



**pennsylvania**

DEPARTMENT OF TRANSPORTATION

# Statewide Crash Analysis and Forecasting

FINAL REPORT

November 20, 2008

By Paul P. Jovanis, Jonathan Aguero  
and Kun-Feng Wu

The Thomas D. Larson  
Pennsylvania Transportation Institute

COMMONWEALTH OF PENNSYLVANIA  
DEPARTMENT OF TRANSPORTATION

CONTRACT # 510401  
PROJECT # 013

PENNSTATE





<b>1. Report No.</b> FHWA-PA-2008-014-510401-013		<b>2. Government Accession No.</b>		<b>3. Recipient's Catalog No.</b>	
<b>4. Title and Subtitle</b> Statewide Crash Analysis and Forecasting				<b>5. Report Date</b> November 20, 2008	
				<b>6. Performing Organization Code</b>	
<b>7. Author(s)</b> Paul P. Jovanis, Jonathan Aguero and Kun-Feng Wu				<b>8. Performing Organization Report No.</b> PTI 2009-11	
<b>9. Performing Organization Name and Address</b> The Thomas D. Larson Pennsylvania Transportation Institute The Pennsylvania State University 201 Transportation Research Building University Park, PA 16802-4710				<b>10. Work Unit No. (TRAIS)</b>	
				<b>11. Contract or Grant No.</b> 510401, Work Order No. 13	
<b>12. Sponsoring Agency Name and Address</b> The Pennsylvania Department of Transportation Bureau of Planning and Research Commonwealth Keystone Building 400 North Street, 6 <sup>th</sup> Floor Harrisburg, PA 17120-0064  The Mid-Atlantic Universities Transportation Center The Pennsylvania State University 201 Transportation Research Building University Park, PA 16802-4710				<b>13. Type of Report and Period Covered</b> Final Report      5/21/2007 – 11/20/2008	
				<b>14. Sponsoring Agency Code</b>	
<b>15. Supplementary Notes</b> COTR: Lydia Peddicord, <a href="mailto:lpeddicord@state.pa.us">lpeddicord@state.pa.us</a> , 717-705-1706					
<b>16. Abstract</b> There is a need for the development of safety analysis tools to allow PennDOT to better assess the safety performance of road segments in the Commonwealth. The project utilized a safety management system database at PennDOT that integrates crash, occupant, vehicle and traffic information in an integrated searchable format (i.e., C-DART). The analyses conducted and models produced in this research should enhance PennDOT's ability to conduct safety analyses, particularly those using C-DART. The list of sites with promise contains a rank ordering of road segments offering the greatest potential for safety improvement. The model containing crash severity levels should give PennDOT additional confidence when combining fatal and severe injury crashes in needed analyses. Lastly, the models including census data have explored the feasibility of using that approach to safety modeling (although additional testing is needed).					
<b>17. Key Words</b> Road safety analysis, safety performance functions, modeling crash severity, sites with promise				<b>18. Distribution Statement</b> No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161	
<b>19. Security Classif. (of this report)</b> Unclassified		<b>20. Security Classif. (of this page)</b> Unclassified		<b>21. No. of Pages</b> 33	<b>22. Price</b>

This work was sponsored by the Pennsylvania Department of Transportation, the Mid-Atlantic Universities Transportation Center, and the U.S. Department of Transportation, Federal Highway Administration. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Federal Highway Administration, U.S. Department of Transportation, the Mid-Atlantic Universities Transportation Center, or the Commonwealth of Pennsylvania at the time of publication. This report does not constitute a standard, specification, or regulation.

# TABLE OF CONTENTS

EXECUTIVE SUMMARY .....	1
I. INTRODUCTION.....	4
II. PROJECT OVERVIEW.....	4
Task 1: Develop Safety Performance Functions for the Commonwealth .....	5
Task 2: Enhance the Analysis to Include Level of Outcome Severity .....	5
Task 3: Conduct Research to Identify Crash Contributors and Their Relationships to Internal and External Data Elements, Integrate Traffic, Population and Demographic Data into the Analysis to Facilitate More Comprehensive Assessment of Crash Indicators .....	6
III. SUMMARY .....	6
REFERENCES .....	8
APPENDIX A: SUMMARY OF MODELING APPROACH AND SAMPLE OUTPUT .....	9
APPENDIX B: MODEL DEVELOPMENT .....	16
APPENDIX C: MODEL FORMULATIONS AND SAMPLE OUTPUT .....	19



## ***Executive Summary***

There is a need for the development of safety analysis tools to enable the Pennsylvania Department of Transportation (PennDOT) to better assess the safety performance of road segments in the Commonwealth. The objective of this project was to conduct analyses of existing PennDOT data to provide PennDOT with tools to better manage road safety, reducing fatalities, injuries, and property damage losses in the Commonwealth. A particular objective was to conduct studies that would produce products of use in PennDOT's Crash Data Analysis and Retrieval Tool (C-DART).

### **Identifying Sites with Promise**

One critical task in the project was the identification of "Sites With Promise" (i.e., SWiPs). A SWiP is a road segment with a crash risk above the mean for comparable segments (those from one of six road classes considered in this research: urban and rural two-lane highways, urban and rural multilane highways, and urban and rural freeways and expressways). Identification of SWiPs included use of segment length and annual average daily traffic (AADT) as predictors.

A list of SWiPs has been provided to PennDOT for potential inclusion in C-DART. The list includes the PennDOT designation for the link (using county, route, and segment numbering) along with the mean and standard deviation of the excess risk. The list is in rank order with the highest-risk segment listed first. This list can be used to identify those road segments that have an elevated risk of a crash and also offer the greatest potential for safety improvement. This potential for improvement arises because the site has an expected crash frequency that is substantially higher than *comparable road segments*, controlling for AADT and segment length. The difference in expected crashes indicates that the site is "less safe" than comparable sites, presumably because of some site, driver, or other characteristics; this poorer safety is an indication that there are likely positive actions that can be taken to improve safety. A safety analyst at PennDOT can use this list of SWiPs to identify high-risk locations in a given district and then use the other analysis capabilities of C-DART to explore differences between the SWiPs and other sites in the comparison group. Such identifications should assist PennDOT in identifying and investing truly high-risk sites.

