Available online at www.sciencedirect.com
ScienceDirect
Journal of Transport Geography 15 (2007) 274–285
JOURNAL OF TRANSPORT GEOGRAPHY
www.elsevier.com/locate/jtrangeo

A GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes

Linhua Li [a,1], Li Zhu [b,2], Daniel Z. Sui [c,*]

[a] Department of Civil Engineering, Texas A&M University, College Station, TX 77843-3136, United States
[b] School of Rural Public Health, Texas A&M University System Health Science Center, TAMU 1266, College Station, TX 77843-1266, United States
[c] Department of Geography, Texas A&M University, College Station, TX 77843-3147, United States

Abstract

This paper develops a GIS-based Bayesian approach for intra-city motor vehicle crash analysis. Five-year crash data for Harris County (primarily the City of Houston), Texas are analyzed using a geographic information system (GIS), and spatial–temporal patterns of relative crash risks are identified based on a Bayesian approach. This approach is used to identify and rank roadway segments with potentially high risks for crashes so that preventive actions can be taken to reduce the risks in these segments. Results demonstrate the approach is useful in estimating the relative crash risks, eliminating the instability of estimates while maintaining overall safety trends. The 3-D posterior risk maps show risky roadway segments where safety improvements need to be implemented. Results of GIS-based Bayesian mapping are also useful for travelers to choose relatively safer routes.

Keywords: GIS; Hierarchical Bayesian modeling; Crash analysis; 3-D visualization of relative risks

1. Introduction

Transportation accidents were the seventh single leading cause of death in the United States (US Department of Transportation [USDOT] and Bureau of Transportation Statistics [BTS], 2001). However, motor vehicle crashes, which account for about 95% of transportation-related deaths and an even higher percentage of transportation injuries, were the leading cause of death for people between the ages of 3 and 33 (National Highway Traffic Safety Administration [NHTSA], 2005). There were 42,815 fatalities and approximately 2,926,000 injuries in the US during 2002, resulting from approximately 6,316,000 police-reported motor vehicle crashes (USDOT and NHTSA, 2004). It was estimated that motor vehicle crashes that occurred in 2000 alone cost $230.6 billion (Blincoe et al., 2002). Deaths, injuries, and property damages due to these crashes are not only a major cause of personal suffering and financial loss to the victims, their families, and friends, but also to society at large. Therefore, motor vehicle crashes have become a major social problem in the US (USDOT and BTS, 2001), and how to improve traffic safety has become a major societal concern in the US (Evans, 2004).

Decades of interdisciplinary research on motor vehicle crashes have revealed that there are generally five major factors affecting traffic safety and efficiency – driver behavior (about 160 million drivers in the US), vehicle types (motorcycles to large trucks), roadway condition (design, capacity, pavement type), traffic characteristics (flow, speed, density, occupancy), and environmental factors (weather, etc.). All these factors interact with each other and influence the occurrences and severity of crashes simultaneously. Among those, driver behavior, affected by variables such as alcohol

* Corresponding author. Tel.: +1 979 845 7154; fax: +1 979 862 4487.
E-mail addresses: lil@geog.tamu.edu (L. Li), LiZhu@srph.tamhsc.edu (L. Zhu), sui@geog.tamu.edu (D.Z. Sui).
[1] Tel.: +1 979 862 6617; fax: +1 979 862 4487.
[2] Tel.: +1 979 458 0079.

0966-6923/$ - see front matter
doi:10.1016/j.jtrangeo.2006.08.005

and drug use, reckless operation of vehicles, failure to properly use occupant protection devices, and fatigue, is a major factor contributing to a high proportion of crashes (USDOT and BTS, 2001). Although driver maneuvers often involve considerable amount of subjective judgment, they are nonetheless the reactions to the roadway condition, traffic situation and other environmental factors. Therefore, roadways and traffic control devices should be designed and built using high standards. Based on limited resources and funding, the question becomes how to determine priorities – when and where (which roadway segments) are more risky than others.

The primary objective of this paper is to develop a geographic information system (GIS)-based Bayesian approach for intra-city motor vehicle crash analysis to estimate relative crash risks and determine the spatial–temporal patterns of risks. The results can help pinpoint risky roadway segments that need special attention from both transportation authorities and drivers. To better capture the risk levels of different road segments, this research differentiates the risks in different directions of the roadways, disaggregates different road types, integrates Bayesian approach to filter the data uncertainty, and uses GIS to visualize the spatial relative crash risks in 3-D views according to different temporal scales.

This paper is organized into five sections. After this introduction, a literature review is provided in Section 2, and then data and methods are described in Section 3. Section 4 presents the results and discussions on the temporal, spatial, and spatial–temporal patterns of intra-city relative crash risks using the GIS-based Bayesian approach and 3-D visualizations. The last section concludes with a discussion of our approach and its implications for future research.

2. Literature review

Motor vehicle crashes have been studied from different spatial and temporal perspectives by different researchers using varied methodologies. For spatial patterns of motor vehicle crashes, several studies examined crashes that occurred in different environmental settings, such as on a specific road section (Skabardonis et al., 1998), road types (Brodsky and Hakkert, 1983), intersections (Golias, 1992; Nicholson, 1985), or corridors (Golob et al., 1990; Okamoto and Koshi, 1989). Additionally, studies on motor vehicle crashes have also been performed at different geographic scales using data aggregated to different administrative units, ranging from census tracts/traffic analysis zones (TAZs) (Levine et al., 1995b; Ng et al., 2002), to cities (Jones et al., 1991), counties (Fridstrøm and Ingebrigtsen, 1991; Jegede, 1988), and state and national levels (Haight and Olsen, 1981).

As for the temporal patterns of motor vehicle crashes, reported studies focused primarily on the fluctuation of the quantity and rate of crashes, injuries, and fatalities according to different temporal scales, such as hourly, daily, monthly, and yearly (El-Sadig et al., 2002; Fridstrøm and Ingebrigtsen, 1991; Levine et al., 1995b,c; USDOT and BTS, 2001). High levels of aggregation (e.g., yearly) could not easily detect the changes of short-term structural variables, while more disaggregated analyses require detailed temporal information that may be hard to obtain.

Extensive studies have also been reported on the relationship between crash characteristics (rate, frequency, fatality, injury, duration, severity, etc.) and related variables, such as weather conditions, geometric design of roads, traffic volume, road density, and driver behaviors using a variety of statistical modeling techniques. Miaou and Lum (1993) found that conventional linear regression models were not appropriate for modeling vehicle crash events on roadways. When the mean and variance of the crash frequencies were approximately equal, the Poisson regression was a more appropriate model for examining the relationship between crashes and influential factors (Miaou, 1994). Overdispersion occurs when observed variance of the data is larger than the predicted variance. When overdispersion was moderate or high, the use of both negative binomial regression and zero-inflated Poisson regression were found to be more appropriate (Miaou, 1994). Zero-inflated probability processes, such as the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models, allow for better isolation of independent variables that determine the relative crash likelihoods of safe vs. unsafe roadways (Lambert, 1992; Shankar et al., 1997).

