

Spatial analysis of fatal and injury crashes in Pennsylvania

Jonathan Agüero-Valverde^a, Paul P. Jovanis^a

^a Department of Civil and Environmental Engineering and Pennsylvania Transportation Institute, Pennsylvania State University, 212 Sackett Building, University Park PA 16802-1408

Abstract

Using injury and fatal crash data for Pennsylvania for 1996-2000, Full Bayes Hierarchical models (with spatial and temporal effects and space-time interactions) are compared to traditional Negative Binomial estimates of annual county-level crash frequency. Covariates include socio-demographics, weather conditions, transportation infrastructure and amount of travel. Full Bayes Hierarchical models are generally consistent with the Negative Binomial estimates. Counties with a higher percentage of the population under poverty level, higher percentage of their population in age groups 0 to 14, 15 to 24, and over 64 and those with increased road mileage and road density have significantly increased crash risk. Total precipitation is significant and positive in the Negative Binomial models, but not significant with Full Bayes. Spatial correlation, time trend, and space-time interactions are significant in the Full Bayes injury crash models. County-level Full Bayes models reveal the existence of spatial correlation in crash data and provide a mechanism to quantify, and reduce the effect of, this correlation. Addressing spatial correlation is likely to be even more important in road segment and intersection-level crash models, where spatial correlation is likely to be even more pronounced.

Keywords

Full Bayes Hierarchical model, spatial correlation, Negative Binomial model, crash risk, weather conditions and crash risk.

1 Introduction

Many factors affecting crashes operate at a spatial scale (e.g. land use policy, demographic characteristics and highway infrastructure functional class). It is therefore reasonable to explore the use of spatial models of crash occurrence to better understand the implications of these policies.

In most roadway accident studies, crashes are grouped in spatial units that range from intersection or road section level to zip code or county level (e.g. Amoros et al., 2003; Miaou et al., 2003; Noland and Oh, 2004; Noland and Quddus, 2004; and MacNab, 2004). One concern with these studies is the effect of spatial correlation (i.e. the spatial dependence among observations) which produces higher variance of the estimates and therefore, underestimated standard errors.

Recent developments in spatial modeling techniques have enabled researchers to investigate important issues related to risk estimation, unmeasured confounding variables, and spatial dependence (Richardson, 1992). An important advantage of spatial models is that spatial effects may reflect unmeasured confounding variables. This is particularly useful for unmeasured confounders that vary in space like weather, population, and others. More important yet, “The methods also facilitate spatial smoothing and data pooling when regions under investigation involve small-population areas”, MacNab, 2004. Here the term ‘small-population areas’ refers to areas that present very few events, given a rare-event phenomenon, for example roadway crashes.

Previous research has dealt with the spatial component of road crashes in different ways. Crashes have been modeled as point events (e.g. Levine et al., 1995 and Jones et al., 1996), while others have modeled road crashes at different area levels, ranging from road sections to local census tracts or counties (e.g. Shankar et al., 1995; Amoros et al., 2003; Miaou et al., 2003; Noland and Oh, 2004; and MacNab, 2004).

Honolulu census tract data have been used (Levine et al., 1995b) in a continuous model for predicting crashes. Analysis at the “ward” (census tract) level has been conducted (Noland and Qudus, 2004) for fatalities, serious injuries, and slight injuries using 4 different categories of predictor variables: land use indicator variables (employment and population density), road characteristics, demographic characteristics (age cohorts), and traffic flow proxies (proximate and total employment). Country-level data for Illinois (Noland and Oh, 2004) were used to estimate the expected number of crashes using infrastructure characteristics and demographic indicators as independent variables in a negative binomial model. Limitations of these studies are the use of proxy variables for traffic flow estimation, and the lack of spatial correlation analysis. An additional paper (Amoros et al., 2003) developed negative binomial models at county level in France that included interactions between road type and county.

Poisson-based Full Bayes Hierarchical models of county-level fatal (K), incapacitating (A), and non-incapacitating (B) injuries were estimated using both frequency and rate for the state of Texas (Miaou et al., 2003). Conditional Auto-Regressive model (CAR) was used to model spatial correlation and Markov Chain Monte Carlo (MCMC) was used to sample the posterior probability distribution. The main limitation of this paper is the use of the surrogate variables: percent of time that the road is wet, sharp horizontal curves, and road-side hazards. These predictor variables were estimated by proportions of crashes. For example, for percent of

time that the road is wet, the variable was estimated by dividing the number of crashes that occurred under wet pavement by the total number of crashes. These estimators are clearly biased in the direction of the effect. Given the poor definition of contributing factors in the model, it is likely that the spatial correlation is overestimated. In a recent paper, Miaou and Song (2005) used the same approach and data in the ranking of sites for engineering safety improvements.

The adoption of the Full Bayes Hierarchical approach by Miaou is an important advance in model estimation and is a departure point for this paper. The purpose of this research is to develop spatial models of road crash frequency for the State of Pennsylvania at the county level while controlling for socioeconomic, transportation- related, and environmental factors. The results from Full Bayes Hierarchical spatial models are compared with the more traditional approach using a Negative Binomial distribution to model crash frequency. Particular attention is paid to the inclusion of weather as a predictor and the search for spatial correlation among neighboring counties.