### **Consideration of Level of Outcome Severity**

An additional challenge in safety management is to find a way to adequately consider level of crash severity in the analysis. Crash frequencies are often treated as a whole, combining crashes of different severities of outcome (as described above). Alternatively, safety analysts may arbitrarily choose outcome levels for inclusion in a study (e.g., fatal and injury crashes only). Penn State researchers developed a sophisticated statistical model that simultaneously estimated the expected crash frequency for each of five levels of outcome severity ranging from fatal to property damage only (Aguero and Jovanis, 2009; 2008; 2007; 2006). Further, the method explicitly considered the correlation between crashes of different severity levels. The more sophisticated model reduced the standard deviation of the crash frequency estimates on the order of 20 percent overall and 40 percent or more for fatal and major injury crashes. This improved precision allows PennDOT safety analysts to be much more confident of the estimates of the mean of each crash outcome. Lists of the most severe outcome road segments have been provided to PennDOT in county, route, segment format for inclusion in C-DART for studies that require specific consideration of crash outcome. One interesting outcome of the analysis (Aguero and Jovanis, 2009) is that fatal and serious injury crashes showed a high correlation, while they were only slightly correlated to moderate, low severity and property damage only outcomes. This supports the concept of PennDOT combining fatal and serious injury crashes together in safety studies.

### **Consideration of Type of Collision and Demographic Factors**

To further illustrate the utility of the Penn State approach, separate analyses were conducted of two specific crash types: single vehicle run-off-road and multiple vehicle head-on and sideswipe crashes. Using a procedure similar to those described above, the team identified SWiPs for each crash type. These have also been provided to PennDOT for use by safety staff interested in these severe outcome crash types.

The Penn State team added census records to link level data in the Harrisburg and Clearfield PennDOT districts. Models were estimated seeking to capture the effect of factors such as area

income and driver age on crash risk. A list of sites was not a required outcome, but the final models are included in Appendix C of this report.

### **Summary**

The analyses conducted and models produced in this research should enhance PennDOT's ability to conduct safety analyses, particularly those using C-DART. The list of SWiPs contains a rank ordering of road segments offering the greatest potential for safety improvement. The model containing crash severity levels should give PennDOT additional confidence when combining fatal and severe injury crashes in needed analyses. Lastly, the models including census data have explored the feasibility of using that approach to safety modeling (although additional testing is needed).

## ***I. Introduction***

There is a need for the development of safety analysis tools to allow PennDOT to better assess the safety performance of road segments in the Commonwealth.

The objective of this project was to conduct analyses of existing PennDOT data to identify sites of elevated crash risk. Additional analyses explore level of injury severity, the ability to estimate changes in numbers of expected crashes as annual average daily traffic (AADT) increases, the ability to identify high-risk crash types, and the development of a crash prediction model that includes traffic, population, and demographic factors. Overall, there is a goal of improved safety management on PennDOT roads so that PennDOT may be able to achieve a reduction in crash fatalities, injuries, and property damage losses.

## ***II. Project Overview***

The project drew upon the availability of a safety management system database at PennDOT that integrates crash, occupant, vehicle and traffic (ADT) information in an integrated searchable format (i.e., C-DART). Penn State has applied this data base using state-of-the-art modeling approaches in a series of studies of the spatial location of “high hazard” sites (1-4). The project was conducted and reports were delivered in a series of tasks, which are summarized in the following sections.

The research team applied the statistical methods in a systematic way throughout the Commonwealth for the following functional classifications of roads:

- Urban and rural two-lane roads,
- Urban and rural multilane Interstates and expressways, and
- Urban and rural multilane roads including arterials.

Maps and lists of the high-risk locations were developed so PennDOT can identify corridors (contiguous roadway segments) for safety improvements. During this initial phase of study, only road segment crashes were analyzed and identified. Intersection and ramp-related crashes were

left for future research. High-risk locations are referred to in this report by the acronym commonly used in national research studies: Sites With Promise or SWiPs.

### **Task 1: Develop Safety Performance Functions for the Commonwealth**

In order to accomplish this task, the Penn State team received the C-DART data from PennDOT and implemented the data in computer labs of The Thomas D. Larson Pennsylvania Transportation Institute (LTI). In addition, the team checked for errors and verified the accuracy of crash, ADT, occupant, and vehicle information for a small sample of crashes. Finally, the team produced estimates of the statewide sites with promise for the functional roadway classifications listed above. Deliverables were provided to PennDOT as lists of sites and maps depicting SWiP road segments. A summary of the modeling approach and sample output are provided in Appendix A.

### **Task 2: Enhance the Analysis to Include Level of Outcome Severity**

The Penn State team enhanced the capability provided by the functions developed in Task 1 by considering the severity of injury outcome (e.g., fatality, injury level). These analyses evaluated several modeling approaches to the problem and produced a list of sites with elevated risk of a crash by outcome.

In addition, PennDOT was interested in a statistical model to forecast crashes given changes in future travel (e.g., changes in the level of fatalities as ADT continues to grow over a 5- to 10-year period). The Penn State team explored alternative modeling approaches to conduct these assessments and delivered a model that included ADT.

Steps required to complete this task included: development of safety performance functions or a similar statistical approach that recognizes level of outcome severity as well as crash occurrence; development of a statewide safety forecasting tool to allow estimates of changes in road safety in future years. Deliverables included a list of sites with “high risk” of severe outcome for each of the six roadway functional classifications and completed development of models capable of estimating changes in expected number of crashes given changes in ADT. These deliverables are summarized along with the model development in Appendix B of the report.

### **Task 3: Conduct Research to Identify Crash Contributors and Their Relationships to Internal and External Data Elements, Integrate Traffic, Population and Demographic Data into the Analysis to Facilitate More Comprehensive Assessment of Crash Indicators**

PennDOT has indicated that it is seeking an automated (i.e., computer-based) procedure for identifying sites that are candidates to be treated along with an identification of the factors contributing to the crashes. The Penn State team conducted this assessment by identifying those road segments with crash types that are overrepresented within a road functional class. For example, for the functional class of two-lane rural roads, the Penn State team developed methods to compare the expected proportion of run-off-road crashes with the mean number experienced by each segment. Two general crash types were considered: single-vehicle run-off-road and multi-vehicle head-on/sideswipe crashes. The research team included data from all single-vehicle, hit-fixed-object crashes that occurred off roadway (right and left); this should provide coverage of the run-off-road crashes. Multi-vehicle head-on/sideswipe crashes are of interest because they are typically high severity. The relation to roadway variable necessary to identify this collision type would be “on travelway” and/or “in median.” Lists of overrepresented segments were delivered to PennDOT with the Task 3 report.