The Bayesian approach has been widely used in statistics and sciences over the past decade. One of the major advantages of the Bayesian approach is its ability to forecast risks accurately even in the presence of sparse data or rare events (Withers, 2002). In the public health area, application of Bayesian methods in disease mapping, risk assessment and prediction are numerous (Besag and Newell, 1991; Wakefield and Morris, 2001; Wakefield et al., 2000; Waller et al., 1997). The ability to incorporate prior knowledge without the restriction of classical distributional assumptions makes Bayesian inference a potent forecasting tool in a wide variety of fields (Withers, 2002). The Bayesian approaches, from empirical Bayes to full Bayes, were also implemented in some crash analysis studies to estimate crash risk and predict crash frequency (Brüde and Larsson, 1988; Hauer, 1992, 2002; Mountain et al., 1996). Recently, GIS, regression analysis, and the Bayesian methods were integrated to better analyze and estimate crash and risk (MacNab, 2004; Miaou and Song, 2005; Ng et al., 2002). However, there is no research incorporating Bayesian approaches to analyze link-based relative crash risks at intra-city level.

This research will use a hierarchical Bayesian approach to spatially estimate relative crash risks by incorporating information from adjacent roadway segments, in order to get a better evaluation of the risks for each road segment. Relative risk maps developed from conventional statistical

models often feature large outlying relative risks in small areas, and hence show high uncertainty. They also fail to catch similarity of relative risks in nearby or adjacent regions, but an appropriately tailored hierarchical Bayesian approach will incorporate spatial assumptions and enable the customary Bayesian "borrow of strength" from the neighboring regions to eliminate the high uncertainty and hence generate a better estimation, called posterior estimation.

GIS has been proven to be a useful tool for mapping and spatial analysis in transportation studies (Miller and Shaw, 2001; Thill, 2000). Many researchers have used GIS to display crash locations on digital maps and perform various spatial analyses (including hot spot analysis) of crashes (Black, 1991; Flahaut et al., 2003; Kam, 2003; Levine et al., 1995a; Petch and Henson, 2000; Steenberghen et al., 2004). Following earlier works on network autocorrelation analysis (Black, 1992; Black and Thomas, 1998), one recent significant advance is the network-constrained approach to conduct spatial point pattern analysis over a network (Yamada and Thill, 2004; Yamada and Thill, forthcoming). In addition, GIS enables researchers to link crash data with travel information, land use, and social-economic information to better capture the relationship between crash occurrence and contributing factors. Researchers could also adjust the width and color of the roadways in GIS to produce 2-D visualization, but so far few studies have explored 3-D visualization techniques for mapping motor vehicle crashes. This research fills in this void by exploring a 3-D visualization approach for mapping relative crash risks along transportation networks.

Using spatially disaggregated data has been another trend in recent crash analyses. Prior to the wide application of GIS, most researchers either used highly aggregated data sets (total number of crashes, total travel distance, etc.) or attempted to disaggregate data based on demographic characteristics (age, sex, race, etc.), severity (fatality vs. injury vs. property-damage-only (PDO); long duration vs. short duration, etc.), or vehicle types (truck vs. car, etc.) rather than the spatial aspects associated with a crash (Brodsky and Hakkert, 1983; Fridstrøm and Ingebrigtsen, 1991; Jones et al., 1991; Miaou, 1994). Each specific crash location has specific contributing factors, requiring spatial disaggregation to determine them.

Trip-based crash analysis was advocated by assuming every trip (route) was unique in terms of crash risk (Kam, 2003). The linear relationship between crashes and distance traveled was refuted by the argument that non-freeways have more crashes per mile driven than freeways. However, since a linear relationship may still exist within each road type (e.g., interstate freeway), disaggregation of the analyses for different road types is a way to circumvent this weakness. Furthermore, we noticed that even for the same roadway, its physical and environmental conditions (e.g., geometry, road condition, and lighting) may change both spatially and temporally. Therefore, we believe that spatial disaggregation (including road type disaggregation) and temporal disaggregation (hourly, weekly and yearly) are necessary and would produce more useful results than previous studies.

When analyzing intra-city level crash patterns, there are no reported studies having disaggregated crashes in the same location based on different directions (e.g., eastbound/westbound, northbound/southbound). The reason for this oversight was probably related to the fact that most crash data did not include this variable or authors thought it was not necessary, especially when dealing with macro level (e.g., county-level, state-level) data. In this paper, we differentiate relative risks in different directions as well as crashes because theoretically opposite directions of roadways may have different risk values due to contrasted crash counts, traffic characteristics (e.g., traffic volumes), roadway conditions (e.g., work zone), and environments (e.g., lighting). Without differentiating directions for motor vehicle crashes, risk values for two directions might have been averaged out, leading to erroneous risk estimation.

This research attempts to fill the gaps in the literature mentioned above by developing an approach to conduct intra-city relative crash risk analysis, in which different directions and road types are disaggregated along both spatial and temporal scales. Bayesian spatial smoothing among adjacent roadway segments was implemented to filter the uncertainty and better capture the real risk trend. The posterior relative crash risks at different temporal scales can then be mapped at the segment level and displayed with 3-D visualization.

3. Data and methods

The overall methodology (Fig. 1) can be subdivided into three parts: (1) data preparation and preliminary processing; (2) Bayesian smoothing and updating; and (3) GIS mapping and visualization of results. ArcMap and ArcScene developed by ESRI were used for GIS operations and mapping, and Bayesian modeling was performed using WinBUGS (Spiegelhalter et al., 2004) and its add-on program GeoBUGS (Thomas et al., 2004) that fits spatial models and produces outputs for maps.

3.1. Data preparation

To illustrate our approach, Harris County, Texas was chosen as the study area. The City of Houston resides primarily in Harris County.

Crash data sets from the Texas Department of Transportation (TxDOT) Traffic Operations Division Crashes Data Files (TRF crash files) were collected and processed for a five-year period (1996–2000). Data for crashes on state-maintained roadways are included in the TxDOT TRF database. Fatal crashes, injury crashes, and PDO crashes where one or more vehicles were towed were recorded in the database. Therefore, crash records include