2 Methodology

2.1 The Poisson and Negative Binomial distributions

When data arise as counts, the Poisson distribution is typically used to model them. Traffic crashes are a clear example of count data; therefore, a Poisson distribution is a useful starting point (see for example Jovanis and Chang, 1986; Shankar et al., 1995). An important characteristic of the Poisson distribution is that its variance is equal to its mean. Several authors (e.g. Shankar et al., 1995 and Noland and Quddus, 2004) have argued that vehicle crashes are

better represented by a Negative Binomial distribution, which is a count distribution generated by a Poisson process with variance greater than the mean (see for example, Hamed and Jaradat, 1998; and Hamed, 1999).

Several goodness-of-fit measures have been proposed for this kind of model including the Poisson R^2 , R_P^2 , and the Freeman-Tukey R^2 , R_{PFT}^2 (Fristrøm et al., 1995). Another measure of goodness-of-fit uses the overdispersion parameter α of the negative binomial model (Miaou, 1996). Negative Binomial models are estimated using R statistical Software (R Development Core Team, 2004)

2.2 Spatial Modelling using Full Bayes Hierarchical approach

Many spatial modelling techniques may be developed within a bayesian approach, because of its flexibility in structuring complicated models, inferential goals, and analysis (Miaou et al, 2003). Bayesian inference has been used in the past in disease mapping and ecological analysis and just recently it has been applied to crash modeling (e.g. Miaou et al., 2003; MacNab, 2004; and Miaou and Song, 2005). For a detail description of Bayesian inference, see Gelman et al. (2003).

The problem of group estimation, namely estimating the parameters of a common distribution thought to underlay a collection of outcomes for similar types of units, has motivated much research in bayesian statistics. One seeks to make conditional estimates of the true outcome rate in each unit of observation (e.g. fatal crashes rate by county), given the parameters of the common density. Such estimation for sets of similar units is known as 'hierarchical modelling' because of its conditioning on higher stage densities (Congdon, 2003). At the first

stage, the observed counts are modeled as a function of area-level summaries such as risk or relative risk. At the second stage, a joint distribution is specified for the collection of these risks as a function of explanatory variables. The second stage distribution depends on unknown parameters and these are assigned a (hyper) prior distribution at the third stage (Wakefield et al., 2000).

In the case of this study, the model developed by Besag et al. (1991) is the base of the formulation used, as shown in Equations 1 and 2:

$$y_i \sim \text{Poisson}(e_i \theta_i) \quad (1)$$

where y_i is the number of Fatal Crashes in county i , θ_i is the risk in county i and e_i is the exposure in county i ; in this case the exposure is the total Daily Vehicle-Miles Traveled (DVMT) by county. DVMT was also included as explanatory variable to account for possible non-linearity between crash frequency and DVMT. This is the first stage in the model. The log risk is modeled as:

$$\log(\theta_i) = \alpha + \mathbf{x}'_i \boldsymbol{\beta} + v_i + u_i \quad (2)$$

where \mathbf{x}_i represents a vector of explanatory variables, or covariates, $\boldsymbol{\beta}$ is a vector of fixed effect parameters, v_i is the uncorrelated heterogeneity or unstructured error and u_i is the correlated heterogeneity or spatial correlation. The last two variables are known as random effects, therefore, this kind of model is commonly known as a Mixture Model where fixed and random effects are combined.

In the third stage, a uniform prior distribution is assigned to α and a highly non-informative normal distribution is assigned to the $\boldsymbol{\beta}$'s with mean 0 and variance 1000 corresponding to vague prior beliefs, given the scale of the covariates. On the other hand the

prior distribution for the uncorrelated heterogeneity (v_i) is $N(0, \tau_v^2)$, where τ_v^2 is the precision (precision = 1/variance) with a prior gamma distribution $Ga(0.5, 0.0005)$ as suggested by Wakefield et al. (2000). This prior distribution, as in case of the β 's, represents a highly non-informative distribution. The definition of the precision corresponds to the fourth stage of the model hierarchy.

2.3 Conditional Autoregressive Model

A pair of areas are considered neighbors if they are adjacent; u_i is the correlated heterogeneity or spatial correlation term reflecting a shared border. For the spatial correlation term, the conditional autoregressive model (CAR) proposed by Besag (1974) is used

$$[u_i | u_j, i \neq j, \tau_u^2] \sim N(\bar{u}_i, \tau_i^2) \quad (3)$$

where

$$\bar{u}_i = \frac{I}{\sum_j w_{ij}} \sum_j u_j w_{ij} \quad (4)$$

and

$$\tau_i^2 = \frac{\tau_u^2}{\sum_j w_{ij}} \quad (5)$$

$w_{ij} = 1$ if i, j are adjacent or 0 otherwise.

As in the unstructured case, the τ_u^2 controls the variability of u and its prior was selected $Ga(0.5, 0.0005)$ as suggested by Wakefield et al. (2000).

2.4 Space-Time Models

Several authors (e.g. Miaou et al., 2003; Knorr-Held and Besag, 1998; Waller et al., 1997; Knorr-Held, 2000; and Bernardinelli et al., 1995) have studied the issue of changes in disease risk over time and space, however, only the work from Bernardinelli et al. presents a clear specification of the time-space interaction term. Extending the model of Bernardinelli to include the effect of covariates:

$$y_{ij} \sim \text{Poisson}(e_{ij}\theta_{ij}) \quad (6)$$

$$\log(\theta_{ij}) = \alpha + \sum_k \beta_k x_{ijk} + v_i + u_i + (\varphi + \delta_i)t_j \quad (7)$$

where y_{ij} is the observed number of crashes for the i th area, $i=1, \dots, N$, and the j th time interval, $j=1, \dots, T$, α is the constant term, x_{ijk} is the k th covariate for the i th area, in the j th time interval, β_k are the regression coefficients, v_i is the uncorrelated heterogeneity, u_i is the spatial correlation, φ is the mean linear time-trend over all areas, t_j is the time interval j , and δ_i is the interaction between time effect and area effect.