In addition, analyses were conducted of approaches to estimate crash risk that include traffic, demographic, and population factors. This assessment was conducted at the district level, developing models for one urban and one rural district. District 2-0 was the rural district; Harrisburg was the urban district. The Task 3 report included a procedure to follow for the identification of roadway segments that have an unusual crash pattern for the six functional classes of roadways and a model of the expected number of crashes (at the road segment level) sensitive to a reasonable initial range of policy variables. Model formulations and sample output are contained in Appendix C.

### ***III. Summary***

Penn State has delivered a series of products to PennDOT during this project that includes:

- Lists of sites and maps depicting SWiP road segments,

- Lists of sites ranked by the severity of injury outcome (e.g., fatality, injury level),
- Lists of sites with excess crash frequency of particular crash types, and
- Models of expected crash frequency that include demographic and socioeconomic variables for the Harrisburg and Centre County districts.

The Penn State team has also responded to additional requests by PennDOT to provide examples of how the data sets and lists developed in this research can be used in safety analyses.

## **References**

Agüero, J., and P. P. Jovanis, “Bayesian Multivariate Poisson Log-Normal Models for Crash Severity Modeling and Site Ranking,” accepted for presentation, Transportation Research Board Annual Meeting, January 2009.

Agüero, J., and P. P. Jovanis, “Analysis of Road Crash Frequency Using Spatial Models,” in press, *Journal of the Transportation Research Board*, January 2008.

Agüero, J., and P. P. Jovanis, “Identifying Road Segments with High Risk of Weather-Related Crashes Using Full Bayesian Hierarchical Models,” CD, *Proceedings of Annual Meeting of Transportation Research Board*, January 2007.

Agüero, J., and P. P. Jovanis, “Spatial Analysis of Fatal and Injury Crashes in Pennsylvania,” *Accident Analysis and Prevention*, Volume 38, Issue 3, pages 618-625, May 2006.

## **APPENDICES**

### ***Appendix A: Summary of Modeling Approach and Sample Output***

Penn State is currently using a state-of-the-art approach to the identification and spatial location of “high hazard” (i.e., referred to in national studies as “site-with-promise” and herein as SWiPs) throughout the Commonwealth for the following functional classifications of roads:

- Urban and rural two-lane roads,
- Urban and rural multilane Interstates and expressways, and
- Urban and rural multilane roads including arterials.

Maps and lists of the high-risk locations were developed so PennDOT can identify corridors (contiguous roadway segments) for safety improvements. During this initial phase of study, only road segment crashes were analyzed and identified. Intersection and ramp-related crashes were left for future research.

Steps accomplished within this task include receipt of data from PennDOT and implementation in computer labs at The Thomas D. Larson Pennsylvania Transportation Institute (LTI); check for errors and verification of accuracy of crash, AADT, occupant, and vehicle information for a small sample of crashes; and production of estimates of the statewide SWiPs for the functional roadway classifications listed above.

#### **Overview of Method**

Full Bayes Hierarchical Models were used to estimate the Safety Performance Functions and the expected Excess Crash Frequency for each segment in the six road classes. This was later used to rank the “sites with promise.” The expected excess crash frequency is defined as the expected number of crashes in a particular segment minus the average number of crashes expected in a group of similar segments. The latter is estimated using the Safety Performance Function.

#### **Statistical Background**

Consider the number of crashes at the  $i$ th segment and  $t$ th time period,  $Y_{it}$ , to be a random variable which is Poisson and independently distributed, when conditional on its mean  $\mu_{it}$ :

$$Y_{it} | \mu_{it} \stackrel{\text{ind}}{\sim} \text{Pois}(\mu_{it}) \quad (1)$$

The expected number of crashes at a site can be defined as the product of the exposure to the risk and the relative risk of a motor-vehicle crash as follows:

$$\mu_{it} = \eta_{it} \rho_i \quad (2)$$

where  $\mu_{it}$  is the expected number of crashes at segment  $i$ , and time period  $t$ ,  $\eta_{it}$  is the exposure function at segment  $i$ , and time period  $t$ , and  $\rho_i$  is the expected crash relative risk at segment  $i$ .

The exposure or Safety Performance Function is defined as:

$$\eta_{it} = \beta_0 V_{it}^{\beta_V} L_{it}^{\beta_L} \quad (3)$$

where  $V_{it}$  is the AADT of segment  $i$ ,  $L_{it}$  is the length of segment  $i$ , and  $\beta_0$ ,  $\beta_V$ ,  $\beta_L$  are parameters of the model.

The relative risk (with respect to the SPF) is defined as:

$$\rho_i = \exp(v_i) \quad (4)$$

where  $v_i$  is an unstructured random effect for segment  $i$  with a normal prior distribution with mean = 0 and variance =  $\sigma_v^2$ . The variance is modeled through the precision parameter  $\tau_v^2 = 1/\sigma_v^2$  with a gamma hyperprior  $\tau_v^2 \sim \text{Gamma}(0.5, 0.0005)$ . The random effects are considered fixed over time, which allows for smoothing of the estimates and controls for regression to the mean bias.

Since FB models are used,  $v_i$  and therefore,  $\rho_i$  are estimated for all crashes for the six different types of segments under study. Unobserved effects can be captured by  $v_i$ , reflecting individual differences between segments. This method allows us to estimate a mean over several years and directly consider changes in ADT over those years.

The full posterior distribution of  $v_i$  is estimated in the FB analysis; consequently, the credible set (confidence interval) for the relative risk is estimated.