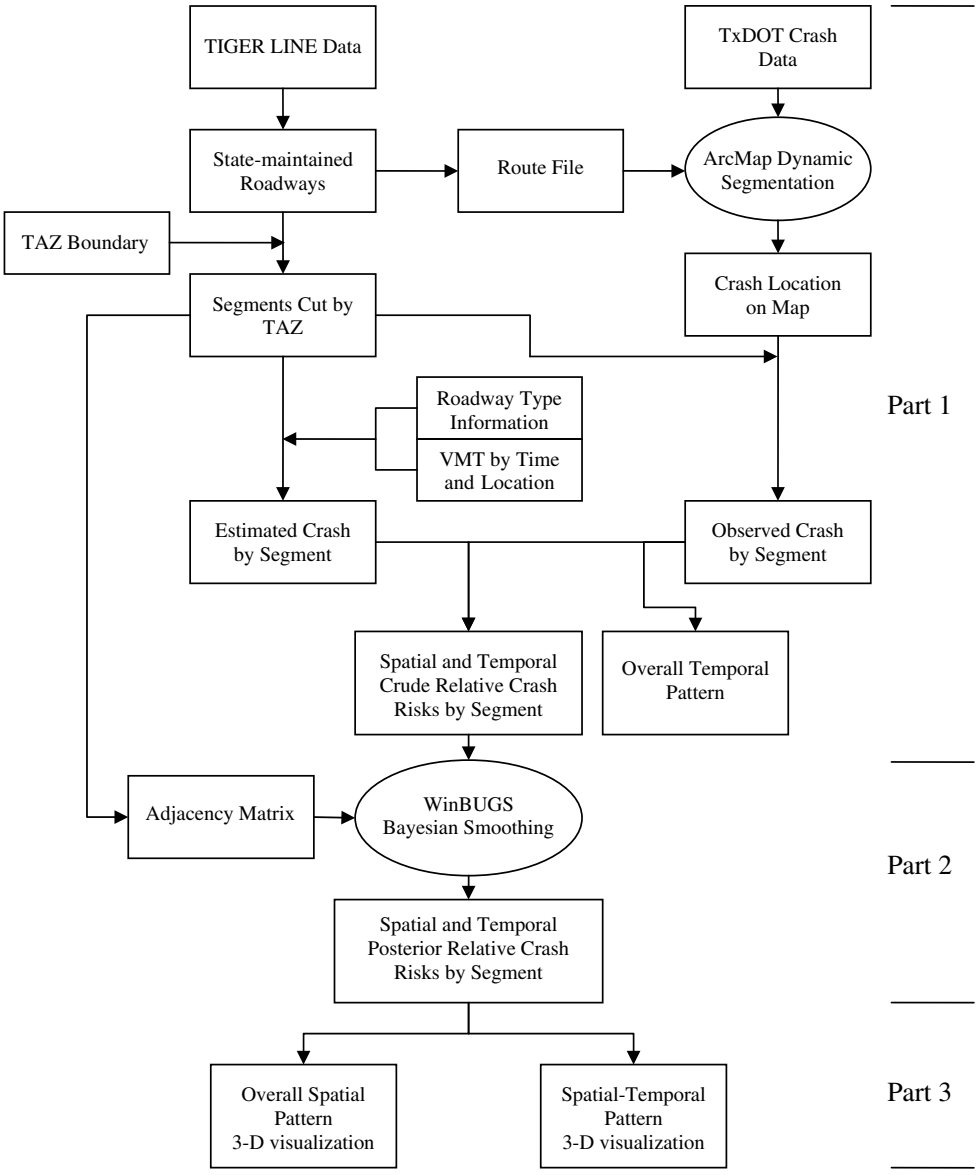

Fig. 1. Overall methodological flow chart.

information on all serious crashes like crash location, severities, time and date, etc. State-maintained roadways can be classified by road types, such as interstate freeway, urban freeway, principal arterial, and minor arterial.

Traffic maps of Harris County from 1996 through 2000 obtained from TxDOTs Houston District Planning Office indicate the annual average daily traffic (AADT) for each highway segment. Annual average daily vehicle miles traveled (VMT) is calculated by multiplying the AADT by the length of each roadway segment.

VMTs for each day of a week and hour of a day were calculated by using annual average daily VMT, daily VMT adjustment factors in a week and hourly VMT adjustment factors in a day, which were obtained from the Texas Transportation Institute (TTI). Hourly directional traffic volumes were adjusted according to hourly directional volumes taken from several traffic count stations[3] on major freeways.

3.2. GIS mapping of crash locations

There are several ways to locate crashes onto digital maps. Crashes can be directly added if the exact geographic references of crash locations, like coordinates, are available. Address geocoding can be conducted when the exact

[3] Hourly directional volume data were obtained from TxDOT Houston District. The traffic count stations continuously recorded traffic volumes on both directions every hour of a day and every day of a week in Harris County. All of the stations are located on major freeways such as I-10, I-45, I-610 loop, US-59 and US-290, thus capturing the major daily commuting flows.

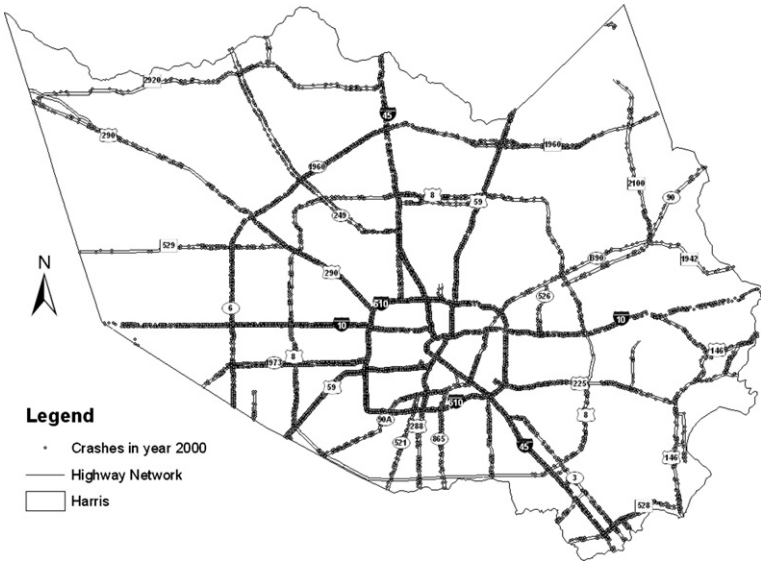

Fig. 2. Identified locations of motor vehicle crashes in year 2000.

address (e.g., street name and number, city, state, zip code) is available. In this research, dynamic segmentation (linear referencing) was performed to locate crashes since the roads on which crashes occurred and their positions relative to the starting points of the routes were known.

The base roadway network shape file was retrieved from TIGER line data.[4] Proper edits (cut, divide, split, merge) were done in accordance to the roadway control-sections. Each roadway section has two parallel links, representing different directions of roadways. Then, routes were created from existing links by giving them corresponding control-section numbers, identifying milepoint directions and adding starting and ending measurements. A Harris County state-maintained roadway network file, which was used as route reference, was created using control-section number as route identifier and milepoint as measurement.

ESRI's ArcMap was used to perform dynamic segmentation. The TRF crash file, which indicates the crash location along the route with control-section number and milepoint, was used as point event table. In ArcMap, the "add route events" function was used to add each crash as a point along the routes on the map layer. The accuracy level of crash location is 0.1 mile based on the crash data, which already provides sufficient accuracy for intra-city level crash analysis.

Approximately 98% of crashes were successfully and correctly located on the maps for each of the five years in the study period. For example, 27,037 out of 27,488 (98.36%) crashes were located in the year 2000 map, as shown in Fig. 2. The remaining 1.64% could not be added due to reasons like unknown control-section number, unknown milepoint, error in the milepoint information, etc. We assume that less than 2% unallocated crashes do not significantly affect the analysis results.

The entire roadway network was divided by 988 TAZs into 1349 segments, with an average length of about 1 mile. Although it is an arbitrary definition of segments, we assume that a segment within a single TAZ has relatively homogeneous road and environmental conditions.

3.3. Hierarchical Bayesian modeling

Bayesian methods infer individual-level parameter estimates by borrowing information from neighbors (Bolstad, 2004; Lee, 2004). Hierarchical Bayesian modeling uses multiple levels of analysis in an iterative way (Carlin and Louis, 1996; Rossi et al., 2006). Unlike conventional statistical inference, which derives the average estimates of parameters, hierarchical Bayesian modeling produces parameter estimates for each analytical unit. It also identifies and flags "extra variance" (Congdon, 2001; Winkler, 2003). In spatial statistics, if there is high uncertainty in a regression model, the result explains only a small amount of variance. But in a hierarchical Bayesian model the unexplained "extra variance" is usually identified as either spatially correlated effects or heterogeneity effects (Best et al., 1999).

Hierarchical Bayesian modeling involves two stages. At the first stage, we specified a likelihood model for the observed crash counts vector based on the relative risk vector[5] of crashes, and then specified a prior model over the

[4] The TIGER Line data are similar to the TxDOT link data we used in the paper. These two data sets overlap quite well. The reason we used TIGER data is that it could be almost perfectly overlaid with the TAZ data. For each segment in the TIGER file, we knew the exact starting and ending milepoint, and we assigned the milepoint to each segment individually. Even if they might contain rather minor positional, geometric or topological inaccuracies, these errors are, at the least for our study area, so small (almost negligible, usually smaller than 0.1 mile according to the sample sites we tested). We have verified manually that each allocated crash is on the segments that have the correct traffic volumes and correct segment length.