This model specification allows for spatiotemporal interactions where temporal trend in crash risk may be different for different spatial locations and may even have spatial structure. However, all temporal trends, including the main temporal trend φ , are assumed to be linear, which may be seen as a restrictive assumption (Lawson et al., 2003).

For comparison of the different bayesian hierarchical models the Deviance Information Criterion (DIC), proposed by Spiegelhalter et al. (2002) is used. It compares the fit and complexity of hierarchical models in which the number of parameters are not clearly defined.

DIC is based on the posterior distribution of the deviance statistic and can be seen as a generalization of the Akaike Information Criterion (AIC).

DIC is defined as an estimate of fit plus twice the effective number of parameters as in Equation 8:

$$DIC = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \quad (8)$$

where $D(\bar{\theta})$ is the deviance evaluated at $\bar{\theta}$, the posterior means of the parameters of interest, p_D is the effective number of parameters in the model, and \bar{D} is the posterior mean of the deviance statistic $D(\theta)$. As with AIC, models with lower DIC values are preferred.

For each Bayesian Hierarchical model two Monte Carlo Markov Chains from different starting points were estimated in order to assure convergence. Depending on the complexity of the models, the first 3000 to 5000 iterations in each chain were removed as burn-ins. Finally, other 5000 iterations were performed for each chain resulting in a sample distribution of 10000 for each parameter. Convergence of the models was monitored by visual examination of MCMC chains, autocorrelation plots, and Gelman-Rubin statistic plots. Full Bayes Hierarchical models are estimated using WinBUGS software (Spiegelhalter et al, 2004)

3 Data Description

Fatal and injury crash data are obtained from PennDOT (Bureau of Highway Safety and Traffic Engineering, 1997-2001). Two different sets of crash models were estimated; those using fatal crashes only as the dependent variable and those using just injury crashes. The spatial distribution of crashes has a very particular form. Figure 1 shows the number of fatal crashes by

county in year 2000 (similar distributions were obtained for other years). It is clear that the highest values are clustered around the cities of Philadelphia in the southeast portion of the state and Pittsburg in the southwest. The fatal crash rate map, Figure 2, presents a very different pattern as high crash rates occur mainly in counties with low to medium number of crashes. This may be due to the fact that fatal crashes are rare events, the counts are very low and therefore, a small increase of say, two or three crashes will significantly increase the rate especially for those counties with low levels of travel (i.e. DVMT).

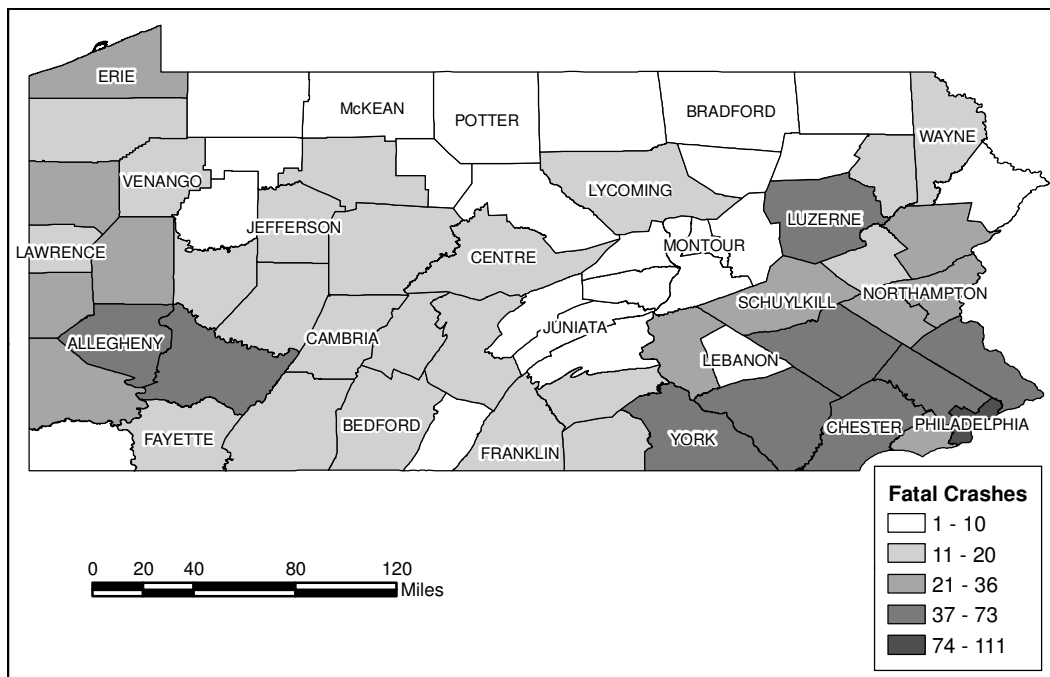


Figure 1 Number of Fatal Crashes by county for the year 2000.

Risk factors, all measured at the county level and obtained for the years 1996-2000, are divided into three main categories: socioeconomic, transportation infrastructure related, and environmental factors. Table 1 summarizes the definitions of each of the variables used to characterize the factors and contains descriptive statistics. Several different variable definitions were considered in exploring alternative model specifications.