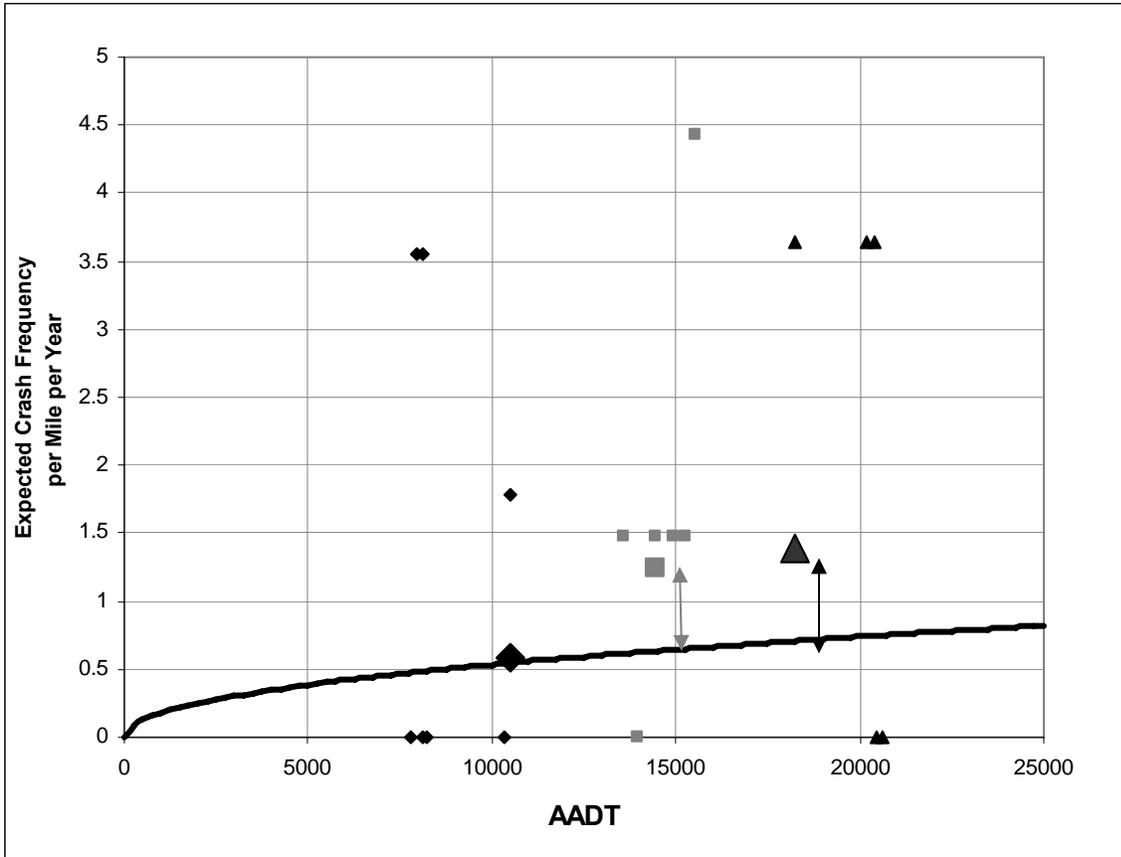
The excess crash frequency can also be estimated using random effects. The excess crash frequency,  $\delta_{it}$ , is defined as the difference between the expected crash frequency at segment  $i$  at time  $t$  and the expected crash frequency of a group of similar sites; for example:

$$\delta_{it} = \eta_{it} * (\exp(v_i) - 1) \quad (5)$$

This can be simplified to:

$$\delta_{it} = \eta_{it} (\rho_i - 1) \quad (6)$$

Figure A-1 illustrates the concept of excess crash frequency as used in the bayesian approach. The black line is the Safety Performance Function. The small squares, triangles, and diamonds represent the observed number of crashes per mile for three different segments. Finally, the bigger symbols represent the expected number of crashes per mile and the arrows measure the expected excess crash frequency for each segment. Additional interpretation and discussion of this model was presented to PennDOT during a briefing in Harrisburg on March 10, 2008. The remaining pages contain summaries of the models developed during Task 2. The list of high-risk sites has been provided on CD, since it is lengthy.



**Figure A-1. Safety Performance Function and expected crash frequency for three segments.**

## Safety Performance Functions

Tables A-1 through A-6 show the SPFs for urban and rural roads.

**Table A-1. Urban Two-Lane SPF.**

				Confidence Interval	
	Mean	Std. Dev.	MC Error	2.50%	97.50%
Constant	-6.1450	0.0765	0.00147	-6.2930	-5.9980
AADT	0.7440	0.0086	0.00017	0.7273	0.7609
Length	0.9998	0.0190	0.00036	0.9628	1.0370
Std. Dev.	0.6796	0.0067	0.00013	0.6664	0.6926
Dbar	Dhat	DIC	pD		
130500	122400	138500	8060		

**Table A-2. Urban Interstate and Freeways SPF.**

				Confidence Interval	
	Mean	Std. Dev.	MC Error	2.50%	97.50%
Constant	-6.8460	0.1535	0.00466	-7.1460	-6.5400
AADT	0.7744	0.0153	0.00048	0.7441	0.8041
Length	0.7874	0.0267	0.00059	0.7346	0.8394
Std. Dev.	0.7289	0.0071	0.00013	0.7152	0.7430
Dbar	Dhat	DIC	pD		
90880	85300	96460	5578		

**Table A-3. Urban Multilane SPF.**

				Confidence Interval	
	Mean	Std. Dev.	MC Error	2.50%	97.50%
Constant	-7.5850	0.4079	0.0081	-8.3830	-6.7780
AADT	0.8788	0.0445	0.0009	0.7913	0.9655
Length	1.0210	0.0490	0.0008	0.9261	1.1180
Std. Dev.	0.7461	0.0225	0.0005	0.7033	0.7911

Dbar      Dhat      DIC      pD  
 13450      12560      14340      892

**Table A-4. Rural Two-Lane SPF.**

				Confidence Interval	
	Mean	Std. Dev.	MC Error	2.50%	97.50%
Constant	-6.2040	0.0381	0.00062	-6.2780	-6.1290
AADT	0.7444	0.0048	0.00008	0.7349	0.7536
Length	0.9321	0.0186	0.00024	0.8956	0.9687
Std. Dev.	0.6210	0.0061	0.00020	0.6092	0.6330