[5] Relative crash risk here is defined as the quotient of observed number of crashes over expected number of crashes, which is estimated by multiplying average crash rate with VMT. When the value of relative risk is greater than 1.0, the segment is riskier than expected.

space of possible relative risks at the second stage. Using software packages such as WinBUGS (Spiegelhalter et al., 2004) or GeoBUGS (Thomas et al., 2004), it is possible to yield a set of posterior means for relative risks given observed crash counts. The set of posterior means of relative risks was then used to create a map to visualize high- or low-risk segments. Crude maps were developed from the likelihood model (the first stage) only, and often feature large outlying relative risks in small areas (where the VMTs are small in this study). Hence, crude maps usually show high levels of uncertainty due to the small sample sizes in the small areas. They also fail to catch similarities of relative risks in nearby or adjacent regions. As demonstrated by recent studies using diverse environmental and public health data (Arató et al., 2006; Mather et al., 2006; Riccio et al., 2006), an appropriately tailored Bayesian approach is capable of incorporating spatial assumptions and help smooth the maps with high variability by borrowing strength from neighbors for those mapping units with small populations. The likelihood model assumes that the observed crash counts Y_{ijt} for road segment i, road type j, and in time period t (hour, day, or year) are conditionally independent Poisson variables given the relative risk. The model can be represented as

$$Y_{ijt}|\mu_{ijt} \sim \text{Poisson}(E_{ijt}\exp(\mu_{ijt})),$$
$$i = 1,\ldots,I;\ \ j = 1,\ldots J;\ \ t = 1,\ldots T, \tag{1}$$

where $\exp(\mu_{ijt})$ is the relative risk for crashes, and μ_{ijt} is log relative risk. In this model, the expected crash count for segment i, road type j, and time period t, E_{ijt}, is proportional to the relevant VMT. In this case, we set $E_{ijt} = R_j n_{ijt}$, where n_{ijt} is the VMT on segment i, road type j in time period t, and $R_j = (\sum_{it} Y_{ijt})/(\sum_{it} n_{ijt})$ is the average crash rate on road type j (i.e., assuming homogeneity of occurrence rate across all segments and time periods within the same road type j). The log-relative risk is modeled as

$$\mu_{ijt} = \beta_{jt} + \alpha x + \theta_{ij}^{(t)} + \phi_{ij}^{(t)} \tag{2}$$

where β_{jt} is an overall intercept for road type j and time t, x is the covariate of road type, α is the corresponding effect of road type, $\theta_{ij}^{(t)}$ is the non-structured heterogeneity (at time t), and $\phi_{ij}^{(t)}$ is the spatially correlated random effect (at time t), respectively. The overall intercept β_{jt} captures the main effect for road type j at time t. Although we could specify a parametric function (e.g., linear or quadratic form) for the time effect, a qualitative form that allows data to reveal the presence of any temporal trend is preferred. The distribution of heterogeneity effects $\theta_{ij}^{(t)}$ is assumed to be exchangeable, i.e., $\theta_{ij}^{(t)} \sim \text{Normal}(0, \tau_t)$, while the spatial effects $\phi_{ij}^{(t)}$ are assumed to follow a conditional autoregressive (CAR) model. According to Besag (1974), the joint distribution of the vector of spatial effects $\boldsymbol{\phi}_t$ (at time t) is proportional to $\exp(-(\lambda_t/2)\boldsymbol{\phi}_\mathbf{t}^\mathbf{T}\mathbf{B}\boldsymbol{\phi}_\mathbf{t})$, i.e., a multivariate normal density with mean $\mathbf{0}$ and covariance matrix $\mathbf{B}^{-1}$. The elements of matrix $\mathbf{B}$ is determined as $\mathbf{B}_{kk} = a_k$ and $\mathbf{B}_{kl} = -a_k\omega_{kl}$, with a_k denoting the number of neighbors of segment k, and ω_{kl} being the elements in the adjacency matrix $\mathbf{W}$. The (k,l)th element in matrix W equals to 1 if two segments k and l are adjacent to each other. To compute the adjacency matrix, each segment is considered to be adjacent to its upstream and downstream segments in the same direction. We assume it is also adjacent to the segment with opposite traveling direction to incorporate the rubbernecking[6] phenomena.

We encourage similarity among the random effects across time by assuming $\tau_t \sim \text{Gamma}(a,b)$ and $\lambda_t \sim \text{Gamma}(c,d)$ based on the conjugate distribution theory. Placing flat (uniform) priors on the main effects β_{jt} and α completes the model specification. It should be noted that the constraints $\sum_{ij}\phi_{ijt} = 0$, at $t = 1,\ldots,T$ must be added to identify the time effects β_{jt}, due to the location invariance of the CAR prior.

We set $a = 0.001$, $b = 0.001$ (i.e., the τ_t have prior mean 1 and standard deviation $\sqrt{1000}$) and $c = 0.01$, $d = 0.01$ (i.e., the λ_t have prior mean 1, standard deviation 10). These are vague priors designed to allow the data to dominate the allocation of excess spatial variability to heterogeneity and clustering. Our Markov Chain Monte Carlo (MCMC) implementation ran two parallel sampling chains for 9000 iterations each, and discarded the first 3000 iterations as pre-convergence "burn-in".

4. Results and discussion

4.1. Temporal analysis

The temporal analysis aims to find out the overall relative risk pattern temporally and its results are a prerequisite for later spatial–temporal analysis targeting spatial patterns in different time periods. Our overall temporal analysis was done without GIS and a hierarchical Bayesian model, because both GIS mapping and spatial smoothing are not necessary. According to our data, the total number of crashes does not display any significant differences by month in the study area, so the temporal analysis only focuses on temporal change of relative crash risks by day of week and hour of day. VMT showed a similar pattern as crash count (Fridays were the highest and Sundays were the lowest). However, Saturdays were the second highest days in terms of crash count, and Mondays through Thursdays had the second highest VMT. Since weekday traffic is mainly work-related, it remains relatively stable from Monday through Thursday. On Fridays not only do people usually drive to work, but they may also travel longer distances to meet with family or friends, or celebrate the weekends. People are more likely to rest at home and get ready for the new week on Sundays. VMT on Saturdays was lower than weekdays and Fridays, but higher than

[6] Drivers tend to look at crashes that occurred in the other direction and may lose concentration on their own driving. This phenomenon will reduce traffic speed, and increase crash risk.

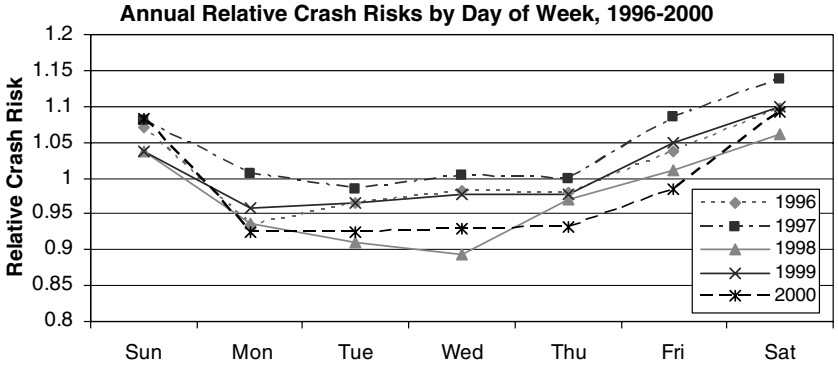

Fig. 3. Relative crash risks by day of week in Harris County.

Sundays. It is probably because there were some work-related traffic, shopping, and party traffic on Saturdays.

