feasible to develop CEMs that for 17 intersection classes predicted the number of crashes for a 1-year period for four crash types: rear-end, angle, injury, and total. The 68 CEMs showed deviance-based pseudo R-square values between 0.07 and 0.74, suggesting that the causal factors explained some, but not all, of the variation in intersection crashes. The CEMs varied by intersection class.

Two actions with regard to crash data analysis may be taken as detailed in this report. First, the eight crash data quality rules developed in this study should be considered for use on a case-by-case basis for studies requiring intersection crash data. Second, when they are collected at the crash scene, the variables that successfully classified rear-end and angle crashes may be given increased attention such that every effort is made to ensure these data elements are accurately recorded.

FINAL REPORT

CAUSAL FACTORS FOR INTERSECTION CRASHES IN NORTHERN VIRGINIA

John S. Miller, Ph.D., P.E.
Associate Principal Research Scientist

Nicholas J. Garber, Ph.D., P.E.
Professor of Civil Engineering
Department of Civil and Environmental Engineering
University of Virginia

Santosh K. Korukonda
Graduate Research Assistant

INTRODUCTION

Intersections represent one of the most complex situations that drivers encounter because of their many conflict points. Drivers must perform a series of complex tasks—maintaining proper lane position; responding to signs, signals, and markings; evading conflicting or adjacent traffic, pedestrians, and bicyclists; and increasing or decreasing their speed as appropriate. At least in part because of this heavy cognitive burden, crashes at intersections accounted for 44% of all reported crashes on the national highway network (Federal Highway Administration [FHWA], 2004); more than one-fifth of all fatalities nationally (FHWA, 2008); and annual societal costs of $40 billion (FHWA, 2002). It has also been reported that for drivers aged 64 and above, 60% of injury crashes and 37% of fatal crashes occur at intersections (Hauer, 1988). According to the Virginia Department of Transportation's (VDOT) CRASHDATA database, there were 154 fatal intersection crashes in Virginia in 2006. For the 6-year period from 2001 through 2006, there were 75,000 crashes at signalized, stop-controlled, and yield-controlled intersections in VDOT's Northern Virginia District (NOVA District) as found in the VDOT CRASHDATA database. It is not surprising, therefore, that Virginia's Strategic Highway Safety Plan (Virginia's Surface Transportation Safety Executive Committee, 2007) lists intersection safety as an emphasis area.

One possible resource for obtaining a better understanding of the causes of intersection crashes is the VDOT CRASHDATA database —a centralized repository containing extensive crash, geometric, and roadway data. With 179 data elements for each crash (e.g., driver age, lane width, and vehicle type as shown in Appendix A), almost 33,000 total crashes (not just intersection crashes) in the NOVA District, and approximately 18 years of data, an extensive amount of information is readily available. To take advantage of this database, however, two challenges must be met.

First, the extent to which these data are reliable needs to be understood. For example, the data quality for an element based on a law enforcement officer's record at the time of a crash

(e.g., weather condition) may not be identical with the data quality of an element based on a roadway inventory (e.g., the lane width). Even data elements that are collected at the same time (e.g., weather and surface condition) may not necessarily have the same level of quality.

Second, this large amount of information must be translated into useable findings given the risk of a crash is a probabilistic process. For example, given that a crash has occurred, is it possible to identify geometric characteristics that will tend to affect rear-end crashes as opposed to angle crashes? Alternatively, is it the case that the random variation is so large that no predictor variables can be identified? If it is feasible to identify key predictor variables, to what extent is it possible to predict the number of rear-end (or angle, injury, or total) crashes as a function of geometric and traffic characteristics over which VDOT exerts some influence?

Meeting both of these challenges can provide a tangible safety benefit: reduced data collection costs when intersection safety countermeasures are evaluated. If reliable data elements can be identified (or if how to render other data elements reliable in certain situations can be ascertained) and used to determine the expected number of crashes, it should be possible to evaluate intersection safety countermeasures without having to collect as much comparison site data.

PURPOSE AND SCOPE

The purpose of this research was twofold: (1) to identify the most beneficial variables in the VDOT CRASHDATA database that can be used to estimate crashes in the NOVA District and, in doing so, (2) o use the VDOT CRASHDATA database to the maximum extent possible given Virginia's investment in the database. The study had three objectives:

1. Determine which data elements in the VDOT CRASHDATA database have adequate data consistency and completeness, and devise rules for working with those data elements for which the data quality in terms of the two dimensions is inadequate.

2. Determine the minimum set of data elements that classifies crash types, i.e., the particular data elements that indicate collisions at an intersection will tend to be rear-end rather than angle crashes and injury rather than non-injury crashes.

3. Develop crash estimation models (CEMs) that predict the number of crashes as a function of intersection characteristics.

The scope of the study was limited to intersection crashes that occurred in the Virginia counties of Fairfax, Prince William, and Loudoun during the period 2000 through 2005.

METHODS

Four tasks were conducted to achieve the study objectives:

1. Conduct a literature review of the relevant literature.

2. Collect, reduce, and assess the quality of the crash data available for Northern Virginia intersections.

3. Develop and evaluate classification trees for classifying crashes as rear-end (or not rear-end), angle (or not angle), and injury (or not injury)

4. Develop CEMs.

Literature Review

Literature relating to the following three areas was identified through the use of TRIS and other search engines and reviewed:

1. *crash data quality*, including limitations of existing datasets and the identification of variables that are essential to safety-related studies

2. *classification trees*, including methods for developing such trees and their practical application

3. *CEMs*, including the mathematical form and specification of critical parameters.

Collection, Reduction, and Assessment of Quality of Crash Data

For every reported crash in Virginia, data elements from the Police Crash Report (Form FR300) are stored in the VDOT CRASHDATA database. These data elements include crash location, severity, driver behavior, vehicle characteristics, and prevailing environmental conditions at the time of the crash. VDOT also maintains the Traffic Monitoring System (TMS), which contains annual average entering volumes for intersections, annual average daily traffic (AADT), and heavy vehicle truck percentages. VDOT's Highway Traffic Records Inventory System (HTRIS) contains, for each link, geometry, traffic control, and operations data. To place these data in a form suitable for analysis, eight steps were followed:

1. Extract crash and roadway data elements.
2. Eliminate dimensional heterogeneity.
3. Create intersection variables.
4. Extract or interpolate intersection-entering volumes.
5. Manually obtain selected geometric data elements.
6. Categorize discrete and continuous data into homogenous bins.
7. Tabulate crashes by collision type, severity type, and intersection characteristics.
8. Document data deficiencies and related solutions.

1. Extract crash and roadway data elements.

Data were extracted from the VDOT CRASHDATA, TMS, and HTRIS databases. The extraction required the creation of 24 queries in the Structured Query Language (SQL) environment. Each year of data had to be extracted individually. Further, because a crash could occur either upstream or downstream of a given node that is the reference for the crash, two queries—one for upstream and one for downstream—were needed. In addition, one query was needed to extract data from the VDOT CRASHDATA and TMS databases and a separate query was needed to extract data from the HTRIS database. Intersection crashes were selected by obtaining crashes within 0.03 mi (about 150 ft) of the intersection.

2. Eliminate dimensional heterogeneity.

The VDOT CRASHDATA database is composed of several linked data tables, such as CrashDocument, CrashVehicle, and CrashPedestrian. When a single crash involves two vehicles, the dimensions of the data are heterogeneous because whereas only one value for each attribute is required from the CrashDocument table (e.g., one intersection location or one speed limit), two values for each attribute are required from the CrashVehicle table (e.g., two sets of driver's ages or two vehicle types). To eliminate this dimensional heterogeneity, the entry from the CrashVehicle table that had an errant driver action was selected. For 3.1% of the two-vehicle crashes and 8.8% of the multiple-vehicle crashes, two or more errant driver actions were listed. In those cases, an entry from the CrashVehicle table was chosen randomly. A similar process was used for crashes with two or more pedestrians.

3. Create intersection variables.

The intersection variables were created from HTRIS. HTRIS data are stored by link rather than intersection. As a consequence, eight variables that classify the intersection by area type (rural versus urban), traffic control (signalized versus stop-controlled), access type (e.g., undivided two-way), administrative roadway type (primary versus secondary), and functional class (e.g., local or arterial) were developed. The latter three variables are repeated for the major and minor approaches. These variables were based on the link characteristics stored in HTRIS. Intersections that had neither signalized control nor stop control, such as interstate ramps and uncontrolled driveways, were not included in the database.

4. Extract and interpolate intersection-entering volumes.

Intersection-entering volumes were obtained from the TMS database. For some years and at some intersections, a traffic volume was unavailable, and in some other cases, the traffic volumes changed dramatically. In these cases, volumes were projected based on previous or later volumes. For example, if data at an intersection were available for 2003 but not for 2000 through 2002, the missing values were projected based on the 2003 volume and an average growth factor (based on 2004 and 2005). A similar approach was applied if the annual volume change was greater than 25% (and the data were thus deemed inconsistent). The decision to exclude volumes that changed by 25% or more represented a tradeoff between the possibility that

a dramatic change was the result of an error and the possibility that such a dramatic change reflected a change in land development.

5. Manually obtain select geometric data elements.

Although some geometric data elements were available in HTRIS, other geometric data elements that are not routinely collected or are part of the roadway inventory that was not fully completed at the start of the project (e.g., number of lanes) had to be obtained manually by examining aerial photographs obtained from Google Maps. These attributes were the number of approaches for an intersection (e.g., three-way versus four-way), number of turn lanes, type of channelization, presence of frontage roads, presence of curb cuts, presence of on-street parking, and number of lanes (1, 2, or more than 2). It was possible for the intersection geometry to change during the study period (2000-2005). Although it was possible to include changes in attribute values for those attributes stored in HTRIS (e.g., pavement width), the fact that photographs from Google Maps reflected only the latest year (2005) meant that the attributes collected manually would not necessarily reflect the earlier intersection geometry if changes to this geometry occurred from 2000 through 2005. In some cases, when it was difficult to view the roadway network, VDOT's GIS Integrator (Figure 1) was used to verify the information from Google Maps.

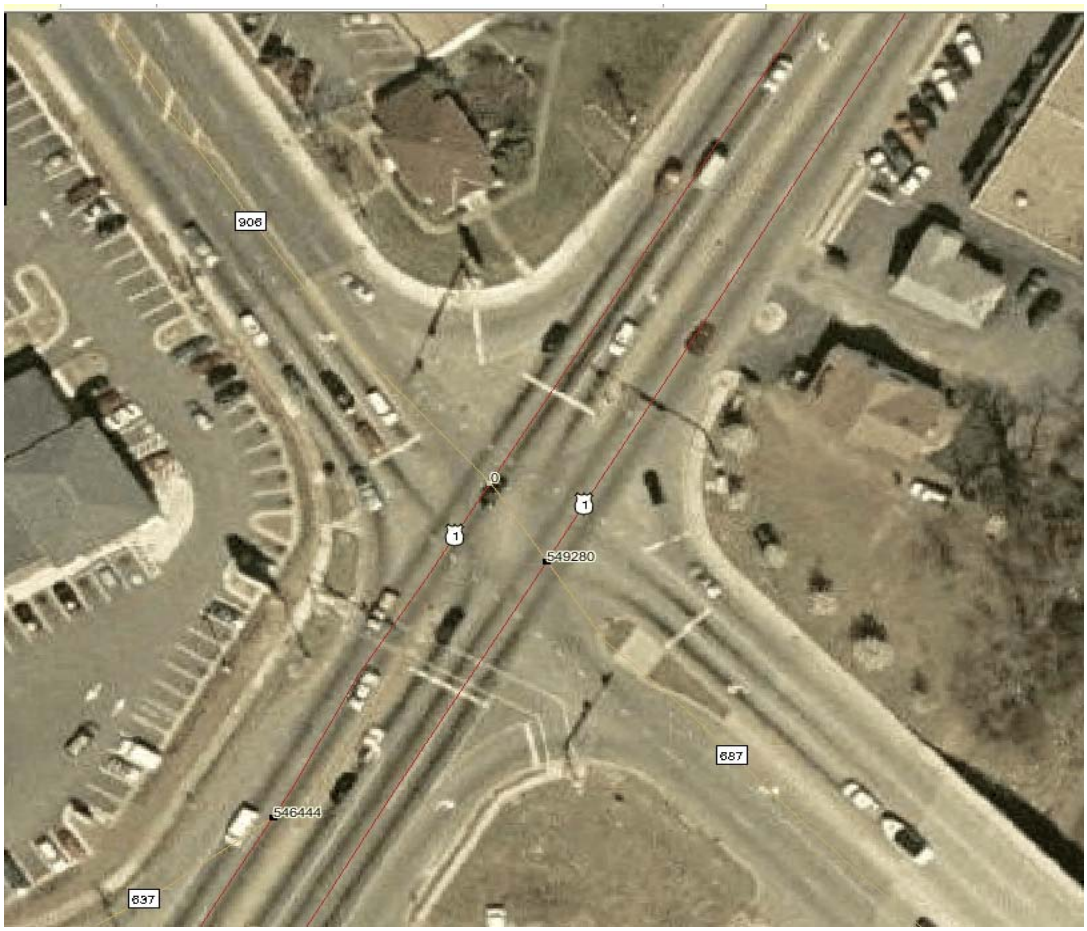

Figure 1. Intersection of Occoquan Road and Jefferson Davis Highway in Prince William County

6. Categorize discrete and continuous data into homogenous bins.

Because the data are a combination of continuous variables and discrete variables, and because some of the discrete variables had numerous bins (categories), it was necessary to transform these data into discrete variables with a comparable number of bins. (Having too many bins for a given variable greatly increases computational time.) For example, the *DRIVERACTION* variable (with 43 bins) was collapsed into 10 bins, and driver age (a continuous variable with hence an infinite number of bins) was collapsed into five categories: <= 19, >19 to 25, >25 to 50, >50 to 65, and >65. This dataset was used to develop the classification trees.

7. Tabulate crashes by collision type and severity type for each intersection.

An aggregate dataset was created by summing crash frequencies by collision type (e.g., rear-end or angle) and severity type (e.g., fatal, injury, or property damage only) for each intersection over the 6-year period. For approximately 2% of the intersections, an intersection-level attribute changed during the period. In those situations, the intersection was represented as two 6-year equivalent data points. For example, if intersection A had one left turn lane and 40 crashes for the first 3 years and two left turn lanes and 35 crashes for the remaining 3 years, the intersection would be represented as two data points: one intersection with one left turn lane and 80 crashes (over a 6-year period) and one intersection with two left turn lanes and 70 crashes (over a 6-year period). This dataset was used to develop the CEMs. Because the annual crash frequencies were added and because the traffic volumes were averaged, the initial CEMs predicted a 6-year crash frequency; however, the final models in Appendix B predict a 1-year crash frequency.

