BAYESIAN HIERARCHICAL POINT-PATTERN-BASED INTENSITY MODEL IN PREDICTION OF HIGHWAY LOSSES

by

Yongping Yan

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

_____	Dr. Edward J. Wegman, Dissertation Director
_____	Dr. Daniel Carr, Committee Member
_____	Dr. Clifton D. Sutton, Committee Member
_____	Dr. David Wong, Committee Member
_____	Dr. William F. Rosenberger, Department Chair
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date: _____	Fall Semester 2013
	George Mason University
	Fairfax, VA

Bayesian Hierarchical Point-Pattern-Based Intensity Model in Prediction of Highway Losses

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Yongping Yan
Master of Science
George Mason University, 2002

Director: Edward J. Wegman, Professor
Department of Statistics

Fall Semester 2013
George Mason University
Fairfax, VA

DEDICATION

This dissertation is dedicated to my family, Lydia, Rebekah, Joshua, Chunqiu, Yaping, and my academic advisor, Dr. Edward J. Wegman. Their encouragement has made this dissertation possible.

ACKNOWLEDGEMENTS

Completing my PhD degree is probably the most challenging activity of my life in a time I need to balance roles as a father of two little kids, a full time employee, and a Ph.D. student, in my long doctoral journey. Fortunately, I have the privilege to study in Statistics Department of George Mason University where excellent professors not only teach and research but also help and encourage.

First and foremost, I wish to thank my advisor, Professor Dr. Edward J. Wegman. He has taught me the very meaning of research and opened that door for me. He is a master in advising and has every skill to inspire, mentor, and support.

I would also like to thank my committee member, Dr. Daniel Carr. He has offered precious guidance in clustering algorithm and visualization methodology.

I also thank my committee member, Dr. Clifton Sutton. Materials from his class, Statistical Inference, were invaluable.

A special thank to my committee member, Dr. David Wong. Because of his wisdom I have chosen a more appropriate data for the application. I also appreciate his help on the GIS software.

I also acknowledge my organization, HLDI, which has funded this research. Especially, I would like to thank Mr. Matthew Moore, vice president of HLDI. Without his consistent support and encouragement I could not finish this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Bayesian Hierarchical Model ..... BHM
Expectation Maximization ..... EM
Finite Mixture Model ..... FMM
Classification and Regress Tree ..... CART
Completely Spatial Randomness ..... CSR
Geographic Information System ..... GIS
Fatality Analysis Reporting System ..... FARS
American Community Survey ..... ACS
Highway Loss Data Institute ..... HLDI
National Highway Traffic Safety Administration ..... NHTSA

ABSTRACT

BAYESIAN HIERARCHICAL POINT-PATTERN-BASED INTENSITY MODEL IN PREDICTION OF HIGHWAY LOSSES

Yongping Yan, Ph.D.

George Mason University, 2013

Dissertation Director: Dr. Edward J. Wegman

Traditional spatial-temporal models either use separable models to separate spatial processes from temporal processes, which often results in a loss of information, or use nonseparable models through the introduction of correlation functions. These functions typically have to be complicated enough to address the real problem and additionally the implementation requires the integral of these functions. In this dissertation, with a focus on contribution to the interdisciplinary area of statistics and GIS (geographic information system), I have developed methods extending EM (expectation-maximization) algorithm to Poisson point processes with incomplete data structure to undercover the underlying components characterizing highway loss events. With component information in the dissertation, I have developed methods that use classification and regression trees along with visualization procedures to identify key features influencing highway loss intensities, and detect key feature patterns of the "hot spot" loss areas. Instead of

examining the correlation between spatial space and temporal space, I have developed methods using a *k*-means based algorithm and specially tailored distance functions to partition the key feature space into homogeneous clusters, and map this partition to the spatial space partition. Then, I have built the Bayesian hierarchical model (BHM) that use the current time point loss information and most recent past loss information to predict the future losses for each cluster. The BHM in this dissertation has a good updating mechanism and is adaptive. Finally, I have successfully applied the methods to 2009-11 FARS (Fatality Analysis Reporting System) data of U.S. Department of Transportation. The application is a good example that methods developed in this dissertation can be widely used on any loss types whose events exhibit a Poisson-point-pattern.

Key words: spatio-temporal model, expectation maximization (EM), Poisson-point-process, Bayesian hierarchical model (BHM)

1. Introduction

Every year, thousands of people died of traffic accidents or got injured and billions of dollars were lost to individuals, institutes, and insurance industry. Efforts and countermeasures that can reduce these losses will definitely benefit the whole of society. This dissertation contributes to this mission by investigating how highway crash events are distributed in spatial and temporal domains. In this chapter, I introduce the dissertation topic, review literature of past research, and set goals for this dissertation.

1.1 Highway Losses and Measurements

1.1.1 Highway Loss Types and Auto Insurance Coverage

Highway losses include deaths, injuries and property damage. The loss of lives is almost always related to injuries and property damage. Injury and property losses sustained on highways are relatively more complicated and a good way to understand highway losses is through auto insurance claim data. Auto insurance (also known as car insurance) is insurance purchased for automobiles. Auto insurance provides protection against losses incurred as a result of traffic accidents and against liability that could be incurred in an accident. There are two insurance systems in United States, the Tort Insurance System and the No-Fault Insurance System. Under Tort Insurance System (White, 2003), a person who

suffers legal damages (loss or injury) as the result of a crash may be able to use tort law to receive compensation from someone who is legally responsible or liable, for those losses or injuries. Generally speaking, tort law defines what constitutes a legal loss or injury and establishes the circumstances under which one person may be held liable for another's loss or injury. In contrast, under No-Fault Insurance System (Insurance Information Institute, 2010) insureds are indemnified for losses by their own insurance company, regardless of fault in the incident generating losses; furthermore insureds are also restricted in the right to seek recovery through the civil-justice system for losses caused by other parties.

Auto insurance coverages can be classified into two categories: coverages against damages to a vehicle and other property, and coverages against injuries to occupants and other people (Insurance Institute of Highway Safety, 2010). The first category includes:

i. **Collision coverage** insures against physical damage sustained in a crash to the insured people's own vehicles if they are at fault. The damage may occur from striking another vehicle or an object such as a tree or pole.

ii. **Property damage liability** coverage insures against the physical damage that at-fault people's vehicles inflict on other vehicles and property.

iii. **Comprehensive coverage** insures against losses from the theft of an insured person's vehicle or vehicle damage for reasons other than crashes. It covers theft, noncrash fire (fire not caused by a collision or vandalism), glass damage caused by rocks and other objects, and other kinds of damage such as from hitting animals, acts of nature, and vandalism.

Injury insurances include:

iv. Personal injury protection coverage insures against medical, hospital, and other expenses for injuries sustained in crashes with insured drivers and other people in their vehicle, regardless of who is at fault in the collision. This coverage is sold in states with no-fault insurance systems. The upper limit of the amount paid to insureds varies by state.

v. Medical payment coverage insures against injuries sustained by insured people in crashes for which they are responsible. It also covers injuries to other occupants in their vehicles. This coverage is only sold in states with a tort insurance system.

vi. Bodily injury liability coverage insures against medical, hospital, and other expenses for injuries that at-fault drivers inflict on occupants of other vehicles or others on the road.

1.1.2 Measurements of Highway Losses

Highway Losses are usually measured by frequency (rate) and severity (size, not applicable to deaths). Two main factors determine auto insurance losses; claim frequency and claims severity. Claim frequency, which is how often claims are filed, is usually measured in claims per 100 insured vehicle years. Claim severity, which is how big the claim payments are, depending on the average loss payment per claim, is measured in dollars. These two factors combine to indicate the average loss payment per insured vehicle year, also known as overall loss. The overall loss is the average cost of insuring a vehicle for one year, excluding administrative costs.

Generalized linear models (GLM), as defined by Nelder and Wedderburn (1972), have been commonly used to quantify the claim frequency and claim severity in the auto insurance industry. The sense of linear lies in the form

$$g(y_i) = x_i'\beta + \epsilon_i \tag{1.1}$$

where y_i is the response variable for the *i*th observation. g is a monotonic differentiable link function, x_i is a column vector of covariates, or explanatory variables, β is a vector of unknown parameters, and ϵ_i is assumed to be independent and identically distributed random variables with zero mean and constant variance. In generalized linear models, the response is assumed to possess a probability distribution of the exponential family. That is the probability density of the response Y for continuous/discrete responses can be expressed as

$$f(y) = exp\{\tfrac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\} \tag{1.2}$$

for some functions a, b, and c that determine the specific distributions. Auto insurance claims are widely accepted to have a Poisson distribution, while claim severities have a Gamma distribution. Maximum likelihood fitting is used to estimate β. In the last decade, two well-developed extensions of GLM were also introduced into auto insurance industry, one is the generalized linear mixed model (GLMM) which first appeared in Laird and Ware (1982), which adds random effects along with fixed effect into the model. The second extension is Generalized Estimating Equations (GEEs), introduced by Liang and Zeger (1986), which targeted handling correlated responses.

1.1.3 Highway Losses and Spatial-Temporal Patterns

The concept of spatial clusters or "hot spots" mainly arose from research in criminal activity and disease incidence. It is well known that crimes tend to cluster in so-called "hot spots" (e.g. convenience stores or bars). Crimes committed by serial criminals often follow established spatial patterns. The "hot spot" phenomenon also exists in highway losses. Based upon the theft claim data from the HLDI (Highway Loss Data Institute) database, for 2003 model year vehicles during time period from 2002 to 2009, the zip code area with the highest theft claim frequency was zip code 48205 of Detroit. This zip code area had theft claim frequencies 24 times that of the national average. The 48205 zip code area is blocks from I-94 and is very close to Canada. An reasonable assumption is the stolen vehicles can be easily transported, either by sea or highway.

In April 2008, HLDI produced an insurance special report on theft losses by county comparing the 2006-07 result with that of 1998-99. Insurance loss results from that report showed theft overall losses (average loss payment per insured vehicle year) increased in the southwest and along the Mexican border. The seven counties with the highest overall theft losses in 2006-07 all border Mexico and they had loss results more than 5 times that of national average. Counties in the Detroit, Miami and New Orleans areas also had theft overall losses much higher than the national average. The report also showed that theft overall losses declined in the New York and Philadelphia regions. Seven of the 10 counties with the highest theft overall losses in 1998-99, all in the New York or Philadelphia metropolitan areas, were no longer among the top 10 in 2006-07.

Another kind of auto insurance loss, animal strikes which are covered under comprehensive coverage, shows not only spatial patterns but also temporal patterns. Animal strike claim frequencies vary with calendar years as well as seasons (HLDI, 2008). In April 2008 HLDI reported that national claim frequencies for animal strikes were lowest in August (3.9 claims per 1,000 insured vehicle years) and highest in November (14.1 claims 1,000 insured vehicle years), and claim frequencies in August were about one quarter of that in November. Three states had the highest November claim frequencies (West Virginia, Pennsylvania, and Kentucky) and two states had very low November claim frequencies (Arizona and Florida). Claim frequencies for West Virginia, Pennsylvania, and Kentucky followed the national seasonality trend. In contrast, there was little variation in claim frequencies for Arizona and Florida.

Predicting and further controlling the death, injury and property damage occurring in the "hot spots" of a hot area benefit both the public and the auto insurance industry. For insurers, pricing insurance premiums in hot areas is of special interest to the cost control and the marginal profit rate. Setting competitive auto insurance premium while minimizing claim loss payments in a specific area, especially an area with heavy vehicle density, is not only essential to an insurer's core competence but also very important to public safety. Traditional claim prediction models like the GLM model, with its extensions GLMM and GEE, and GAM (Generalized Additive Models, Hastie and Tibshirani, (1986 and 1990)) have limited ability in dealing with this problem.

1.2 Introduction of Spatial-Temporal Models

Spatial-temporal models arise from analysis of data collected across time as well as space. A typical example is the climate data collected from a network of meteorological stations, at regular intervals, say every week, over decades. The observed data at each monitor typically are not independent but form a time series. At each time point the data collected from all monitors construct a spatial structure and therefore spatial dependence must be taken into consideration.

1.2.1 Initial Spatial-Temporal Analysis of Acid Rain in New York

One early paper on spatial-temporal statistics was published by Bilonick and Nichols (1983). The authors analyzed the rainfall data from 22 stations in or near New York state, collected from 1965 to 1979. Variables measured included acidity (pH), and concentrations of sulfates, nitrates, and calcium, as well as the amount of rainfall, in milliequivalents per liter (meq l^{-1}). The data were summarized into monthly values at each station to perform a time series analysis to determine whether an increasing trend existed over the time period for the variables measured.

- $H_{x,t},$ total deposition of hydrogen ion in month t at location x.
- $S_{x,t},$ total deposition of sulfate in month t at location x.
- $N_{x,t},$ total deposition of nitrate in month t at location x.
- $C_{x,t},$ total deposition of calcium in month t at location x.

These data were then aggregated across stations into monthly temporal data:

$$H_t = \frac{1}{n_{H,t}} \sum_x H_{x,t}$$

$$S_t = \frac{1}{n_{S,t}} \sum_x S_{x,t}$$

$$N_t = \frac{1}{n_{N,t}} \sum_x N_{x,t}$$

$$C_t = \frac{1}{n_{C,t}} \sum_x C_{x,t}$$

where $n_{H,t}$ denotes the number of stations reporting hydrogen ion deposition in month t, and similarly for $n_{S,t}, n_{N,t}$, and $n_{C,t}$. The authors applied ARIMA model to the four time series and concluded "... there is no evidence for a long-term change in the mean level of acidity. The observed patterns in the hydrogen ion data can be completely explained in terms of a stationary ARIMA model."

In contrast with the initial temporal analysis looking at the trend over time, Bilonick (1983) applied a pure spatial analysis on the same data to determine whether a spatial pattern existed on the monthly precipitation $P_{x,t}$ series and deposition $D_{x,t}$ series. Bilonick created monthly semivariograms and aggregated them across months, and then chose the parametric spherical model to fit the empirical semivariograms, separately for $P_{x,t}$ and $D_{x,t}$. Kriging point average estimates and corresponding mean squared errors were derived based on the fitted semivariograms and aggregated to get block average estimates, with each block covering an area of 80 km^2. Concentration of H^+ was defined as D/P, and approximation was used for the variance of the ratio. The resulting maps predicted a "weak tendency" that concentrations decreased in moving from West to East, and suggested this tendency was barely significant.

Bilonick (1985) continued the work done in the above analysis and extended it into a fully spatial-temporal analysis. This time the variable of interest was either sulfate concentration, measured in milligrams per liter ($mg\ l^{-1}$), or sulfate

deposition, measured in kilograms per hectare per year ($kg \;\; ha^{-1}y^{-1}$). For $D_{x,t}$ series, a space-time semivariogram was computed of the form $\gamma(h,t)$, defined by

$$2\gamma(h,t) = E\{\mathrm{Z}_{x1,t1} - Z_{x2,t2} \,:\, ||x_1 - x_2|| = h, |t_1 - t_2| = t\} \qquad (1.3)$$

where $||x_1 - x_2||$ denotes the Euclidean distance between x_1 and x_2, and it is assumed that (1.3) is stationary and isotropic in space and stationary in time.