The plot of relative crash risk by day of week (Fig. 3) shows that although different years had different risks, the risks in each year followed a similar "U" shape pattern. Saturdays had the highest overall relative crash risk in a week, which was over 1. The reason is that Saturday has the moderate crash counts, but low VMTs. Sunday and Friday were the second and third high-risk days, respectively. In each year, most Mondays through Thursdays had similar risk values, which were below 1, implying that they were safer than usually expected.

Relative crash risks also fluctuated from year to year. The relative risk in 1997 was higher than that in the other years and year 2000 had a relatively low risk.

The hourly analysis of relative crash risks was performed separately for each of the following four categories: weekdays (Monday through Thursday), Friday, Saturday, and Sunday, since each category shows a specific hourly pattern. Crash counts and VMT had a different daily pattern in those four categories. In general, the daily crash counts were consistently related to changes of daily VMT, but different patterns were shown in different categories. Crash counts on weekdays and Fridays had two peaks (7:00 a.m.–9:00 a.m. and 3:00 p.m.–7:00 p.m.), which were similar to the VMT distribution. It was also noticed that the crash counts on Saturdays and Sundays had two peaks (1:00 a.m.–3:00 a.m. and 1:00 p.m.–5:00 p.m.), but VMT did not peak in the early morning.

The plot of the relative crash risk by hour of day (Fig. 4) showed that different categories had very similar trends in spite of having different counts and VMT distributions. Generally, early morning (1:00 a.m.–4:00 a.m.) had the highest risk during the day, with weekends being less safe than weekdays. The highest risk (over 4.0) was found between 2:00 and 3:00 a.m. on Sundays and Saturdays. Risk in this same time period on weekdays and Fridays was also higher than 1.0. It was also noticed that relative risks were also higher than 1 from 10:00 p.m. to 4:00 a.m. for all categories. The contributing factors may include the dark environment, lack of good lighting, tired drivers, or even drunk drivers. The plot also shows that on weekdays and Friday, morning (risk < 1) was safer than the afternoon (risk > 1) although morning and afternoon

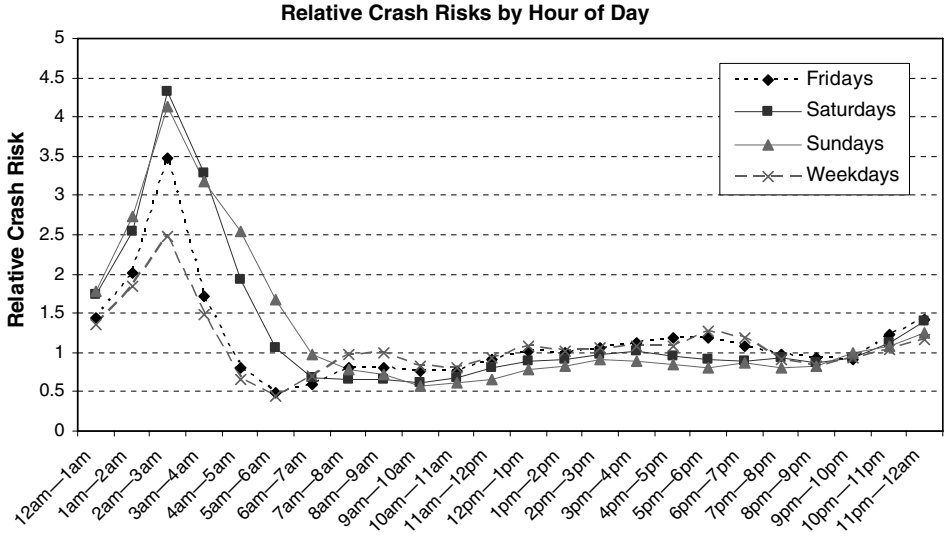

Fig. 4. Relative crash risk by hour of day in Harris County.

had almost the same VMT. The lowest risk for Mondays through Fridays existed at 5:00 a.m.–6:00 a.m., but for Sundays and Saturdays the lowest-risk period lagged several hours.

4.2. Spatial analysis

The identification of risky roadway segments requires spatial analysis using GIS and hierarchical Bayesian modeling. It is common knowledge among transportation researchers that non-freeways have more crashes per mile driven than freeways, but the linear relationship between number of crashes and VMT may still exist within each particular road type. Therefore, the assumption that crash frequency is proportional to the VMT within each road type is appropriate as long as road type disaggregation is performed. The comparison of average crash rates on road type j in the study area during the study period was calculated by using $R_j = (\sum_{it} Y_{ijt})/(\sum_{it} n_{ijt})$ and shown in Fig. 5.

Fig. 5 shows that principal arterials were, from all different road types, the least safe road type in this study area. It also shows that interstate and urban freeways had lower crash rates than principal and minor arterials. This could be due to the fact that in interstate and urban freeways, the large VMT in the denominator makes quotients small, even when there is high crash frequency in the numerator. Principal arterials had the highest rate due to their moderate VMT and large crash counts, which may be related to numerous traffic signals, conflicting traffics, and pedestrian interference. Minor arterials had moderate crash counts and low VMT values, therefore moderate risk. This figure suggests that drivers should choose urban freeways or interstate freeways from a traffic safety perspective, because these two types have lower crash rates relative to travel distance.

In this paper, all the relative risk maps are shown in a 3-D view with the height representing value of relative risks. This is an attempt of visualizing relative crash risks in a new way. A 3-D map has not only the power of a 2-D map, but also the expression of the third dimension. A 3-D map can display risk information for different road segments better than a 2-D map.

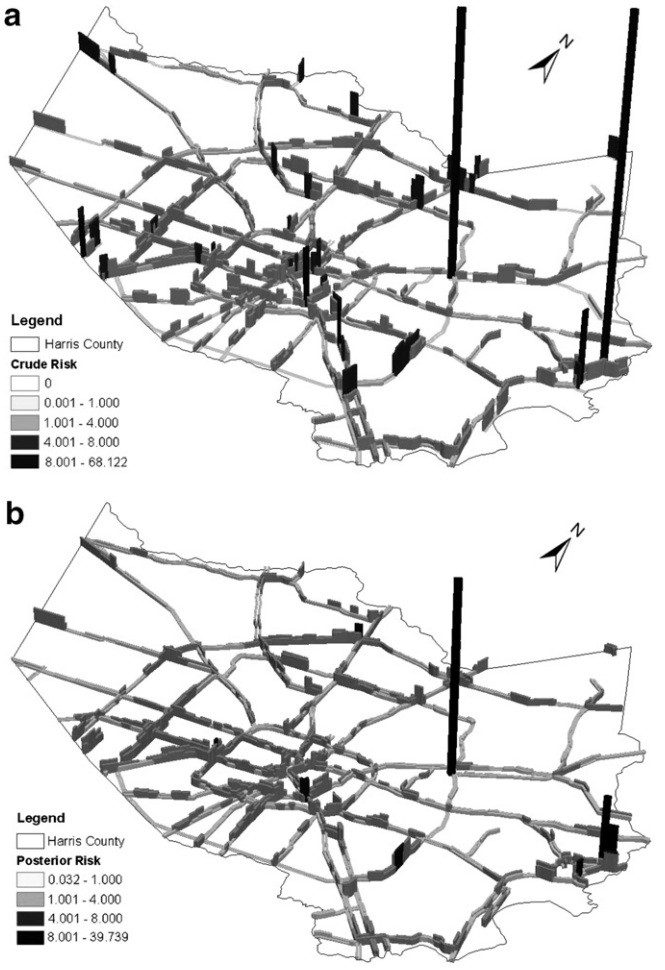

Fig. 6. Crude and posterior relative crash risks in Harris County, TX 1996–2000. (a) Crude relative crash risk, 1996–2000 and (b) posterior relative crash risk, 1996–2000.