8. Document data deficiencies and related solutions.

The first seven steps, summarized in Figure 2, led to a single 6-year database (2000-2005) containing 72,218 crashes occurring at more than 6,000 intersections. In the course of creating this database, data deficiencies such as missing or contradictory data elements and rules that could overcome these deficiencies were identified and documented.

Development and Evaluation of Classification Trees

Classification trees were used to identify the minimum set of data elements for classifying crash types. Four steps comprised this process.

1. Develop classification trees.
2. Induce rules regarding the necessary variables from the classification trees.
3. Assess the accuracy of the classification trees.
4. Identify the most important variables with regard to the rules.

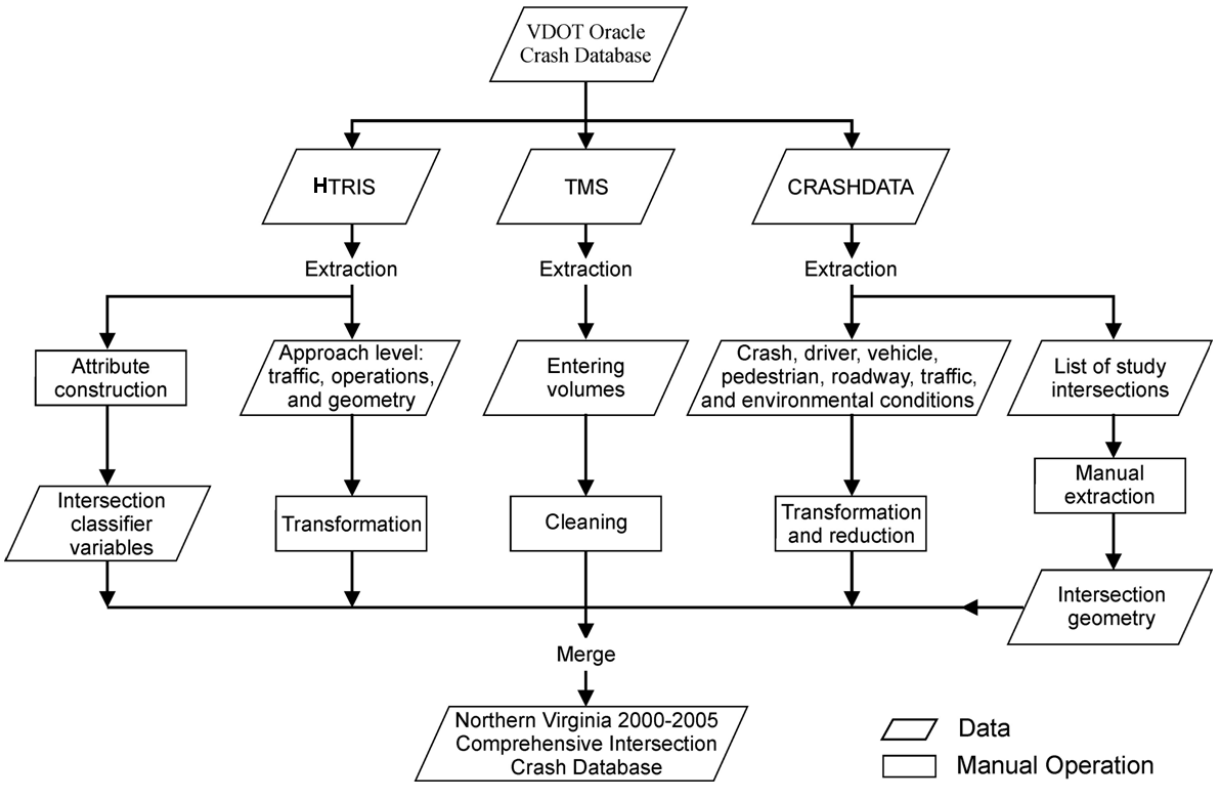

Figure 2. Summary of Data Collection and Reduction. VDOT Oracle Crash Database = VDOT's Crash Report Database.

As discussed previously, the reason for developing the classification trees was to determine if a minimum set of variables could classify crashes. If so, this may lead to two decisions if effective trees can be developed: (1) focus future data collection on just those minimal variables, and/or (2) use those minimal variables in developing CEMs, thereby ignoring the rest of the dataset.

1. Develop classification trees.

In this report, a classification tree (Han and Kamber, 2001) refers to a hierarchical set of decisions that predicts whether a given crash is a specific crash type. For each of 17 intersection types, three classification trees were developed to predict whether or not a given crash was rear-end, angle, or injury. As a consequence, a total of 51 trees were developed. Rear-end and angle crashes were chosen because these were believed to indicate the predominant types of crashes occurring at intersections. For example, a cursory review of crashes in Loudoun County for the period 2006 through 2008 where traffic control was designated by a stop sign or traffic signal in VDOT's Crash Report Database indicated that 41% of crashes were rear-end and 45% were angle. Injury crashes were chosen because these were believed to be a useful indication of risk of personal harm, as opposed to fatal crashes, which tend to be so few in number that it is difficult to use them to assess the impact of a wide variety of causal factors.

The Gini impurity index (Shmueli et al., 2007) was used to split nodes, and the optimal tree size was identified through the minimal-cost complexity cross-validation pruning algorithm

proposed by Breiman et al. (1984). Purity denotes the extent to which node splitting results in members of each resulting category having the same value of the dependent variable. For example, with a simple classification tree consisting of only one node (e.g., driver action) and one dependent variable (e.g., whether the crash type is rear-end or not rear-end), driver action is used to split the crashes into two categories: rear-end and not rear-end. To say that purity was achieved means that the values of driver action used to split the data were chosen such that, to the extent possible, one category of crashes was entirely rear-end and the other category was not rear-end. However, if purity were the only goal when a classification tree was created, the result would be a complex tree that classified every crash correctly in the training set. Thus, in the creation of such trees, two competing goals are considered: purity (e.g., the accuracy with which training data are classified) and simplicity (e.g., the ability to create a classification tree with a relatively modest number of nodes and branches). Breiman et al. (1984) offered a method for considering both goals, and this method is incorporated in the software used for this project to develop classification trees.

Figure 3 shows the classification tree for rear-end crashes at rural three-way signalized intersections on four-lane roads. In Figure 3, each node contains an identification number (ID), the number of data points that belong to the node (N), a histogram representing the frequency of rear-end crashes, and a *1* or *0* signifying whether the tree represents rear-end or non rear-end crashes. For example, the top node indicates that there were 475 crashes in the dataset; these crashes are initially partitioned based on the driver action as follows: 274 points where the driver action has a value of 1, 4, 6, 7, or 10 and 201 data points where the driver action has a value of 0, 2, 3, 5, or 8. Each node similarly divides crashes into rear-end and non–rear-end; e.g., with respect to the aforementioned 274 points (or crashes since each crash is one point), two sub-datasets are created: one with 260 crashes (if APP_ALIGNMENT = 1, 4, 0, or 2) and one with 14 crashes (if APP_ALIGNMENT = 1, 3, or 8). The utility of these subdivisions becomes evident when rules are induced as shown in Step 2.

2. Induce rules from the classification trees.

Rules were induced from each tree to predict rear-end, angle, and injury crashes. For example, in Figure 3, Rule 1 may be traced by starting with the top node (node 0) and working down to the leftmost node (node 8). Rule 1 is:

IF	(DRIVER_DRAC = 4, 6, 7, 10, or 1)	AND
	(APP_ALIGNMENT = 1, 4, 0, or 2)	AND
	(VEH_SPEED <= 50.5000)	AND
	(CR_TRCONTROL = 5, 6, 7, 0, 1, or 4)	

THEN the crash is of type rear-end.

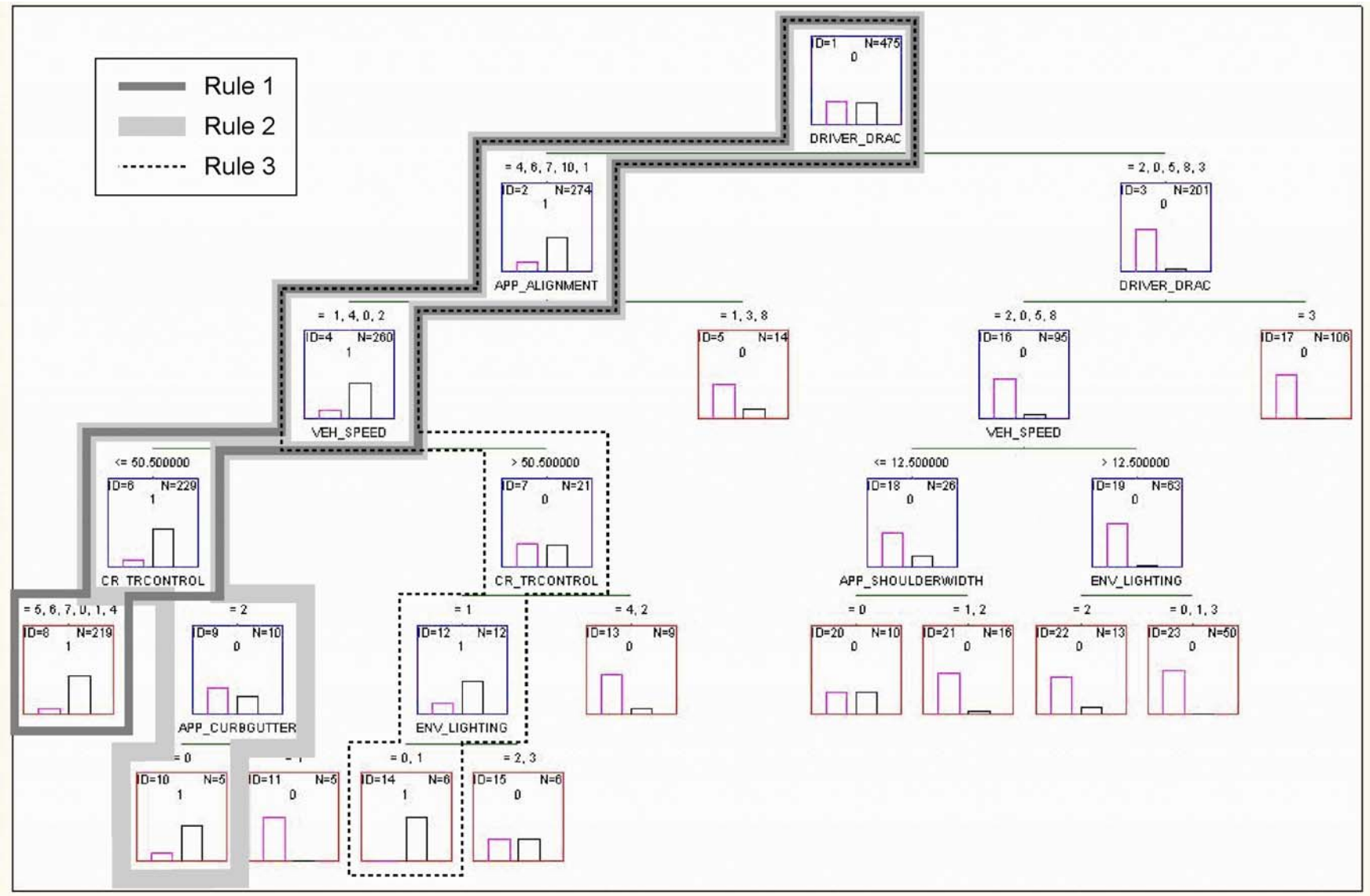

Figure 3. Classification Tree for Rear-End Crashes at Rural Signalized 3-Way Intersections on 4-Lane Roads. Rules are heuristics for classifying crashes as rear-end based on select attributes. For example, Rule 1 indicates that a crash is of the type rear-end if the crash has the attributes of CR_TRCONTROL, VEH_SPEED, APP_ALIGNMENT, and DRIVER_DRAC shown in the figure.

A qualitative translation of Rule 1 is as follows:

IF (Driver action is following too close; failure to maintain proper control, improper lane change, or overcorrection; driver inattention; avoiding pedestrians, animals, or vehicles, or other improper action; or exceeded speed limit, exceeded safe speed.)
AND (Alignment is horizontal level; curved grade; curve straight and level or straight grade.)
AND (Vehicle speed is less than or equal to 50.5 mph.)
AND (Traffic control is no passing lanes; yield sign; others [e.g., not stop sign and not slow or warning sign]; none; traffic signal; or traffic lanes marked.)

THEN the crash is of type rear-end.

Note that in Figure 3, two additional rules can be induced for predicting rear-end crashes starting at node 10 (shown as Rule 2) and node 14 (shown as Rule 3).

3. Assess the accuracy of the classification trees.

The classification trees were developed using only 75% of the available crashes; these comprised the "training" dataset. The accuracy of these trees was determined by using the trees to classify crashes for the remaining 25% of the crashes—hence the "test" dataset. For example, whereas the classification tree in Figure 3 classified approximately 88% of crashes correctly in the training dataset, it classified approximately 84% of crashes correctly in the test dataset.

4. Identify the most important variables.

The importance of each rule in a tree may be determined from its *gain*, which is the number of data points classified by the rule. For example, the gain from Rule 1 is 219, since the rule properly classifies 219 data points as shown in node 8. By contrast, the gain from Rule 2 is 6 (as shown in node 14), and the gain from Rule 3 is 5 (as shown in node 10). Thus, Rule 1 is the most important rule (since it has the largest gain) and Rule 3 is the least important rule (since it has the smallest gain). Note further that the variable *ENV_LIGHTING* appears only in Rule 2 and that the variable *APP_CURBGUTTER* appears only in Rule 3. For this example only, the variable *ENV_LIGHTING* (which contributes to Rule 2 with a gain of 6) is more important than the variable *APP_CURBGUTTER* (which contributes to Rule 3 with the smaller gain of 5). In practical terms, if one of these two variables had to be dropped, *APP_CURBGUTTER* (and hence lose a rule with a gain of 5) rather than *ENV_LIGHTING* (and hence lose a rule with a gain of 6) would be the preferred drop. For each tree, the 10 most important variables were identified to develop the minimal set of variables for classifying crashes.

The rules also show the influence of each variable on crash classification. For example, the main reason Rule 1 has a substantially higher gain than the other two rules is simply that a large proportion of the rear-end crashes in this particular dataset occurred on facilities with vehicle speeds less than or equal to 50.5 mph (see node 6 with 229 crashes) than on roads with speeds more than 50.5 mph (see node 7 with 21 crashes). (Generally, these speeds are integers and are estimated by law enforcement officers after the crash; thus a speed might be recorded as

50 mph or 55 mph but not 50.5 mph. Thus the "equal to 50.5" is noted here merely for the sake of completeness.)