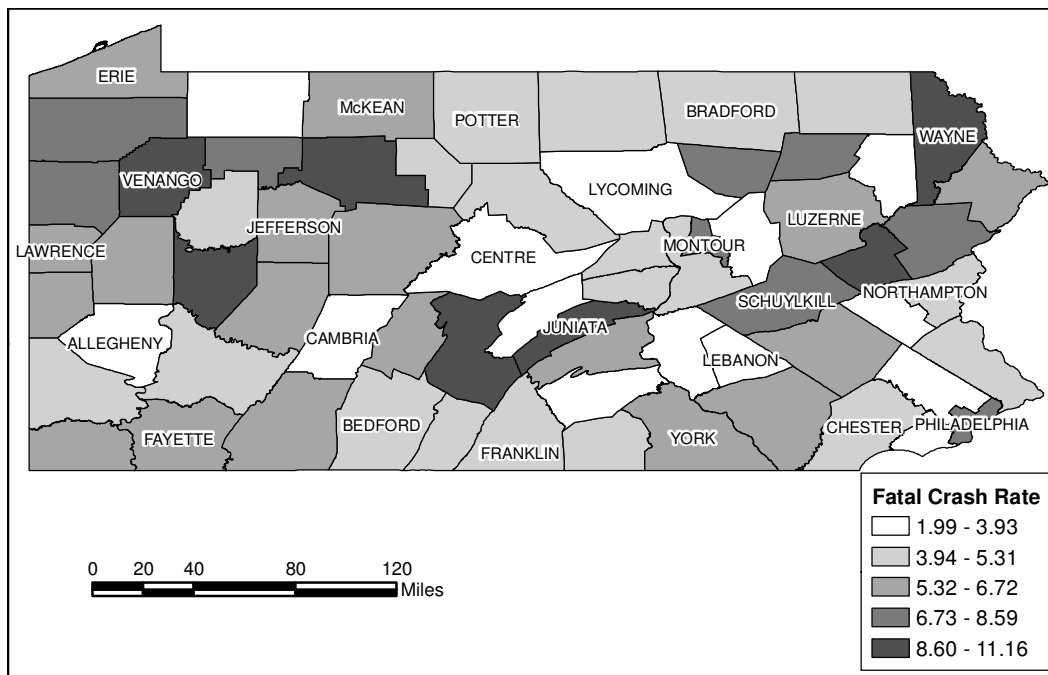


Figure 2 Fatal Crash Rate by county for the year 2000 (crashes by million DVMT)

Socioeconomic factors include county-level summaries of age, sex, level of poverty (all from U.S. Census Bureau, 2004a-2004c) and level of drunk driving (Pennsylvania Uniform Crime Reporting System, 2004). Shinar (1978), Evans (1991 and 2004), and Kam (2003) have suggested factors including age, sex, and personality to explain crash risk. Age is characterized as the percentage of persons younger than 15 (proxy for young pedestrians), between 15 and 24 (proxy for young drivers), and percent of persons over 65 (proxy for senior drivers) with the percentage of population between 25 and 65 used as a baseline. Note, these variables correspond to the total population rather than the population of drivers, but these are the only data publicly available. Area deprivation has been found to be positively related to crash occurrence (e.g. Chichester et al., 1998; Abdalla et al., 1997; and Noland and Quddus, 2004) and the percentage of persons under poverty is used as a surrogate of area deprivation in this paper. In the case of drunk driving, authors like Evans (1991) and Gary et al. (2003) have found association between

alcohol consumption and crash risk. The number of DUI arrests in a county is used to relate alcohol consumption to driving and crash frequency (Pennsylvania Uniform Crime Reporting System, 2004).

Table 1 Summary statistics of variables included in the models.

Type	Variables	Min.	Median	Mean	Max.	Stand. Dev.
Crashes	Fatal Crash Frequency	0	14	20.59	137	19.94
	Injury Crash Frequency	33	600	1328.85	12370	1951.17
	Fatal & Injury Crash Frequency	35	615	1349.44	12489	1969.85
	Fatal Crash Rate (by million DVMT)	0	5.5	5.84	16.51	2.39
	Injury Crash Rate (by million DVMT)	136.64	254.75	270.83	813.46	90.74
	Fatal & Injury Crash Rate (by million DVMT)	142.06	261.45	276.67	821.28	90.84
Socioeconomic Factors	Pop. under 15 years (%)	14.87	19.69	19.51	22.64	1.62
	Pop. between 15 & 24 (%)	9.36	12.2	13.04	30.04	2.87
	Pop. 65 years & older (%)	9.86	15.97	16.2	21.97	2.29
	Males (%)	46.46	48.74	48.95	55.71	1.3
	Pop. Under Poverty (%)	4.4	10.8	10.66	23.8	3.11
	DUI Arrests Frequency	18	314	581.29	4231	750.89
	DUI Rate (by 1000 pop)	0.87	3.38	3.55	12.7	1.48
Transportation-Related Factors	DVMT (millions)	0.16	2.33	4.09	25.63	4.46
	Federal Aid DVMT (%)	42.65	81.93	79.34	89.12	8.37
	Miles of Road (Thousands)	0.31	1.68	1.78	5.67	0.95
	Federal Aid Miles of Road (%)	13.78	21.98	22.59	36.66	4.18
	Miles of Road Density (mi/mi ²)	0.78	2.45	2.96	16.83	2.28
Environment-related Factors	Mean Total Precipitation (in)	32.14	42.11	44.57	68.19	7.64
	Mean Number of Rainy Days	20.44	27.69	29.4	43.32	5.07
	Mean Total Snowfall (in)	2.44	35.76	40.5	127.09	24.42
	Mean Number of Days with Snow	2.49	36.16	41.96	114.4	26.15