Dbar      Dhat      DIC      pD  
 241300      228300      254200      12920

**Table A-5. Rural Interstate and Freeways SPF.**

				Confidence Interval	
	Mean	Std. Dev.	MC Error	2.50%	97.50%
Constant	-6.2060	0.2473	0.0046	-6.6850	-5.7210
AADT	0.6751	0.0255	0.0005	0.6247	0.7243
Length	0.7914	0.0601	0.0010	0.6752	0.9085
Std. Dev.	0.6140	0.0103	0.0002	0.5938	0.6341

Dbar      Dhat      DIC      pD  
 46620      44030      49200      2581

**Table A-6. Rural Multilane SPF.**

				Confidence Interval	
	Mean	Std. Dev.	MC Error	2.50%	97.50%
Constant	-7.0550	0.5055	0.0091	-8.0520	-6.0740
AADT	0.8133	0.0585	0.0011	0.7001	0.9291
Length	0.8927	0.0851	0.0013	0.7269	1.0600
Std. Dev.	0.6158	0.0381	0.0010	0.5437	0.6919

Dbar      Dhat      DIC      pD  
 5134      4850      5418      283.7

## Appendix B: Model Development

### Severity of Outcome Models

The following models incorporate the severity of outcome of the crash. The levels of severity included are those used by PennDOT: deaths, major injuries, moderate injuries, minor injuries, and property damage only (PDO). There are two analyses; in the first one, the response variable is the number of crashes classified by severity according to the maximum level of severity observed in that crash. In the second analysis, the response variable is the number of persons injured by severity of injury. For this analysis the count of crashes is used for PDO, since there are no injured persons in this type of crash.

For the crash count models, the number of crashes is Poisson distributed:

$$y_{ijt} \sim \text{Poisson}(\theta_{ijt}) \quad (7)$$

where  $y_{ijt}$  is the observed number of crashes in segment  $i$  of the type severity  $j$  at time  $t$  (in years), and  $\theta_{ijt}$  are the expected Poisson crash rate for segment  $i$  of severity  $j$  at time  $t$ . The Poisson rate is modeled as a function of the covariates following a log-normal distribution, as shown in Equation 8:

$$\log(\theta_{ijt}) = \beta_{0j} + \beta_{Aj} \ln(\text{aadt}_{ijt}) + \beta_{Lj} \ln(\text{length}_{ijt}) + v_i \quad (8)$$

where  $\beta_{0j}$  is the intercept for severity  $j$ ,  $\beta_{Aj}$  is the coefficient for AADT for severity  $j$ ,  $\beta_{Lj}$  is the coefficient for segment length for severity  $j$ , and  $v_i$  captures the heterogeneity among segments. Now, one can assume that the coefficients for each severity type are independent, and therefore have the following prior distributions:

$$\begin{aligned} \beta_{0j} &\sim N(0, 0.001)_{,j=1:5} \\ \beta_{Aj} &\sim N(0, 0.001)_{,j=1:5} \\ \beta_{Lj} &\sim N(0, 0.001)_{,j=1:5} \end{aligned} \quad (9)$$

A more reasonable assumption is that the number of crashes per severity level are positively correlated to each other, i.e. the higher the number of fatal crashes the higher the number of injury crashes. For these models, correlated priors in the coefficients are estimated using multivariate normal priors:

$$\begin{aligned}
 \boldsymbol{\beta}_0 &\sim \text{MN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
 \boldsymbol{\beta}_A &\sim \text{MN}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) \\
 \boldsymbol{\beta}_L &\sim \text{MN}(\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L)
 \end{aligned} \tag{10}$$

where  $\boldsymbol{\mu}_A$  is a vector of zeroes  $\boldsymbol{\mu}_A = (0, 0, 0, 0, 0, 0)$  and  $\boldsymbol{\Sigma}_A$  is the variance-covariance Matrix with a hyper-prior defined by:

$$\boldsymbol{\Sigma}_A^{-1} \sim \text{Wishart}(\mathbf{R}, n) \tag{11}$$

where  $\boldsymbol{\Sigma}_A^{-1}$  is a symmetric positive definite matrix, R is the scale matrix =

$$\begin{vmatrix}
 0.1 & 0.005 & 0.005 & 0.005 & 0.005 \\
 0.005 & 0.1 & 0.005 & 0.005 & 0.005 \\
 0.005 & 0.005 & 0.1 & 0.005 & 0.005 \\
 0.005 & 0.005 & 0.005 & 0.1 & 0.005 \\
 0.005 & 0.005 & 0.005 & 0.005 & 0.1
 \end{vmatrix}$$

and n is the degrees of freedom = 5.

For ranking of sites, the cost of the crash can be used. For this work, the costs associated with crashes were obtained from the 2006 Pennsylvania Crash Facts and Statistics report, published by PennDOT. In particular, excess crash cost was used. Excess crash cost is defined in Equation 12:

$$\delta - cost_{it} = \sum_{j=1}^5 cost_j \delta_{ijt} \tag{12}$$

where cost  $j$  is the average cost associated with a crash of severity  $j$  and  $\delta_{ijt}$  is the expected excess crash frequency for segment  $i$ , of the severity  $j$ , at time  $t$  defined in Equation 13:

$$\delta_{ijt} = e^{\beta_{0j} + \beta_{Aj} \ln(aadt_{ijt}) + \beta_{Lj} \ln(length_{ijt})} (e^{v_i} - 1) \quad (13)$$

## **Appendix C: Model Formulations and Sample Output**

### **Background**

Task 3 of WO 13 required the estimation of models of crash risk that include socio-demographic attributes for two districts. In addition, the Penn State team developed specific SPFs for particular crash types.

### **Model Development**

The authors explored two possibilities of two-level models in highway safety to include demographic and population factors in the mix. For this analysis we defined the first level unit to be the road segment and the second level unit to be the census tract. This formulation enabled the inclusion of socio-demographic information from census tracts in the PennDOT district of interest with the crash, roadway, and traffic attributes from C-DART. Using a variance components model we have:

$$y_{ikt} \sim \text{Poisson}(\theta_{ikt}) \quad (14)$$

where  $y_{ikt}$  are the observed number of crashes in segment  $i$  and census track  $k$  at time  $t$  (in years), and  $\theta_{ikt}$  is the expected Poisson rate for segment  $i$  and census track  $k$  at time  $t$ . The Poisson rate is modeled as a function of the covariates following a log-normal distribution as shown in Equation 2:

$$\log(\theta_{ikt}) = \mathbf{X}_{ikt} \boldsymbol{\beta} + \mathbf{W}_k \boldsymbol{\gamma} + u_k + v_{ik} \quad (15)$$

where  $\mathbf{X}_{ikt}$  is the vector of covariates for segment  $i$  and census track  $k$  at time  $t$ ,  $\boldsymbol{\beta}$  is the vector of coefficients for the segment level fixed effects,  $\mathbf{W}_k$  is the vector of covariates for census track  $k$ ,  $\boldsymbol{\gamma}$  is the vector of coefficients for census track level fixed effects,  $u_k$  are the random effects across census tracks, and  $v_{ik}$  are the random effects across segments.