Development of Crash Estimation Models

CEMs were developed that predicted intersection angle, rear-end, injury, and total crashes over a 1-year period for 17 classes of intersections. (Initially a 6-year period was chosen because it reduces the number of intersections with zero crashes and the correlation of dependent variables, both of which may adversely affect model estimation; however the CEM in Appendix B gives a 1-year crash frequency.) The functional form of the CEM is given by Equation 1.

$$\hat{y}_i = \text{Expected number of crashes at site } i = a(\text{Volume})^b \exp^{(XB)} \quad \text{[Eq. 1]}$$

where

a, b, and **B** = parameters
volume = total entering daily volume of the intersection
X = a vector of intersection attributes found to be statistically significant.

To estimate a, b, and **B**, negative binomial generalized linear models were used with a log link function and the fit of these models was judged by their deviance R-squared (R^2_{DEV}) measures.

A justification for this goodness-of-fit measure is given in the literature (Cameron and Windmeijer, 1996; McCullagh and Nelder, 1989). In classical linear models, maximum likelihood estimation finds the parameters (a, b, and **B**) that minimize the difference between the predicted crashes $\hat{y}_i$ and the actual crashes y_i. Such linear models presume that the errors are normally distributed with zero mean and a constant variance; further, the actual crashes y_i are normally distributed with a constant variance. Accordingly, maximum likelihood estimation for linear models seeks to minimize Equation 2. Negative binomial models, however, do not presume that y_i follow this normal distribution; instead, it is presumed that y_i follow the negative binomial distribution as shown in Equation 3 (SAS Institute Inc., 2009), where μ is the expected value of the response variable and k is the dispersion parameter. Accordingly, maximum likelihood estimation with the negative binomial model seeks to minimize Equation 4 (Hardin and Hilbe, 2007).

$$\sum_i (y_i - \hat{y}_i)^2 \quad \text{[Eq. 2]}$$

$$f(y) = \frac{\Gamma(y + 1/k)}{\Gamma(y+1)\Gamma(1/k)} \frac{(k\mu)^y}{(1+k\mu)^{y+1/k}} \quad \text{[Eq. 3]}$$

$$\sum_i \{y_i \log(k\hat{y}_i) - (y_i + 1/k)\log(1 + k\hat{y}_i) + \log(\Gamma(y_i + 1/k)/\Gamma(y_i + 1)\Gamma(1/k))\} \quad \text{[Eq. 4]}$$

For these types of models, Cameron and Windmeijer (1996) suggested evaluating the goodness of fit based on Equation 5.

$$R^2_{DEV} = \frac{l(\hat{y}) - l(\bar{y})}{l(y) - l(\bar{y})} \quad \text{[Eq. 5]}$$

In Equation 5, $l(\bar{y})$, $l(\hat{y})$, and $l(y)$ are the log-likelihood values of the null, current, and saturated models. The null model is equivalent to taking a guess in that all of the independent variables are set to zero; the current model is the one for which the goodness of fit is to be evaluated, and the saturated model consists of the same number of parameters as the number of data points and in effect simply replicates the observed dependent values.

SAS software was used to implement Equation 4, which yielded the parameters for the CEMs in Equation 1 and the goodness-of-fit measures in Equation 5.

RESULTS AND DISCUSSION

Literature Review

Crash Data Quality

It has been shown that multiple variables are needed to explain crash causation, including those related to roadway geometry (Campbell and Knapp, 2005; Wang et al., 2003), traffic volume (Wang et al., 2003), and the vehicle and driver factors (Kaysi and Abbany, 2007). For example, Yan et al. (2005) used a relative accident involvement ratio to measure crash propensity and the significance of driver, vehicle, and roadway factors. Abdel-Aty et al. (2006) found that an increase in the number of lanes led to an increase in crash frequency, where it appears that the number of lanes represented intersection complexity. Improper driver behavior was found to be the principal cause of crashes at urban signalized intersections in Riyadh, Saudi Arabia (Al-Ghamdi, 2003). A model that predicted aggressive driver behavior included binary variables such as whether the driver's age was less than 26 and whether the vehicle was a sports car (Kaysi and Abbany, 2007). Thus, the literature shows that a wide range of highway, geometric, vehicle, and driver variables may be needed for studies of crash causation.

The literature also highlights specific data deficiencies, such as the need for more detailed crash data and location information (Al-Ghamdi, 2003), the lack of integrated crash and geometric data (Campbell and Knapp, 2005), the problem of underreporting crashes (Kumara et al., 2003), crash models where important variables were omitted (Lord et al., 2005; Oh et al., 2003), and the use of poorly measured or surrogate variables (Oh et al., 2003). The manner in which the studies have been conducted also emphasizes the need for disaggregate crash, traffic, and geometric data, as was suggested by Wong et al. (2007). Thus, the literature suggests that data quality is a national issue and not unique to Virginia. The challenge of integrating data from disparate datasets (Campbell and Knapp, 2005) also appears applicable to Virginia's situation of gathering data from disparate databases.

Classification Trees

The literature suggests that classification trees may hold promise for safety-related studies. Kim et al. (2007) demonstrated that rural intersection crash data are hierarchical in structure, categorizing explanatory variables at two levels: crash and intersection. Classification trees have been implemented for the purposes of analyzing intersection crashes (Keller et al., 2006); relating injury levels to driver, vehicle, environmental, and crash variables (Chang and Wang, 2006; Tesema et al., 2005); and identifying freeway rear-end crashes (Pande and Abel-Aty, 2005).

Extensive literature is devoted to the different steps required to create an effective classification tree (Breiman et al., 1984; Han and Kamber, 2001; Shmueli et al., 2007). When a large number of variables is present, a key challenge is deciding when to split the nodes. Although multiple approaches are feasible (Han and Kamber, 2001), one technique is the Gini impurity index, which is given as Equation 6 (Shmueli et al., 2007).

$$I = 1 - \sum_{i=1}^{m} p_i^2 \qquad \text{[Eq. 6]}$$

where

$$p_i = \frac{s_i}{s} \qquad \text{[Eq. 7]}$$

For example, if there are 60 rear-end crashes and 40 angle crashes, m = two distinct classes, $s_1 = 60$, $s_2 = 40$, $s = 100$, $p_1 = 0.60$, and $p_2 = 0.40$.

When a sample is split using a given attribute (e.g., driver action), the reduction in impurity is given by Equation 8:

$$\text{Reduction in impurity} = I - \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} \left(1 - \sum_{i=1}^{m} p_{ij}^2 \right) \qquad \text{[Eq. 8]}$$

Classification tree algorithms applying the Gini impurity index select attributes for each node split by maximizing the reduction in impurity shown in Equation 8.

In practice, Equations 6 through 8 are not implemented by hand but rather with software. Although splitting the nodes is a critical decision in tree building, it is not the only decision: another step is postpruning, where a full tree is developed and then selected branches are removed (Breiman et al., 1984; Han and Kamber, 2001). The reason for this step is that classification trees are prone to overfitting: in the extreme, a tree that accurately predicted every rear-end crash in Figure 3 would have a large number of nodes and would reflect the noise in the training set such that the tree would not be valid for a subsequent class of data. Thus, testing the accuracy of these trees on a different dataset is a necessary step to ensure the tree is valid.

Crash Estimation Models

In developing CEMs, the state of the art is to assume that crashes follow the negative binomial distribution. It has been suggested that when modeling crashes, multiple linear regression is not appropriate (Jovanis and Chang, 1986) and Poisson regression (Joshua and Garber, 1990; Miaou and Lum, 1993) is preferred. This is because linear models, which assume the normal distribution, are not appropriate for crash frequencies that are not normally distributed (Miaou and Lum, 1993). However, the Poisson assumption also has limitations: although valid for cases where a crash probability is quite low, it may be the case that the variance is greater than the mean and hence the Poisson assumption is not appropriate (Lord et al., 2004; Miaou and Lum, 1993). In response to this large variance, one option is to explore zero-inflated Poisson (ZIP) models (Miaou, 1994), although others have suggested that the appearance of a large number of zeroes in crash data that initially made ZIP models attractive is the result of data problems, such as a high percentage of missing crashes or the grouping of low- and high-risk crash sites, such that ZIP models should not be used (Lord et al., 2005). In response, models based on the assumption that crashes follow a negative binomial distribution have gained popularity, with many studies recommending or using models based on the negative binomial distribution (Abdel-Aty and Radwan, 2000; Garber et al., 2005; Miaou, 1994; Poch and Mannering, 1996; Shankar et al., 1995). The literature (Cameron and Windmeijer, 1996; McCullagh and Nelder, 1989) provides guidance on how to evaluate such models based on the negative binomial distribution.

Variations in modeling persist, however. Kumara et al. (2003) applied a random-effect negative binomial model to identify intersection crash causal factors. However, while modeling crash frequencies with a multiyear panel of cross-sectional data, Ulfarsson and Shankar (2003) reported that negative multinomial models performed better than negative binomial and random-effects negative binomial models.

Further, negative binomial models use a dispersion parameter (which explains the extent to which the variance is greater than the mean and hence that the negative binomial, rather than the Poisson, distribution is appropriate). The literature suggests a constant dispersion parameter across sites and time periods should not be assumed as this limits the predictive power of the crash models (Miaou and Lord, 2003). The aforementioned data quality problem of omitting important variables also affects this dispersion parameter (Mitra and Washington, 2007).

Guiding Principles From the Literature Review

The literature review suggests four principles that should guide an analysis of Virginia intersections:

1. In terms of data quality, a wide variety of disaggregate data elements may be required; thus, every effort should be made to link, where possible, crash, geometric, vehicle, and driver data.

2. Unless Virginia data are different from those of other states, the quality of some data elements will be a limitation; thus, procedures are needed to assess the quality of the data.

3. To the extent that intersection crashes are hierarchical in nature as has been suggested (Kim et al., 2007), it should be feasible to use regression trees to identify the most important rules and, by extension, the most important variables, for determining intersection crash causality. Before such trees can be used, however, it is essential that they be tested on a dataset different from that used to create them to ensure the trees reflect true underlying relationships and not noise within the data.

4. The form of the CEM should be one that is based on the negative binomial distribution, but the model should allow for different dispersion parameters rather than presuming a constant value. Because it is believed that Poisson models are appropriate for variation by time at a given site, it appears likely that dispersion parameters will need to vary by site as opposed to varying by year. Thus, it is reasonable to develop CEMs for specific intersection classes rather than specific intersection years.

Completeness and Consistency of Crash Data

Data Elements With Adequate Completeness and Consistency

Two dimensions of data quality were considered: completeness (i.e., whether a value for the data element is consistently given) and consistency (i.e., if the data element is completed, whether the value used is consistent with the manner in which the variable is defined). The complete list of variables examined in this fashion from the CRASHDATA and HTRIS databases, in addition to information with regard to whether they passed the accuracy and completeness test, is provided in Appendix A in Tables A1 through A4.

Some variables in these databases, such as *INTERSECTIONTYPE* (in the CrashDocument table), are not routinely completed, whereas others, such as *WORKZONE*, were available only after September 2003. Neither variable passed the completeness test. Other variables, such as *ACCIDENTCITY* and *ACCIDENTCOUNTY* (in the CrashDocument table), did not pass the consistency test as there are some instances where both a county and city were identified and these locations were inconsistent. Variables that did not pass either test were not necessarily discarded, as is discussed later.

Overall, Table 1 suggests that of the 179 data elements in the four tables, 103 (58%) showed no problems with regard to completeness or consistency. Seventeen (9%) of these data elements had no data, and 33 (18%) had missing data. The single largest reason for data being missing was imperfections with the geographic referencing system for locating crashes. For 6 of the data elements (3%), additional categories are available for crashes occurring after September 2003, and for 20 (11%), the data elements themselves were available only for crashes occurring after September 2003.

Table 1. Summary of Data Completeness and Consistency in the Four Crash Database Tables (Tables A1-A4 in Appendix A)[a]

Status of Data Element	CrashDocument Table	CrashVehicle Table	CrashPedestrian Table	EyRoad Table	Total
No problems	21	22	7	53	103
No data	5	3	1	8	17
Incomplete data	[b]24	1	0	8	33
New categories after 9/03	3	3	0	0	6
Available after 9/03	5	15	0	0	20
Total	58	44	8	69	179

[a]Each cell shows the number of data elements or variables. For example, 21 variables in the CrashDocument table have no issues with regard to data completeness and consistency as noted by the authors.
[b]For the CrashDocument table, these 24 data elements denote the following: no data for crashes without reference nodes (12 data elements), incomplete data for crashes without reference nodes (1), inappropriate reference system (3), other incomplete data (5), and data inconsistencies (3).

Eight Rules for Working with Data Elements

Eight rules for using data elements with inadequate data quality were developed. The primary reason for the rules was that for some of the weaker data elements, the quality of the data element was not so poor that it had to be discarded. For example, in the aforementioned instance of both a county and city being noted, of 33,201 total crashes (not just intersection crashes) in the VDOT CRASHDATA database that occurred in the NOVA District for year 2005, 770 (about 2%) had both a city and a county filled in. With certain caveats, some data elements may be used for certain studies. (For example, a study relating seat belt use to injury prevention in a given city may not be adversely affected if the node is not known.) A secondary reason for the rules is that they illustrate how to obtain data for specific intersections.

Rules 1 through 5 are solutions that analysts can implement to improve the quality of the data in their specific study. Rules 6 thought 8 are observations that analysts should consider as they decide which time period and which data elements a given study should include. The eight rules are as follows:

1. Use the *PHYSICALJURISDICTION* variable rather than the *CITY* or *COUNTY* variable to determine the jurisdiction where a crash is located.

2. Use the *TRAFFICCONTROL* variable rather than the *INTERSECTIONTYPE* variable to determine whether an intersection is signalized.

3. Manually extract volumes for intersections that contain a one-way street rather than using the IntersectionEnteringVolume function.

4. Use the *NODE* and *OFFSET* variables from the CrashIntersection table rather than from the CrashDocument table.

5. Create intersection variables based on link variables as necessary.

6. Recognize that 20 variables will not have crash data until sometime after September 2003 and/or that these data may be available in another location than those studied here.

7. Recognize that new categories were added for six existing variables in September 2003.

8. Recognize that the node variable is incomplete for about one-third of the crashes.

1. *Use the PHYSICALJURISDICTION variable rather than the CITY or COUNTY variable to determine the jurisdiction where a crash is located.*

For some crashes, both the *CITY* and *COUNTY* variables contain a value. This might happen either because a reporting officer entered the name of both the city and the surrounding county on the FR300 or because a data entry error occurred while the electronic database was updated. Having both *CITY* and *COUNTY* coded, however, might lead to confusion regarding whether the crash should be located in the city or the county. Although a nested query can be developed to identify crashes as being located in cities if and only if the city field is not null, a more straightforward approach is simply to merge the CrashJurisdiction table (which includes the *PHYSICALJURISDICTION* variable) and the CrashDocument table and then use the *PHYSICALJURIDICTION* variable. (When both a county and a city were indicated for a given crash, the FR300 was examined to ascertain the correct location of the crash.)