The sample semivariogram was computed using the Methods of Moments where pairs of observations ($\mathrm{Z}_{x1,t1}$, $Z_{x2,t2}$) were grouped into bins according the values of h and t. The author proposed a parametric model after examining the space-time semivariogram graphically by form

$$\gamma = \gamma_0 + \gamma_P + \gamma_S + \gamma_L \qquad (1.4)$$

where

$$\gamma_0 = C_{0,}$$

$$\gamma_P = C_p\{1 - 0.5\, cos(\tfrac{2\pi t}{365})\},$$

$$\gamma_S = \begin{cases} 0 & \text{if } t = 0 \\ C_S\{3t(2a_S)^{-1} - 2^{-1}t_3 a_S^{-3}) & \text{if } 0 < t < a_S \\ C_S & \text{if } t \geq a_S \end{cases}$$

$$\gamma_L = k_L h$$

- γ_0 is the nugget effect,

- γ_P represents the periodic effect and t is measured in days, consideration of seasonality is reflected in this term,

• γ_S is interpreted as the aperiodic residual effect,

• γ_L is modeled as linear in h. This component reflects the spatial part of the variogram.

Parameter estimates were C_0 = 100, C_p = 300, C_S = 150, a_S = 30, k_L = 1.0. The fitted semivariograms are shown in Figure 1.1 and Figure 1.2.

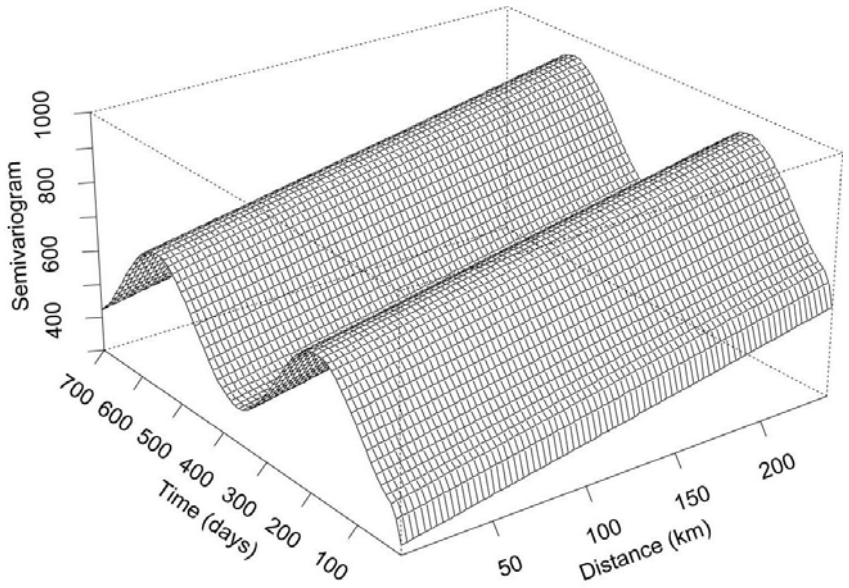


Figure 1.1 Fitted semivariogram from equation (2.1).Bilonick (1985)

Point and block calculation procedures similar to that in the pure spatial analysis were applied to the collected data to predict the sulfate deposition. The resulting maps for each year from 1966 to 1975 showed clear differences in the spatial

pattern of deposition from year to year, but no evidence of overall temporal trend was found, though the deposition showed a small peak around 1972. Figure 1.2 illustrates the temporal variogram for a single location. The inclusion of a periodic effect in equation (1.4) is distinctly reflected in the map, indicating seasonality in the data.

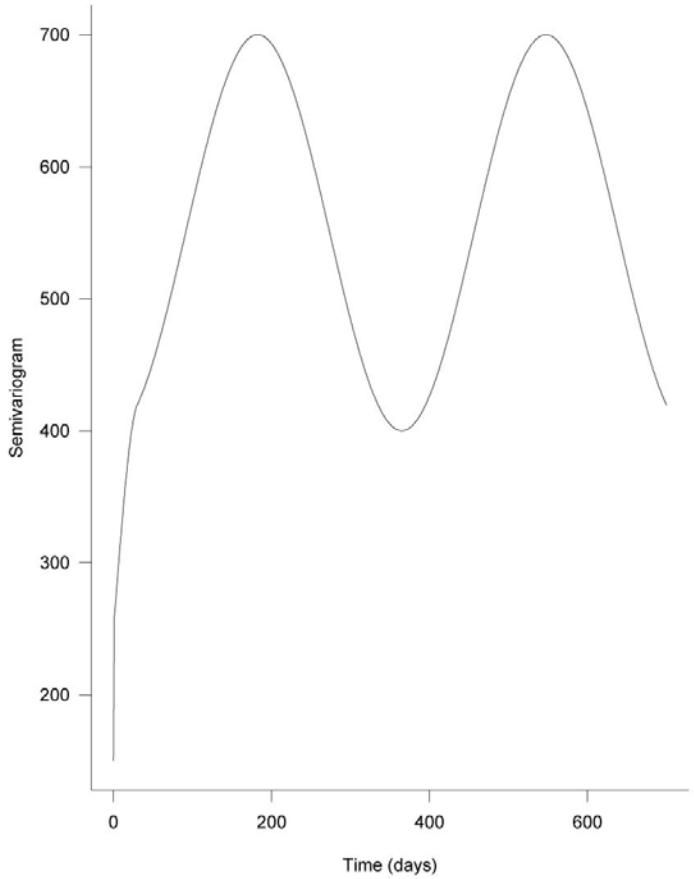


Figure 1.2 The temporal variogram for a single location, $\gamma(0,t)$. Bilonick (1985)

Biloncik (1988) applied "indicator kriging" (Cressie 1993, pp.281-283) to the previous spatial-temporal model on other data collected from 35 stations in the states of New York, Pennsylvania, West Virginia, Virginia, Ohio, Indiana, Kentucky, Illinois plus Ontario (Canada), in the time period from July 1982 to September 1984. The fitted semivariogram was composed of a pure temporal semivariogram term, and another term that was effectively of a "geometrically anisotropic" form in the space and time variables. The estimated spatial-temporal variogram can be used to construct maps of the estimated median, as well as other quantiles, of hydrogen ion (H^+), the main variable of interest.

1.2.2 Egbert and Lettenmaier's Multivariate Space-Time Model

Egbert and Lettenmaier (1986) introduced a rather general class of multivariate space-time models based upon the analysis of National Atmospheric Deposition Program (NADP) data, produced from weekly observations of 10 ionic species of the monitoring network. The featured data exhibited spatial dependence for both the long-term and the short-term averages, plus seasonality, for multiple components.

Egbert and Lettenmaier divided each year into four 3-month seasons and fitted the following basic temporally stationary model to the data in each season,

$$Z_{st}^{p}(x) = W_{st}^{p}(x) + Y_{s}^{p}(x) + M^{p}(x) \qquad (1.5)$$

where $Z_{st}^{p}(x)$ denotes the *p*th component of the observed process in year s, week t and location x ($1 \leq p \leq P, 1 \leq s \leq S, 1 \leq t \leq T$); $W_{st}^{p}(x)$ is the weekly variation in year s, week t and location x; $Y_{s}^{p}(x)$ is the yearly variation in year s and location x; $M^{p}(x)$ represents the long-term effect. Assumption are,

- means of $W_{st}^{p}(x)$ and $Y_{s}^{p}(x)$ are 0,

- $E\{W_{st}^{p}(x)W_{st}^{p'}(x')\} = k_{W}^{pp'}(x,x')$ is a smooth function of either $x-x'$ for stationary case or $|x-x'|$ for stationary and isotropic case,

- $E\{W_{st}^{p}(x)W_{s't'}^{p'}(x')\} = \delta_{ss'}k_{W}^{pp'}(x-x',t-t')$, in which $\delta_{ss'}$ is the Kronecker delta function, $\delta_{ss'} = (1 \text{ if } s=s', 0 \text{ otherwise})$,

- $E\{Y_{s}^{p}(x)Y_{s'}^{p'}(x')\} = \delta_{ss'}k_{Y}^{pp'}(x-x')$, in which $\delta_{ss'}$ is the defined the same as above,

- $E\{M^{p}(x)\} = \mu^{p}$,

- $\frac{1}{2}E[\{M^{p}(x)-M^{p}(x')\}\{M^{p'}(x)-M^{p'}(x')\}] = \gamma^{pp'}(x-x')$.

Thus, the W and Y processes in different years are uncorrelated. Two scenarios are discussed in fitting the model, with the first case assuming time-independent weekly effects, $k_{W}^{pp'}(x,t) = \begin{cases} k_{W}^{pp'}(x) & \text{for } t \neq 0, \\ 0 & \text{for } t = 0. \end{cases}$

The second case assumed temporal autocorrelation, but Egbert and Lettenmaier assumed weekly independence after all lags greater than some T_0, and derived a series of equations of the form

$$E\{\sum_{\mathrm{s=1}}^{\mathrm{S}}\sum_{t'=1}^{T-t}(Z_{st'i}^{p} - \overline{Z}_{si}^{p})(Z_{st'i}^{p'} - \overline{Z}_{si}^{p'}\} = \sum_{t'=0}^{T}a_{ii'tt'}^{pp'}k_{W}^{pp'}(x_i - x_{i'},t)$$

for $0 \leq t \leq T_0$, where $\overline{Z}_{si}^{p} = \frac{\sum_{t} Z_{sti}^{p}}{T_{si}}$, T_{si} is the number of observed data points at x_i in year s, x_i is the ith sampling location. Egbert and Lettenmaier developed estimation techniques in analogous to the three-way analysis of variance to

estimate $k_W^{pp'}$, $k_Y^{pp'}$, $\gamma^{pp'}$ and other parameters like $a_{ii'tt'}^{pp'}$, details of the exploitation can be found in their paper.

Egbert and Lettenmaier applied this method to the data collected from a network of 51 sites in the northeast U.S. in years 1980 and 1981. The data were subdivided into four seasons and three variables were considered, pH, precipitation and sulfate acidity. Some of the main findings were:

- Some mild temporal autocorrelation was seen in the precipitation data, but no autocorrelation was found in other two variables.

- Little "yearly" effect was found once masked by weekly and long-term effects.

- Seasonal effect was strong for spatial ranges.

- Spatial correlation for sulfate concentrations were stronger than for pH.

1.2.3 Empirical Orthogonal Functions

Cane et al. (1996) implemented a reduced dimension space-time dynamic model using Kalman filter via empirical orthogonal function basis functions, in simulating tropical Pacific sea level by linear wind driven models. The method of empirical orthogonal function (EOF) analysis is a decomposition of a data set in terms of orthogonal basis functions. The ith basis function is chosen to be orthogonal to the basis functions from the first through $i-1$, and to minimize the residual variance. It is the same as performing a principal components (PC) analysis on the data, except that the EOF method finds both time series and spatial patterns. The basis functions are typically found by computing the eigenvectors of the covariance matrix of the data set. The dimension reduction in

Cane et al. (1996) made the calculation highly feasible. Cane et al. (1996) compared the reduced state space filter with a full grid point Kalman filter using the same dynamic model and concluded that results were not inferior to the full grid point filter even when the reduced filter retained only nine EOFs from 297 time series.

1.2.4 Stein's Spatial Processes Model

Stein (1986) proposed a model with the form

$$z(x,t) = m(x) + \mu(t) + e(x,t) \tag{1.6}$$

in which $z(x,t)$ are nk observed space-time point values $(x_{i,}t_j)$, $i = 1, ..., n$, $j = 1, ..., k$, $\mu(t)$ are fixed time constants with unknown values $\mu(t_1), \ldots, \mu(t_k)$ while $m(x)$ and $e(x,t)$ are both random processes satisfying

$$E\{e(x,t)\} = 0\,,$$

$$\tfrac{1}{2}E[\{e(x,t)\text{-}e(x',t)\}^2] = \gamma\,(x - x')\,,$$

$$E[\{e(x,t_i)\text{-}e(x',t_i)\}\{e(x'',t_j)\text{-}e(x''',t_j)\}] = 0 \quad \text{for } i \neq j,$$

$$E\{m(x)\} = 0,$$

$$\tfrac{1}{2}E[\{m(x)\text{-}m(x')\}^2] = \eta\,(x - x')\,.$$

Thus the model includes two noise processes with zero mean, one spatial process $m(x)$ with known semivariogram η independent of time point, and another spatial process $e(x,t)$ that is generated independently at each time point with known semivariogram γ.

Based upon this framework Stein simplified functions of the model into prediction of 2 problems:

1. Predict $z(x_0, t_\beta)$ for any arbitrary $x_0 \notin \{x_1, ..., x_n\}$ for any observed time point t_β, $t_\beta \in \{t_1, ..., t_k\}$.

2. Predict differences of spatial averages across different time points

$\frac{1}{|R|} \int_R \{z(x, t_\alpha) - z(x, t_\beta)\}\, dx,$ where $|R|$ denotes the area of R.

In the solution of the first problem Stein proved that the optimal predictor of $z(x_0, t_\beta)$, $\widehat{z}(x_0, t_\beta)$, is a linear combination of $z(t_\beta)$, the vector of observations at time t_β, and the vector of time-averaged responses, $\overline{z} = \frac{1}{k}\sum_{i=1}^{k} z(t_i)$.

For problem 2, Stein showed that the optimal kriging solution is a function of only pairwise differences, $z(x_i, t_\alpha) - z(x_i, t_\beta)$, $i = 1, \ldots, n$. He also pointed out the predictor based upon the pairwise differences was superior to the alternative solution in which both

$\frac{1}{|R|} \int_R \{z(x, t_\alpha) dx$ and $\frac{1}{|R|} \int_R \{z(x, t_\beta) dx$ were predicted respectively.

1.2.5 Cressie and Huang's Covariance Function Approach

Cressie and Huang (1999) proposed a generic approach, the nonseparable, spatio-temporal stationary covariance function, which generalized the separable space-time covariance structure of Matern (1986) used in pure spatial processes.