Now the random effects themselves are normally distributed. For the effects across tracks we have:

$$u_k \sim N(0, \tau_u) \quad (16)$$

where  $\tau_u$  has a hyperprior:

$$\tau_u \sim \text{Gamma}(0.001, 0.001) \quad (17)$$

For the effects across segments:

$$v_{ik} \sim N(0, \tau_v) \quad (18)$$

where  $\tau_v$  has a hyperprior:

$$\tau_v \sim \text{Gamma}(0.001, 0.001) \quad (19)$$

### Model Results

The multi-level District 8 model in Table C-1 is a two-level specification. The response variable is the number of crashes that occurred on each segment during 2003 to 2006, and the explanatory variable includes road segment characteristics at level 1 and census tract at level 2. The slope coefficients also vary randomly at the road segment level by functional classes: urban/rural 2-lane, urban/rural interstate, and urban/rural multi-lane.

One of the differences between multilevel models and standard multiple regression is the multilevel model has two random variables,  $u_k$ , a segment-level random variable, and  $v_{ik}$ , a tract-level random variable. Standard multiple regression has only one random variable, called the error term. The correlation between two segments in the same track, which is referred to as the variance component variance (VPC), or intra-level2-unit correlation, is given by

$$VPC = \frac{\sigma_{u_k}}{\sigma_{u_k} + \sigma_{v_{ij}}}$$

This is the between-tract variance over the total variance. The higher the value of VPC, the more similar two segments from the same tract are, compared to two segments picked at random from the population. The VPC in this model is 0.23, indicating a strong clustering effect of the tract. Clearly, in the presence of clustering, the assumption of independent observations in standard multiple regression is wrong, which can lead to incorrect inference. Fitting a model that does not recognize the presence of clustering creates serious problems such as underestimation of the standard error of regression coefficients.

AADT and length of segment by different road types are inherently significant and positive, reflecting the fact that higher vehicle miles traveled increases the likelihood of crashes. Moreover, tracts with both higher population density and higher population have a significantly increased likelihood of crash occurrence. Though census tract variables of poverty proportion and income greatly improve the goodness-of-fit, they are not significant in this model.

<b>Table C-1. Multi-level district 8 model</b>								
			<b>Percentile</b>					
	<b>Mean</b>	<b>SD</b>	<b>5%</b>	<b>10%</b>	<b>90%</b>	<b>95%</b>		<b>Sig</b>
VPC	0.23	0.02	0.19	0.20	0.26	0.27		*
Intercept_Urban two-lane	-3.64	0.31	-4.06	-4.00	-3.18	-3.09		*
Intercept_Urban interstate	-3.79	0.42	-4.48	-4.38	-3.27	-3.13		*
Intercept_Urban multi-lane	-6.49	0.89	-7.90	-7.71	-5.39	-4.98		*
Intercept_Rural two-lane	-3.70	0.29	-4.22	-4.14	-3.34	-3.28		*
Intercept_Rural interstate	-2.59	0.58	-3.49	-3.29	-1.78	-1.67		*
Intercept_Rural multi-lane	-6.63	1.38	-8.76	-8.23	-4.67	-4.17		*
AADT_Urban two-lane	0.68	0.02	0.64	0.65	0.71	0.72		*
AADT_Urban interstate	0.65	0.03	0.61	0.62	0.69	0.70		*
AADT_Urban multi-lane	1.00	0.09	0.84	0.87	1.12	1.15		*
AADT_Rural two-lane	0.69	0.01	0.67	0.68	0.70	0.70		*
AADT_Rural interstate	0.50	0.05	0.41	0.43	0.55	0.57		*
AADT_Rural multi-lane	0.94	0.14	0.66	0.73	1.10	1.15		*
Length_Urban two-lane	1.03	0.05	0.96	0.98	1.09	1.11		*
Length_Urban interstate	0.88	0.07	0.77	0.79	0.96	0.99		*
Length_Urban multi-lane	1.15	0.12	0.96	1.00	1.30	1.35		*
Length_Rural two-lane	0.89	0.04	0.82	0.83	0.94	0.96		*
Length_Rural interstate	0.71	0.13	0.50	0.55	0.89	0.94		*
Length_Rural multi-lane	0.57	0.19	0.26	0.33	0.81	0.89		*
Females_age under15	-0.03	0.01	-0.04	-0.04	-0.02	-0.02		*
Females_age 15-17	-0.06	0.02	-0.10	-0.09	-0.04	-0.02		*
Females_age 18-24	0.00	0.01	-0.02	-0.02	0.01	0.01		
Females_age 25-34	-0.01	0.01	-0.03	-0.03	0.01	0.01		
Females_age 50-59	-0.02	0.01	-0.04	-0.04	-0.01	-0.01		*
Females_age 60-69	0.03	0.01	0.01	0.01	0.05	0.05		*
Females_age above70	-0.03	0.01	-0.04	-0.04	-0.02	-0.02		*

<b>Table C-1. Multi-level district 8 model (continued).</b>								
			<b>Percentile</b>					
	<b>Mean</b>	<b>SD</b>	<b>5%</b>	<b>10%</b>	<b>90%</b>	<b>95%</b>		<b>Sig</b>
Males_age under15	-0.01	0.01	-0.03	-0.03	0.00	0.00		
Males_age 15-17	0.02	0.05	-0.05	-0.04	0.09	0.11		
Males_age 18-24	-0.02	0.01	-0.04	-0.04	-0.01	-0.01		*
Males_age 25-34	-0.01	0.01	-0.03	-0.03	0.00	0.00		
Males_age 50-59	-0.05	0.01	-0.07	-0.07	-0.03	-0.02		*
Males_age 60-69	-0.09	0.02	-0.13	-0.11	-0.05	-0.05		*
Males_age above70	0.02	0.01	0.00	0.00	0.03	0.04		*
Population	0.10	0.01	0.08	0.09	0.11	0.12		*
Population density	57.90	20.64	22.91	31.90	83.72	91.18		*
Poverty proportion	-0.14	0.55	-1.05	-0.85	0.54	0.74		
Income	0.01	0.02	-0.02	-0.01	0.03	0.04		
				Dbar	Dhat	DIC	pD	
				84130	79679	88580	4450	