2. *Use the TRAFFICCONTROL variable rather than the INTERSECTIONTYPE variable to determine whether an intersection is signalized.*

The *INTERSECTIONTYPE* variable is defined by categories that are not mutually exclusive as shown in Table 2. For example, the values for *INTERSECTIONTYPE* for a signalized T-intersection crash should be both *1* and *3;* however, only one code can be specified in the database. Thus, the absence of a *1* does not guarantee the intersection is not signalized.

To determine whether an intersection uses a signal, the *TRAFFICCONTROL* variable, which specifies the type of traffic control present on the intersection approach where a crash occurred, may be used. As shown in Table 3, the intersection is signalized if at least one crash has the code “03” for *TRAFFICCONTROL.* Theoretically, there are two situations where this

Table 2. Possible Values for the *INTERSECTIONTYPE* Variable

INTERSECTIONTYPE	Description
0	Crossover in median not at intersection
1	Signalized Intersection
2	Crossing (All crossroads at grade regardless of intersecting angle)
3	"T" (Leg enters between 80 degree and 100 degree angle)
4	Branch (One leg enters at angle other than "T" angle)
5	Offset (All offset intersections when offset does not exceed 150 feet)
6	5 way or more
7	Major channelization (Include traffic circle)
8	Interchange (Grade separation of intersection leg)
9	Not stated or not applicable.

Table 3. Possible Values for the *TRAFFICCONTROL* Variable

TRAFFICCONTROL	Description
01	No Traffic Control
02	Officer or Watchman
03	Traffic Signal
04	Stop Sign
05	Slow or Warning Sign
06	Traffic Lanes Marked
07	No Passing Lanes
08	Yield Sign
09	One Way Road or Street
10	Railroad Crossing with Markings and Signs
11	Railroad Crossing with Signals
12	Railroad Crossing with Gate and Signals
13	Other
14	Pedestrian Crosswalk
15	Reduced Speed—school zone
16	Reduced Speed—work zone
17	Special Corridor

rule could incorrectly indicate that a signalized intersection is unsignalized: (1) at a signalized intersection where there was also a yield control for right turning vehicles and the officer indicated that traffic control consisted only of the yield sign and (2) where a traffic signal was installed at some point during the study period.

3. *Manually extract volumes for intersections that contain a one-way street rather than using the IntersectionEnteringVolume function.*

For most intersections, the AADT entering an intersection may be extracted from the TMS database using the IntersectionEnteringVolume function of the PkgCrashRate package. (This query is implemented in SQL and uses the intersection node plus applicable dates.) If the intersection includes a one-way street, however, which was the case with 2.7% of the study intersections, this function will return a null value. In such a situation, another method for obtaining the volume should be used, such as obtaining the volume from VDOT's TMS database without using this function.

4. *Use the NODE and OFFSET variables from the CrashIntersection table rather than from the CrashDocument table.*

Intersection crashes can be referenced in two ways: by *NODE* in the CrashDocument table or by *NODE* in the CrashIntersection table. In the CrashDocument table, crashes that are offset from a node are referenced to the node lying to the immediate west (for an east-west link) or south (for a north-south link); in the CrashIntersection table, the crashes are referenced to the nearest node regardless of direction. Thus, the reference nodes of the CrashDocument table might not refer to the intersection nearest the crash location. In such cases, the *NODE* values in the two tables do not match, which was the case for 18.6% of all crashes in the study dataset.

For example, Figure 4 shows 2004 crashes in the proximity of nodes 704392 (node A) and 546262 (node B), which are located on Route 640, an east-west link.

As expected from Figure 4, the CrashDocument table will reference crash 1 to node A and crash 4 to node B as these crashes do not have an offset but rather occurred directly at the node. However, crashes 2, 3, and 5 do have an offset as they occurred between nodes. The CrashDocument table will reference crashes 2 and 3 to node A because node A is the closest node located west of these crashes. The CrashDocument table will reference crash 5 to node B, again because node B is the closest node located west of crash 5.

However, the CrashIntersection table will reference only crash 1 to node A, whereas crashes 2 through 5 are referenced to node B. The reason is that the CrashIntersection table uses the closest nodes to an intersection regardless of whether that node is located east or west of the crash.

If one tries to identify intersection crashes by using node offset information from the CrashDocument table, crash 2 will appear far away from its reference node (node A) and will not be considered an intersection crash. To avoid this issue, when data for intersection crashes are extracted, the CrashDocument and CrashIntersection tables should be merged and only *NODE* and *OFFSET* values from the CrashIntersection table used for identifying intersection crashes. It should be noted that other intersection-level variables in the CrashDocument table, such as *INTERSECTIONTYPE* and *NODETYPE,* do not necessarily pertain to the intersection nearest to the crash and thus could not be used in this study.

Figure 4. Representation of Intersection Crashes

5. Create intersection variables based on link variables as necessary.

The HTRIS database contains link variables such as rural or urban, primary or secondary, functional class (e.g., local, collector, arterial, or freeway), signalized versus stop-controlled, and facility type (one-way, undivided two-way, divided with partial access control, and divided with total access control). However, a similar classification is not available for intersections. The solution was to use these link variables as a surrogate for constructing intersection variables as shown in Table 4.

Table 4. Creation of Intersection Variables From Link Variables

Intersection Variable	Method to Create Intersection Variable
INT_RURALURBAN: Classifies location of intersection as rural or urban	The rural or urban link variable (*RURALURBAN*) of 1 of the downstream or upstream links from the link inventory table (EYROADXX) is examined. If the link is classified as rural (*RURALURBAN* = 1), the intersection is classified as rural.
INT_CLASSIFICATION: Classifies intersection as primary or secondary [a]	The route prefix values (*ROUTEPREFIX)* of all upstream and downstream links from the link inventory table (EYROADXX) are queried. If at least 1 of the links is a primary road (e.g., *ROUTEPREFIX* = US or SR or FR), the intersection is classified as primary.
INT_FUNCTIONALCLASS: Specifies functional class of intersection as local, collector, arterial, or freeway	The functional classification (*FUNCTIONALCLASS)* of all downstream and upstream links from the link inventory table (EYROADXX) is queried. and the highest classification among them is recorded. The ascending order of functional class as used in this study is local (lowest classification), collector, minor arterial, principal arterial, and urban freeway or expressway (highest classification).
INT_SIGNALIZATION: Classifies intersection as signalized or stop-controlled	Since the link inventory tables (EYROADXX) do not have the traffic control information for links, the *TRAFFICCONTROL* link variable from the CRASHDOCUMENT table is used as a surrogate for classifying intersections on the basis of signalization. The *TRAFFICCONTROL* attribute values of all upstream and downstream crashes for each intersection are obtained through a query. The intersection is classified as signalized if at least 1 of the crashes at the intersection showed the traffic control to be a signal (*TRAFFICCONTROL*= 3); otherwise, the intersection is classified as stop-controlled. (The possibility exists that a signalized intersection might be incorrectly classified as stop-controlled if all crashes specify a traffic control measure other than the traffic signal that was also present at the intersection. This can occur if the additional traffic control measure, such as a yield sign at 1 of the signalized right turn approaches, was cited by the police officer as being a contributor to the crash.)
INT_MINOR_CLASSIFICATION: Indicates whether minor road is a primary or a secondary road	The route prefix values (*ROUTEPREFIX*) of all upstream and downstream links from the link inventory table (EYROADXX) are queried. If at least 1 of the links is a secondary road (as indicated by the variable *ROUTEPREFIX* containing the number corresponding to a county, e.g., 29 (Fairfax County), the minor approach is classified as secondary; otherwise, it is classified as primary.
INT_MINOR_FUNCTIONALCLASS: Specifies functional class of minor intersection approach	The functional classification (*FUNCTIONALCLASS)* of all downstream and upstream links from the link inventory table (EYROADXX) are queried, and the lowest classification among them is recorded. The ascending order of functional classes is local (lowest), collector, arterial, and freeway (highest).
INT_MAJOR_FACILITY: Specifies facility type of major intersection approach as 1-way, undivided 2-way, divided 2-way, or full access controlled divided road[a]	The facility types (*FACILITY)* of all downstream and upstream links from the link inventory table (EYROADXX) are queried, and the highest *FACILITY* type observed is recorded as the *INT_MAJOR_FACILITY*.
INT_MINOR_FACILITY: Specifies facility type of minor intersection approach	The facility types (*FACILITY)* of all downstream and upstream links from the link inventory table (EYROADXX) are queried, and the lowest *FACILITY* type observed is recorded as the *INT_MINOR_FACILITY*.

[a]Note that the higher volume approach is the major approach and the lower volume approach is the minor approach. A primary road is one that has a route number below 600 (e.g., Route 29); a secondary road is one that has a route number above or equal to 600 (e.g., Route 729).

6. *Recognize that 20 variables will not have crash data until sometime after September 2003 and/or that these data may be available in another location than those studied here.*

Five of these variables (*WORKZONE, WORKERSPRESENT, VDOTPROPERTY, DMVSURFACETYPE*, and *TRAFFICCONTROLWORKING)* are in the CrashDocument table. Fifteen of these variables (*DRIVERAIRBAG, PASSENGERAIRBAG, DRIVEREMSTRANSPORT, PASSENTEREMSTRANSPORT, DRIVERSAFETYEQUIPMENT, DRIVERDISTRACTION, ALCHOHOLDETERMINATION, DRUGUSE, EMERGENCYVEHICLETYPE, EMERGENCYVEHICLESTATUS, OVERSIZE, CARGOSPILL, OVERRIDE, UNDERRIDE,* and *VEHICLECMVHAZINDICATOR*) are in the CrashVehicle table. One reason is that in September 2003, the FR300 was updated to include these new variables (except for the variable *SafetyEquipment*, where a new category 8—child safety seat—was added to the FR300).

DRIVERSAFETYEQUIPMENT, which appears in the CrashVehicle table, has a blank value for crashes before January 1, 2004. For crashes occurring in 2004 and later, *DRIVERSAFETYEQUIPMENT* has a value; however, this value does not appear to correspond to the FR300. For example, for one crash for which the FR300 was examined, the motorcycle driver was wearing a helmet (hence code 6 according to the FR300) but the *DRIVERSAFETYEQUIPMENT* field shows a value of 5. Because a pattern could not be discerned for the other values of *DRIVERSAFETYEQUIPMENT*, it is not recommended at this time that this variable be used.

However, since this study was completed, the variable *SAFETYEQUIPMENT* was added to two new tables: CrashPerson and CrashInjury. In those tables, the value of *SAFETYEQUIPMENT* generally appears to correspond to the FR300 with one exception regarding the use of code 8. For crashes that occurred prior to January 1, 2004 (when the September 2003 revised FR300 would have taken effect), code 8 appears to mean "unknown." For crashes that occurred January 1, 2004, or later, code 8 correctly matches the definition shown on the September 2003 FR300, i.e., the use of a booster seat.

The recommended practice for determining restraint use as captured by the *SAFETYEQUIPMENT* variable is summarized as follows:

- Do not use the variable *DRIVERSAFETYEQUIPMENT* in the CrashVehicle table.
- For crashes occurring on or after January 1, 2004, use the *SAFETYEQUIPMENT* variable in the CrashInjury and CrashPerson tables.
- For crashes occurring prior to January 1, 2004, recognize that a code of 8 shown in the *SAFETYEQUIPMENT* variable in the CrashInjury and CrashPerson tables does not indicate the use of a booster seat. Instead, the code of 8 for such crashes prior to January 1, 2004, indicates that the safety equipment is unknown.

7. Recognize that new categories were added for six existing variables in September 2003.

Three of these variables are in the CrashDocument table (*TRAFFICCONTROL, ALIGNMENT,* and *SURFACECONDITION*), and three are in the CrashVehicle table (*VISIBILITYOBSTRUCTION, DRIVERACTION,* and *FIXEDOBJECT*). Similar to Rule 6, the reason for the new categories being in the database is that the FR300 was updated in September 2003.

8. Recognize that the node variable is incomplete for about one-third of the crashes.

The variable *NODE* is critical for locating intersection crashes, but for 35.7% of crashes in the entire database, there is no entry for *NODE* (i.e., *NODE* = 999999). For example, of 33,201 total crashes (not just intersection crashes) in the VDOT CRASHDATA database that are noted to occur in the NOVA District for year 2005, 2955 (about 9%) had a node value of 999999. A very small portion (40 crashes) was located in a city or town; the others were coded as being in the counties of Arlington (1,062 crashes), Fairfax (693 crashes), Loudoun (830 crashes), or Prince William (330 crashes). For those crashes, it was noted that attributes that would logically rely on node information, such as impact zone, shoulder width, and functional class, were not known. These attributes are ROUTEPREFIX, ROUTENUMBER, ROUTESUFFIX, NODEOFFSET, NODETYPE, SURFACETYPE, SURFACEWIDTH, SHOULDERWIDTH, FACILITY, INTERSECTIONTYPE, IMPACTZONE, SYSTEM, and FUNCTIONALCLASS.

In the absence of reference node information, it is not possible to pinpoint the location of crashes through querying. Instead, such crashes must be located manually. This manual procedure entails four steps:(1) identify their document numbers in the VDOT CRASHDATA database, (2) obtain the image of the corresponding FR300 from VDOT's Crash Report Database, (3) review the location description provided by the law enforcement officer, and (4) use VDOT's GIS Integrator to locate the intersection and extract the *NODE* value from GIS Integrator. This method is manually intensive and infeasible for a study that deals with thousands of crashes but may be feasible for safety studies using smaller datasets.

Option for Future Work

Although the eight rules are targeted toward crash analysts, there remains an option for database designers: in the future, consider adding intersection variables to the existing node inventory table.

Several geometric attributes that are directly relevant to intersection-based analyses, such as number of left, through, and right lanes; channelization (e.g., none, median only, painted islands, and raised islands); presence of frontage roads, curb cuts, and on-street parking; etc., are not maintained in the link inventory tables. These variables are essential for intersection studies such as this one. In this effort, such intersection information was obtained directly through a manual data collection effort using aerial photographs from VDOT's GIS Integrator. A more comprehensive solution would be to add such information to the node inventory tables.