Transportation-related factors included daily vehicle miles traveled and amount of infrastructure (PennDOT, 2004). Vehicle-Miles Traveled is often used as indication of exposure to crash risk (Miaou et al., 2003; Fristrøm et al., 1995; Jovanis and Chang, 1986) along with the number of miles of different functional classes per county (Noland and Quddus, 2004). The number of miles of infrastructure in each county was included in the model to estimate the effect of different levels of transportation infrastructure supply in the expected crash rate. The

functional categories were aggregated in two groups: Federal Aid (i.e. Interstate, other Freeways or Expressways, other Principal Arterials, Minor Arterials and Major Collectors), and Non-Federal Aid Roads (i.e. Minor Collectors and Local Roads). Even though the amount of travel (DVMT) was included into the model as an offset term, the proportion of travel occurring on Federal Aid roads was included in the models as a separate variable to account for the possible differences in design and operation levels between federal aid and non-federal aid roads where the former are expected to have higher levels.

Environmental factors included descriptors for level of rain and snow (National Climatic Data Center, 2004). Environmental factors have been investigated in the past by Jovanis and Chang (1986), Shankar et al. (1995), Fristrøm et al. (1995), and Edwards (1996) among others. The studies found positive correlation between weather and crash frequency. The amount of precipitation, snowfall, and number of rainy days and days with snow are analyzed in this study. Hundreds of weather stations are used to generate estimated surfaces for each variable and then the variables are summarized at county level for inclusion in the database. Spatial trend and correlation of weather variables with elevation are included into the models to improve their predictive power and fit. For more details on estimation of weather variables see Aguero-Valverde, 2005.

4 Results

A series of Negative Binomial regression models are used in initial data analysis and to provide a comparison set for Full Bayes (FB) models to follow. In each model, all variables listed in Table 1 are the starting point; variables are removed if they have significance levels

above 0.10. Table 2 presents the Negative Binomial (NB) model of fatal crashes. For this model three transportation related variables are significant: DVMT, infrastructure mileage, and Percentage of travel on federal-aid roads. The coefficient for DVMT is negative which implies that the fatal crash rate decrease with DVMT or, in other words, the fatal crash frequency increase with DVMT at a decreasing rate. Total mileage is significant and positively correlated with fatal crash frequency while the percentage of travel in federal-aid roads is negatively correlated. Higher design standards for federal-aid roads may be responsible for this reduction on fatal crash risk.

Table 2 Negative Binomial model of Fatal Crashes.

VARIABLE	Estimate	Std. Error	z value	Pr(> z)
DVMT	-0.033	0.005	-6.46	1.08E-10
INFRASTRUCTURE MILEAGE	0.060	0.026	2.26	0.024
DVMT ON FED. AID ROADS (%)	-0.006	0.003	-2.30	0.021
PERSONS UNDER POVERTY (%)	0.039	0.004	9.22	2.0E-16
PERSONS 0 -14 (%)	0.051	0.009	5.52	3.4E-08
TOTAL PRECIPITATION	0.014	0.004	3.74	0.000
1997	0.290	0.083	3.49	0.000
1998	0.238	0.077	3.11	0.002
1999	0.236	0.070	3.38	0.001
2000	0.274	0.071	3.86	0.000
Dispersion Parameter α	0.021	0.006	3.67	0.000

Residual deviance: 351.71 on 325 d.f.

AIC: 1980.6

2 x log-likelihood: -1958.611

$R_p^2 = 0.962$

$R_{FT}^2 = 0.967$

$R_K^2 = 0.968$

Several socioeconomic variables are significant in the model. The variable representing area deprivation, percent of population under poverty, as well as persons less than 15 years are highly significant and positive. These findings are consistent with the expectation of increases in crash risk with increases in area deprivation and number of young persons (under 15).

Within the environmental variables, the only one that is significant is total precipitation. Several facts may contribute to this, among them that total precipitation and rainy days as well as total snowfall and snowy days are highly correlated (0.93 and 0.88 respectively).

Fixed time effects are also significant and positive, which means that the expected number of fatal crashes increased for years 1997 to 2000 compared to the base year of 1996. Overall goodness of fit for the model is excellent for all utilized fit statistics. The dispersion parameter is positive and highly significant, indicating the presence of overdispersion which is common with models estimating the expected number of crashes.

For the negative binomial injury crash models (Table 3) the original set of independent variables form the starting point. Fatal and injury models differ in the number of significant predictors, 6 for the fatal crash model and 8 for the injury model without including the fixed time effects that are significant in both models.

For the injury model, four transportation-related variables are significant: DVMT, infrastructure mileage, mileage density, and percentage of federal-aid roads. As in the fatal crash model, DVMT is negative suggesting an increasing risk of injury crashes at a decreasing rate. The next two variables represent the amount of infrastructure in absolute and relative terms, respectively, and are both positively correlated with injury crashes, as expected. The third variable measures the percentage of roads under federal aid and the positive coefficient indicates that increases in the variable value increase the injury crash risk. Federal aid roads in our study have generally higher speeds; non-federal aid roads have low operational speeds which may generally reduce the probability of injury or fatality.

In the case of socioeconomic variables, percentage of persons under 15 is significant and with the same sign as for Fatal crashes. In addition, the surrogates for young drivers and

pedestrians (persons between 15 and 24) and elderly drivers and pedestrians (persons 65 and over) are significant while in the Fatal Crash model those variables are not found to be significant.

Table 3 Negative Binomial model of Injury Crashes.