Cressie and Huang constructed a stationary spatio-temporal covariance function from Bochner's theory (Bochner, 1955) of the form

$$C(\boldsymbol{h},\mu) = \int\int e^{i(\boldsymbol{h}^T\boldsymbol{\omega}+\mu\tau)} g(\boldsymbol{\omega},\tau)\, d\boldsymbol{\omega} d\tau \tag{1.7}$$

in which $\boldsymbol{h}$ is a d-dimensional vector serves as a spatial lag while μ is a scalar time lag, and $g(\boldsymbol{\omega},\tau)$ is the spectral density of the covariance function C, where $\boldsymbol{\omega}$ is d-dimensional and τ is scalar. $C(\,\cdot\,;\,\cdot\,)$ is further assumed to be integrable, then

$$g(\boldsymbol{\omega},\tau) = \tfrac{1}{2\pi} \int e^{-i\mu\tau} h(\boldsymbol{\omega};\mu)\, d\mu \tag{1.8}$$

where

$$\begin{aligned} h(\boldsymbol{\omega};\mu) &\equiv (\tfrac{1}{2\pi})^d \int e^{-i\boldsymbol{h}^T\boldsymbol{\omega}} C(\boldsymbol{h};\mu)\, dh \\ &= \int e^{i\mu\tau} g(\boldsymbol{\omega},\tau) d\tau, \end{aligned}$$

by assuming that

$$h(\boldsymbol{\omega};\mu) = \rho(\boldsymbol{\omega};\mu) k(\boldsymbol{\omega}), \tag{1.9}$$

which satisfies the following two conditions:

(C1) For each $\boldsymbol{\omega} \in \mathcal{R}^d$, $\rho(\boldsymbol{\omega};\,\cdot\,)$ is a continuous autocorrelation function, $\int \rho(\boldsymbol{\omega};\mu) d\mu < \infty$ and $k(\boldsymbol{\omega}) > 0$.

(AC) $\int k(\boldsymbol{\omega}) < \infty$.

Then (1.8) can be written as

$$g(\boldsymbol{\omega},\tau) \equiv \tfrac{1}{2\pi}\, k(\boldsymbol{\omega}) \int e^{-i\mu\tau} \rho(\boldsymbol{\omega};\mu)\, d\mu \; > 0$$

by (AC). Furthermore by

(C2) $\int\int g(\boldsymbol{\omega},\tau)\,d\boldsymbol{\omega}d\tau = \int k(\boldsymbol{\omega}) < \infty.$

Thus (1.7) becomes

$$C(\boldsymbol{h},\mu) = \int\int e^{i\boldsymbol{h}^T\boldsymbol{\omega}}\rho(\boldsymbol{\omega};\mu)\,k(\boldsymbol{\omega})d\boldsymbol{\omega}, \tag{1.10}$$

where $k(\boldsymbol{\omega})$ is the spectral density of a pure spatial process and $\rho(\boldsymbol{\omega};\mu)$ is a valid temporal autocorrelation function in μ for each given $\boldsymbol{\omega}$.

Cressie and Huang developed seven models based upon the covariance function structure built above and here I demonstrate one of them. The others are of similar forms.

Model 1. Let

$$\rho(\boldsymbol{\omega};\mu) = exp\{-||\boldsymbol{\omega}||^2\mu^2/4\}exp\{-\delta\mu^2\}; \qquad \delta > 0,$$

and $k(\boldsymbol{\omega}) = \exp\{\text{-}c_0||\boldsymbol{\omega}||^2/4\}; \quad c_0 > 0.$

The construction of $\rho(\boldsymbol{\omega};\mu)$ and $k(\boldsymbol{\omega})$ satisfies condition of (C1) and (C2), furthermore, from (1.10) and Matern (1960, p.17),

$$C(\boldsymbol{h};\mu) \propto \frac{1}{(\mu^2+c_0)^{d/2}}exp\{-\frac{||\boldsymbol{h}||^2}{(\mu^2+c_0)}\}exp\{-\delta\mu^2\}; \quad \delta > 0,$$

is a continuous spatio-temporal covariance function in $\mathcal{R}^d \times \mathcal{R}$. As $\delta \to 0$, the above formula evolves to

$$C^0(\boldsymbol{h};\mu|\boldsymbol{\theta}) = \frac{\sigma^2}{(a^2\mu^2+1)^{d/2}}exp\{-\frac{b^2||\boldsymbol{h}||^2}{(a^2\mu^2+1)}\},$$

where $\boldsymbol{\theta} = (a, b, \sigma^2)^T$, $a \geq 0$ is the scaling parameter of time and $b \geq 0$ is the scaling parameter of space, and $\sigma^2 = C^0(\mathbf{0}; 0|\boldsymbol{\theta}) > 0$, here c_0 is set to 0.

1.3. Point-Pattern-Based Spatial-Temporal Transition Density Model

According to Diggle (2003, p.1), a spatial point pattern is a set of locations, irregularly distributed within a designated region and presumed to have been generated by some form of stochastic mechanism. Diggle (2003, p.42) further defined spatial point process as a stochastic mechanism which generates a countable set of events x_i in the plane. Stationarity and isotropy are often assumed for these processes, which means all properties of the processes are invariant under translation, and invariant under rotation. It should be noted that these two assumptions do not rule out the random heterogeneity in the modeling. The basic hypothesis for a spatial point pattern is complete spatial randomness (CSR), Diggle (2003, p.6), which asserts that the number of events in any planar region A follows a Poisson distribution with mean $\lambda|A|$, and the given n events x_i are an independent random sample from the uniform distribution on A.

1.3.1 Spatial Clustering in Point Processes

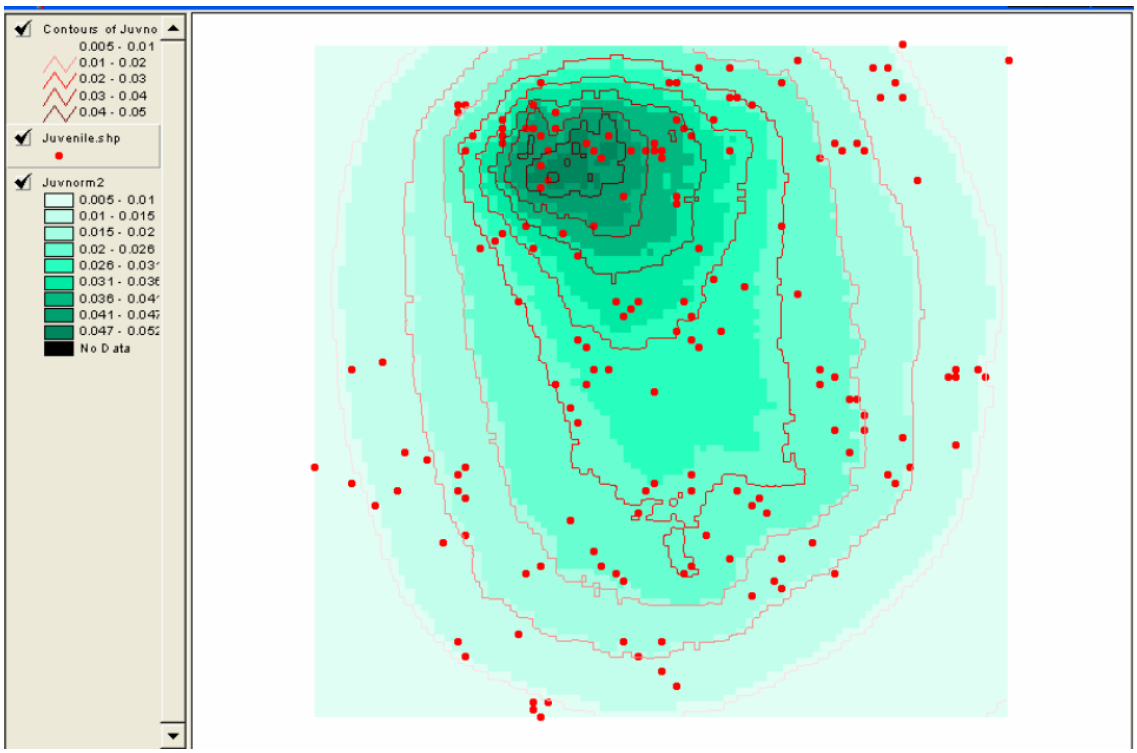


Figure 1.3 Normal density kernel with contour lines of Cardiff juvenile delinquents. Anselin (2003)

A spatial cluster, also know as a "hot spot", is a common phenomenon of point processes in fields like epidemiology and criminology. The normal density kernel with contour lines of Cardiff juvenile delinquents shown in Figure 1.3 is a good illustration of this concept. The plot was created by Anselin (2003) using Ned

Levine's CrimeStat2.0 software package. The spatial clustering formed a good basis for the prevalent spatial forecasting. A widely used method is the Spatial and Temporal Analysis of Crime program (STAC), which clusters crime points within ellipses (Block,1995). Levine (1998) demonstrated the kernel density estimation method shown in Figure 1.3, which extended STAC in a more sophisticated way. Many researchers have investigated spatial decision making by criminals and in their models spatial attributes or features (e.g. distance to a road, type of residential community) serve as predictors to forecast criminal incident. The underlying assumption is that the likelihood of a criminal incident at a specified location is based upon the history of the same type of incident and independent spatial features.

1.3.2 Definition of Brown and Liu's Point-Pattern-Based Density Model

Liu and Brown (2003) proposed a point-pattern-based transition density model derived from the theory of point patterns (Diggle, 1983). Their model extends crime clustering methods by incorporating offender's preferences in crime site selection. The model represents criminal preferences as the functional relationship between demographic, economic, social, victim, and spatial attributes and measure of criminal activity. Liu and Brown gave a formal description of their forecast model.

Denote the locations and times of criminal incidents as $(\boldsymbol{s}_1, t_1), (\boldsymbol{s}_2, t_2), \ldots,$ $t_0 - 0 < t_1 < t_2 < \cdots,$ where $\boldsymbol{s}_i$ is the two-dimensional location of incident i of a given type of crime and t_i is the corresponding time of incident i, $f_1, f_2, \ldots, f_p$ are p measurable features that are believed to be associated with the occurrences of the incidents, $\boldsymbol{x}_i$ is the feature vector consisting of values for p elements at time

t_i. Taken together, $\{\boldsymbol{x}_{\boldsymbol{s},t} \in \mathcal{X} : \boldsymbol{s} \in D, t \in T\}$ formed a marked space-time shock point process (Cressie, 1993), where $t, \boldsymbol{s}$, and $\boldsymbol{x}_{\boldsymbol{s},t}$ are all random quantities defined within study horizon $T \subset \mathcal{R}^+$, a study region $D \subset \mathcal{R}^2$, and a feature space $\mathcal{X} \subset \mathcal{R}^p$, respectively. The reason that the point process is classified as a shock point process instead of a survival process is the events are considered instantaneous.

The measurement of interest is the density of the process, which is the likelihood that a criminal incident occurs within a study region at the future time given the times, locations and features of the past criminal incident of the same type and bounded by the same region and time range. Liu and Brown (2003) defined the transition density in the following equation,

$$\psi_n(\boldsymbol{s}_{n+1}, t_{n+1} | D_n, T_n, \boldsymbol{\mathcal{X}}_n) \equiv \lim_{\nu(d\boldsymbol{s}_{n+1})dt_{n+1} \to 0} \frac{Pr\{N(ds_{n+1}, dt_{n+1})=1 | D_n, T_n, \boldsymbol{\mathcal{X}}_n\}}{\nu(d\boldsymbol{s}_{n+1})dt_{n+1}} \tag{1.11}$$

where $T_n = \{t_1, t_2, \ldots, t_n\}$; $D_n = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\}$, in which $\boldsymbol{s}_i = \{\boldsymbol{s}_{i1}, \boldsymbol{s}_{i2}\}$; $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, in which $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip}\}'$; $\boldsymbol{s}_{n+1}$ and t_{n+1} are the location and time of a future crime incident; $\nu(d\boldsymbol{s}_{n+1})$ is the Lebesgue measure of the infinitesimal region $d\boldsymbol{s}_{n+1}$; $N(ds_{n+1}, dt_{n+1})$ is the count of crime incidents within $d\boldsymbol{s}_{n+1}$ and the infinitesimal time interval dt_{n+1}.

1.3.3 Brown and Liu's Point-Pattern-Based Transition Density Model

1.3.3.1 Model Search Method

Many factors are believed to be related to criminal preferences. Liu and Brown initiate their model by specifying triplet (F, c, s) to reduce the dimension of the

features to form a key feature space, where F is the initial feature set, c is a criterion function defined for subsets of F, and s is a subset search procedure. To measure the cohesiveness of a point pattern observed in the independent variable or defined subspace, they produced an inter-event distance $d_{ij,}$ which is a distance between event i and j in the feature subspace defined by the feature subset to be evaluated, and then transformed into similarity s_{ij} as follows.

$$s_{ij} = \frac{1}{1+\alpha d_{ij}} \tag{1.12}$$

where $\alpha = 1/\overline{d}$ and $\overline{d}$ is the average inter-event distance. Distance refers to differences in value of an independent variable. They further define the Gini index as,

$$g_{ij} = 4s_{ij}(1 - s_{ij}) \tag{1.13}$$

for a data set of n events, the average Gini index is suitable to measure cohesiveness:

$$I_g = \frac{2\sum_{i=1}^{n-1} \sum_{j=1+1}^{n} g_{ij}}{n(n-1)}. \tag{1.14}$$

The smaller the value the I_g index is, the higher the level of point pattern cohesiveness or the better the set of features that define the point pattern. The authors evaluate I_g for each individual feature and select a subset of features based upon the I_g scores. Before the actual calculation of I_g scores, a ratio of r_k is examined, in case the feature values for a large sample of locations uniformly chosen over the study region, called a prior feature data set, are available.

$$r_k = \frac{\max\limits_{x_{ik}, x_{jk} \in E_k} |x_{ik} - x_{jk}|}{\max\limits_{x_{ik}, x_{jk} \in P_k} |x_{ik} - x_{jk}|} \tag{1.15}$$

where E_k and P_k are the event and the prior feature data sets for feature f_k, respectively. If the ratio is sufficiently small, I_g won't be calculated for feature f_k. Otherwise, adjusted $I_g^{(k)}$ is calculated,

$$\text{Adjusted } I_g^{(k)} = \frac{I_g(E_k)}{I_g(P_k)} \tag{1.16}$$

where $I_g(E_k)$ and $I_g(P_k)$ are the I_g scores for f_k over the event feature data set and the prior feature data set. $I_g(P_k)$ is a indicator of how the prior distribution of f_k deviates from the uniform distribution, and is designed to adjust $I_g(E_k)$.

1.3.3.2 The Transition Density Model

Liu and Brown (2003) develop the transition density model defined in equation (1.11) in a multi-step componentization and then estimated the corresponding components. The model is schematically represented in Figure 1.4. In the process of componentization Liu and Brown first separated spatial and temporal transitions as follows,

$$\psi_n(\boldsymbol{s}_{n+1}, t_{n+1} | D_n, T_n, \boldsymbol{\mathcal{X}}_{n)} =$$

$$\psi_n^{(1)}(\boldsymbol{s}_{n+1} | D_n, \boldsymbol{\mathcal{X}}_n, T_n, t_{n+1}) \cdot \psi_n^{(2)}(t_{n+1} | T_n), \tag{1.17}$$

the standard Bayesian decomposition of $\psi_n^{(2)}(t_{n+1} | D_{n,} \boldsymbol{\mathcal{X}}_n, T_n)$ is simplified to $\psi_n^{(2)}(t_{n+1} | T_n)$ based upon the assumption that any inherently temporal features (e.g., seasonality and holiday/non holiday) that are categorized as time instants are excluded because this models deals with a short time period (e.g. one week or

a few weeks). Also according to Cressie (1993), temporal transition of the marked space-time shock point process is assumed not to depend on its spatial transition.