Identification of Minimum Set of Crash Causal Factors

The minimum set of variables that identify crash causal factors was determined through developing *classification trees*, so named because they identified the factors that determined whether or not a crash was any one of the following: a rear-end crash, an angle crash, or an injury crash. (Note that the rear-end and angle categories are mutually exclusive as a crash cannot be both rear-end and angle. However, the injury crash can also be rear-end, angle, or neither.) Because there were three crash types (angle, rear-end, and injury) and 17 intersection types (e.g., urban stop-controlled four-way intersections on four-lane roads, urban signalized four-way intersections on multi-lane roads, etc.), a total of 51 trees were developed.

Three types of results are discussed: the size of the classification trees, the variables that are the important predictors of crash types (rear-end versus not rear-end, angle versus not angle, and injury versus not injury) in these trees, and the relative accuracy of these trees with regard to their ability to predict crash types. These results may be used to identify a minimum set of crash causal factors, although the utility of such a set has practical limitations as discussed.

Size of Classification Trees

Each of the 51 trees has multiple branches that may be used to classify a crash as being a crash type. The tree for classifying whether crashes at rural signalized three-way intersections on four-lane roads are rear-end or not may be used as an example. The three branches of this tree, shown previously in Figure 3, are as follows:

- *Branch 1*: A crash is likely to be a rear-end crash as opposed to not being a rear-end crash if it meets the following four criteria:

    1. Driver action is code 4 (following too close), 6 (failure to maintain proper control, improper lane change, or overcorrection), 7 (driver inattention), 10 (avoiding pedestrians, animals, vehicles, or other improper action), or 1 (exceeded speed limit or exceeded safe speed).
    2. Approach alignment is code 1 (horizontal level curve), 4 (straight hill crest), 0 (straight and level), or 2 (straight grade).
    3. Vehicle speed is less than or equal to 50.5 mph.
    4. Traffic control is code 5 (no passing lanes), 6 (yield sign), 7 (other), 0 (none), 1 (traffic signal), or 4 (traffic lanes marked).

- *Branch 2*. A crash is likely to be a rear-end crash if it meets the following five criteria:

    1. As with branch 1, driver action is code 4 (following too close), 6 (failure to maintain proper control, improper lane change, or overcorrection), 7 (driver inattention), 10 (avoiding pedestrians, animals, vehicles, or other improper action), or 1 (exceeded speed limit or exceeded safe speed).
    2. As with branch 1, approach alignment is code 1 (horizontal level curve), 4 (straight hill crest), 0 (straight and level), or 2 (straight grade).

3. As with branch 1, vehicle speed is less than or equal to 50.5 mph.
4. Contrary to branch 1, traffic control is code 2 (stop sign).
5. Curb and gutter is code 0 (no curb and gutter).

- *Branch 3*. A crash is likely to be a rear-end crash if it meets the following five criteria:

1. As with branches 1 and 2, driver action is code 4 (following too close), 6 (failure to maintain proper control, improper lane change, or overcorrection), 7 (driver inattention), 10 (avoiding pedestrians, animals, vehicles, or other improper action), or 1 (exceeded speed limit or exceeded safe speed).
2. As with branches 1 and 2, approach alignment is code 1 (horizontal level curve), 4 (straight hill crest), 0 (straight and level), or 2 (straight grade).
3. Contrary to branches 1 and 2, vehicle speed is greater than 50.5 mph.
4. As with branches 1 and 2, traffic control is code 1 (traffic signal).
5. Environmental lighting is code 0 (daylight) or 1(dawn or dusk).

As discussed previously, the gain differs for each rule: the gain for Rule 1 is large, i.e., 219 (it properly classifies 219 rear-end crashes); the gains for Rules 2 and 3 are smaller, i.e., 6 and 5, respectively.

The three branches show that capturing just six variables allows one to classify a rear-end crash versus a non–rear-end crash at this intersection. The six variables that comprise the three branches are driver action, approach alignment, vehicle speed, traffic control, curb and gutter, and environmental lighting. (Returning to the initial reason for developing the classification trees, the question becomes whether a similar set of variables could successfully classify crashes at other types of intersections other than the type shown in Figure 3; if so, then a case could be made that a minimum set of variables has been identified. If such a minimum set exists, it may be used to prioritize data collection efforts or determine which variables are most productive for estimating crash risk.)

Not all trees used the same variables or had the same number of branches. In general, the number of branches for each tree varied from 2 (e.g., rear-end crashes at rural stop-controlled three-way intersections on four-lane roads) to 15 (e.g., angle crashes at urban signalized three-way intersections on four-lane roads). The number of rules induced from each tree varied depending on the size of the tree and the structure of the dataset.

Variables That Are the Important Predictors of Crash Types in Classification Trees

The classification tree procedure also ranks the variables (used in the rules) based on their importance regarding their ability to predict crash types during the training process. In the case of injury crashes at urban stop-controlled four-way intersections on a four-lane road, there were four branches. From these, the 10 most important predictors for injury crashes at an urban stop-controlled four-way intersection on a four-lane road are vehicle speed, traffic volume, driver action, lighting, alignment, weather, shoulder width, surface condition, whether the driver's visibility is obstructed, and traffic control. (Although the traffic control for the intersection is

known [e.g., stop-controlled], the officer may indicate other forms of traffic control that may be in the vicinity of, but not supersede, the stop sign. These include none, traffic signal, stop sign, slow or warning sign, traffic lanes marked, no passing lanes, yield sign, or other.)

It is also possible to compare the important predictors. from this tree to the important predictors of the other 50 trees (e.g., the trees for angle or rear-end crashes at the same intersection type and the trees for angle, rear-end, or injury crashes at the remaining 16 intersection types). For example, the approach alignment variable is used to classify all three crash types (angle, rear-end, and injury) at urban stop-controlled intersections on four-lane roads. Given 17 intersection types and three crash types, a variable that was needed to classify all crash types at all intersections would appear at some point in all 51 trees.

Figure 5 shows the number of trees in which each variable was deemed an important predictor. Not surprisingly, vehicle speed was the most common variable, appearing as an important predictor in 49 of the 51 trees. Other highly important predictors included driver action (48 trees), alignment of the approach (47 trees), lighting and weather (44 trees each), traffic control (42 trees), driver visibility (40 trees), and traffic volume (38 trees).

Although the variables in Figure 5 gradually decrease in the number of trees cited from left to right, there is a discontinuity at the surface condition variable. The variable immediately to the left—shoulder width—appears as an important predictor in 35 trees, whereas the surface condition variable appears in 28 trees followed immediately by vehicle type, which appears in 18 trees. At this discontinuity, there is also an important shift in the types of crashes for which the variable is a useful predictor. With regard to the variables to the left of vehicle type, each is

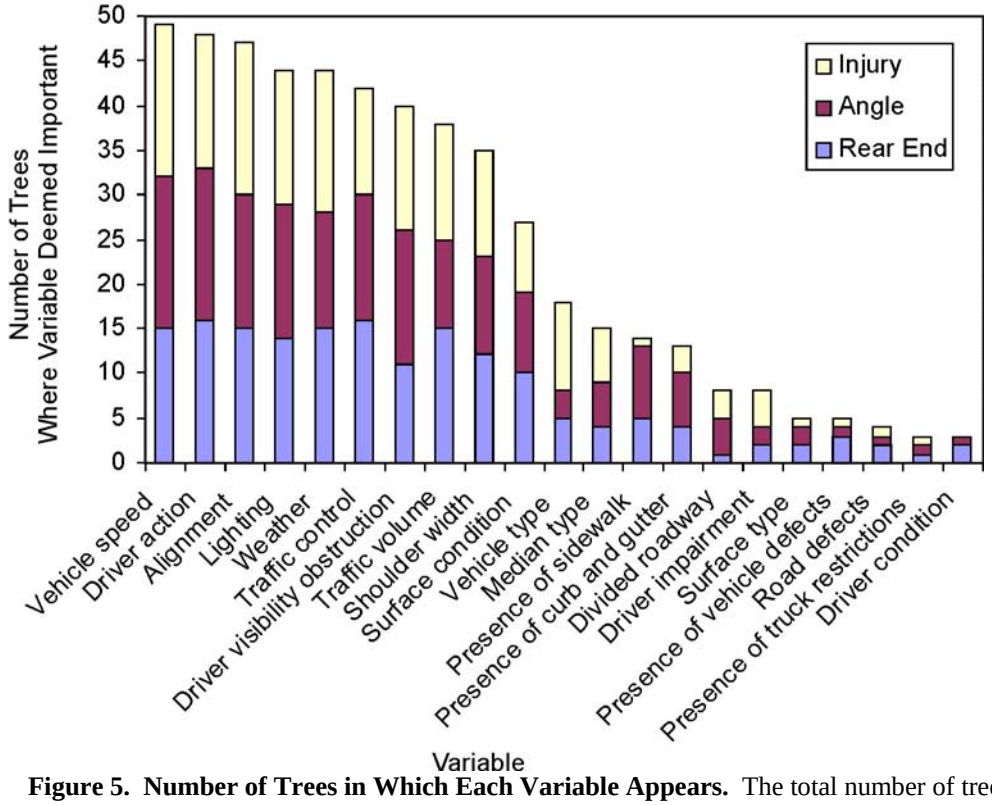

Figure 5. Number of Trees in Which Each Variable Appears. The total number of trees was 51.

generally used in equal proportions for all three crash types; e.g. the surface condition variable was useful for predicting rear-end crashes (10 trees), angle crashes (9 trees), and injury crashes (8 trees). Immediately to the right of the surface condition variable, however, there are two variables where these proportions are not similar. Although the vehicle type variable was used in 10 trees for injury type classification, it was used in only 5 trees for rear-end crash classification and in only 3 trees for angle crash classification. The other discontinuity was the presence of a sidewalk variable: used in 5 rear-end crash trees, 8 angle crash trees, and only 1 injury crash tree. These discontinuities show the potential exists to delineate essential variables from less important ones. A more rigorous approach that takes advantage of these discontinuities may be considered but only after the accuracy of these classification trees is evaluated.

Accuracy of Classification Trees With Regard to Their Ability to Predict Crash Types

The classification trees were developed using about three-fourths of the dataset. To assess their accuracy, they were applied to the remaining one-fourth of the dataset, the error rates are shown in Table 5. For example, consider the cell for rural three-way stop-controlled two-lane intersections with respect to rear-end crashes in Table 5. According to that cell, for such intersections on two-lane roads, the rear-end tree did not correctly classify 13.24% of the training crashes (classifying either a non–rear-end crash as a rear-end crash or a rear-end crash as a non–rear-end crash). When the tree is applied to new data from which the tree was not developed, i.e., the testing set data, the error rate increased slightly to16.03%.

Classification trees for rear-end crashes had average error rates of 12.83% and 16.20% for training and testing samples, respectively. For angle crash classification trees, the average error rates were 9.01% (training) and 12.21% (testing). Injury (including fatalities) crash classification trees were less accurate, with error rates of 39.78% and 38.63% for training and

Table 5. Predictive Accuracy of Classification Trees

Intersection Class				Error Rate					
				Rear-end		Angle		Injury and Fatal	
				Training	Testing	Training	Testing	Training	Testing
Rural	3-way	Stop-controlled	2-lane	13.24%	16.03%	4.36%	8.71%	42.03%	42.86%
			4-lane	11.94%	22.64%	10.29%	15.09%	40.19%	33.96%
		Signalized	2-lane	12.26%	18.10%	7.53%	11.21%	31.25%	41.38%
			4-lane	12.07%	15.69%	8.59%	13.07%	51.67%	33.33%
	4-way	Stop-controlled	2-lane	14.71%	14.75%	8.33%	14.75%	32.65%	54.10%
			4-lane	7.22%	9.26%	10.56%	18.52%	28.06%	27.57%
		Signalized	2-lane	12.39%	13.01%	8.42%	12.33%	36.18%	34.25%
			4-lane	12.89%	19.05%	8.24%	13.23%	37.29%	42.06%
Urban	3-way	Stop-controlled	2-lane	14.52%	14.59%	10.06%	11.18%	38.91%	34.72%
			4-lane	14.80%	18.52%	8.02%	11.35%	37.09%	42.83%
		Signalized	2-lane	13.09%	20.40%	8.91%	11.04%	43.27%	39.29%
			4-lane	15.45%	15.37%	9.58%	10.19%	43.98%	38.96%
	4-way	Stop-controlled	2-lane	11.56%	13.07%	11.22%	13.91%	39.21%	35.74%
			4-lane	12.08%	17.61%	9.45%	12.20%	38.99%	40.18%
		Signalized	2-lane	12.78%	15.88%	11.19%	11.07%	47.21%	38.28%
			4-lane	13.23%	14.92%	9.99%	10.25%	46.88%	38.84%
			Multi-lane	13.96%	16.44%	8.43%	9.49%	41.41%	38.42%

testing samples, respectively. The poor performance of injury crash classification trees might be attributed to the heterogeneity in injury and fatal crashes, as they include all collision types. (For example, an injury crash might be a rear-end collision, an angle collision, or a collision with a fixed object.) Thus, the higher accuracy for angle or rear-end crashes may be attributable to each of these datasets being more homogeneous than the injury crash dataset.

Error Rate Obtained From Chance Alone

To determine whether the error rates shown in Table 5 indicate good performance, they were compared with the error rate that would result from chance alone. Note that this error rate is not necessarily 50%. For example, it is generally the case that most crashes are not injury crashes.

To determine the error rates based on chance alone, a query of select Northern Virginia intersections for the time period 2006 through 2007 in the VDOT CRASHDATA database was executed. The query indicated that the percentage of crashes that were rear-end, angle, and injury and fatal, respectively, were 40.23%, 34.08%, and 34.29%. With regard to only injury and fatal versus non-injury and fatal crashes, because a minority of crashes were injury and fatal crashes (34.29%), an analyst who had no other information could simply guess that each crash was a non-injury and fatal crash. This guess would yield an error rate of 34.29%. By extension, without any information such as that provided by the classification trees, the error rate attributable to chance alone will be 40.23% (for rear-end crashes), 34.08% (for angle crashes), and 34.29% (for injury crashes).