To provide some insight into the implications of parameter estimation results, elasticities are computed to determine the marginal effects of the independent variables (also refer to Shankar et al., 1995; Shankar et al., 1996; Ulfarsson, 2001; Washington et al., 2003). Elasticities provide an estimate of the impact of a variable on the expected frequency and are interpreted as the effect of a 1 percent change in the variable on the expected frequency  $\lambda_i$ . Elasticity of frequency in log-linear relation ( $\log \lambda = \alpha + \beta x$ ) is defined as

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik} \quad 1$$


---

<sup>1</sup> In log-linear relation  $\log \lambda = \alpha + \beta x$ ,  $\frac{1}{\lambda} \frac{\partial \lambda}{\partial x} = \beta$ ,  $E = \frac{\partial \lambda}{\partial x} \frac{x}{\lambda} = \beta \lambda \times \frac{x}{\lambda} = \beta x$

Where  $E$  represents the elasticity,  $x_{ik}$  is the value of the  $k$ th independent variable for observation  $i$ ,  $\beta_k$  is the estimated parameter for the  $k$ th independent variable, and  $\lambda_i$  is the expected frequency for observation  $i$ . Note that elasticities are computed for each observation  $i$ . It is common to report a single elasticity as the average elasticity over all  $i$ . Note that elasticity of frequency in log-log relation ( $\log \lambda = \alpha + \beta \log x$ ) is defined as

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k \quad 2$$

For example, a 1 percent increase in the AADT on an urban, two-lane road causes an average 0.68 percent increase in crash frequency and a 1 percent increase in the population causes an

---

<sup>2</sup> In log-log relation  $\log \lambda = \alpha + \beta \log x$ ,  $\frac{\partial \lambda}{\lambda} = \beta \frac{\partial x}{x}$ ,  $E = \frac{\partial \lambda}{\partial x} \frac{x}{\lambda} = \beta \frac{\lambda}{x} \times \frac{x}{\lambda} = \beta$

average 0.1 percent increase in crash frequency, as shown in Table C-2. It must be noted that the elasticity on multi-lane roadways is always greater than elasticities of two-lane roadways and interstates. It's important to note that we can't interpret the elasticities for length of segments, since it is defined by geometric characteristics.

### **District 2-0 Model Results**

Results for District 2-0 are summarized in Table C-3.

**Table C-2 District 8-0 Variable Elasticity**

	<b>Coefficients</b>	<b>Elasticity</b>	<b>Sig</b>
AADT_Urban two-lane	0.68	0.68	*
AADT_Urban interstate	0.65	0.65	*
AADT_Urban multi-lane	1.00	1.00	*
AADT_Rural two-lane	0.69	0.69	*
AADT_Rural interstate	0.50	0.50	*
AADT_Rural multi-lane	0.94	0.94	*
Length_Urban two-lane	1.03	1.03	*
Length_Urban interstate	0.88	0.88	*
Length_Urban multi-lane	1.15	1.15	*
Length_Rural two-lane	0.89	0.89	*
Length_Rural interstate	0.71	0.71	*
Length_Rural multi-lane	0.57	0.57	*
Females_age under15	-0.03	-0.58	*
Females_age 15-17	-0.06	-0.23	*
Females_age 18-24	0.00	0.00	
Females_age 25-34	-0.01	-0.13	
Females_age 50-59	-0.02	-0.23	*
Females_age 60-69	0.03	0.11	*
Females_age above70	-0.03	-0.38	*
Males_age under15	-0.01	-0.21	
Males_age 15-17	0.02	0.09	
Males_age 18-24	-0.02	-0.18	*
Males_age 25-34	-0.01	-0.13	
Males_age 50-59	-0.05	-0.59	*
Males_age 60-69	-0.09	-0.56	*
Males_age above70	0.02	0.17	*
Population	0.10	0.10	*
Population density	57.90	0.06	*
Poverty proportion	-0.14	-0.01	
Income	0.01	0.01	

**Table C-3 District 2-0 Model**

	Mean	SD	Percentile				Sig
			5%	10%	90%	95%	
VPC	0.17	0.04	0.12	0.13	0.22	0.24	*
Intercept_Urb two-	-2.03	1.14	-3.63	-3.37	-0.35	0.21	.
Intercept_Urb inters	-3.95	2.16	-7.24	-6.35	-0.94	0.01	.
Intercept_Urb multi-	-0.93	2.69	-5.39	-4.53	2.70	3.56	
Intercept_Rural two-	-2.86	1.03	-4.20	-3.98	-1.22	-0.85	*
Intercept_Rural	-2.75	1.40	-4.55	-4.28	-0.58	0.12	.
Intercept_Rural	-2.49	2.21	-6.06	-5.35	0.42	1.15	
AADT_Urban two-	0.57	0.05	0.49	0.51	0.63	0.65	*
AADT_Urban	0.77	0.18	0.50	0.55	1.00	1.08	*
AADT_Urban multi-	0.47	0.26	0.04	0.14	0.81	0.89	*
AADT_Rural two-	0.66	0.02	0.64	0.65	0.68	0.69	*
AADT_Rural	0.65	0.10	0.47	0.50	0.77	0.79	*
AADT_Rural multi-	0.61	0.22	0.24	0.32	0.89	0.96	*
Length_Urban two-	0.91	0.10	0.75	0.78	1.05	1.09	*
Length_Urban	0.97	0.18	0.68	0.74	1.21	1.28	*
Length_Urban	1.25	0.24	0.86	0.94	1.56	1.66	*
Length_Rural two-	0.88	0.06	0.78	0.81	0.96	0.99	*
Length_Rural	1.23	0.21	0.90	0.97	1.50	1.57	*
Length_Rural multi-	1.32	0.44	0.65	0.78	1.89	2.07	*
Females_age	-0.01	0.03	-0.07	-0.06	0.01	0.02	
Females_age 17	-0.06	0.07	-0.16	-0.14	0.04	0.06	
Females_age 24	-0.02	0.02	-0.07	-0.06	0.01	0.01	
Females_age 34	-0.03	0.03	-0.08	-0.07	0.01	0.02	
Females_age 59	0.03	0.05	-0.04	-0.03	0.10	0.11	
Females_age 69	-0.02	0.05	-0.10	-0.08	0.04	0.06	
Females_age	0.01	0.02	-0.01	-0.01	0.04	0.05	
Income	-0.13	0.09	-0.30	-0.25	-0.02	0.01	.
Males_age under15	0.01	0.02	-0.03	-0.02	0.04	0.05	
Males_age 17	0.00	0.05	-0.09	-0.07	0.06	0.08	
Males_age 24	0.01	0.03	-0.03	-0.02	0.05	0.05	
Males_age 34	-0.01	0.03	-0.05	-0.04	0.03	0.04	
Males_age 59	-0.03	0.04	-0.09	-0.08	0.02	0.03	
Males_age 69	0.01	0.04	-0.05	-0.04	0.07	0.08	
Males_age above70	-0.04	0.03	-0.09	-0.08	0.00	0.01	