Fig. 6 shows the distribution of overall crude and posterior relative crash risks in Harris County during the period 1996 through 2000. The posterior relative risk map (Fig. 6b) clearly shows the characteristic Bayesian smoothing of the crude relative risks (Fig. 6a). In particular, no segment was assigned a risk of exactly zero, and the high

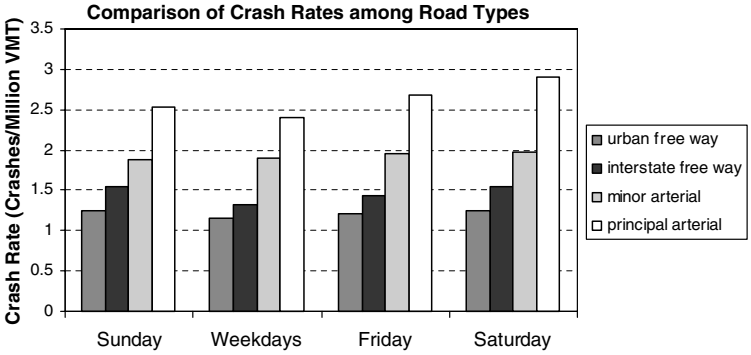

Fig. 5. Crash rates by road types in Harris County.

risks on the segments with low VMT were substantially reduced. However, the observed high risks in the interchange of I-610 and US-59 remained high, as the method properly recognized the much bigger sample sizes in this area.

Transportation authorities as well as motorists need to pay more attention to the safety of those segments with high relative risks. The posterior relative risk map shows there were several risky roadway segments in this area after the uncertainties were filtered. The US-90 at Beltway-8 and the east part of HY-146 were the two least safe segments in Harris county. Some segments such as I-610 at US-59 and US-90 at I-45 also displayed very high risk values. The posterior relative risk map also shows that different directions of some roadways may have different risk levels. For example, the eastbound segments on I-10 west outside Beltway 8 (Sam Houston Toll Way), whose relative risk was more than 1, were riskier than the westbound segments which have relative risks below 1. This reinforces the need to consider direction differentiation in the analysis of relative crash risks.

4.3. Spatial–temporal analysis

The objective of the spatial–temporal analysis is to show the variation of spatial distribution of relative crash risks by hour of day and by day of week. The spatial analysis by day of week was conducted on different days (Sunday, weekdays, Friday, and Saturday) and different years (1996–2000). Almost all the posterior maps showed similar patterns. Fig. 7 shows some examples. Similar relative risk distributions suggest the stability of posterior results of Bayesian spatial smoothing and the explicitness of high-risk segments no matter which day or year is analyzed.

The spatial analysis by hour of day is a high-level disaggregated analysis, which needs very detailed temporal information about crash and hourly VMT information for every roadway segment. The crude relative risks for several short roadway segments were extremely high because of the very small VMT on those segments. The Bayesian approach showed its ability again in smoothing the spatial relative risks in all segments. Fig. 8a shows the posterior relative risk distribution in Harris County between 2:00 and 3:00 a.m. on Saturdays. The figure shows that high risks were spread all over the study area, from downtown to suburban areas, and from interstate freeways to minor arterials in this period, and that relative risks were much higher than at other times. The same period on Sundays had a similar pattern of relative risks.

Fig. 8b and c show the posterior relative risk at morning peak (7:00 a.m.–8:00 a.m.) and afternoon peak (5:00 p.m.–6:00 p.m.) on weekdays. Figures show that relative risks at the afternoon peak were higher and distributed more broadly than those at the morning peak, implying that people were more negligent or more aggressive when driving after work than before work since the VMT at morning peak is the same as the one at afternoon peak.

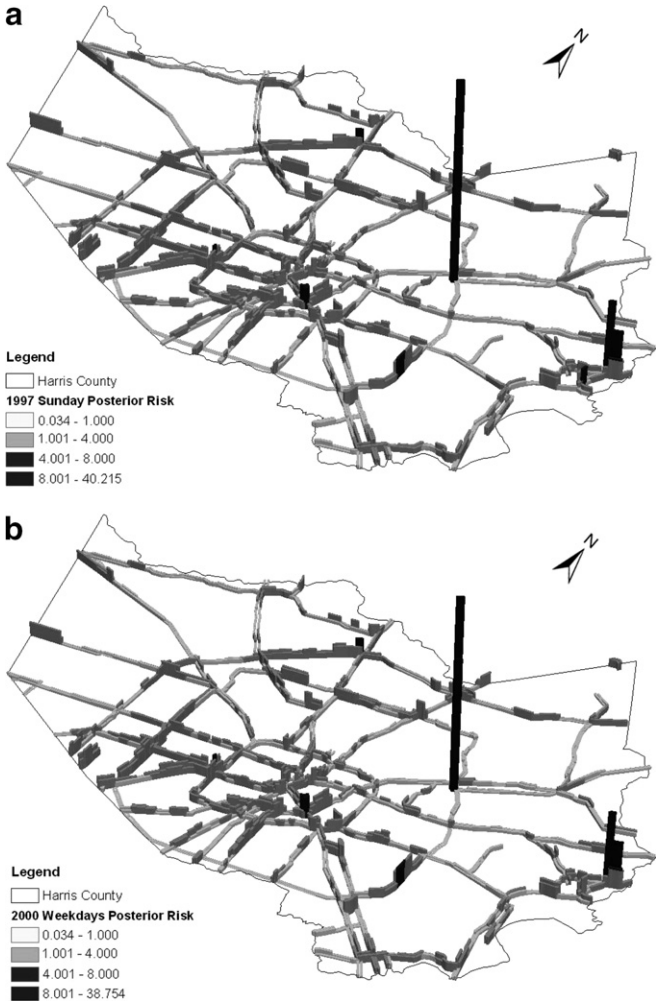

Fig. 7. Posterior relative crash risks in Harris County by week of day. (a) Posterior relative crash risk, Sunday in 1997 and (b) posterior relative crash risk, weekdays in 2000.

In weekend afternoons, relative crash risks were moderate across the county. The typical high-risk segments were similar to those in weekdays, but the number of high-risk segments in the morning was less.

The figures also show that some roadway segments have similar relative risk levels at both directions. However, over 30% of the roadways have statistically significant different risk values for different directions based on a two sample *t*-test. This is strong evidence supporting the need to differentiate directions to evaluate relative crash risks. For segments with different risk values in each direction, only improving one direction with risk over 1 all the time is a wise idea rather than improving both directions if the other half has risk value less than 1. Therefore, transportation authorities may be able to save safety improvement funds by differentiating roadway directions in risky roadway identification.