Comparing Error Rates From Chance Alone and Classification Trees

In general, the error rates for rear-end and angle crashes for all intersection types shown in Table 5 were significantly lower than what would have been obtained from chance alone. For example, with regard to the 16.03% testing error for rural stop-controlled three-way intersections on two-lane roads for rear-end crashes in Table 5, Equation 9 shows that this error is significantly different from the error of 40.23% that would result from chance alone since the confidence interval given by Equation 9 (13.64% to 18.42%) does not include 40.23%.

$$\text{Error rate of } 16.03\% \pm 1.96\sqrt{\frac{(\text{Error rate of } 16.03\%)(1 - \text{Error rate of } 16.03\%)}{\text{Sample size of } 905 \text{ crashes}}}$$

$$= 13.64\% \text{ to } 18.42\% \qquad \text{[Eq. 9]}$$

By contrast, except for one intersection, the error rate for classifying injury and fatal crashes is not significantly lower than the error rate one would expect to result from chance alone. For example, Table 5 shows a 42.86% testing error for rural stop-controlled three-way intersections on two-lane roads. Equation 10 shows that the 95% confidence interval for this rate, based again on an estimated testing sample size of 905 crashes, is 39.64% to 46.08%. This error rate is not only higher than the 34.29% error rate expected from chance alone, it is also significantly higher than the error rate obtained from chance alone.

$$\text{Error rate of } 42.86\% \pm 1.96\sqrt{\frac{(\text{Error rate of } 42.86\%)(1 - \text{Error rate of } 42.86\%)}{\text{Eample size of } 905 \text{ crashes}}}$$

= 39.64% to 46.08% [Eq. 10]

In only one instance—tural four-way stop-controlled intersection with a four-lane road—was the testing error in Table 5 for injury and fatal crashes (in this case 27.57%) significantly lower than the 34.29% error that would have resulted from chance alone. For five of the intersection groups (rural three-way stop-controlled intersection with a four-lane road; rural three-way signalized intersection with a four-lane road; rural four-way signalized intersection with a two-lane road; urban three-way stop-controlled intersection with a two-lane road; and urban four-way stop-controlled intersection with a two-lane road), there was no significant difference between the injury error rate in Table 5 and the error rate resulting from chance alone. For the remaining intersections, the injury error rate shown in Table 5 was significantly higher than the 34.29% error rate that would have resulted from chance alone. Thus, the injury classification results do not suggest the classification trees as developed in this study are useful for identifying causal factors for injury crashes.

It may be possible for additional research to improve the usefulness of classification trees for intersection crashes through other methods not explored in this study. For example, one approach would be to apply a two-step procedure where crashes were first classified as rear-end or angle and then as injury or non-injury. A second approach would be to redo the classification trees with a new dataset, taking advantage of variables that were not readily available previously, such as safety equipment (which indicates whether or not the driver was using a seat belt).

A third approach would be to use the results of CEMs to combine certain intersection types where the geometric differences were not reliable predictors of intersection crashes.

A Minimum Set of Crash Causal Factors

The relatively large injury error rates in Table 5 suggest that the use of classification trees as in this study (where crashes were classified as injury or non-injury regardless of crash type) is simply not a reliable method for determining whether a crash will result in an injury. However, it appears reasonable to use the trees to identify causal factors for classifying rear-end and angle crashes.

For rear-end and angle trees only, Figure 5 may be considered to indicate a total "tree factor space" of 340, based on 17 rear-end trees, 17 angle trees, and 10 factors per such tree. The first few factors on the horizontal axis of Figure 5 account for a relatively high proportion of the tree factor space. For example, 6 factors (vehicle speed, driver action, alignment, lighting, weather, and traffic control) are cited as important predictors 182 times—more than 50% of the 340-unit tree factor space. Thus although there is a total of 21 variables on the horizontal axis of Figure 5, 6 account for 54% of the tree factor space, 10 account for 81%, and 13 account for almost 90%.

To identify the minimum set of crash causal factors, one approach is to identify the number of variables in Figure 5 that occupies the percentage of tree factor space equivalent to the

accuracy rate shown in Table 5. For example, since Table 5 shows an average error rate of 16.20% for rear-end crashes (a slightly lower rate of 12.21% for angle crashes), one approach would be to identify the variables in Figure 5 that account for 84% of the space. In this case, only 12 variables account for precisely 85.88% of the space: vehicle speed, driver action, alignment, lighting, weather, traffic control, driver visibility obstruction, traffic volume, shoulder width, surface condition, vehicle type, and median type.

Utility of Minimum Set of Crash Causal Factors

Although a set of crash causal factors can be developed, the utility of the set as a criterion for taking any action is limited for two reasons. First, some judgment is necessary regarding the demarcation of the most important and less important variables in Figure 5. Indeed, the case could be made that all variables shown in Figure 5 are useful in some capacity. Second, the selection of these variables is based solely on classifying crashes as rear-end or angle rather than injury crashes.

Thus, the minimum set of crash causal factors have utility as a diagnostic tool in that given that a crash will occur, the 12 variables noted previously should help identify whether the crash will be a rear-end or angle crash. In that sense, the factors may prove useful for delineating the conditions that make a given site more prone to rear-end crashes or angle crashes. They also suggest that specific variables stored in the VDOT CRASHDATA database may be used to make some types of predictions; i.e., although crashes are probabilistic, the random variation does not obscure all trends. However, it does not appear justifiable to use this list of 12 variables as a way to prioritize which data elements should be collected in the future for all crash types. Because the trees did not serve their purpose of identifying which variables should be used, they are not presented in this report.

Development of Crash Estimation Models

CEMs were developed for all 17 intersection types and four crash categories: angle, rear-end, injury, and total. The approximate number of intersections used for each intersection class was about 6,752, as shown in Table 6. In a few cases, intersection characteristics changed during

Table 6. Number of Intersections by Class

Intersection Class				No. of Intersections	No. of Data Points
Rural	3-way	Stop-controlled	2-lane	378	382
			4-lane	39	43
		Signalized	2-lane	39	40
			4-lane	38	47
	4-way	Stop-controlled	2-lane	89	89
			4-lane	59	65
		Signalized	2-lane	30	31
			4-lane	56	64
Urban	3-way	Stop-controlled	2-lane	3,059	3,079
			4-lane	372	382
		Signalized	2-lane	199	217
			4-lane	325	338
	4-way	Stop-controlled	2-lane	1,028	1,036
			4-lane	293	297
		Signalized	2-lane	190	204
			4-lane	685	718
			Multi-lane	73	77

the study period, in which case the intersection was represented as two data points (with the crash data normalized to reflect a 6-year period). As a consequence, the number of data points is slightly greater than the number of intersections, as shown in Table 6.

For example, the total number of crashes at a rural signalized three-way intersection with a two-lane road is computed via Equation 11:

$$TC = 0.000346(\text{Volume})^{0.9409} \cdot \exp\left((0.5354\,\text{int\_func}) - (0.549\,\text{channelization}) - (0.1568\,\text{left turn lanes})\right) \quad \text{[Eq. 11]}$$

where

TC = total annual crashes
Volume = total daily approach volume, which is the sum of volume entering the intersection from all four links (or three links for a three-way intersection)
int_func = 0 if major road is a local or a collector road and 1 if major road is an arterial road
channelization = 0 if no channelization or only medians on approaches, 1 if painted or raised islands
left turn lanes = sum of exclusive left turn lanes on all approaches.

Although Equation 11 may be used to estimate the total number of crashes at a signalized rural three-way intersection with a two-lane road, inspection of the parameters in the equation offers several insights into factors affecting crash risk at such intersections. From left to right in Equation 11, the volume exponent of 0.9409 suggests that crashes are slightly inelastic to volume: a doubling of traffic volume will result not in twice as many crashes but rather in $2^{0.9409}$ = 1.92 times as many crashes. The positive coefficient for the int_func variable suggests that crash risk is increased if the major road is an arterial facility as opposed to a local or collector facility. As expected, total crashes are reduced by the presence of painted or raised islands (see the negative coefficient for the channelization variable) and exclusive left turn lanes (see the negative coefficient for the left turn lanes variable).

The CEMs, dispersion parameters, and goodness-of-fit statistics are presented in Tables B2 through B5 in Appendix B for all 68 CEMs. Three observations pertaining to the volume exponent, the impact of the remaining explanatory variables, and the goodness of fit of the CEMs can be made regarding the CEMs.

Volume Exponent

For most of the 68 CEMs, the volume exponent was usually below 1.0: this was the case for 16 of the 17 angle crash models and 16 of the 17 total crash models. For rear-end and angle crash models, the single exception—where the volume exponent exceeded 1.0—was rural three-way signalized intersections with a two-lane road (angle crashes) and rural four-way signalized intersections with a two-lane road (total crashes).

Of the 17 injury crash models, the volume exponent was also below 1.0 except in three instances: rural three-way signalized intersections with a two-lane road (consistent with the volume exponent for angle crashes), rural four four-way signalized intersections with a two-lane road (consistent with the volume exponent for rear-end crashes), and urban four-way signalized intersections with a multi-lane facility.

There was one crash type (rear-end) where, in the various intersection models, the volume exponent was evenly distributed above and below 1.0. For rear-end crash models, the volume exponent was below 1.0 for slightly less than one-half (8) of the 17 crash types. The rear-end volume exponent was above 1.0 for all three intersections noted previously where the injury, total, or angle crash exponent was above 1.0 as well as for six other intersection types: rural four-way stop-controlled intersections with a two-lane road, rural three-way stop-controlled intersections with a four-lane road, rural three-way signalized intersections with a four-lane road, urban three-way stop-controlled intersections with a two-lane road, urban four-way stop-controlled intersections with a two-lane road, and rural three-way stop-controlled intersections with a two-lane road.

Impact of Remaining Explanatory Variables

In addition to volume, 15 explanatory variables were considered in developing the GLMs as shown in Table B1 in Appendix B. Four related either to functional class (local or collector versus arterial) or administrative class (primary or secondary) of the intersection approaches. Four variables related to the number of turning lanes, such as exclusive left turn lanes, exclusive right turn lanes, and lanes where different turning movements were combined. Two variables addressed whether either or both approaches were divided, one addressed whether frontage roads were present, and one addressed whether painted or raised islands were employed at the intersection. The three remaining variables concern the presence of on-street parking, the number of lanes at the intersection, and the number of curb cuts.

Table 7 indicates the number of CEMs—of a total of 68—where a change in the variable reduces, does not affect, or increases the number of crashes. For example, consider Equation 11. A change in the number of left turn lanes from 0 to 1 will reduce the number of total crashes at a rural signalized three-way intersection with a two-lane road. By contrast, consider the model shown in Table B2 (Appendix B) for angle crashes at rural three-way stop-controlled intersections with four-lane roads. That model, reprinted as Equation 12, shows that the coefficient for left turn lanes (left turn lanes) is above zero with a value of 0.5132.

$$\text{Total crashes} = 0.01322\,(\text{Volume})^{0.8952}\,X$$
$$\exp\left(0.8952\,\text{int\_primsec} + 1.242\,\text{minor\_primsec} - 0.9234\,\text{rtlanes} + 0.5132\,\text{left turn lanes}\right)$$

[Eq. 12]

Thus, Equation 11 represents one model where increasing the number of left turn lanes decreases crashes, whereas Equation 12 represents a model where increasing the number of left turn lanes increases crashes. Overall, Table 7 shows that an increase in left turn lanes from 0 to

Table 7. Impact of Changing Each Explanatory Variable on Number of Crashes[a]

Change in Variable (Relative to Base Condition)	Models Where Change		
	Decreases crashes	Has no impact	Increases crashes
Major road is secondary (rather than primary)[b]	12[b]	27[b]	29[b]
Minor road is secondary (rather than primary)	12	48	8
Both approaches are primary (rather than at least 1 being secondary)	0	64	4
Major road is arterial (rather than local or collector)	19	39	10
Painted or raised islands (rather than no channelization or only medians on approaches)	22	38	8
Frontage roads are present (rather than absent)	4	55	9
On-street parking observed (rather than not observed)	8	50	10
Both approaches are undivided (rather than at least 1 being divided)[c]	14	29	25
Major approach is divided (rather than being undivided)	0	63	5
1 exclusive right turn lane (rather than none)	16	40	12
1 exclusive left turn lane (rather than none)	26	28	14
1 lane where through, right, and left turns are allowed (rather than none)	22	43	3
1 lane where both right and left turns are allowed (rather than none)	4	57	7
Intersection of 4-lane road with another 4-lane (or more) road (rather than a 2-lane road)	7	56	5
1 curbcut (rather than no curbcut)	10	46	12

[a]The total number of models always sums to 68 as there are 17 intersection types and 4 crash types per intersection type.

[b]For example, when the major road is changed to a secondary facility from a primary facility, 12 of the models show a decrease in crashes, 29 show an increase, and 27 are not affected. Computationally, this is done by changing the value of the variable *INT_PRIMSEC* (see Table B1 in Appendix B) from 0.0 (meaning the major road is primary) to 1.0 (meaning the major road is secondary) and then determining whether the quantity shown in the "Intersection Factors" column of Tables B2 through B6 is positive, negative, or unchanged.

1 reduces the number of crashes for 26 models, increases the number of crashes for 28 models, and has no effect on crashes for 14 models.

Consistency Among Signs of Explanatory Variables

Table 7 has a few results that were to be expected. For example, of the 30 models where channelization was included as a variable, most (22) showed that using painted or raised islands reduced crashes compared to the few (8) where such channelization increased crashes. Having an exclusive left turn lane (rather than not having such a lane) tended to reduce crashes. Further, for 9 of the 10 angle crash models that included a variable indicating whether both approaches are undivided, changing this variable from at least one approach being divided to having both approaches being undivided increased crashes.

However, Table 7 also shows several results that initially were surprising: e.g., having a major approach that is divided, rather than undivided, shows an increase in 5 of the models (and no decreases). Similarly, the presence of a lane where through, right, and left turns are allowed tends to decrease, rather than increase, crashes. Finally, for 6 of the 10 rear-end crash models that included a variable indicating whether both approaches are undivided, changing this variable

from at least one approach being divided to both approaches being undivided decreased crashes, contrary to what one would expect from removing a physical barrier.

There are two possible explanations for these non-intuitive results. One is that changes to some variables that should reduce crash risk may in some situations have an adverse impact because of the interaction effect with other intersection characteristics. For example, the existence of left turn and right turn lanes at an intersection could lead to additional weaving that otherwise would not have occurred.

The second explanation is that in some cases, there may be variables that are associated with increased crash risk but not causing this increased crash risk. For example, Equation 12 shows a positive coefficient of 0.5132 for left turn lanes. Barring the interaction possibility noted, it is likely not the case that the construction of left turn lanes results in additional crash risk given that channelization should reduce crashes. Instead, in this situation, it may be the case that the left turn lanes are a surrogate for an increase in conflict points resulting from left turning demand, which would increase crashes relative to a situation where few left turns are made. Similarly, there may be variables that are associated with decreased crash risk but not causing this decreased crash risk. For example, with regard to the models for estimating crashes on two-lane roads at a stop-controlled intersection, where the presence of on-street parking has a negative coefficient, it may be the case that two-lane facilities with on-street parking have characteristics relative to facilities without on-street parking that reduce crashes, such as lower speeds, less cut-through traffic, and greater driver vigilance. In short, the variables may sometimes reflect association rather than causation.