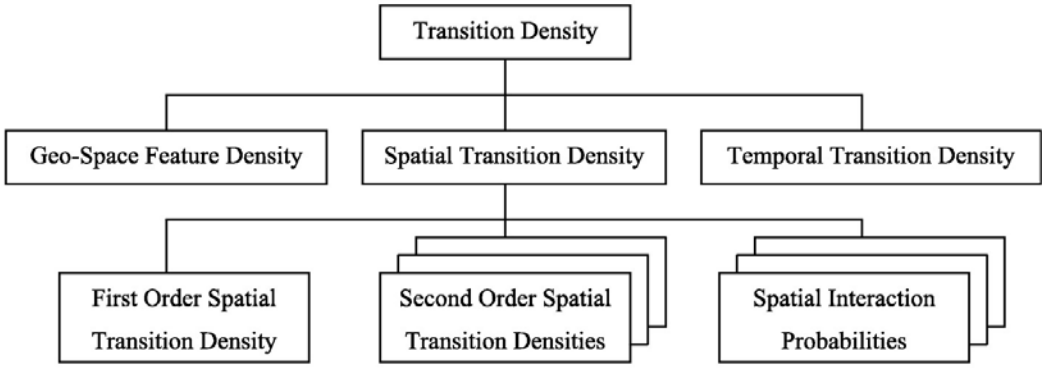


Figure 1.4 Components of the transition density model, Liu and Brown (2003)

The second step of the componentization is to model the spatial transition density $\psi_n^{(1)}(\boldsymbol{s}_{n+1}|D_n, \boldsymbol{\mathcal{X}}_n, T_n, t_{n+1})$, in other words, decide the likelihood of future events occurring at certain locations based upon past site selection preferences. The site selection preferences are defined by a distinct clustering pattern into key feature space. Liu and Brown decomposed the key feature space $\boldsymbol{\mathcal{X}}$ into C disjoint continuums $\{\boldsymbol{\mathcal{X}}^{(j)} : j = 1, 2, \ldots, C\}$ in relation to some underlying clustering pattern, which defines the set of preferences. Accordingly, $\boldsymbol{\mathcal{X}}_n$ is partitioned into C disjoint subsets $\{\boldsymbol{\mathcal{X}}_n^{(j)} : j = 1, 2, \ldots, C\}$ where $\boldsymbol{\mathcal{X}}_n^{(j)} \subset \boldsymbol{\mathcal{X}}^{(j)}$, $\{\boldsymbol{\mathcal{X}}_n^{(j)} : j = 1, 2, \ldots, C\}$ where $\boldsymbol{\mathcal{X}}_n^{(j)}$ also defines the corresponding partition of

D_n and T_n, $\{D_n^{(j)} : j = 1, 2, \ldots, C\}$, $\{T_n^{(j)} : j = 1, 2, \ldots, C\}$, locations and times of past events. Based upon the partition, Liu and Brown (2003) further define the transition density as the following:

$$\psi_n^{(1)}(\boldsymbol{s}_{n+1}, t_{n+1} | D_n, T_n, \boldsymbol{\mathcal{X}}_{n)} =$$

$$\alpha \cdot \psi_n^{(11)}(\boldsymbol{x}_{n+1} | \boldsymbol{\mathcal{X}}_n) \cdot \sum_{j=1}^{C} \psi_n^{(12)}(\boldsymbol{s}_{n+1} | D_n^{(j)}, T_n^{(j)}, t_{n+1})$$

$$\times Pr(\boldsymbol{x}_{n+1} \in \boldsymbol{\mathcal{X}}^{(j)} | \boldsymbol{\mathcal{X}}_n^{(j)}) \qquad (1.18)$$

where $\boldsymbol{x}_{n+1}$ is the feature vector at location $\boldsymbol{s}_{n+1}$, $\psi_n^{(11)}(\boldsymbol{x}_{n+1} | \boldsymbol{\mathcal{X}}_n)$ is called the first-order spatial transition density (i.e., first-order effects), which is the event intensity at $\boldsymbol{x}_{n+1}$ in the key feature space, $\psi_n^{(12)}(\boldsymbol{s}_{n+1} | D_n^{(j)}, T_n^{(j)}, t_{n+1})$, $j = 1, 2, ...C$, are called the second order spatial transition density (i.e., second-order effects). $Pr(\boldsymbol{x}_{n+1} \in \boldsymbol{\mathcal{X}}^{(j)} | \boldsymbol{\mathcal{X}}_n^{(j)})$ is the probability the next feature vector falls in the same continuum of the key feature space $\boldsymbol{\mathcal{X}}^{(j)}$ as $\boldsymbol{\mathcal{X}}_n^{(j)}$ did, and α is a normalizing factor.

In theory a spatial pattern can be regarded as the result of first-order effects coupled with second-order effects. Equation (1.18) models first-order effects as event intensity in key feature space instead of in geographic space, and this is the key point differentiating it from the traditional "hot spot" model in which event intensity is the expected number of accumulated events at alterative sites. The same site selection preferences are assumed to persist at t_{n+1} and will be captured by feature space event density.

Liu and Brown model second-order effects in geographic space which only examines spatial interaction among events in the same feature space cluster

because these events are initiated with the same set of preferences. To deal with the uncertainty associated with assigning a new event to a specific cluster they weigh second-order effects pertaining to individual clusters by the probabilities that quantify this uncertainty (i.e., the spatial interaction probabilities). Technically speaking, the overall process is partitioned into C sub-processes based upon the partitioned $D_n^{(j)}$, the consequent geographic partition defined by $\boldsymbol{\mathcal{X}}_n^{(j)}$ in the process of feature space partition, and the weighted average of the second-order effects of C thinned point processes in geographic space is calculated.

The model presented in equation (1.18) is based upon the assumption that event locations follow a homogeneous Poisson point process and are hence uniformly and independently distributed in geographic space. However, this complete randomness does not necessarily hold true in feature space due to the form of the mapping from $\boldsymbol{s}_{n+1}$ to $\boldsymbol{x}_{n+1}$ and the possible inherent randomness of $\boldsymbol{x}_{n+1}$. In equation (1.19), a new item is introduced to adjust this nonuniformity,

$$\begin{aligned}\psi_n^{(1)}(\boldsymbol{s}_{n+1}, t_{n+1}|D_n, T_n, \boldsymbol{\mathcal{X}}_{n)} = \beta \cdot (1/\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})) \\ \cdot \psi_n^{(11)}(\boldsymbol{x}_{n+1}|\boldsymbol{\mathcal{X}}_n) \cdot \sum_{j=1}^{C} \psi_n^{(12)}(\boldsymbol{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1}) \\ \times Pr(\boldsymbol{x}_{n+1} \in \boldsymbol{\mathcal{X}}^{(j)}|\boldsymbol{\mathcal{X}}_n^{(j)})\end{aligned} \tag{1.19}$$

where $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$ denotes the probability density function of $\boldsymbol{x}_{n+1}$ given a prior probability density function of $\boldsymbol{s}_{n+1}$ over the study region D. $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$ is called the geographic-space feature density and β is a normalizing factor. By including the reciprocal of $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$, individual locations with certain

feature values that are more typical than others in the study region are adjusted lower so that all locations are put on equal footing.

It should be noted that $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$ does not depend on event feature $\boldsymbol{\mathcal{X}}_n$ while $\psi_n^{(11)}(\boldsymbol{x}_{n+1}|\boldsymbol{\mathcal{X}}_n)$ does. When $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$ is uniformly distributed, the model in equation (1.19) reduces to that in (1.18) and the model in (1.18) is used when $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$ is unknown. Liu and Brown (2003) implemented the estimation of individual components in equation (1.17) to (1.19) using the following four steps:

(1). Partition the event features into the best number of clusters (C) .

(2). Estimate $\psi_n^{(11)}(\boldsymbol{x}_{n+1}|\boldsymbol{\mathcal{X}}_n)$ and $Pr(\boldsymbol{x}_{n+1} \in \boldsymbol{\mathcal{X}}^{(j)}|\boldsymbol{\mathcal{X}}_n^{(j)})$ in the key feature space.

(3). Estimate $\psi_n^{(12)}(\boldsymbol{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1})$ in the partitioned geographic space.

(4). Estimate $\kappa_n(\boldsymbol{x}_{n+1}|\boldsymbol{s}_{n+1})$ where appropriate and feasible.

1.3.4 Component Estimation of Liu and Brown's Model

1.3.4.1 Partition Event Feature Data

Liu and Brown (2003) applies a hierarchical clustering algorithm to a data set of size n and generates a succession of n partitions $P_0, P_1, \ldots, P_{n-1}$, where $P_0, P_1, \ldots, P_{n-1}$ contains $n, n-1, \ldots, 1$ cluster(s), respectively. It merges the two "closest" clusters in P_j to P_{j+1}. The stop rule is a revision of Mojena (1977) and the revised rule stops the merging clusters and select the first partition P_j satisfying

$$\alpha_{j+1} > \overline{\alpha_j} + k \cdot s_{a_j} \qquad (1.20)$$

where a_j is the shortest pair cluster distance in the partition P_j and $\overline{\alpha_j}$ and s_{a_j} are the mean and unbiased standard deviation of $a_0,\ a_1, \ldots,\ a_j,$ and k equals 1.25 according to Milligan and Cooper (1985).

1.3.4.2 Estimate First-Order Spatial Transition Density and Spatial Interaction Probabilities

Liu and Brown (2003) considers two classes of models for estimating the first-order spatial transition density and the corresponding spatial interaction probabilities. The first class is *finite mixture distributions,* which has the form

$$\widehat{f}(\boldsymbol{x};\boldsymbol{\pi},\boldsymbol{\Theta}) = \sum_{j=1}^{C} \pi_j f_j(\boldsymbol{x};\boldsymbol{\theta}_j) \tag{1.21}$$

where $\pi_j > 0,\ j = 1, \ldots, C,\ \pi_1 + \pi_2 + \cdots + \pi_C = 1,\ \boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_C]'$, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_C]$. $f_j(\boldsymbol{x};\boldsymbol{\theta}_j)$ is the jth component density with the set $\boldsymbol{\theta}_j$ parameters and $\pi_1, \pi_2, \ldots, \pi_C$ are *mixing weights* and $\boldsymbol{\Theta}$ is the collection of *component parameters*. Gaussian mixture models (GMM) are used for continuous feature space and Latent Class Models (LCM) (see Everitt, 1984) are used for discrete feature space. The Expectation-Maximization (EM) algorithm is used to quantify the parameters $[\pi_1, \pi_2, \ldots, \pi_C]'$ and $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_C]$.

Liu and Brown (2003) also applied the non-parametric techniques called *filtered kernel estimators* (FKE) (see Marchette et al., 1996) which takes the form

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} \frac{\rho_j(\boldsymbol{x}_j)}{|\boldsymbol{H}_j|} K(\boldsymbol{H}_j^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)) \tag{1.22}$$

where $K(\cdot)$ is a kernel function, $\boldsymbol{H}_j,\ j = 1, 2, \ldots, C,$ are C $p \times p$ nonsingular *local bandwidth matrices* and $\rho_j(\boldsymbol{x}),\ j = 1, 2, \ldots, C,$ satisfying

$$0 \leq \rho_j(\boldsymbol{x}) \leq 1 \text{ and } \sum_{j=1}^{C} \rho_j(\boldsymbol{x}) = 1$$

for all $\boldsymbol{x}$, are *filtering functions*. Liu and Brown assume that the kernel function $K(\cdot)$ is the standard multivariate Gaussian density function. The filtering functions $\rho_j(\boldsymbol{x})$ are prior weights over variations of local smoothness. The local bandwidth matrices $\boldsymbol{H}_j$ contain posterior parameters settings that enforce localized smoothness. $\boldsymbol{H}_j$ = diag $[h_{j1}, h_{j2}, \ldots, h_{jp}]$, $j = 1, 2, \ldots, C$, where h_{jl} $(j = 1, 2, \ldots, C;\ l = 1, 2, \ldots, p)$ is a local bandwidth for the lth dimension $[\boldsymbol{x}]_l$ of the jth region of support. Two assumptions of this *filtered product kernel (FPK)* estimators are (1) All dimensions are mutually independent and (2) Kernel functions follow a multivariate Gaussian distribution. Liu and Brown (2003) derives the filtering functions based upon the data $\{\boldsymbol{x}_i; i = 1, 2, \ldots, n\}$, which has been partitioned into C clusters. $\Omega_1, \Omega_2, \ldots, \Omega_C$.

• Let the indicator function $\mathbf{1}_{\{\boldsymbol{x} \in \Omega_j\}}$ be 1 if $\boldsymbol{x} \in \Omega_j$ and 0 otherwise. Set

$$\rho_j(\boldsymbol{x}) = \mathbf{1}_{\{\boldsymbol{x} \in \Omega_j\}}, j = 1, 2, \ldots, C. \tag{1.23}$$

The FPK estimators with the filtering functions defined in (1.23) are termed as *weighted product kernel (WPK)* estimators. Denote n_j as the number of data points in Ω_j, the local bandwidths are estimated by the following solution,

$$\widehat{h}_{jl} = (\tfrac{4}{p+2})^{1/(p+4)} \widehat{\sigma}_{jl} n_j^{-1/(p+4)} \quad j = 1, 2, \ldots, C; l = 1, 2, \ldots, p \tag{1.24}$$

where $\widehat{\sigma}_{jl}$ is the standard deviation of the lth variable $[\boldsymbol{x}]_l$ estimated from the unidimensional local data set $\{[\boldsymbol{x}_i]_l; \boldsymbol{x}_{i \in \Omega_j}\}, j = 1, 2, \ldots, C$.

Spatial interaction probabilities correspond to either finite mixture or filtered kernel estimators and are based upon the local structures specified by these

estimators. The corresponding spatial interaction probabilities for finite mixture distributions are given as

$$Pr\{\boldsymbol{x}_{n+1} \in \boldsymbol{\mathcal{X}}^{(j)} | \boldsymbol{\mathcal{X}}_n^{(j)}\} = \pi_j f_j(\boldsymbol{x}_{n+1}; \boldsymbol{\theta}_j) / f(\boldsymbol{x}_{n+1}; \boldsymbol{\pi}, \boldsymbol{\Theta}),\ j = 1, 2, \ldots, C. \tag{1.25}$$

In case a filtered kernel estimator is used, spatial interaction probabilities take the form

$$Pr\{\boldsymbol{x}_{n+1} \in \boldsymbol{\mathcal{X}}^{(j)} | \boldsymbol{\mathcal{X}}_n^{(j)}\} = \widehat{f}_j(\boldsymbol{x}_{n+1}) / \widehat{f}(\boldsymbol{x}_{n+1}), j = 1, 2, \ldots, C \tag{1.26}$$

where

$$\widehat{f}_j(\boldsymbol{x}_{n+1}) = \tfrac{1}{n} \sum_{i=1}^{n} \tfrac{\rho_j(\boldsymbol{x}_j)}{|\boldsymbol{H_j}|} K(\boldsymbol{H}_j^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)), j = 1, 2, \ldots, C. \tag{1.27}$$

1.3.4.3 Estimate Second-Order Spatial Transition Densities

To estimate second-order spatial transition densities, Liu and Brown adapt two models developed by Fiksel (1984) to their case based on two additional assumptions. First, event initiators favor geographically closer location for the next event, and second, event initiators tend not to wait long before they act again. The first model, known as the order model, is described below.