It should be noticed that the same risk values on different road segments does not mean the same impact after a crash occurred. For example, during the morning peak,

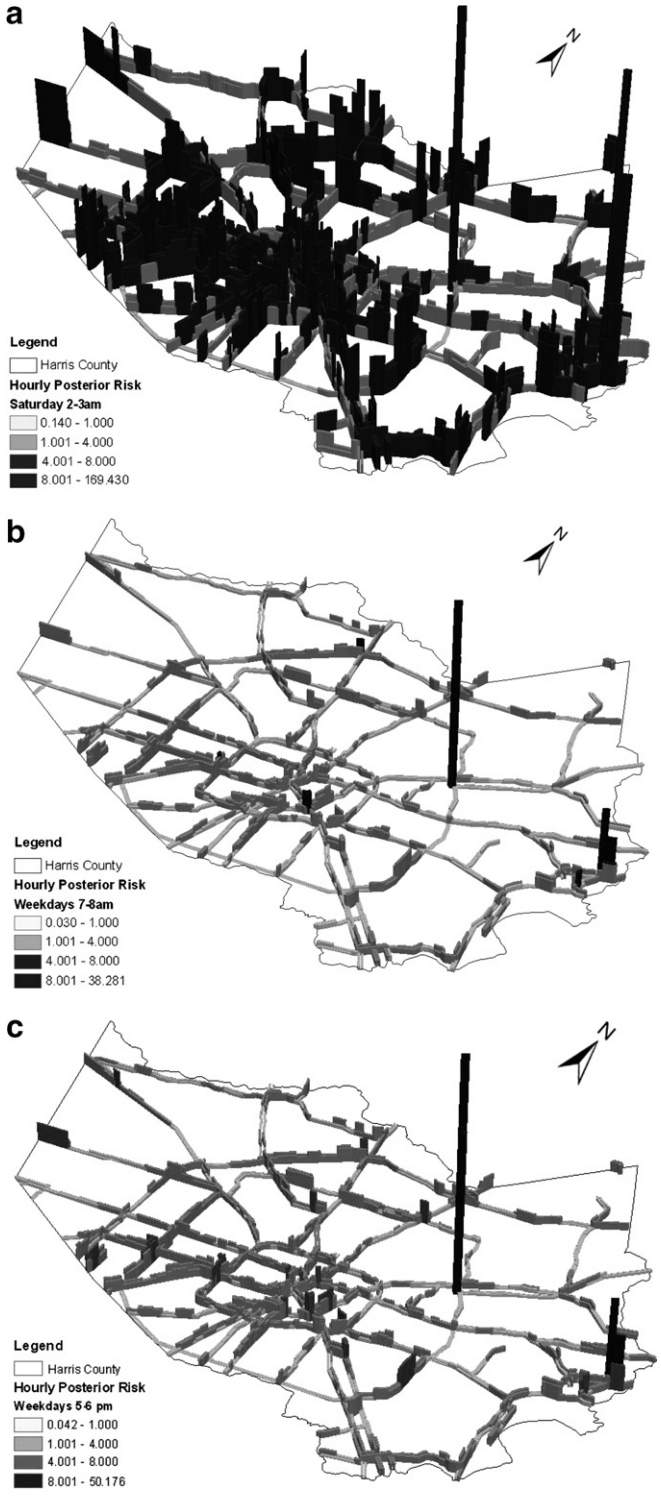

Fig. 8. Posterior relative crash risks in Harris County by hour of day. (a) Posterior relative crash risk, 2:00–3:00 a.m., Saturdays, (b) posterior relative crash risk, 7:00–8:00 a.m., weekdays and (c) posterior relative crash risk, 5:00–6:00 p.m., weekdays.

one lane blockage due to a crash in an inbound lane will cause more congestion, delay, and idle emissions than one on an outbound lane if the capacity of the remaining inbound lanes cannot accommodate the whole inbound traffic. Furthermore, it would take more time for emergency personnel to clear the crashed vehicles from a congested lane than from a non-congested lane.

5. Summary and conclusion

The research aim of this paper was to develop a GIS-based Bayesian approach to perform link-based relative crash risk analysis at the intra-city level. The research approach targets the spatial and temporal patterns of intra-city motor vehicle crashes. To better capture the real safety indications, this paper differentiates traffic directions, disaggregate different road types, integrates hierarchical Bayesian approach to filter the uncertainties based on the spatial assumptions and presents the results in 3-D visualizations. The results of spatial-temporal crash risk patterns can be used to determine risky roadway segments that need attention by transportation agencies as well as motorists.

Compared to conventional statistical inference models, which derive point and interval estimates for parameters, hierarchical Bayesian modeling can produce full inference in parameters by taking into account heterogeneity effects, spatial autocorrelation, and covariate effects. A hierarchical Bayesian approach is effective for updating spatial crude risks with high variability and uncertainty, and extracts realistic safety tendencies from noisy crash data. Small mapping units like road segments are necessary for discerning possible risk factors, but the use of small units can cause unstable risk estimates due to small sample sizes and hence maps with high variability. Smoothing is therefore helpful in visualizing possible spatial patterns. The Bayesian modeling approach we applied here demonstrated that the approach is behaving as expected since the relative risk map based on the posterior estimates preserves the high-risk segments for areas with high traffic volumes while smoothing out variability in low traffic areas due to random noise.

Differentiating directions in relative risk analysis can also avoid averaging effects between opposite directions and generates more accurate results for each direction. Currently the US Department of Transportation (DOT) normally selects dangerous segments and acts to improve safety on both directions. If the relative risk is lower than 1 in one direction and larger than 1 in the other, only improving the safety in one direction may be all that is necessary. Therefore, direction differentiation, which is a key feature of our approach, can potentially be used as a more effective and efficient method to allocate safety improvement funds.

While this paper has filled in several gaps in the literature on the spatial–temporal analysis of motor vehicle crashes, many areas still require future research. Our analysis of relative crash risk does not consider the severity of the crash. Theoretically, the distribution of relative risks of severe crashes (e.g., crashes with a fatality or multiple injuries) is not always the same as those of property damage only (PDO). Risk maps help to determine where and

when each roadway segment has a high relative crash risk, but this approach does not answer why some segments are riskier or what can be done to improve their safety. Therefore, contributing factors must be determined, not only for the whole county area, but also for some specific high-risk roadway segments due to the fact that each road segment has its own traffic volumes, road characteristics, and environmental conditions.

Acknowledgements

The authors would like to thank (without implicating) Texas Department of Transportation (TxDOT) for the crash data, Mr. Emmanuel Samson at TxDOT Houston District for the traffic data, TTI for the VMT data, and Dr. Dominique Lord at Texas Transportation Institute (TTI) and Dr. Ned Levine at Houston-Galveston Area Council (H-GAC) who kindly provided constructive comments on an earlier draft. Thanks are also due to Jose Gavinha who provided extensive editorial comments on the final draft. Two anonymous reviewers have also provided constructive comments that have substantially improved the overall quality of this paper. We take full responsibility for any remaining errors.

References

Arató, M.N., Dryden, I.L., Taylor, C.C., 2006. Hierarchical Bayesian modelling of spatial age-dependent mortality. Computational Statistics & Data Analysis, doi:10.1016/j.csda.2006.02.007.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society, Series B 36, 192–236.

Besag, J., Newell, J., 1991. The detection of clusters in rare diseases. Journal of the Royal Statistical Society, Series A 154, 143–155.

Best, N.G., Waller, L.A., Thomas, A., Conlon, E.M., Arnold, R.A., 1999. Bayesian models for spatially correlated diseases and exposure data. In: Bernardo, J.M. et al. (Eds.), Bayesian Statistics 6. Oxford University Press, Oxford.

Black, W.R., 1991. Highway accidents: a spatial and temporal analysis. Transportation Research Record 1318, 75–82.

Black, W.R., 1992. Network autocorrelation in transport network and flow systems. Geographical Analysis 24 (3), 207–222.

Black, W.R., Thomas, I., 1998. Accidents in Belgium's motorways: a network autocorrelation analysis. Journal of Transport Geography 6 (1), 23–31.

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., Spicer, R., 2002. The economic impact of motor vehicle crashes, 2000, Report DOT HS 809 446. National Highway Traffic Safety Administration, Washington, DC.

Bolstad, W.M., 2004. Introduction to Bayesian Statistics. John Wiley & Sons, New York.