VARIABLE	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.257	0.498	4.53	5.96E-06
DVMT	-0.031	0.007	-4.36	1.29E-05
INFRASTRUCTURE MILEAGE	0.147	0.024	6.14	8.08E-10
MILEAGE DENSITY	0.089	0.008	11.66	2.00E-16
MILEAGE OF FED. AID ROADS (%)	0.016	0.003	5.38	7.35E-08
PERSONS 0 -14 (%)	0.060	0.013	4.61	4.02E-06
PERSONS 15 -24 (%)	0.024	0.007	3.55	0.000
PERSONS 65 AND OVER (%)	0.030	0.009	3.40	0.001
TOTAL PRECIPITATION	0.010	0.003	3.80	0.000
1997	0.204	0.060	3.41	0.001
1998	0.136	0.055	2.46	0.014
1999	0.107	0.053	2.00	0.046
2000	0.111	0.054	2.05	0.040
Dispersion Parameter α	0.030	0.003	11.93	0.000

Residual deviance: 338.78 on 322 d.f.

AIC: 4191.1

2 x log-likelihood: -4163.099

$R_p^2 = 0.981$

$R_{FT}^2 = 0.983$

$R_K^2 = 0.974$

The Total Precipitation effect is also positive in the injury model. The fixed time effects are positive and significant for injury crash models. Their coefficient values, however, decrease from 0.20 in 1997 to 0.11 in 2000.

4.1 Full Bayes Hierarchical Models

Several Full Bayes (FB) fatal crash models are tested. The basic FB fatal crash model includes the variables found to be significant in the Negative Binomial models as well as fixed time effects and uncorrelated heterogeneity. Fixed time effects are found not significant and as a result, they are excluded from further models. When a model including unstructured heterogeneity and spatial correlation is estimated, the overdispersion of the model is due to unstructured heterogeneity and therefore fatal crashes do not indicate significant spatial correlation (for more see Aguero-Valverde, 2005).

Finally, a FB model with uncorrelated heterogeneity and without fixed time effects is estimated (see Table 4). DVMT, Poverty, and persons under 15 are significant and very similar to the coefficients from the NB model. This FB model presents the lowest value of DIC; which means that, among the models estimated in this work, this is the model that best fits the data. Given the fact that there is no evidence of spatial correlation or time dependence in the Fatal Crash model, no further models accounting for space-time interactions are estimated.

Table 4 Full Bayes model of Fatal Crashes with uncorrelated heterogeneity (v_i)

VARIABLE	Estimate		95% Credible set	
	Mean	SD	2.50%	97.50%
DVMT	-0.0397	0.0104	-0.0600	-0.0197
INFRASTRUCTURE MILEAGE	0.0466	0.0496	-0.0504	0.1439
DVMT ON FED. AID ROADS (%)	0.0030	0.0033	-0.0034	0.0096
PERSONS UNDER POVERTY (%)	0.0273	0.0076	0.0119	0.0420
PERSONS 0 -14 (%)	0.0662	0.0128	0.0414	0.0913
TOTAL PRECIPITATION	-0.0006	0.0016	-0.0037	0.0026
σ_v^2	0.0337	0.0091	0.0190	0.0545
	\bar{D}	$D(\bar{\theta})$	DIC	pD
	1842.0	1794.0	1890.0	48.32

A FB Injury Crash model including fixed time effects and uncorrelated heterogeneity or random error was tested next, then a model that replaces the unstructured error term in each country with a CAR error term, or spatial correlation. From these models it was evident that temporal and spatial effects were significant and therefore, the space-time model shown in equation 7 is estimated (without including the unstructured error term). The model includes a spatially correlated term (u_i), a time trend effect (ϕ), and an interaction term between time effect and area effect (δ_i) as presented in Table 5.

Table 5 Full Bayes model of Injury Crashes with spatial correlation (u_i), time trend (ϕ), and space x time interactions (δ_i)

VARIABLE	Estimate		95% Credible set	
	Mean	SD	2.50%	97.50%
(Intercept)	3.110	0.764	1.593	4.582
DVMT	-0.092	0.013	-0.118	-0.067
INFRASTRUCTURE MILEAGE	0.323	0.051	0.227	0.425
MILEAGE DENSITY	0.131	0.022	0.086	0.174
MILEAGE OF FED. AID ROADS (%)	0.026	0.007	0.013	0.040
PERSONS 0 -14 (%)	0.047	0.020	0.008	0.088
PERSONS 15 -24 (%)	0.015	0.012	-0.008	0.039
PERSONS 65 AND OVER (%)	0.013	0.016	-0.018	0.044
TOTAL PRECIPITATION	5.86E-05	2.27E-04	-3.85E-04	5.01E-04
σ_u^2	0.176	0.038	0.115	0.263
ϕ	-0.015	0.004	-0.022	-0.008
σ_δ^2	6.66E-04	2.15E-04	3.31E-04	0.001164

\bar{D}	$D(\bar{\theta})$	DIC	pD
3261.0	3159.0	3363.0	102.1

From the estimates of the coefficients in Table 5, there are several important points to highlight. First, all transportation-related coefficients are significant and fairly close to the values in the NB model. In the case of age cohorts, persons-under-15 is significant, positive, and very similar to the NB estimate. The coefficients of the other two cohorts are not significant. As in the FB Fatal model, the precipitation effect is non-statistically significant for the Injury model.

The time trend effect is significant and negative for the FB injury crash model. This means that controlling for all the other variables in the model, the number of injury crashes decrease over time. This is consistent with previous FB models for injury crashes (not presented here); but more important yet, if one examines the NB model for injury crashes, it is evident that the coefficients for the fixed time effects generally decrease with time after a positive ‘jump’ from 1996 (base year) to 1997. Note that the coefficient estimate for the time trend may be biased if the underlying time trend for the population is not linear; however, with only 5 years of data this is very difficult to assess.