Suppose there are m data units in cluster j. Let $D_n^{(j)} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m\}$, $T_n^{(j)} = \{t_1, t_2, \ldots, t_m\}$, and $t_1 < t_2 < \cdots < t_m$ and $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m$ be ordered according to $t_1, t_2, \ldots, t_m$. Liu and Brown postulate the following function for the second order spatial transition density for cluster j

$$\psi_n^{(12)}(\boldsymbol{s} | D_n^{(j)}, T_n^{(j)}, t) = \psi_m(\boldsymbol{s} | \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m) = \tfrac{\lambda^2}{2\pi m} \sum_{i=1}^{m} e^{-\lambda ||\boldsymbol{s} - \boldsymbol{s_i}||} \tag{1.28}$$

where t and $\boldsymbol{s}$ are the time and location of a future event's occurrence respectively, $t > t_m$ and $||\boldsymbol{s} - \boldsymbol{s_i}||$ is the distance from that future event's location $\boldsymbol{s}$ to an older event location $\boldsymbol{s}_i$ $(i = 1, 2, \ldots, m)$. In this model, only the temporal order of the events is considered.

The second model is called the instant model and it incorporates the values of the time series $t_1, t_2, \ldots, t_m$. Based upon this model Liu and Brown postulate that the second-order spatial transition density takes the form

$$\begin{aligned}\psi_n^{(12)}(\boldsymbol{s}|D_n^{(j)}, T_n^{(j)}, t) &= \eta_m(\boldsymbol{s}|\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m, t_1, t_2, \ldots, t_m, t) \\ &= \frac{\lambda^2}{2\pi \sum_{i=1}^{m} e^{-\tau(t-t_i)}} \sum_{i=1}^{m} e^{-\lambda||\boldsymbol{s}-\boldsymbol{s_i}||-\tau(t-t_i)}. \qquad (1.29)\end{aligned}$$

A maximum likelihood method is used to estimate the parameters of λ and τ in (1.28) and (1.29).

1.3.4.4 Estimate Geographic-Space Feature Density

To estimate the geographic-space feature density, when appropriate and feasible, generally requires sampling over the study region. Liu and Brown obtain feature values for sample locations chosen uniformly and independently over the study region and then fit a density function to the sample applying either the finite mixture or the filtered kernel method.

1.4 Point-Pattern-Based Hierarchical Bayesian Intensity Model

1.4.1 Limitations of Liu and Brown 's Model

In section 1.3, I discussed Liu and Brown' successful theory framework of the point-pattern transition density model. They established the procedure to decompose the big model into components and then implemented the component estimates. In addition, they applied their model to a sample of crime data, which included 579 commercial and residential "breaking and entering" incidents in Richmond, VA, between July 1,1997 and August 31,1997, and demonstrated its superiority over the traditional "hot spot" model. Although their model is complete and capable, it is still not enough to reach the goals this dissertation set to resolve.

The first limitation arises from the partition of key feature space in Liu and Brown's model. Liu and Brown decomposed the key feature space $\boldsymbol{\mathcal{X}}$ into C subspaces, which defined the consequent partition of location space D_n. In other words, the model neither considered nor recorded any geographic characteristic or information of the sub feature space in the partition, and therefore by nature, it "lost" the geographic information in the process and tended to be incapable of detecting of any geographic pattern, if it exists. One can easily imagine if one or more subspaces spanned the whole horizontal dimension or vertical dimension of the study region the model would lose the ability to detect any geographic pattern in that dimension.

Liu and Brown's model is designed to deal with short time periods, within a week or a few weeks. Their model did consider any inherently temporal features, like seasonality, day of week variation, etc. Liu and Brown also assumed that for a

typical space-time point process the temporal transition is independent of its spatial transition. In Liu and Brown's model, the temporal transition density $\psi_n^{(2)}(t|T_n)$ is invariant over all locations within the study region at any given instant t_{n+1}. As a result they did not estimate this component because they narrowed their goal to forecasting only the relative transition density in the study region at any future time point. Although in this dissertation, I agree with the invariance assumption, I mainly target forecasting annual auto insurance losses in the region of interest based upon years of legacy data, and therefore I must encompass the temporal component and the relevant temporal features.

Liu and Brown's model seems adapted to small study regions (in their application, Richmond, VA). For highway losses we hope to predict the local density, usually annually, as accurately as possible. This may require the partition of study regions as micro as possible, but as for the geographic pattern, I need them to be identified at a much more macro level, e.g., county or even state level. In the study of geographic patterns, I am looking to determine whether the auto insurance claim frequencies in subregions exhibit significantly high or low values. Obviously Liu and Brown's model laid a solid foundation for density estimation, but left the geographic pattern detection blank.

In addition to the detection of a geographic pattern of annual auto insurance claim frequencies, examining the evolution of geographic pattern is possible since HLDI data span 10 years. The geographic pattern of annual auto insurance claims changes over time and hence the interest of how to measure this pattern shift. Again the method to measure pattern shift, and the corresponding visualization of pattern evolution are far beyond Liu and Brown's model. Hopefully this work can at least address the problem.

Besides the gaps mentioned above between the goals set by this dissertation and the coverage of Liu and Brown's model, the computation of large datasets is another concern. Stratified by parameters used in the estimation of density models, HLDI auto insurance loss data could easily reach 100 million records, upon which Liu and Brown's model may either be inefficient or not viable.

1.4.2 Bayesian Hierarchical Point-Pattern-Based Intensity Model

Inspired by Liu and Brown's work, in this dissertation I design and build a spatial-temporal Bayesian hierarchical model (BHM) aimed at predicting intensities of highway losses whose spatial process follows a Poisson-point-pattern. The proposed model has following functions,

Undercover latent subpopulations

The dissertation develops methods that can undercover latent distribution components of highway loss events whose spatial process is characterized as Poisson point process, but with an incomplete data structure. Methods developed should be able to determine the finite mixture structure of the underlying Poisson point process, and estimate the posterior probability from which subpopulation an observation arises.

Identify key features having influence on highway losses

This dissertation also develops methods that can identify key features having great influence over highway losses by filtering out irrelevant/uncritical ones from a large pool of features.

Detect key feature patterns corresponding to "hot spot" areas

Methods developed in this dissertation can also detect key feature patterns corresponding to "hot spot" areas where loss event intensities are classified to be highly risky.

Partition key feature space and study area

The dissertation also develops an algorithm that can partition key feature space to detect homogeneous clusters and map this partition to the study area allowing highway losses to be measured over clusters.

Predict future losses

This dissertation also develops a BHM model that can be practically used to predict future highway losses based upon information of current losses and most recent past losses. The methods can also detect and visualize the evolution of "hot spot" geographic patterns with time.

1.4.3 Analogue of terms in this dissertation to traditional spatial statistics

"Features" in this dissertation means independent variables could influence highway losses and more traditionally they are "attributes" in the spatial statistics. In spatial statistics, "feature" means geometric objects, such as points, lines, and polygons. "Distances" in this dissertation are defined in feature space instead of in geographic space unless otherwise stated.

2. Theory of Point-Pattern Spatio-Temporal Model for Highway Loss

In Chapter 2, a mixture model is proposed to model highway loss incidents in the study area D_s by extending the expectation maximization (EM) algorithm to this new field. The proposed strategy can identify subpopulations of highway loss incidents and use the Random Forest algorithm to identify features having key influences on the distribution of highway loss incidents. It also quantifies the importance level of each selected key feature. Then the patterns of the key feature vectors associated with highway loss "hot spots" can be detected.

2.1 Spatio-Temporal Process of Highway Losses

2.1.1 Highway Loss Incidents

Denote a series of highway loss incidents as $(\boldsymbol{s}_1, t_1),\ (\boldsymbol{s}_2, t_2), \ldots, (\boldsymbol{s}_m, t_m)$, where $0 < t_1 < \cdots < t_{j-1} < t_j < t_{j+1} < \cdots < t_m$, $\boldsymbol{s}_j$ is the two-dimensional location of incident j of a given type of loss and t_j is the corresponding time of the jth incident, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m$ are m p-dimension measurable feature vectors that are believed to be associated with the occurrences of the incidents, where $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ is a realization of the feature vector corresponding to incident $(\boldsymbol{s}_j, t_j)$, then $\{\boldsymbol{x}_{\boldsymbol{s},t} \in \mathcal{X} : \boldsymbol{s} \in D_{\boldsymbol{s}}, t \in T_s\}$ form a space-time shock point process (Cressie, 1993), where $t, \boldsymbol{s}$, and $\boldsymbol{x}_{\boldsymbol{s},t}$ are all random quantities defined

within study horizon $T_s \subset \mathcal{R}^+$, a study region $D_{\boldsymbol{s}} \subset \mathcal{R}^2$, and a feature space $\mathcal{X} \subset \mathcal{R}^p$, respectively. Highway loss events are considered instantaneous so this point process is classified as a shock point process instead of a survival process. Furthermore, I assume that this point process is simple, which means at a given time point, almost surely, either no incident or a single incident occurs at any point on $D_{\boldsymbol{s}}$. The study region $D_{\boldsymbol{s}}$ is partitioned into n disjoint geographic cells $\{c_1, c_2, \ldots, c_n\}$. On 2-D space, a cell is a polygon. Typically, a cell can be a census tract, a zip code area, or a grid defined in geographic information systems (GIS). Incidents' locations, times, and associated attributes can be studied at cell level, or higher level, according to the study interest.

2.1.2 Spatio-Temporal Process of Highway Losses

Highway loss incidents typically form stochastic spatio-temporal process, I denote a spatio-temporal process model as $Y(\boldsymbol{s}; t : \boldsymbol{s} \in D_{\boldsymbol{s}}, t \in T_s)$, where the study time range

$T_s \subset \mathcal{R}^+$, the study region $D_{\boldsymbol{s}} \subset \mathcal{R}^2$. A spatio-temporal process can be aggregated (sliced) into a pure spatial process, $Y(\boldsymbol{s} : \boldsymbol{s} \in D_{\boldsymbol{s}})$, or a temporal process. I write a temporal process model as $Y(t : t \in T_s)$, it can be either a point process thus T_s is a random set made up of randomly occurring time points of events on $[0, \infty)$, or a temporal discrete-time process thus $T_s = \{0, 1, 2, \ldots\}$.

On a subset of $A \subset D_{\boldsymbol{s}} \subset \mathcal{R}^2$, I define a stochastic highway loss spatial point process Z, where A is a 2-dimensional Lebesgue measurable with defined area. Let $Z(A)$ denote the number of loss events in A and $Z(\cdot)$ the counting process defined on the set of Lebesgue measurable subsets of $D_{\boldsymbol{s}}$, with furthermore

assumption that $D_{\boldsymbol{s}}$ is bounded and $Z(A)$ is finite for all $A \subset D_{\boldsymbol{s}}$, the expected number of events $E(Z(A))$ is given by an intensity function $\lambda(\boldsymbol{s})$ defined on A.

Let $\boldsymbol{s}$ be a location $\boldsymbol{s} \in D_{\boldsymbol{s}}$, let $d_{\boldsymbol{s}}$ be a small region located at $\boldsymbol{s}$ with area $|d_{\boldsymbol{s}}|$, then the first-order intensity function of the Poisson point process $Z(\cdot)$ is defined as,

$$\lambda(\boldsymbol{s}) \equiv \lim_{|d_{\boldsymbol{s}}| \to 0} E(Z(d_{\boldsymbol{s}}))/|d_{\boldsymbol{s}}|, \qquad \boldsymbol{s} \subset D_{\boldsymbol{s}}, \tag{2.1}$$

provided the limit exists. Hence,

$$E(Z(A)) = \int_A \lambda(\boldsymbol{s}) d_{\boldsymbol{s}}, \qquad A \subset D_{\boldsymbol{s}}.$$

An infinitesimal interpretation of $\lambda(\boldsymbol{s})\,|d_{\boldsymbol{s}}|$ is,

$$\lambda(\boldsymbol{s})\,|d_{\boldsymbol{s}}| \approx P\,(Z(d_{\boldsymbol{s}}) = 1).$$

When Z exhibits completely spatial randomness (CSR) it is a homogeneous Poisson point process. Whenever $\lambda(\boldsymbol{s}) \equiv \lambda^0$, which is a constant. The number of events over A follows a Poisson distribution

$$Z(A)|\lambda^0 \sim Poisson\,(\lambda^0\,|A|), \qquad A \subset D_{\boldsymbol{s}},$$

where $\lambda^0 > 0$ is the parameter of the Poisson point process, and $|A|$ is the area of A.

If $Z(\cdot)$ keeps independence for disjoint sets but $\lambda(\cdot)$ varies over $D_{\boldsymbol{s}}$, the Poisson point process becomes an inhomogeneous one,

$$Z(A) \sim Poisson\left(\int_A \lambda(\boldsymbol{x}) d\boldsymbol{x}\right), A \subset D_{\boldsymbol{s}}. \tag{2.2}$$

To further extend the highway loss spatial point process to a spatio-temporal point process, I define a bounded subset $D_{\boldsymbol{s},t}$ of $\mathcal{R}^2 \times \mathcal{R}$ and define $D_{\boldsymbol{s},t} = D_{\boldsymbol{s}} \times [0, T]$, where $T \in T_s \subset \mathcal{R}^+$ and T is the largest time. Let $A \subset D_{\boldsymbol{s},t}$ and $Z(A)$ be number of events in A then $\{Z(A) : A \subset D_{\boldsymbol{s}} \times [0, T]\}$ characterized the spatio-temporal point process.

The whole spatio-temporal point process of highway losses can be thought as a temporal process of a spatial point process, and Cressie (2011) defined the conditional intensity function of the above spatial-temporal point process as

$$\psi(\boldsymbol{s};t) \equiv \lim_{\substack{|d_{\boldsymbol{s}}|\to 0 \\ d_t \to 0}} \frac{E(Z(d_{\boldsymbol{s}};d_t)\,|\mathcal{H}_t)}{\nu(d\boldsymbol{s})d_t}. \tag{2.3}$$

provided the limit exists. In (2.3), $\boldsymbol{s} \in D_{\boldsymbol{s}}$ and $d_{\boldsymbol{s}}$ is a small region located at $\boldsymbol{s}$ with area $|d_{\boldsymbol{s}}|$, $t \in T_s$ and d_t is a small time interval at t. $\mathcal{H}_t$ contains all the history information of spatial-temporal point process up to the time point t. $\psi(\boldsymbol{s};t)$ is the frequency with which events occurs at $(\boldsymbol{s}; t)$.