Note that the R^2_{dev} does not quantify the extent to which a given variable causes, as opposed to being correlated with, a given level of crash risk. (In Equation 4, the R^2_{dev} simply provided a score, between 0 and 1, indicating the relative strength of the model. Strength is measured as a proportion, with the denominator being the difference in likelihood between a saturated model that has 100% accuracy and a null model that contains only an intercept and the numerator being the difference in maximum likelihood between the model developed herein and the same null model.) Thus it can be said that the rear-end crash model for rural four-way signalized intersections with a two-lane road ($R^2_{dev} = 0.73$) is stronger than that for urban four-way signalized intersections with a two-lane road ($R^2_{dev} = 0.39$). However, other information (such as recognition that a facility may not have been converted from being undivided to divided) is necessary to determine whether correlation or causation explains the role of the variable in the model.

Consistency Among Magnitudes of Explanatory Variables

Visual inspection of the coefficients in the CEMs suggests that even when the variables have the same sign, the impact on crashes may differ from one model to the next. For example, consider the nine models that predict angle crashes and show that changing both approaches to undivided will increase crash risk. Table 8 shows that this crash risk increase may range from 43% to 154% depending on the specific intersection being modeled. Although a few of the intersection types have visually similar increases (e.g., a stop-controlled intersection with a four-lane road shows a 73% to 77% increase regardless of whether it is urban or rural), the increase in

Table 8. Impact of Making Both Approaches Undivided on Angle Crash Risk

Intersection Type	Crash Increase
Urban 3-way stop-controlled intersection with a 4-lane road	43%[a]
Urban 3-way signalized intersection with a 2-lane road	51%
Urban 4-way signalized intersection with a 4-lane road	53%
Urban 4-way signalized intersection with a 2-lane road	64%
Urban 3-way stop-controlled intersection with a 2-lane road	68%
Urban 4-way stop-controlled intersection with a 4-lane road	73%
Rural 4-way stop-controlled intersection with a 4-lane road	77%
Urban 4-way stop-controlled intersection with a 2-lane road	78%
Rural 3-way signalized intersection with a 4-lane road	154%

[a]For example, for urban 3-way stop-controlled intersections with a 4-lane road, the variable *BOTHUNDIV* has a coefficient of 0.3584. Changing the value of this variable from 0 (at least 1 approach is divided) to 1 (both approaches are undivided) raises the risk of angle crashes from exp(0) = 1 to exp(0.3584) = 1.43, a 43% increase.

crash risk more than triples when the lowest increase in Table 8 is compared to the highest increase. If the intersections with the most and least crash increase in Table 8 are removed, the increase in crash risk still varies by about one-half when the remaining highest and lowest increases are compared.

Summary of Consistency Among Explanatory Variables

Table 8 shows the variation that results even when the sample of CEMs is limited to those having the same sign for a given variable. For one CEM not shown in Table 8 (urban four-way signalized intersection with a multi-lane road), making both approaches undivided reduced angle crash risk by 28%. Thus, the impact of this variable on angle crashes differs by road type and ranges from -28% (for urban four-way signalized intersections with a multi-lane road) to +154% (for a rural three-way signalized intersection with a four-lane road).

Further, whereas making both approaches undivided increased crash risk for nine angle crash CEMs and reduced crash risk for just one angle crash CEM, this same change increased crash risk for four rear-end crash CEMs and reduced crash risk for six CEMs. Thus, variables have variation by crash type and within those crash types by intersection class—even in cases where, such as in Table 8, the variable has the same sign.

Goodness of Fit of CEMs

For the 68 CEMs provided in Appendix B, the dispersion parameter k varied between 0.119 and 3.337, suggesting that dispersion is not constant for each model. Although there was some overdispersion and hence the negative binomial distribution was appropriate, the low value of k in some cases suggested that the Poisson assumption would also have been appropriate. This is because when k approaches zero, the Poisson model can be used (Miaou, 1993). For consistency, however, this study used the negative binomial distribution for all models. The lack of constant dispersion is anecdotally supported by the differences in the signs of the coefficients for the explanatory variables noted previously.

The deviance-based pseudo R-square measure (R^2_{dev}) values ranged from 0.07 to 0.74, with an average value of 0.356. In practical terms, a value above 0 means the model offers some

improvement over the case of not having a model, although there is no statistical criterion for accepting or rejecting a given model based solely on the deviance-based pseudo R-square measure.

It is possible to perform a test of statistical significance—the likelihood ratio test (Koppelman and Bhat, 2006)—to determine whether all coefficients in a model are necessary; the CEM such as that shown in Equation 11 would be compared to a more restricted version where certain parameters are set equal to zero. Such a test will detect whether the model requires the parameters in the CEMs, an outcome that is analogous to the model building process used herein. However, once the models are developed, there is no universal criterion for using (R^2_{dev}) alone to reject or accept a model, except to say that values closer to 1.0 are better than values closer to 0.0.

Model Adequacy for Predicting Crash Risk

Appendix C shows further testing with the model for one class of intersections: urban four-way signalized intersection with a four-lane road. The results show that there was no significant difference between the performance of the CEMs with 2004 through 2005 data (which were used to build the models) and 2006 through 2007 data (which were not used to build the models). This suggests the CEMs may be used for other time periods. However, the performance of the particular model examined in Appendix C does not suggest that it can be used to predict crashes on an intersection-by-intersection basis, since on average the number of actual crashes was approximately 3.5 times the number of predicted crashes. That is, the particular model examined showed a 77% average percentage error when trying to predict crashes at a specific intersection.

Model Adequacy for Identifying High-Risk Crash Locations

The models may have some utility for identifying high-risk crash locations. Using data from 2006 through 2007—i.e., data that were not used to build the models—the same intersections as noted previously were classified as high, medium, and low risk based on the number of crashes predicted by the models. Then, this classification was repeated using actual crash data from 2006 through 2007. Table 9 shows that the approach has some promise: e.g., 52% of the intersections that were high risk (based on actual crash data) were classified as high risk (based on the models).

Table 9. Comparison of Crash Risk Based on CEMs and 2006-2007 Crash Data

Actual Crash Risk	Predicted Crash Risk Based on CEM		
	Low	**Medium**	**High**
Low	55%[a]	22%	24%
Medium	36%	39%	25%
High	8%	40%	52%

[a]For example, according to actual 2006-2007 crash data, there were 93 intersections between percentile 0 (which was 1 actual crash) and percentile 33.33 (which was 12 actual crashes). Of these 93 intersections, 51 were classified by the *model* as being between percentile 0 (which was 0.82 predicted crashes) and percentile 33.33 (which was 3.466 predicted crashes). Accordingly, 55%, based on 51/93, is reported in Table 9.

Because this approach was tested for only one subset of one class of intersections, additional research is needed to determine whether this approach merits wider implementation. The use of CEMs is advantageous because they can, in theory, allow one to identify the numerous attributes that may influence crash risk. However, as shown in Table 10, CEMs did not provide a better estimation of crash risk than the use of volume alone.

Table 10. Comparison of Crash Risk Based on Volumes and 2006-2007 Crash Data

Actual Crash Risk	Predicted Crash Risk Based on Volume Alone		
	Low	**Medium**	**High**
Low	63%[a]	26%	11%
Medium	31%	52%	17%
High	4%	22%	74%

[a]For example, and consistent with Table 9, according to actual 2006-2007 crash data, there were 93 intersections between percentile 0 (which was 1 actual crash) and percentile 33.33 (which was 12 actual crashes). Of these 93 intersections, 59 were classified by the volume as being between percentile 0 (which was a daily entering volume of 2,216) and percentile 33.33 (which was a daily entering volume of 30,039). Thus, Table 10 shows a percentage of 63% based on 59/93.

CONCLUSIONS

- *Most crash data have an acceptable level of quality, at least in terms of data completeness and consistency.* Of the 179 data elements examined in Appendix A, 103 showed no data quality problems with regard to completeness or consistency. For the remaining 76, problems noted were inconsistency (e.g., a location of a crash might be designated as both a city and a county), incompleteness (e.g., a value might be missing because of difficulties with the crash referencing system), and availability (e.g., some data elements have been added to the VDOT CRASHDATA database but only for crashes occurring after September 2003).

- *For the data elements where data are imperfect, eight rules can enable some use of these data.* Five of the rules are ways analysts can improve data quality for a specific study. One example of such a rule is to use the *PHYSICALJURISDICTION* variable rather than the *CITY* or *COUNTY* variable to determine where a crash is located. Three of the rules are caveats that analysts should consider as they decide which time period and which data elements are necessary for a given study. An example is that 20 of the variables shown are not available until after September 2003, and thus in some cases it may be productive to consider such crashes only as part of a study.

- *No minimum set of variables for classifying crashes as injury or non-injury could be reliably identified based on the methods used in this study*. Although classification trees were developed for such an approach, the average error rate across all intersection types when such trees were applied to testing data was 38.63%. Given that injury crashes account for 35.55% of all crashes, statistical testing showed that for 16 of the 17 intersection classes, the error rate from the classification tress was not significantly lower than the error rate based on chance alone.

- *A minimum set of variables for classifying crashes as rear-end or angle can be identified.* These trees gave average error rates of 12.21% (angle) and 16.20% (rear-end). For all 17 intersection classes, the error rates were significantly lower than the error rates that would have resulted from chance alone (i.e., 35.99% for angle and 38.65% for rear-end crashes). Although 21 variables were used to develop these trees fully, much of this tree space (86%) is composed of 12 variables: vehicle speed, driver action, alignment, lighting, weather, traffic control, driver visibility obstruction, traffic volume, shoulder width, surface condition, vehicle type, and median type. These variables do not necessarily indicate causality; rather, they indicate the extent to which, given a crash has occurred, a determination can be made that the crash is rear-end or angle.

- *Crash risk is not uniformly proportional to volume*. For angle, injury, and total crashes, CEMs tended to have a volume exponent less than 1, meaning that a doubling of volume will increase crash risk but by less than 100%. For rear-end crashes, the results were mixed, with slightly more than one-half of the models suggesting that a doubling of volume will increase rear-end crash risk by more than 100%.

- *Models are not transferable across different intersection types or different crash types.* Examination of the 68 crash models showed substantive differences in the sign of the coefficients; e.g., having both approaches be undivided tends to increase the crash risk for angle crashes but reduces the crash risk for rear-end crashes. Even when the signs are consistent, the coefficients may differ substantially; e.g., of the nine angle crash models studied where making the approaches undivided increased risk, the increase varied between 43% and 154% depending on the particular intersection type modeled.

RECOMMENDATIONS

1. *VDOT district engineering staff, researchers, consultants, or other persons who use crash data should consider the following eight rules for improving data quality if their study relies on any of the 86 data elements shown to have data problems in Tables A1 through A4 in Appendix A:*

    - Use the *PHYSICALJURISDICTION* variable rather than the *CITY* or *COUNTY* variable to determine the jurisdiction where a crash is located.

    - Use the *TRAFFICCONTROL* variable rather than the *INTERSECTIONTYPE* variable to determine whether an intersection is signalized.

    - Manually extract volumes for intersections that contain a one-way street rather than using the IntersectionEnteringVolume function.

    - Use the *NODE* and *OFFSET* variables from the CrashIntersection table rather than from the CrashDocument table.

- Create intersection variables based on link variables as necessary.

- Recognize that 20 variables will not have crash data until sometime after September 2003 and/or that these data may be available in another location than those studied here.

- Recognize that new categories were added for six existing variables in September 2003.

- Recognize that the node variable is incomplete for about one-third of the crashes.

2. *If resources permit law enforcement to improve the quality of only a limited number of data elements in* the *VDOT CRASHDATA database, the more important predictive variables from Figure 5 should be given a higher priority since they were most important for classifying angle versus rear-end crashes.* This recommendation may be implemented through communications between the users of crash data—VDOT and the Department of Motor Vehicles—and the providers of these crash data—local law enforcement and the Virginia State Police—via the FR300. The nine most important predictive variables are vehicle speed, driver action, alignment, lighting, weather, traffic control, driver visibility obstruction, traffic volume, and shoulder width. The formal names for eight of these variables, from Appendix A, are *VEHICLESPEED, DRIVERACTION, ALIGNMENT, LIGHTING, WEATHER, TRAFFICCONTROL, VISIBILITYOBSTRUCTION,* and *SHOULDERWIDTH.* (*Traffic volume* was generally acquired from the TMS database using the IntersectionEnteringVolume function of the PkgCrashRate package). Although these variables were useful for classifying rear-end versus angle crashes, other criteria, such as the ability to classify intersection crashes, may alter the variables listed in this recommendation.

   The utility of this recommendation is that it can help identify those variables that classify a given crash as rear-end or angle. Its limitation is that the variables that can be identified as critical based on this study reflect only rear-end and angle crashes at intersections and not injury crashes.

3. *VDOT analysts should consider the CEMs provided in Appendix B as a way of identifying higher risk crash locations if other methods for identifying them are infeasible.* This study showed that other approaches for identifying high-risk crash locations, such as those that use volume alone, may prove as productive as this approach; thus, further research is needed before this recommendation is implemented. Further, the CEMs have not been tested on intersections outside Northern Virginia.

OPTIONS FOR FURTHER RESEARCH

- *Regarding Recommendation 1, VDOT may wish to consider adding geometric attribute variables to the existing node inventory table.* Without such variables (e.g., number of left, through, and right lanes; channelization; presence of frontage roads, curb cuts, and on-street parking), data must be collected manually. The benefits of such an initiative would need to

be considered against the costs of collecting and maintaining such data as part of the node inventory table.

- *Regarding Recommendation 2, an additional area of research for classifying crashes as injury or non-injury may be to modify the approach used in this study up to three ways:* (1) develop injury classification trees for rear-end crashes apart from angle crashes (rather than for all crashes); (2) use a more recent dataset that includes driver restraint usage; and (3) consider combining certain intersection types where the CEMs suggest the 17 intersection types are not reliable classifiers of injury versus non-injury crashes.

- *Regarding Recommendation 3, VDOT traffic engineering staff may wish to determine the extent to which the CEMs provided in Appendix B constitute an improvement over safety performance functions (SPFs) based solely on volumes and are transferable to areas outside Northern Virginia.* After this study was under way, a separate research effort (VTRC, 2010) was initiated to develop SPFs that estimate intersection crashes solely as a function of volume. Such SPFs do not include the other independent variables shown in Appendix B and are being developed exclusively for applying SafetyAnalyst software (VTRC, 2010). If Recommendation 3 is adopted, these staff may wish to compare the performance of models based solely on volumes (VTRC, 2010) with the performance of models based on volumes and other factors (Appendix B). Further, district traffic engineering staff may wish to consider testing selected CEMs from Appendix B with data from other locations. Although the result may be that recalibration of the CEMs is necessary, it is possible that the CEMs may prove applicable elsewhere to the extent that different geographic conditions, such as variation in traffic volume, are reflected in the models. In this endeavor, there may be opportunities to use data elements that were added to the FR300 in September 2003 (see Appendix A) in the development of these CEMs.