The Poisson extra-variation due to spatial correlation (u_i) is significantly greater than the variation due to the space-time interaction. The variance of u (σ_u^2) is equal to 0.176 while the variance of δ (σ_δ^2) is equal to 0.00066. However, the fact that the variance of δ is different from zero indicates that the space-time interaction term is significant.

5 Conclusions

There is no evidence of spatial correlation in fatal crashes; however, spatial correlation was found to be significant in injury crashes. The variance of the spatially correlated term (σ_u^2) is significant in the Full Bayes injury crash model, which implies that some Poisson extra-variation in the data can be explained by spatial correlation.

Results concerning the effects of the covariates on fatal and injury crash risk are mostly consistent in the direction and magnitude for Negative Binomial and Full Bayes models. In general, highly significant variables in the NB models are also significant in the FB models. On

the other hand, variables just marginally significant in the NB models are generally non-significant in the FB models. Because the FB models address spatial correlation and take into consideration all sources of uncertainty, the authors believe the FB models more accurately associate covariates with crash risk and are better suited for this type of data.

Fatal Crash models have fewer significant variables compared to Injury models. A main reason for this is the small number of events for fatal crashes, roughly a fatal crash for every thousand injury crashes. Given this difference of three orders of magnitude, it is expected that the random variation of the fatal crash data to be small compared to the variation explained by the covariates in the injury models. This may be also a sample size issue; with a larger sample size the noise or random variation is likely to be reduced and therefore, a higher portion of the variation can be explained by the covariates.

6 Recommendations for Future Research

Crash models at the county-level have several advantages over other types of crash models; one of the most important ones is the availability of transportation and socioeconomic data at county level. With the increase in quality and quantity of Geographic Information Systems (GIS) data available in the Internet, public organizations may be able to incorporate additional land-use covariates in crash models, exploring land use-transportation interactions and their effects on crash risk.

Now that data on poverty and other socioeconomic variables are available for the years 2001 and 2002 the database can be increased by 40%. This will have two important benefits: first, more data will improve the explanatory power of the models especially in the fatal crash

models; and secondly, two more years of data will help to estimate more precisely the time effect for both fatal and injury models.

Given the existence of spatial correlation, at least for injury models, it is expected that spatial correlation plays a more important role at smaller spatial scales. Full Bayes models with spatial correlation may be even more useful at road section level where units are smaller and closer and therefore, the probability of spatial correlation is higher. Using Full Bayes models with spatial correlation is a natural next step in research work.

Area deprivation (in the form of poverty percent in the case of this work) is consistently correlated with crash risk. More research on the causes of this correlation and on the causal-effect relationship between these phenomena is needed. The relationship likely requires some level of spatial aggregation (e.g. county or census tract), so analyses at the segment level are unlikely. This limitation implies that more studies at county or census tract level are necessary in order to shed some light on the issue.

From the methodological standpoint, Full Bayes Hierarchical models are more effectively developed by starting with simple specifications, building to more complex models. The output from the simpler models can be used as initial values in the Monte Carlo Markov Chains (MCMC). Two or more MCMCs with different initial values for all the parameters are recommended in order to assure convergence to a consistent value. This method allows the modeling to verify output from each model and generally be in greater control of the FB model building process.

7 References

- Abdalla, I.M., Robert, R., Derek, B., McGuicagan, D.R.D., 1997. An investigation into the relationships between area social characteristics and road accident casualties. *Accident Analysis and Prevention*, Vol. 29 (5), 583–593.
- Aguero-Valverde, J., 2005. *Spatial Models of County-Level Roadway Crashes for Pennsylvania*. MS Thesis. The Pennsylvania State University
- Amoros, E., Martin, J.L., Laumon, B., 2003. Comparison of road crashes incident and severity between some French counties. *Accident Analysis and Prevention*, vol. 35. pp 537-547.
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian Analysis of Space-Time Variation in Disease Risk. *Statistics in Medicine*. Vol. 14. pp 2433-2443.
- Besag, J., 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B*. Vol. 36 No. 2. pp 192-236.
- Besag, J., York, J., Mollié, A., 1991. Bayesian Image Restoration with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics*. 43, pp 1-59
- Bureau of Highway Safety and Traffic Engineering, 1997. *1996 Pennsylvania Crash Facts and Statistics*, Pennsylvania Department of Transportation.
- Bureau of Highway Safety and Traffic Engineering, 1998. *1997 Pennsylvania Crash Facts and Statistics*, Pennsylvania Department of Transportation.
- Bureau of Highway Safety and Traffic Engineering, 1999. *1998 Pennsylvania Crash Facts and Statistics*. Pennsylvania Department of Transportation.
- Bureau of Highway Safety and Traffic Engineering, 2000. *1999 Pennsylvania Crash Facts and Statistics*. Pennsylvania Department of Transportation.
- Bureau of Highway Safety and Traffic Engineering, 2001. *2000 Pennsylvania Crash Facts and Statistics*, Pennsylvania Department of Transportation.
- Chichester, B.M., Gregan, J.A., Anderson, D.P., Kerr, J.M., 1998. Associations between road traffic accidents and socio-economic deprivation on Scotland's west coast. *Scot. Med. J.* 43 (5), 135–138.
- Congdon, P., 2003. *Applied Bayesian Modelling*, John Wiley & Sons.