2.2 Finite Mixture Models and Highway Loss Incidents

The use of mixture models can be traced back more than one century. Pearson (1894) fitted a mixture of two normal distributions with different means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 in proportions π_1 and π_2 to some biological data. From the 1980s to the 1990s, the advent of high-speed computers and the maximum likelihood estimation made mixture models practical. The Dempster et al. (1977) paper on the expectation–maximization (EM) algorithm, the McLachlan and Basford (1988) paper and the McLachlan and Peel (2001) paper on the use of EM algorithm for the fitting of finite mixture model cleared main theoretical and practical obstacles blocking the use of finite mixture models.

2.2.1 Basic Definition

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n, where $\boldsymbol{Y}_j$ is a p-dimensional random vector with probability function $f(\boldsymbol{y_j})$ on $\mathcal{R}^p$. Let $\boldsymbol{Y}$ = $(\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$, where $\boldsymbol{Y}_j^T$ denotes the transpose of $\boldsymbol{Y}_j$, and thus $\boldsymbol{Y}$ is the entire sample which is a n-tuple of points in $\mathcal{R}^p$. We use $\boldsymbol{y}_j$ to denote a realization of $\boldsymbol{Y}_j$ so that $\boldsymbol{y}$ = $(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ is an observed random sample. Suppose $f(\boldsymbol{y}_j)$ originates from multiple distributions and can be written in the form

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j), \tag{2.4}$$

where $f_i(\boldsymbol{y_j})$ is a probability density function and

$$0 \leq \pi_i \leq 1 \quad (i = 1, \ldots, g) \text{ and}$$
$$\sum_{i=1}^{g} \pi_i = 1, \qquad g \in \mathcal{Z} \text{ and } g > 2.$$

The nonnegative π_i are called the mixing proportions or weights. $f_i(\boldsymbol{y}_j)$ is the ith component density of the mixture and $f(\boldsymbol{y}_j)$ is a g-component finite mixture density; its corresponding distribution function $F(\boldsymbol{y}_j)$ is referred to as a g-component finite mixture distribution.

A mixture model can be viewed as a probabilistic model for representing the presence of subpopulations within an overall population. In the context of parametric methodology, I have to determine the following estimates to fit a finite mixture model: the number of components; the weight of each component; and parameters of each component. To identify from which components observations are generated, let g be fixed and $\boldsymbol{Z}_j$ be a g-dimensional component vector, whose ith element, Z_{ij} = $(\boldsymbol{Z}_j)_i$, valued in 1 or 0 according to whether $\boldsymbol{Y}_j$ is generated from the ith component in the mixture. Thus $\boldsymbol{Z}_j$ follows a multinomial distribution consisting of one draw on g categories with probabilities $\pi_1, \ldots, \pi_g$,

$$\boldsymbol{Z}_j \sim \text{Multinomial}(1, \boldsymbol{\pi}) \tag{2.5}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^T$.

2.2.2 Component Parameters and the Likelihood Function

Suppose $f_i(\boldsymbol{y}_j)$ belongs to some parametric family and I specify the component density functions as $f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i)$ where $\boldsymbol{\theta}_i$ are unknown parameters of the ith component in the mixture. I rewrite the probability density function of the mixture in (2.4) as

$$f(\boldsymbol{y}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i) \ , \tag{2.6}$$

where $\boldsymbol{\Psi}$ consists of all unknown parameters in the mixture model

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T$$

and $\boldsymbol{\theta}_i$ are the parameters of the corresponding family of the ith component.

In most cases, component densities belong to the same parametric family. Thus the mixture density in (2.6) has the form

$$f(\boldsymbol{y}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\boldsymbol{y}_j;\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i \in \Theta$ and Θ denotes parameter space of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$.

Assume $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{Y}_j|\boldsymbol{\Psi})$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are observed values of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from a parametric family $\mathcal{P}(\boldsymbol{Y}_j|\boldsymbol{\Psi})$, based upon probability density function of (2.6), corresponding likelihood function is in the form

$$\boldsymbol{L}(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\boldsymbol{y}_j;\boldsymbol{\Psi}) \tag{2.7}$$

and log likelihood function is

$$\log \boldsymbol{L}(\boldsymbol{\Psi}) = \sum_{j=1}^{n} log f(\boldsymbol{y}_j;\boldsymbol{\Psi})$$

$$= \sum_{j=1}^{n} log\{\sum_{i=1}^{g} \pi_i f(\boldsymbol{y}_j;\boldsymbol{\theta}_i)\}. \quad (2.8)$$

2.2.3 Incomplete Data Structure

To identify the component of an observation is a task of EM algorithm. In the EM framework, a random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is defined as incomplete because their associated component indicators $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ remain unknown, or in terms of observability, unobserved. Thus $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ are observed data and

$\boldsymbol{y_c} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T$

are defined as complete data vector where

$\boldsymbol{z} = (\boldsymbol{z_1^T}, \ldots, \boldsymbol{z_n^T})$.

The log likelihood function of the complete data structure can be written as

$$\log \boldsymbol{L_c}(\boldsymbol{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{\log \pi_i + \log f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i)\} \quad (2.9)$$

Under Bayesian definition, π_i in (2.4) can be viewed as the prior probability that the observation belongs to the ith component of the mixture $F(\boldsymbol{Y}_j|\boldsymbol{\theta}_i)(i = 1, ..., g)$, denote τ_i as the posterior probability that the observation belongs to the ith component of the mixture

τ_i = pr $(z_{ij} = 1|\boldsymbol{y}_j)$

$$= \pi_i f_i(\boldsymbol{y}_j) / f(\boldsymbol{y}_j) \ (i = 1, ..., g; j = 1, \ldots, n) \quad (2.10)$$

2.2.4 Component Parameter Estimate by Use of Direct Approach

McLachlan and Krishnan (1997) gave a theoretical direct approach to estimate unknown parameters in (2.6) where $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\xi}^T)^T$ contains all unknown parameters in the mixture model and $\boldsymbol{\xi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)^T$ are all parameters known *a priori* to be distinct $(\pi_g = 1 - \pi_1 - \pi_2, \ldots, -\pi_{g-1})$. Theoretically, the computation of the maximum likelihood estimator (MLE) of $\boldsymbol{\Psi}$ is equivalent to solving the likelihood equation,

$$\partial \log \boldsymbol{L}(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = 0. \tag{2.11}$$

McLachlan and Krishnan (1997, Section 1.4) detailed the manipulation so that the MLE of $\boldsymbol{\Psi}$, $\widehat{\boldsymbol{\Psi}}$, satisfies

$$\hat{\pi}_i = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \widehat{\boldsymbol{\Psi}})/n \quad i = (1, \ldots, g) \tag{2.12}$$

and

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \widehat{\boldsymbol{\Psi}}) \partial \log f_i(\boldsymbol{y}_j; \widehat{\boldsymbol{\theta}}_i) / \partial \boldsymbol{\xi} = \mathbf{0} \tag{2.13}$$

where

$$\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \Big/ \sum_{l=1}^{g} \pi_l f_l(\boldsymbol{y}_j; \boldsymbol{\theta}_l). \tag{2.14}$$

Unfortunately equation (2.13) is not always solvable and thus limits the use of direct theoretical approach in many applications.

2.3 EM Framework on Finite Mixture Model Fitting

2.3.1 Definition of the EM Algorithm

The Dempster et al. (1977) paper demonstrated that solving the equations of (2.12) and (2.13) formed an iterative computation solution whereby for the *p*th

round of estimate $\boldsymbol{\Psi}^{(p)}$ of $\boldsymbol{\Psi}$ in the right-hand side of these equations, a new estimate $\boldsymbol{\Psi}^{(p+1)}$ can be computed for $\boldsymbol{\Psi}$ and the $\boldsymbol{\Psi}^{(p+1)}$ can be substituted into the right-hand side equations to produce $\boldsymbol{\Psi}^{(p+2)}$ and so on until p is big enough and $\boldsymbol{\Psi}^{(p)}$ converge.

Dempster et al. (1977) defined the EM algorithm first on regular exponential families starting with two sample spaces $\mathcal{Y}$ and $\mathcal{X}$ and a many-to-one mapping $\boldsymbol{X} \longrightarrow \boldsymbol{Y}(\boldsymbol{X})$ from $\mathcal{X}$ to $\mathcal{Y}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random variables from exponential families. The incomplete-data $\boldsymbol{y}$, which were a realization from $\boldsymbol{Y}$ were observed while corresponding complete-data $\boldsymbol{x}$ cannot be observed but only indirectly through $\boldsymbol{y}$. Let $f(\boldsymbol{x}|\boldsymbol{\Psi})$ be density functions for the family depending on parameter $\boldsymbol{\Psi}$ and its corresponding incomplete-data specification $f(\ldots|\ldots)$ is related to $f(\boldsymbol{x}|\boldsymbol{\Psi})$ by

$$g(\boldsymbol{y}|\boldsymbol{\Psi}) = \int_{\mathcal{X}(\boldsymbol{Y})} f(\boldsymbol{x}|\boldsymbol{\Psi}) d_{\boldsymbol{x}} \tag{2.15}$$

and $f(\boldsymbol{x}|\boldsymbol{\Psi})$ has the form

$$f(\boldsymbol{x}|\boldsymbol{\Psi}) = b(\boldsymbol{x}) \exp\left(\boldsymbol{\Psi}\boldsymbol{t}(\boldsymbol{x})^T\right)/a(\boldsymbol{\Psi}) \tag{2.16}$$

where $\boldsymbol{\Psi}$ denotes a 1 x r vector parameter, and $\boldsymbol{t}(\boldsymbol{x})$ a 1 x r vector of complete-data sufficient statistics, and $\boldsymbol{\Psi}$ is restricted to an r-dimension convex set Ω.

Suppose $\boldsymbol{\Psi}^{(p)}$ denotes the current value of $\boldsymbol{\Psi}$ after p cycles of the algorithm, the next cycle can be described in two steps,

E-*step*: Estimate the complete-data sufficient statistics $\boldsymbol{t}(\boldsymbol{x})$ by finding

$$\boldsymbol{t}^{(p)} = E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{y}, \boldsymbol{\Psi}^{(p)}) \tag{2.17}$$

M-*step*: Determine $\boldsymbol{\Psi}^{(p+1)}$ as the solution of the following equations

$$E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\Psi}) = \boldsymbol{t}^{(p)} \tag{2.18}$$

Equation (2.18) actually defines the MLE estimator of $\boldsymbol{\Psi}$ given $\boldsymbol{t}^{(p)}$ in (2.17). The sufficient statistics computed from an observed $\boldsymbol{x}$ from (2.16). It should be noted in case of regular exponential family, maximizing $\log f(\boldsymbol{x}|\boldsymbol{\Psi})$ is equivalent to maximizing

$$\log \boldsymbol{\Psi} \boldsymbol{t}(\boldsymbol{x})^T - \log a(\boldsymbol{\Psi}).$$

To explain why repeated application of the E-steps and M-steps leads to the value $\boldsymbol{\Psi}^*$, which maximizes likelihood

$$\boldsymbol{L}(\boldsymbol{\Psi}) = \log g(\boldsymbol{y}|\boldsymbol{\Psi}) \tag{2.19}$$

for exponential family, Dempster et al. (1977) introduced the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\boldsymbol{\Psi}$,

$$k(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Psi}) = f(\boldsymbol{x}|\boldsymbol{\Psi}) / g(\boldsymbol{y}|\boldsymbol{\Psi}) \tag{2.20}$$

in order to rewrite (2.19) into the form

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\Psi}) &= \log f(\boldsymbol{x}|\boldsymbol{\Psi}) - \log k(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Psi}) \\ &= -\log a(\boldsymbol{\Psi}) + \log a(\boldsymbol{\Psi}|\boldsymbol{y}), \end{aligned} \tag{2.21}$$

where

$$a(\boldsymbol{\Psi}|\boldsymbol{y}) = \int_{\mathcal{X}(\boldsymbol{Y})} b(\boldsymbol{x}) \exp(\boldsymbol{\Psi} \boldsymbol{t}(\boldsymbol{x})^T) d_{\boldsymbol{x}} \tag{2.22}$$

and

$$a(\boldsymbol{\Psi}) = \int_{\mathcal{X}} b(\boldsymbol{x}) \exp(\boldsymbol{\Psi} \boldsymbol{t}(\boldsymbol{x})^T) d_{\boldsymbol{x}}. \tag{2.23}$$

To maximize (2.21), I need to set the derivative of (2.21) to 0, which produces

$$\mathrm{d}\boldsymbol{L}(\boldsymbol{\Psi}) = E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\Psi}) + E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\Psi}, \boldsymbol{y}) = 0. \tag{2.24}$$

Solving equation (2.24) leads to d$\boldsymbol{L}(\boldsymbol{\Psi})$ = 0 at $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{*}$. Thus, if the algorithm converges, then

$$\boldsymbol{\Psi}^{(p)} = \boldsymbol{\Psi}^{(p+1)} = \boldsymbol{\Psi}^{*}.$$

Dempster et al. (1977) then extended the above defined E-step and M-step to a more general case, a curved exponential family for which $\boldsymbol{\Psi}$ lies in a curved submanifold Ω_0 of the r-dimension convex set Ω.

In this case the E-step (2.18) remains the same, but the M-step becomes:

Determine $\boldsymbol{\Psi}^{(p+1)}$ to be a value of $\boldsymbol{\Psi}$ in Ω_0 which maximizes

$$\log \boldsymbol{\Psi}\boldsymbol{t}^{(p)}(\boldsymbol{x})^{T} - \log a(\boldsymbol{\Psi}).$$

Dempster et al. (1977) further extended above definition of EM algorithms to all densities by introduction of a new function

$$Q(\boldsymbol{\Psi}^{'}|\boldsymbol{\Psi}) = E(\log f(\boldsymbol{x}|\boldsymbol{\Psi}^{'})|\boldsymbol{y}, \boldsymbol{\Psi}) \tag{2.25}$$

under assumptions a) $f(\boldsymbol{x}|\boldsymbol{\Psi}) > 0$ a.e. in $\mathcal{X}$ for all $\boldsymbol{\Psi} \in \Omega$. b) $Q(\boldsymbol{\Psi}^{'}|\boldsymbol{\Psi})$ exists for all pairs of $(\boldsymbol{\Psi}^{'},\boldsymbol{\Psi})$. Then from pth cycle to $(p+1)$st cycle

E-*step*: Compute $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$. $\tag{2.26}$

M-*step*: Choose $\boldsymbol{\Psi}^{(p+1)}$ on Ω that maximizes $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$.

It should be noted $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(p)})$ must be computed for all $\boldsymbol{\Psi} \in \Omega$.