**Table C-3 District 2-0 Model (continued).**

			Percentile				
	<b>Mean</b>	<b>SD</b>	<b>5%</b>	<b>10%</b>	<b>90%</b>	<b>95%</b>	<b>Sig</b>
Population	-0.06	0.09	-0.20	-0.18	0.06	0.08	
Population density	31.16	28.25	-	-5.02	67.31	77.75	
Poverty proportion	0.01	0.01	-0.01	0.00	0.02	0.02	.
sd.u	0.25	0.02	0.22	0.22	0.29	0.30	
sd.v	0.56	0.02	0.53	0.54	0.58	0.59	
sigma2.u	0.07	0.02	0.04	0.05	0.09	0.10	
sigma2.v	0.32	0.02	0.28	0.29	0.34	0.35	
tau.u	15.80	3.81	10.46	11.36	20.85	22.76	
tau.v	3.17	0.20	2.86	2.92	3.43	3.51	

Dbar    Dhat    DIC    pD  
 34933   33232   36634   1701

For the rural District 2-0, we see significance for intercepts, ADT and length for each of the facility types. Many fewer socio-demographic variables are significant. Poverty proportion and income both increase crash risk with reductions in socio-demographic status. These results are similar to those found at the county level in an earlier paper.

### Models of Specific Crash Types

One additional aspect of Task 3 was the development of SPFs for specific crash types. These are summarized for single and multi-vehicle run-off-road crashes in Tables C-4 and C-5, respectively. These models were derived by searching for run-off-road events as a crash type within the PennDOT crash data. Specifically, for the functional class of two-lane rural roads, the Penn State team included data from all single-vehicle, hit-fixed-object crashes that occurred off roadway (right and left). Multi-vehicle, head-on/sideswipe crashes are of interest because they are typically high severity. The relation to roadway variable necessary to identify this collision type would be “on travelway” and/or “in median.”

Once these data searches were complete, the SPFs were developed.

**Table C-4. Single Vehicle Run-off-road Crash SPF**

	<b>Mean</b>	<b>SD</b>	<b>MC_error</b>	<b>2.5%</b>	<b>97.5%</b>
Intercept	-5.6917	0.0469	7.54E-04	-5.7842	-5.6001
AADT	0.6101	0.0058	9.21E-05	0.5987	0.6215
Length	1.0341	0.0240	3.63E-04	0.9870	1.0814
sd(v)	0.7390	0.0081	2.92E-04	0.7232	0.7550
$\sigma^2v$	0.5461	0.0124	4.45E-04	0.5222	0.5709

Dbar	Dhat	DIC	pD
178282	166706	189858	11576.1

934 Significant excess crash risk segments

**Table C-5. Multi-Vehicle Head on and Sideswipe Crash SPF**

	<b>Mean</b>	<b>SD</b>	<b>MC_error</b>	<b>2.5%</b>	<b>97.5%</b>
Intercept	-10.1091	0.1224	0.00204	-10.3528	-9.8720
AADT	0.8983	0.0145	0.00017	0.8702	0.9269
Length	0.8862	0.0553	0.00066	0.7784	0.9945
sd(v)	0.7379	0.0265	0.00196	0.6842	0.7865
$\sigma^2v$	0.5453	0.0392	0.00289	0.4679	0.6191

Dbar      Dhat      DIC      pD  
 39009.8    36949.3    41070.3    2060.48  
 44 Significant excess crash risk segments

Tables C-4 and C-5 illustrate models of single-vehicle run-off-road and multi-vehicle run-off-road crash events. Note that 934 and 44 excess crash segments were identified; fewer segments are indicated for multi-vehicle crashes because they are more rare and have a higher variance, and thus fewer are significantly different from the mean for any specific ADT level.

**References**

Agüero, J., and P. P. Jovanis, “Spatial Analysis of Fatal and Injury Crashes in Pennsylvania,” *Accident Analysis and Prevention*, Volume 38, Issue 3, pp. 618-625, May 2006.  
 Browne, W. J., and J. Rasbash (2004), “Multilevel Modelling.” In Bryman, A., and Hardy, M. (Eds.), *Handbook of Data Analysis*, pp. 459-479. London: Sage Publications.

## **End Note: Details of Bayes Estimation**

Models were estimated using the open source software Open BUGS. For the models, 3,000 iterations were discarded as burn-in; an additional 5,000 iterations were used to obtain summary statistics of the posterior distribution of parameters. Convergence was assessed by visual inspection of the Monte Carlo Markov chains. Furthermore, the number of iterations was selected such that the Monte Carlo error for each parameter in the model would be less than 15 percent of the value of the standard deviation of that parameter.

The hierarchical modeling structure (Full Bayes) produces what are called 5 percent and 95 percent credible set estimates instead of the confidence intervals normally produced in frequentist estimation. Parameters that have 5 percent to 95 percent credible set values that do not include 0 are generally accepted as “significant.” The last column contains a “\*” for a significant.