- Edwards, J.B., 1996. Weather-related road accidents in England and Wales: a spatial analysis. *Journal of Transport Geography*, vol. 4.
- Evans, L., 1991. *Traffic safety and the driver*. Van Nostrand Reinhold, New York.
- Evans, L., 2004. *Traffic Safety*. Science Serving Society.
- Fristrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the variation in Road Accident Counts. *Accident Analysis and Prevention*. Vol. 27 pp 1-20.
- Gary, S.L.S., Aultman-Hall, L., McCourt, M., Stamatiadis, N., 2003. Consideration of driver home county prohibition and alcohol-related vehicle crashes. *Accident Analysis and Prevention*. Vol. 35 pp 641-648.
- Gelman, A., Carlin, J., Stern, H.S., Rubin, D.B., 2003. *Bayesian Data Analysis*, Chapman & Hall/CRC, second edition.
- Hamed, M. 1999, Analysis of Pedestrians' Behavior at Pedestrian Crossings, *Safety Science*, Vol.38, No.1, pp.63-82.
- Hamed, M., Jaradat and Easa, 1998. Analysis of Commercial Mini-Bus Accidents, *Accident Analysis and Prevention*, Vol. 30, No. 5, pp. 555-567.
- Jones, A.P., Langford, I.H., Bentham, G., 1996. The Application of K-function Analysis to the Geographical Distribution of Road Traffic Accident Outcomes in Norfolk, England, *Social Science and Medicine*, vol. 42, No 6, pp 879-885.
- Jovanis, P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. *Transportation Research Record* 1068. pp 42-51.
- Kam, B.H., 2003. A disaggregate approach to crash rate analysis. *Accident Analysis and Prevention*. Vol. 35. pp 693-709.
- Knorr-Held, L., 2000. Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk. *Statistics in Medicine*. Vol. 19. pp 2555-2567.
- Knorr-Held, L., Besag, J., 1998. Modelling Risk from a Disease in Time and Space. *Statistics in Medicine*. 17, pp 2045-2060.
- Lawson, A.B., Browne, W.J., Vidal Rodeiro, C.L., 2003. *Disease Mapping with WinBUGS and MLwiN*. John Wiley and Sons.
- Levine, N., Kim, K.E., Nitz, L.H., 1995. Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns. *Accident Analysis and Prevention*, Vol. 27, No 5, pp 663-674.

- Levine, N., Kim, K.E., Nitz, L.H., 1995b. Spatial Analysis of Honolulu Motor Vehicle Crashes: II. Zonal Generators. *Accident Analysis and Prevention*, Vol. 27, No 5, pp 675-685.
- MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. *Accident Analysis and Prevention*, Vol. 36 no. 6, pp 1091-1028.
- Miaou, S.P., 1996. Measuring the Goodness-of-Fit of Accident Prediction Models. FHWA-RD-96-040, Federal Highway Administration, Washington, D.C.
- Miaou, S., Song, J.J., Mallick, B.K., 2003. Roadway Traffic Crash Mapping: A Space-Time Modeling Approach. *Journal of Transportation and Statistics*, Vol. 6, No 1. pp 33-57.
- Miaou, S., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*, Vol. 37 No. 4, pp 699–720.
- National Climatic Data Center, 2004. NNDC Climate Data Online. National Oceanic and Atmospheric Administration. Web page:
<http://www.ncdc.noaa.gov/oa/climate/stationlocator.html>. Visited 07/15/2004.
- Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. *Accident Analysis and Prevention*, Vol. 36. pp 525-532.
- Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road casualties in England. *Accident Analysis and Prevention*, Vol. 36 no. 6, pp 973-984.
- Pennsylvania Department of Transportation, 2004. Highway Statistics. Web page:
<http://www.dot.state.pa.us/> visited 7/13/2004
- Pennsylvania Uniform Crime Reporting System of the State Police, 2004. Pennsylvania State Police. Web Page: <http://ucr.psp.state.pa.us/UCR/ComMain.asp>. Visited 7/13/2004.
- R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL <http://www.R-project.org>.
- Richardson, S., 1992. Statistical Methods for geographical correlation studies. In: Elliott, P., Cuzick, J., English, D., Stern, R. (Eds.), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford University Press, London, pp. 181–204.
- Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention*, vol. 27. pp 371-389.
- Shinar, D., 1978. *Psychology on the Road, the Human Factor in Traffic Safety*. John Wiley and Sons. USA.

Spiegelhalter, D., Best, N., Carlin, B.P., Linde, A., 2002. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society B*. Vol. 64, Part 4, pp. 583–639.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2004. WinBUGS User Manual version 2.0. MRC Biostatistics Unit, Cambridge, UK.

US Census Bureau, 2004. Population Estimates Datasets. Web page
<http://www.census.gov/popest/datasets.html>, visited 10/18/2004.

US Census Bureau, 2004b. Decennial Census Datasets American FactFinder. Web page
http://factfinder.census.gov/servlet/DatasetMainPageServlet?_program=DEC&_lang=en, visited 10/18/2004.

US Census Bureau, 2004c. Small Area Income & Poverty Estimates. Web page
<http://www.census.gov/hhes/www/saipe/>, visited 10/18/2004.

Wakefield, J.C., Best, N.G., Waller, L., 2000. *Bayesian Approaches to Disease Mapping, on Spatial Epidemiology: Methods and Applications*, Oxford University Press.

Waller, L.A., Carlin, B.P., Xia, H., Gelfan, A.E., 1997. Hierarchical Spatio-Temporal Mapping of Disease Rates. *Journal of the American Statistical Association*, Vol. 92. pp 607-617.