Dempster et al. (1977) also proved that the generalized EM algorithm likelihood $\boldsymbol{L}(\boldsymbol{\Psi})$ is non-decreasing on each iteration, and $Q(\boldsymbol{\Psi}^{(p+1)}|\boldsymbol{\Psi}^{(p)}) > Q(\boldsymbol{\Psi}^{(p)}|\boldsymbol{\Psi}^{(p)})$ on any iteration (strictly increasing). In addition, they discussed the convergence of the generalized EM algorithm and also demonstrated methods to calculate the rate of EM convergence.

2.3.2 Finite Mixture Models under the EM Framework

Although Section 4.3 of Dempster et al. (1977) discussed the application of EM in finite mixture models, McLachlan and Krishnan (1997) made substantial effort to tailor EM algorithm for finite mixture models. McLachlan and Krishnan applied EM to finite mixture models by treating the z_{ij} defined in (2.9) as missing data the iteration proceeds in two steps. E (expectation) and M (maximization)

$$\text{E-}\textit{step}\text{: } Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}(\log \boldsymbol{L_c}(\boldsymbol{\Psi})\,|\boldsymbol{y}). \tag{2.27}$$

The E-step computes the expectation of the complete-data log likelihood $\log \boldsymbol{L_c}(\boldsymbol{\Psi})$ given the observed data $\boldsymbol{y}$ and the current estimates of $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}^{(k)}$. Here $\boldsymbol{\Psi}^{(k)}$, calculated from the kth EM iteration, serves as $\boldsymbol{\Psi}$. From (2.9), $\boldsymbol{L_c}(\boldsymbol{\Psi})$ is linear in z_{ij}, so the calculation of the $E_{\boldsymbol{\Psi}^{(k)}}(\log \boldsymbol{L_c}(\boldsymbol{\Psi})\,|\boldsymbol{y})$ is equivalent to

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}) &= \Pr_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij}=1|\boldsymbol{y}\} \\ &= \tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}) \end{aligned} \tag{2.28}$$

thus (2.14) transforms to

$$\begin{aligned} \Pr_{\boldsymbol{\Psi}^{(k)}}(z_{ij}=1|\boldsymbol{y}_j) &= \pi_i f_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})\,/f(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}) \\ &= \pi_i^{(k)} f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i^{(k)})\,/\sum_{l=1}^{g}\pi_l^{(k)} f_l(\boldsymbol{y}_j;\boldsymbol{\theta}_l^{(k)})\,, \end{aligned} \tag{2.29}$$

$i=1,\ldots,g; j=1,\ldots,n.$

$\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})$ is the posterior probability that the jth observation $\boldsymbol{y}_j$ is generated from the ith component of the mixture given the observed $\boldsymbol{y}_j$ and the current estimate $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}^{(k)}$. After kth iteration of the EM algorithm, (2.27) can be rewritten to

$$Q(\boldsymbol{\Psi}\,|\boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})\,\{\log\pi_i + \log f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i)\}. \tag{2.30}$$

On the first iteration we assign $\boldsymbol{\Psi}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_{g-1}^{(0)}, (\boldsymbol{\xi}^{(0)})^T)^{\boldsymbol{T}}$ and

$$Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(0)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(0)}) \{\log\pi_i + \log f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i)\}$$

M-*step*: Maximize $Q(\boldsymbol{\Psi}^{(k)}|\boldsymbol{\Psi})$ of (2.30).

The maximization of $Q(\boldsymbol{\Psi}^{(k)}|\boldsymbol{\Psi})$ on the $(k+1)$th iteration with respect to $\boldsymbol{\Psi}$ over its parameter space Ω globally will determine the updated $\boldsymbol{\Psi}^{(k+1)}$.

$$\boldsymbol{\Psi}^{(k+1)} = \arg\max_{\boldsymbol{\Psi}} \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}) \{\log\pi_i + \log f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i)\} \qquad (2.31)$$

For finite mixture models the updating of $\pi_i^{(k+1)}$, the mixing proportions, is independent of the updating of $\boldsymbol{\xi}^{(k+1)}$, the density parameters of components. McLachlan and Krishnan (1997) showed that parallel to MLE estimator of $\pi_i = \sum_{j=1}^{n} z_{ij}/n \, (i = 1, \ldots, g)$,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}) \, /n \, (i = 1, \ldots, g). \qquad (2.32)$$

Updating $\boldsymbol{\xi}$ in the $(k+1)$th iteration from $\boldsymbol{\xi}^{(k)}$ to $\boldsymbol{\xi}^{(k+1)}$ needs to maximize $\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}) \{\log\pi_i + \log f_i(\boldsymbol{y_j};\boldsymbol{\theta_i})\}$ in (2.31) by solving the equation

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}) \, \partial \log f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i) \, / \, \partial\boldsymbol{\xi} = \mathbf{0} \qquad (2.33)$$

which often takes close form. Dempster et al. (1977) had shown that the incomplete-data likelihood values $L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)})$ and the repeating of the above E-step and M-step leads to convergence of $\boldsymbol{\Psi}^{(k)}, \boldsymbol{\Psi}^{(k)} \rightarrow \boldsymbol{\Psi}^{(k+1)} \rightarrow \boldsymbol{\Psi}^{*}$, when k reaches some value.

2.4 Extension of EM on Highway Loss Incidents

2.4.1 Finite Mixture Model and Highway Loss Incidents

We defined the point process of highway loss incidents in 2.1.1, and the counting process of the Poisson point process in 2.1.2 and partitioned the study region $D_{\boldsymbol{s}}$ into n disjoint geographic cells $\{c_1, c_2, \ldots, c_n\}$. Assume complete spatial randomness (CSR) for $c_j\,(j = 1, \ldots, n)$, at the specific time point $t \in T_s$. Let Y be a discrete random variable, and let $Y\big|_{c_j}$ be the number of events observed on c_j, and $|c_j|$ be the area of c_j, and λ_{c_j} be the underlying intensity dominating the Poisson counting process on c_j,

then

$$Y\big|_{c_j} \sim \ Poisson\,(\lambda_{c_j}\,|c_j|)$$

and the probability mass function of $Y\big|_{c_j}$ is

$$f(y_j;\,\lambda_{c_j}) \ = \Pr\,(Y\big|_{c_j} = y_j) \ = \frac{(\lambda_{c_j}\,|c_j|)^{\,y_j}}{y_j!}\,e^{-(\lambda_{c_j}\,|c_j|)}. \tag{2.34}$$

Here y_j is a nonnegative integer and λ_{c_j} is found in a 1-dimension parameter space $\mathcal{R}^+$. Although n is large many of the λ_{c_j}s may be the same. So I let $\lambda_{c_j} \in \{\lambda_1, \lambda_2, \ldots, \lambda_g\}$ where $g \ll n$. I now draw a sample according to Y. The sample is a spatial sample and will come from known c_js. However, I do not necessarily know which λ_i is appropriate.

Because all components come from same family of distribution, for Poisson counting process of highway loss incidents, the finite mixture model of (2.6) can be rewritten to

$$f(y_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g}\pi_i f(y_j;\lambda_i)$$

$$= \sum_{i=1}^{g}\pi_i \frac{(\lambda_i|c_j|)^{y_j}}{y_j!}e^{-(\lambda_i|c_j|)}$$

where c_j is the associated area on which y is observed. The log likelihood function of the complete data structure can be written as

$$\log \boldsymbol{L_c}(\boldsymbol{\Psi}) =$$
$$\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{\log\pi_i + y_j\log(\lambda_i|c_j|) - \log(y_j!) - \lambda_i|c_j|\}. \qquad (2.35)$$

To apply finite mixture model on highway loss incidents, three problems need to be resolved.

1. to determine the number of components.
2. to estimate the mixing weight for each component.
3. to estimate parameters for each component.

Let g be the unknown number of components of the study subject, and let $\boldsymbol{\Psi} = (\pi_1,\ldots,\pi_{g-1},\boldsymbol{\xi}^{\boldsymbol{T}})^{\boldsymbol{T}}$ contains all unknown parameter in the mixture model and

$\boldsymbol{\xi} = (\lambda_1,\ldots,\lambda_g)^T$ are all parameters known *a priori* to be distinct for all components. The whole process of applying the finite mixture model to highway loss incidents is simply the process to decide $\boldsymbol{\Psi}$ given each of the component follows a Poisson distribution.

2.4.2 EM Algorithm with Known Number of Components

Assume the finite mixture model for the study subject has g components and g is known here. The EM algorithm on highway loss incidents also composes the E-step and the M-step, from kth iteration to $(k+1)$th iteration,

E-*step*: Rewrite the function in (2.30)

$$Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n}$$

$$\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})\ \{\log\pi_i + (y_j\log(\lambda_i|c_j|) - \log(y_j!) - \lambda_i|c_j|)\} \tag{2.36}$$

where

$$\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}) = \pi_i^{(k)}\frac{(\lambda_i|c_j|)^{y_j}}{y_j!}e^{-(\lambda_i|c_j|)}/(\sum_{l=1}^{g}\pi_l^{(k)}\frac{(\lambda_l|c_j|)^{y_j}}{y_j!}e^{-(\lambda_l|c_j|)})$$

$$= \pi_i^{(k)}(\lambda_i|c_j|)^{y_j}e^{-(\lambda_i|c_j|)}/(\sum_{l=1}^{g}\pi_l^{(k)}(\lambda_l|c_j|)^{y_j}e^{-(\lambda_l|c_j|)}) \tag{2.37}$$

M-*step*: Maximize the function in (2.36).

The maximization of $Q(\boldsymbol{\Psi}^{(k)}|\boldsymbol{\Psi})$ on the $(k+1)$th iteration will determine the updated $\boldsymbol{\Psi}^{(k+1)}$.

$$\boldsymbol{\Psi}^{(k+1)} = \arg\max_{\boldsymbol{\Psi}} \sum_{i=1}^{g}\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})$$

$$\{\log\pi_i + y_j\log(\lambda_i|c_j|) - \log(y_j!) - \lambda_i|c_j|\}. \tag{2.38}$$

Here the parameter space for $\pi_i^{(k+1)}$ is g-dimensional space (0,1] $\times$ (0,1] $\times \cdots \times$ (0,1] and $\sum_{i=1}^{g}\pi_i = 1$. We further assume that intensities of the Poisson point processes studied are all finite and parameter space Ω for $\boldsymbol{\xi}^{(k+1)}$ is on g-dimensional space $(0,\ldots,0) \cup \mathcal{R}^+ \times \mathcal{R}^+ \cdots \times \mathcal{R}^+$ and Ω is a closure.

The updating of $\pi_i^{(k+1)}$, the mixing proportions, is independent of updating $\boldsymbol{\xi}^{(k+1)}$, the mass function parameters of components.

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}) \, / n \, (i = 1, \ldots, g). \tag{2.39}$$

Updating $\boldsymbol{\xi}$ in the $(k+1)$th iteration from $\boldsymbol{\xi}^{(k)}$ to $\boldsymbol{\xi}^{(k+1)}$ needs to maximize $\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}) \, \{\log\pi_i \, + \log f_i(\boldsymbol{y}_j;\boldsymbol{\theta}_i)\}$ in (2.36) by solving g equations

$$\sum_{j=1}^{n} \tau_i(y_j; \boldsymbol{\Psi}^{(k)})(\tfrac{y_j}{\lambda_i} - |c_j|) \, = 0 \tag{2.40}$$

which often takes closed form. Solutions from (2.40) could be local or global minima, or local or global maxima. When (2.40) does not take closed form or fails to produce global maxima, I define a closure $\Omega^0 \subset \Omega$, $\Omega^0 \, = [0, \, \lambda_{max}] \times [0, \, \lambda_{max}] \times \cdots \times [0, \, \lambda_{max}]$, where λ_{max} be the maximum intensity of the study subject, then we maximize (2.36) on the closure Ω^0.

2.4.3 EM Algorithm when the Number of Components is Known

2.4.3.1 Criterion in Determining the Number of Components

Brooks et al. (2003) showed a complete scheme of Reversible Jump Markov Chain Monte Carlo (RJMCMC) which can be used to determine the number of components in the mixture model. However, just as Robert and Casella (2011) said, the implementation of a complex algorithm like RJMCMC is somewhat of an overkill for the comparison of a few models. For highway loss incidents studied in this dissertation, the finite mixture model is used in a clustering context to identify subpopulations (groups) rather than to model unknown distributional shapes such as skewness and kurtosis, thus the number of components remains unknown. For a given sample there may not be a one-to-one correspondence between the mixture components and the groups even if we have the sample and know the specification of the parametric family of its underlying distributions.

Fitting finite mixture models using maximum likelihood methods may result in multiple models each having a different number of components. The goal of deciding number of components is to find the smallest value of g, g_0, to fit the g-component mixture model while being able to differentiate each of the components from others and getting good performance in terms of likelihood. McLachlan and Peel (2001) named g_0 as the order of mixture model.

2.4.3.2 Prior Information of Components

If no prior information is available about the component distributions, then nonparametric methods of detecting number of modes might be more appropriate for a given sample. According to the literature, the relationship between number of modes and number of components has not been determined completely. Miguel A. Carreira-Perpiñan (1999) proved that the number of modes cannot be more than the number of components in mixing Gaussian distributions and that they are contained in the convex hull of the component centroids. In my case, I begin with detecting the number of modes of observed samples in expectation that it will help to define the range of number of components and thus makes computation in fitting mixture model more efficient.

Let y_j be the observed number of highway loss incidents on corresponding geographic cell c_j and let $|c_j|$ be area of c_j, $j = 1, 2, \ldots, n$. I assume CSR on c_j and define the observed intensity $\lambda_j^{'}$ as

$$\lambda_j^{'} = \frac{y_j}{|c_j|}. \tag{2.41}$$

Thus $\lambda_1^{'}, \lambda_2^{'} \ldots, \lambda_n^{'}$ form a random univariate sample of size n. I use kernel density estimation (KDE) to investigate the multimodality of this sample. Silverman (1986) describes the kernel density estimator as

$$\widehat{f}_h(\lambda^{'}) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{(\lambda^{'}-\lambda_j^{'})}{h}) \tag{2.42}$$

where $K()$ is a symmetric probability density function satisfies the condition

$$\int_{-\infty}^{+\infty} K(x)dx = 1,$$

and h is the bandwidth or smoothing parameter. The selection of h is key in kernel density estimation. When h is chosen too small undersmoothing occurs and spurious fine structure becomes visible. On the other hand, when h is chosen too large, oversmoothing occurs so that multimodality of the distribution is obscured. Jones et al. (1996) showed in case of a standard normal kernel function rescaled by h and if $h \to 0$ and $nh \to \infty$, the optimal bandwidth (asymptotic mean integrated squared error) was

$$h_{\text{AMISE}} = \{\frac{1}{2\sqrt{\pi}n\int(f'')^2}\}^{1/5}$$

which needs approximation methods. Sheather (2004) suggested a number of bandwidth based around a "center point" bandwidth and recommends the Sheather-Jones plug-in bandwidth be used due to its overall good performance. In my case, although the density of $\lambda_1^{'}, \lambda_2^{'}...\lambda_n^{'}$ is highly skewed and is at the boundary, the Sheather-Jones plug-in bandwidth might have some bias, it is still good enough to detect the number of modes to help the determination of number of components.