Brodsky, H., Hakkert, A.S., 1983. Highway crash rates and rural travel densities. Accident Analysis and Prevention 15 (1), 73–84.

Brüde, U., Larsson, J., 1988. The use of prediction models for eliminating effects due to regression-to-the-mean in road accident data. Accident Analysis and Prevention 20 (4), 299–310.

Carlin, B.P., Louis, T.A., 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall/CRC, New York.

Congdon, P., 2001. Bayesian Statistical Modelling. John Wiley & Sons, Chichester, UK.

El-Sadig, M., Norman, J.N., Lloyd, O.L., Romilly, P., Bener, A., 2002. Road traffic accidents in the United Arab Emirates: trends of morbidity and mortality during 1977–1998. Accident Analysis and Prevention 34 (4), 465–476.

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

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. Accident Analysis and Prevention 35 (6), 991–1004.

Fridstrøm, L., Ingebrigtsen, S., 1991. An aggregate accident model based on pooled, regional time-series data. Accident Analysis and Prevention 23 (5), 363–378.

Golias, J.C., 1992. Establishing relationships between accidents and flows at urban priority road junctions. Accident Analysis and Prevention 24 (6), 689–694.

Golob, T.F., Recker, W.W., Levine, D.W., 1990. Safety of freeway median high occupancy vehicle lanes: a comparison of aggregate and disaggregate analyses. Accident Analysis and Prevention 22 (1), 19–34.

Haight, F.A., Olsen, R.A., 1981. Pedestrian safety in the United States: some recent trends. Accident Analysis and Prevention 13 (1), 43–55.

Hauer, E., 1992. Empirical Bayes approach to the estimation of unsafety: the multivariate regression method. Accident Analysis and Prevention 24 (5), 457–477.

Hauer, E., 2002. Estimating safety by the empirical Bayes method. Transportation Research Record 1784, 126–131.

Jegede, F.J., 1988. Spatio-temporal analysis of road traffic accidents in Oyo State, Nigeria. Accident Analysis and Prevention 20 (3), 227–243.

Jones, B., Janssen, L., Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention 23 (4), 239–255.

Kam, B.H., 2003. A disaggregate approach to crash rate analysis. Accident Analysis and Prevention 35 (5), 693–709.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.

Lee, P.M., 2004. Bayesian Statistics: An Introduction, third ed. Arnold, London.

Levine, N., Kim, K., Nitz, L., 1995a. Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. Accident Analysis and Prevention 27 (5), 663–674.

Levine, N., Kim, K., Nitz, L., 1995b. Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. Accident Analysis and Prevention 27 (5), 675–685.

Levine, N., Kim, K., Nitz, L., 1995c. Daily fluctuations in Honolulu motor vehicle accidents. Accident Analysis and Prevention 27 (6), 785–796.

MacNab, Y., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Accident Analysis and Prevention 36 (6), 1019–1028.

Mather, F.J., Chen, V.W., Morgan, L.H., Correa, C.N., Shaffer, J.G., Srivastav, S.K., Rice, J.C., Blount, G., Swalm, C.M., Wu, X., Scribner, R.A., 2006. Hierarchical modeling and other spatial analyses in prostate cancer incidence data. American Journal of Preventive Medicine 30 (2), S88–S100, Suppl. 1.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binominal regressions. Accident Analysis and Prevention 26 (4), 471–482.

Miaou, S.-P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25 (6), 689–709.

Miaou, S.-P., Song, J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis and Prevention 37 (4), 699–720.

Miller, H.J., Shaw, S.-L., 2001. Geographic Information Systems for Transportation: Principles and Applications. Oxford University Press, New York, NY.

Mountain, L., Fawaz, B., Jarrett, D., 1996. Accident prediction models for roads with minor junctions. Accident Analysis and Prevention 28 (6), 695–707.

National Highway Traffic Safety Administration, 2005. Motor vehicle traffic crashes as a leading cause of death in the United States, 2002. Traffic Safety Facts Research Note. NHTSA Washington DC. <http://www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/RNotes/2005/809831.pdf> (accessed 09/27/05.).

Ng, K., Hung, W., Wong, W., 2002. An algorithm for assessing the risk of traffic accident. Journal of Safety Research 33 (3), 387–410.

Nicholson, A.J., 1985. The variability of accident counts. Accident Analysis and Prevention 17 (1), 47–56.

Okamoto, H., Koshi, M., 1989. A method to cope with the random errors of observed accident rates in regression analysis. Accident Analysis and Prevention 21 (4), 317–332.

Petch, R.O., Henson, R.R., 2000. Child road safety in the urban environment. Journal of Transport Geography 8 (3), 197–211.

Riccio, A., Barone, G., Chianese, E., Giunta, G., 2006. A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. Atmospheric Environment 40 (3), 554–566.

Rossi, P.E., Allenby, G.M., McCulloch, R., 2006. Bayesian Statistics and Marketing. Wiley, Hoboken, NJ.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability process: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829–837.

Skabardonis, A., Chira-Chavala, T., Rydzewski, D., 1998. The I-880 field experiment: effectiveness of incident detection using cellular phones. California PATH Research Report, UCB-ITS-PRR-98-1.

Spiegelhalter, D.J., Thomas, A., Best, N., Lunn, D., 2004. WinBUGS User Manual Version 1.4.1. Medical Research Council Biostatistics Unit, Cambridge.

Steenberghen, T., Dufays, T., Thomas, I., Flahaut, B., 2004. Intra-urban location and clustering of road accidents using GIS: a Belgian example. International Journal of Geographical Information Science 18 (2), 169–181.

Thill, J.-C. (Ed.), 2000. Geographic Information Systems in Transportation Research. Pergamon, New York, NY.

Thomas, A., Best, N., Lunn, D., Arnold, R., Spiegelhalter, D.J., 2004. GeoBUGS User Manual Version 1.2. Medical Research Council Biostatistics Unit, Cambridge.

US Department of Transportation, Bureau of Transportation Statistics, 2001. Transportation Statistics Annual Report 2000, BTS01-02. Washington, DC. <http://www.bts.gov/publications/tsar/2000/index.html> (accessed 10/10/04.).

US Department of Transportation, National Highway Traffic Safety Administration, 2004. Traffic Safety Facts 2002: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. Washington DC. <http://www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSFAnn/TSF2002Final.pdf> (accessed 03/12/05.).

Wakefield, J.C., Morris, S.E., 2001. The Bayesian modeling of disease risk in relation to a point source. Journal of the American Statistical Association 96 (453), 77–91.

Wakefield, J.C., Best, N.G., Waller, L., 2000. Bayesian approaches to disease mapping. In: Elliott, P., Wakefield, J.C., Best, N.G., Briggs, D.G. (Eds.), Spatial Epidemiology: Methods and Applications. Oxford University Press, pp. 104–127.

Waller, L., Carlin, B., Hong, X., Gelfand, A., 1997. Hierarchical spatio-temporal mapping of disease rates. Journal of the American Statistical Association 92, 607–617.

Winkler, R.L., 2003. An Introduction to Bayesian Inference and Decision, second ed. Probabilistic Publishing, Boston, MA.

Withers, S.D., 2002. Quantitative methods: Bayesian inference, Bayesian thinking. Progress in Human Geography 26, 553–566.

Yamada, I., Thill, J.-C., 2004. Comparison of planar and network K-functions in traffic accident analysis. Journal of Transport Geography 12, 149–158.

Yamada, I., Thill, J.-C., (forthcoming). Local indicators of network-constrained clusters in spatial point patterns. Geographical Analysis.