COSTS AND BENEFITS ASSESSMENT

Table 11 summarizes the costs and benefits of implementing the three study recommendations. The potential costs are the personnel hours required and the uncertainty associated with implementing the recommendations in various situations. The potential benefits are improved crash data, increased efficiency in acquiring such data, or both, which may ultimately reduce crash risk.

For example, the benefit of implementing Recommendation 1—the eight rules for data quality—is an ability to use a greater portion of the crash data than would otherwise be the case. The value of this benefit depends on the application. For applications for which a large amount of data is readily available, the benefit of additional data might be very small. However, for specialized applications for which only a small amount of data is available, being able to maximize the portion of those data that are useable provides a large benefit to the analyst. The costs of applying five of the eight rules are minimal, but the costs of applying the rules that require specialized queries can vary substantially, depending on the data required.

Table 11. Summary of Benefits and Risks Associated with Implementing Each Recommendation

	Recommendation 1	Recommendation 2	Recommendation 3
Deliverable	Eight rules for working with data elements that have imperfect consistency and completeness.	A minimal set of variables to predict rear-end and angle crashes.	68 crash estimation models (CEMs) for 17 intersection types.
Implementation	Make practitioners aware of the rules since the rules can increase the utility of available data.	If resources become available to improve data quality, consider focusing on variables noted in this recommendation, especially to the extent that rear-end and angle crashes are the focus of the analysis.	In Northern Virginia, use the CEMs to support countermeasure evaluation. Outside Northern Virginia, recalibrate the CEMs or test their transferability to other areas.
Potential Benefits	Greater use of existing data.	Better data quality for critical intersection-related data elements.	A technique, i.e., CEMs, that may be used to identify high-risk crash locations.
Potential Risk	None.	Data elements besides those listed in the recommendation may have utility for other applications and thus may be more important than recognized in this study.	It is not known if the CEMs may be extended to intersections outside Northern Virginia. CEMs should not replace the safety performance functions being developed for SafetyAnalyst (VTRC, undated).
Priority	*High:* There are no risks associated with implementing the recommendation, and it can improve the utility of crash data.	*Medium:* Recommendation has merit for rear-end and angle crashes, but other criteria, such as injury crashes, may influence which variables are targeted for improvement.	*Low:* Recommendation offers a potential approach to prioritize intersections, but it is not known if this approach constitutes an improvement over the use of volume alone.

ACKNOWLEDGMENTS

A VDOT steering committee composed of Hari Sripathi of VDOT's Northern Virginia District (chair), Steve Edwards and Stephen Read of VDOT's Traffic Engineering Division, and Gene Arnold and Mike Fontaine of VTRC provided insights at the inception of this research. Data collection was led by Lewis Woodson with assistance from Chase Buchanan, Stergios Gousios, Qun Liu, Trieu Nguyen, Njeri Kamatu, Phillip Haas, and Griselle Rivera, all of VTRC. Linda Evans edited this document.

REFERENCES

Abdel-Aty, M.A., and Radwan, E.A. Modeling Traffic Accident Occurrence and Involvement. *Accident Analysis and Prevention,* Vol. 32, No. 5, 2000, pp. 633-642.

Abdel-Aty, M., Salkapuram, H., Lee, C., and Brady, P.A. A Simplistic, Practical Approach to Identify Traffic Crash Profiles at Signalized Intersections. *ITE Journal*, April 2006, pp. 28-33.

Al-Ghamdi, A.S. Analysis of Traffic Accidents at Urban Intersections in Riyadh. *Accident Analysis and Prevention*, Vol. 35, No. 5, 2003, pp. 717-724.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees.* Chapman and Hall, Boca Raton, FL, 1984.

Cameron, A.C., and Windmeijer, F.A.G. R-squared Measures for Count Data Regression Models With Applications to Health Care Utilization. *Journal of Business and Economic Statistics*, Vol. 14, No. 2, 1996, pp. 209-220.

Campbell, J.R., and Knapp, K.K. Geometric Categories as Intersection Safety Evaluation Tools. In *Proceedings of the 2005 Mid-Continent Transportation Research Symposium.* Ames, IA, August, 2005.

Chang, L.Y., and Wang, H.W. Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Techniques. *Accident Analysis and Prevention*, Vol. 38, No. 5, 2006, pp. 1019-1027.

Federal Highway Administration. National Agenda for Intersection Safety, Washington, DC, 2002. http://safety.fhwa.dot.gov/intersection/resources/intersafagenda/. Accessed June 16, 2009.

Federal Highway Administration. *The National Intersection Safety Problem*. Washington, DC, 2004. http://safety.fhwa.dot.gov/intersections/interbriefing/01prob.htm. Accessed January 22, 2009.

Federal Highway Administration. *Intersection Safety*. Washington, DC, 2008. http://safety.fhwa.dot.gov/intersections/. Accessed January 22, 2009.

Garber, N.J., Miller, J.S., Yuan, B., and Sun, X. *The Safety Impacts of Differential Speed Limits on Rural Interstate Highways.* FHWA-HRT-05-042. Federal Highway Administration, McLean, VA, 2005. http://www.tfhrc.gov/safety/pubs/05042/05042.pdf. Accessed March 23, 2009.

Hardin, J.W., and Hilbe, J.M. *Generalized Linear Models and Extensions*, 2nd ed. StataCorp LP, College Station, TX, 2007.

Han, J., and Kamber, M. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, San Francisco, 2001.

Hauer, E. The Safety of Older Persons at Intersections. *Transportation in an Aging Society. Improving Mobility and Safety for Older Persons, Vol.2.* TRB Special Report 218.

Transportation Research Board of the National Academies, Washington, DC, 1988, pp. 194-252.

Joshua, S.C., and Garber, N.J. Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models. *Transportation Planning and Technology,* Vol. 15, 1990, pp. 41-58.

Jovanis, P., and Chang, H. Modeling the Relationship of Accidents to Miles Traveled. In *Transportation Research Record: Journal of the Transportation Research Board,* No. 1068. Transportation Research Board of the National Academies, Washington, DC, 1986, pp. 42-51.

Kaysi, I.A., and Abbany, A.S. Modeling Aggressive Driver Behavior at Unsignalized Intersections. *Accident Analysis and Prevention,* Vol. 39, No. 4, 2007, pp. 671-678.

Keller, J., Abdel-Aty, M., and Brady, P.A. Type of Collision and Crash Data Evaluation at Signalized Intersections. *ITE Journal,* February 2006, pp. 30-39.

Kim, D.G., Lee, Y., Washington, S., and Choi, K. Modeling Crash Outcome Probabilities at Rural Intersections: Application of Hierarchical Binomial Models. *Accident Analysis and Prevention,* Vol. 39, No. 1, 2007, pp. 125-134.

Koppelman, F., and Bhat, C. *A Self Instructing Course in Mode Choice Modeling: Multinomial and Logit Models*. Federal Transit Administration, Washington, DC, 2006. http://www.ce.utexas.edu/prof/bhat/COURSES/LM_Draft_060131Final-060630.pdf. Accessed June 8, 2010.

Kumara, S.S.P., Chin, H.C., and Weerakoon, W.M.S.B. Identification of Accident Casual Factors and Prediction of Hazardousness of Intersection Approaches. In *Transportation Research Record: Journal of the Transportation Research Board,* No. 1840. Transportation Research Board of the National Academies, Washington, DC, 2003, pp. 116-122.

Lord, D., Washington, S.P., and Ivan, J.N. *Statistical Challenges With Modeling Motor Vehicle Crashes: Understanding the Implications of Alternative Approaches.* Center for Transportation Safety, Texas Transportation Institute, College Station, 2004.

Lord, D., Washington, S.P., and Ivan, J.N. Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention,* Vol. 37, 2005, pp. 35-46.

McCullagh, P., and Nelder, J.A. *Generalized Linear Models*, 2nd ed. Chapman & Hall, New York, 1989.

Miaou, S.P. *The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions*. Oak Ridge National Laboratory, Oak Ridge, TN, 1993.

Miaou, S.P. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis and Prevention,* Vol. 26, No. 4, 1994, pp. 471-483.

Miaou, S.P., and Lum, H. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention,* Vol. 25, No. 6, 1993, pp. 689-709.

Miaou, S.P., and Lord, D. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. In *Transportation Research Record: Journal of the Transportation Research Board,* No. 1840. Transportation Research Board of the National Academies, Washington, DC, 2003, pp. 31-40.

Mitra, S., and Washington, S. On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. *Accident Analysis and Prevention,* Vol. 39, No. 3, 2007, pp. 459-468.

Oh, J., Lyon, C., Washington, S., Persaud, B., and Bared, J. Validation of FHWA Crash Models for Rural Intersections: Lessons Learned. In *Transportation Research Record: Journal of the Transportation Research Board,* No. 1840. Transportation Research Board of the National Academies, Washington, DC, 2003, pp. 41-49.

Pande, A., and Abdel-Aty, M. Identification of Rear-End Crash Patterns on Instrumented Freeways: A Data Mining Approach. In *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems,* Vienna, Austria, 2005, pp. 337-342.

Poch, M., and Mannering, F. Negative Binomial Analysis of Intersection-Accident Frequencies. *Journal of Transportation Engineering,* Vol. 122, No. 2, 1996, pp. 105-113.

SAS Institute Inc. *SAS/STAT ® 9.2 User's Guide*, 2nd ed. Cary, NC, 2009. http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf. Accessed June 10, 2010.

Shankar, V., Mannering, F., and Barfield, W. Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. A*ccident Analysis and Prevention,* Vol. 27, No. 3, 1995, pp. 371-389.

Shmueli, G., Patel, N.R., and Bruce, P.C. *Data Mining for Business Intelligence.* Wiley-Interscience, Hoboken, NJ, 2007.

Tesema, T.B., Abraham, A., and Grosan, C. Rule Mining and Classification of Road Traffic Accidents Using Adaptive Regression Trees. *International Journal of Simulation Systems, Science and Technology,* Vol. 6, Nos. 10 and 11, September 2005, pp. 80-94.

http://ducati.doc.ntu.ac.uk/uksim/journal/Vol-6/No.10-11/cover.htm. Accessed March 24, 2009.

Ulfarsson, G.F., and Shankar, V.N. Accident Count Model Based on Multiyear Cross-Sectional Roadway Data With Serial Correlation. In *Transportation Research Record: Journal of the Transportation Research Board,* No. 1840. Transportation Research Board of the National Academies, Washington, DC, 2003, pp. 193-197.

Virginia Transportation Research Council. Project Detail: Development of Safety Performance Functions for Intersections in Virginia: Phase II. Charlottesville, 2010. http://vtrc.virginiadot.org/ProjDetails.aspx?Id=441. Accessed August 31, 2009.

Virginia's Surface Transportation Safety Executive Committee. *Virginia's Strategic Highway Safety Plan*: *2006-2010.* Richmond, 2007. http://www.virginiadot.org/info/resources/Strat_Hway_Safety_Plan_FREPT.pdf. Accessed February 20, 2008.

Wang, Y., Leda, H., and Mannering, F. Estimating Rear-End Accident Probabilities at Signalized Intersections: Occurrence-Mechanism Approach. *Journal of Transportation Engineering*, Vol. 129, No. 4, July-August 2003, pp. 377-384.

Wong, S.C., Sze, N.N., and Li, Y.C. Contributory Factors to Traffic Crashes at Signalized Intersection in Hong Kong. *Accident Analysis and Prevention,* Vol. 39, No. 6, 2007, pp. 1107-1113.

Yan, X., Radwan, E., and Abdel-Aty, M. Characteristics of Rear-End Accidents at Signalized Intersections Using Multiple Logistic Regression Model. *Accident Analysis and Prevention,* Vol. 37, No. 6, 2005, pp. 983-995.

APPENDIX A

ADEQUACY OF SPECIFIC VARIABLES

Two databases are the focus of this appendix: the VDOT CRASHDATA database and HTRIS.

The VDOT CRASHDATA database contains numerous tables, three of which are shown in Tables A1, A2, and A3, respectively: the CrashDocument table (Table A1), the CrashVehicle table (Table A2), and the CrashPedestrian table (Table A3). These tables contain information about the crash (e.g., the route where the crash occurred as per Table A1); the vehicle and drivers involved in the crash (e.g., the driver's age as per Table A2); and, if a pedestrian was involved, information about the pedestrian (e.g., the pedestrian's age as per Table A3). For a given crash, these three tables may be linked through the *DOCUMENTNUMBER*, which is a unique number assigned to each crash as reported on the Police Crash Report (Form FR300). Variables for which the data quality was adequate have no entry in the fourth column.

Table A1. Variables in the CrashDocument Table

Variable	Type[a]	Source[b]	Problems with Data Quality
DOCUMENTNUMBER	NUMBER	FR300	
ROUTEPREFIX	CHAR	FR300	No data for crashes without reference nodes
ROUTENUMBER	CHAR	FR300	No data for crashes without reference nodes
ROUTESUFFIX	CHAR	FR300	No data for crashes without reference nodes
NODE	CHAR	FR300	Inappropriate reference system. Use instead the NODE variable from the CrashIntersection table
NODEOFFSET	NUMBER	FR300	Inappropriate reference system. Use instead the NODE variable from the CrashIntersection table.
NODETYPE	CHAR	HTRIS	Inappropriate reference system. Use instead the NODE variable from the CrashIntersection table.
CRASHDATE	DATE	FR300	
CRASHHOUR	NUMBER	FR300	
DISTRICT	CHAR	FR300	
COUNTY	CHAR	FR300	Inconsistent: conflicts with *ACCIDENTCITY*. Use instead the *PHYSICALJURISDICTION* variable from the CrashJurisdiction table.
SURFACETYPE	CHAR	HTRIS	No data for crashes without reference nodes
SURFACEWIDTH	NUMBER	HTRIS	No data for crashes without reference nodes
SHOULDERWIDTH	NUMBER	HTRIS	No data for crashes without reference nodes
LANECOUNT	NUMBER	HTRIS	No data for crashes without reference nodes
FACILITY	CHAR	HTRIS	No data for crashes without reference nodes
INTERSECTIONTYPE	CHAR	HTRIS	Inconsistent categories, no data for crashes without reference nodes, and incomplete data for the rest
TRAFFICCONTROL	CHAR	FR300	New categories added after September 2003.[c]
ALIGNMENT	CHAR	FR300	New categories added after September 2003.[c]
WEATHER	CHAR	FR300	
SURFACECONDITION	CHAR	FR300	New categories added after September 2003.[c]
ROADDEFECT	CHAR	FR300	
LIGHTING	CHAR	FR300	
COLLISIONTYPE	CHAR	FR300	
IMPACTZONE	CHAR	FR300	Incomplete data