2.4.3.3 The Sequence of EM Algorithms of Model Fitting Computation

I assume kernel density estimation (KDE) has determined a given sample has m modes. Based upon m and the sample size, I suggest the numbers of components between g_L, the, lower bound, and g_U, the upper bound. To be conservative I suggest

$$g_L = min\,(2, \lfloor 0.5m \rfloor),\ g_U = max\,(\sqrt{n}, \lfloor 1.5m \rfloor) \tag{2.43}$$

I run a sequence of $g_U - g_L + 1$ fittings looking for the one with best fitting from the sequence to decide the order of the mixture model. The procedure of the model finding algorithm is described as below,

Inputs: Poisson mixture model defined by $\sum_{i=1}^{g}\pi_i f(y_j;\,\lambda_i)$

Constants: $\varepsilon \leftarrow 10^{-8}$

Control parameters: $k_{max} \leftarrow 500$, $g_L \leftarrow min\,(2, \lfloor 0.5m \rfloor)$,

$g_{\mathrm{U}} \leftarrow max\,(\sqrt{n}, \lfloor 1.5m \rfloor)$

for $g = g_L, \ldots, g_U$

set initial

$k \leftarrow 0$

$\boldsymbol{\Psi}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_{g-1}^{(0)}, (\lambda_{\mathbf{1}}, \ldots, \lambda_{\boldsymbol{g}})^T)^T$ and

$Q(\boldsymbol{\Psi}\,|\boldsymbol{\Psi}^{(0)}) =$

$\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(0)})\,\{\log\pi_i + (y_j\log(\lambda_i\,|c_j\,|) - \log(y_j!) - \lambda_i\,|c_j\,|)\}$

repeat E-*step*: rewrite Q function in (2.30)

M-*step*: $\boldsymbol{\Psi}^{(k+1)} = \underset{\boldsymbol{\Psi}}{\arg\max}\sum_{i=1}^{g}\sum_{j=1}^{n}Q(\boldsymbol{\Psi}\,|\boldsymbol{\Psi}^{(\mathrm{k})})$ in (2.38)

$\boldsymbol{\Psi}^{(k+1)} = \underset{\boldsymbol{\Psi}}{\arg\max}\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})$

$k \leftarrow k+1$

until $L^{(k+1)} - L^{(k)} < \varepsilon$ or $k = k_{max}$

output convergence status (yes or no)

$\boldsymbol{\Psi}^{(*)}$ composed of $2g - 1$ parameters

$\log L^{(*)}$ or $\log L^{(k+1)}$ and other model fitting results

end for

Return

2.4.3.4 Model Selection and Decision of Order

The repeated EM algorithms in section 2.4.3.3 produced a sequence of modeling results composed of $g_U - g_L + 1$ elements. Each element of the sequence corresponds to a specific g taking values from g_L to g_U. To decide the order g_0, of the mixture model, we want to maximize the likelihood resulted from the mixture model while making the model as simplified as possible. To achieve this balance, a concept of penalization was introduced such that as the likelihood increases with the addition of a component to a mixture model, the likelihood (log likelihood) is penalized by the subtraction of a term. This term measures the complexity of the mixture model and often is a function of the numbers of parameters used in the model.

Akaike (1974) developed a method of model selection following the above concept and named it AIC (Akaike's Information Criterion). AIC selects the model that minimizes

$$-2\log L(\widehat{\mathbf{\Psi}}) + 2d \tag{2.44}$$

in which the first term measure the lack of fitting and the second term serves to penalize the model complexity. $L(\widehat{\mathbf{\Psi}})$ is the likelihood corresponds to parameters $\widehat{\mathbf{\Psi}}$ estimated by a model and d is the total number of parameters in the model. Smaller AIC means better model performance. Hurvich and Tsai (1989) proposed another criterion AICC (AIC with a correction) which takes the form

$$\text{AICC} = \text{AIC} + \frac{2d(d+1)}{n-d-1} \tag{2.45}$$

where n is the sample size and $d = 2g - 1$. Thus, AICC applied greater penalty for extra parameters than AIC.

I use AICC as the primary criterion in evaluating the sequence of models produced in algorithms described in section 2.4.3.3. Compared with AICC, AIC is more prone to overfitting which leads to including more components in a mixture model. It should be noted that the Likelihood Ratio Test (LRT) is not used here although it can produce exact p-values in model comparison. First, LRT demands more complicated MCMC sampling in Bayesian framework; second, catching the component with high incidents intensity is of interest in this dissertation. Minor overfitting is not considered a disadvantage here.

2.5 Key Feature Space Formation and Hot-Spot Key Feature Patterns

Let the finite mixture model

$$f(y_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g_0} \pi_i \frac{(\lambda_i |c_j|)^{y_j}}{y_j!} e^{-(\lambda_i |c_j|)}$$

be the selected mixture model in section 2.4.3.4, and g_0 be the order of the mixture model and $\{(\widehat{\pi}_1, \widehat{\lambda}_1), (\widehat{\pi}_2, \widehat{\lambda}_2), \ldots, (\widehat{\pi}_{g_0}, \widehat{\lambda}_{g_0})\}$ be estimated parameters for each of the g_0 components. Based upon these and the observed sample, I will prepare input data to identify features having key influences on highway loss incidents and detect key feature patterns associated with hot-spots.

2.5.1 Feature Dimension Reduction

Many factors are believed to influence highway losses. A recent hot topic in highway safety is texting-while-driving, which has been proven to be associated

with many cases of fatal crashes. Many states have passed laws banning texting-while-driving to curb this dangerous and burgeoning distraction. Although it is widely accepted that texting-while-driving bans are intuitive countermeasures, the effectiveness of these laws has been debated in highway safety community, largely because of difficulty enforcement of such laws. Police often complain that lack of economic or human resources, difficulty to discern whether use of cellphone is texting (typically illegal) or dialing (legal in some states), and other factors, undermined their enforcementability.

Texting-while-driving bans are a good example to show the complication of factors related to highway losses, which include deaths, injuries and property damage. It is well known that demographic factors, social and economic factors, and legislative factors contribute to highway losses at certain spatial levels. Previous studies also have shown that roadway related factors, weather related factors, vehicle related factors and driver related factors are also related to highway losses. In addition, many of these factors may confound or interact with each other, making prediction of highway losses even more difficult.

The high dimension of the multivariate data often bring side effects in the process of prediction: first, many variables create noise in the process of prediction and mask real discriminators; second, many variables exhaust the computation resources and sometimes even make real-time systems infeasible. If data mining is the first process of highway losses prediction, reduction of dimensions should be resolved first. In other words, I need to identify the key predictors first.

2.5.2 Initial Screen by Visualization

A three-phase procedure was applied to first classify the available feature set into two subsets. One is a subset whose elements are homogeneous through a study time period or at least their variation during the study time period are trivial so that an independence from time can be assumed. Then, the feature subset not sensitive to time is reduced to a dimension in order to make the further computation feasible. Then quantitative solutions will be applied to further suppress the dimension to form the final subset, the key feature subset. We initiate the second process using a triplet (F, v_1, F_1), where F is the initial non-time-sensitive feature set, v_1 is a visualization screen procedure, and F_1 is the subset of F which is composed of elements chosen from F in the procedure. Data in F come with the form $(x_{j1}, x_{j2}, \ldots, x_{jp})$ and data in F_1 have the form of $(x_{j1}, x_{j2}, \ldots, x_{jl})$ where $p > 0$, $l > 0$, and $p > l$.

In visualizing highway loss data, parallel coordinate plots (PCP) served as the primary phase one tool in dimension reduction. The concept of PCP was invented by the French mathematician d'Ocagne (d'Ocagne, 1885). Wegman (1990) discussed the parallel coordinates geometry and demonstrated statistical interpretations, which laid the foundation for applications of parallel coordinates. Parallel coordinates maps a set of points on a line in a p-dimensional Cartesian coordinate system to a set of polylines (or curves) in parallel coordinates all intersecting at $n - 1$ points, thus overcomes the limitation that scatter diagrams do not generalize readily beyond three dimensions. Wegman (1990) implemented parallel coordinates in the way that n axes are drawn in parallel, and a vector $(x_{j1}, x_{j2}, \ldots, x_{jp})$ is created by plotting x_1 on axis 1, x_2 on axis 2, and so on

through x_p on axis p. These p points are joined by a broken line which intersects with each axis, thus a point in the p-dimensional orthogonal coordinate system is transformed to a set iof polylines in 2-dimensional coordinate system. Although in the transformation some loss of information is expected, structures such as linear or nonlinear features, clustering, and outliers can be detected.

Figure 2.1 is an example which represents 25-dimensional ZIP level census data in a parallel coordinates plot based upon Maryland 2000 census data. Wegman (2003) demonstrated the implementation of variable selection and dimension reduction using Brush-Tour, Tour-Prune, color design and other strategies. Moustafa (2011) further explored density estimation techniques to overcome the visual cluttering limitations inherent in the plot and discussed the duality theorem and its usability in identifying patterns visually or by automatic means. In another article published in the same time period, Moustafa discussed space transformed visualization (STV) techniques for visualizing multivariate data, which empowers the discovery of correlated records, clusters and outliers based on the curve's intersections, gaps and isolations, respectively. By visualizing highway loss data structure using above techniques, elements of F_1 can be decided.

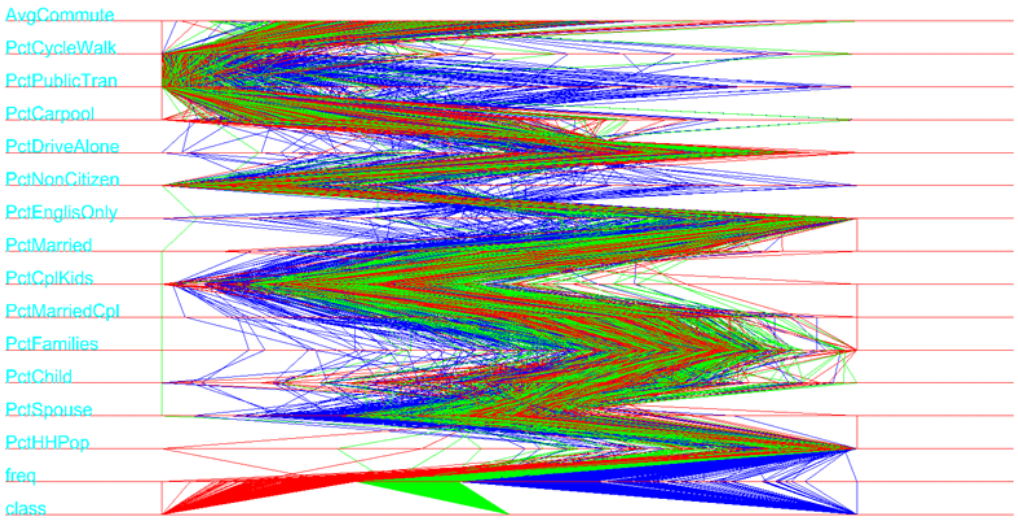


Figure 2.1 Demonstration of Crystal Vision parallel coordinate plot using ZIP level Maryland 2000 census data

2.5.3 Feature Selection via Classification and Regression Trees

Although in phase 1 of variable selection a large number of predictors were excluded, there are still difficulties in deciding the key features. First, to avoid the case that a real predictor is accidentally eliminated, phase 1 tends to be conservative so still a large number of predictors have been produced. Second, complex interactions or patterns may exist in the data. For example, percent of population with high school education or higher is correlated with per capita income, and both variables could influence highway losses of a county.

(F_1, v_2) is defined as the phase 2 variable selection procedure where v_2 is the feature search procedure which produces final key features. Classification and

Regression Trees, commonly referred as CART (CART is a registered trademark of California Statistical Software, Inc.), serves as the tool in this process.

2.5.3.1 Highway Loss Data Input to CART

We sort component estimates $\{(\widehat{\pi}_1, \widehat{\lambda}_1), (\widehat{\pi}_2, \widehat{\lambda}_2), \ldots, (\widehat{\pi}_{g_0}, \widehat{\lambda}_{g_0})\}$ by value of $\widehat{\lambda}_i$ in ascending order and denote the new set $\{(\widehat{\pi}_{(1)}, \widehat{\lambda}_{(1)}), (\widehat{\pi}_{(2)}, \widehat{\lambda}_{(2)}), \ldots, (\widehat{\pi}_{(g_0)}, \widehat{\lambda}_{(g_0)})\}$, $\widehat{\pi}_{(i)}$ is the mixture weight corresponding to the component with estimated parameter of $\widehat{\lambda}_{(i)}$, thus $\widehat{\lambda}_{(g_0)} = \max\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{g_0}\}$, and $\widehat{\lambda}_{(1)} = \min\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{g_0}\}$.

Define λ_α as a threshold such that if highway loss incidents intensity over c_j having underlying intensity greater than λ_α, then c_j are classified as "hot spots" and they are cells considered to have high risk of highway losses. Setting of λ_α often takes form of percentiles of $\lambda_1', \lambda_2', \ldots, \lambda_n'$, which are the observed highway loss incidents intensity for the given sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $\alpha \in [0, 1]$. For example, $\lambda_{0.95}$ is the 95th percentile of $\lambda_1', \lambda_2', \ldots, \lambda_n'$.

Suppose a number, g_h, of components having estimated intensity higher than or equal to λ_α and $g_h \in \{0, 1, \ldots, g_0 - 1\}$, let $I_h = \{\widehat{\lambda}_{(h)}, \ldots, \widehat{\lambda}_{(g_0)}\}$ and $I = \{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{g_0}\}$ respectively denote the high-risk intensity set and the whole intensity set of components and $I_h \subset I$. If an observation is identified from the component corresponds to $\widehat{\lambda}_{(i)} \in I_h$, then I say this observation should be highly risky.

First I form a categorical response variable z used to label the source component of an observation so

$$z_j = \arg\max_i \hat{\pi}_{(i)}(\hat{\lambda}_{(i)}|c_j|)^{y_j} e^{-(\hat{\lambda}_{(i)}|c_j|)} / (\sum_{l=1}^{g_0} \hat{\pi}_{(l)}(\hat{\lambda}_{(l)}|c_j|)^{y_j} e^{-(\hat{\lambda}_{(l)}|c_j|)}) \quad (2.46)$$

I classify z as categorical. It has g_0 levels of values : $1, 2, \ldots, g_0$, and forms a sample $((\boldsymbol{X}_1, z_1), (\boldsymbol{X}_2, z_2), \ldots, (\boldsymbol{X}_n, z_n)$ whose jth observation takes the form $(x_{j1}, x_{j2}, \ldots, x_{jl}, z_j)$ and $(x_{j1}, x_{j2}, \ldots, x_{jl})$. These are the selected features resulting from the initial screen in section 2.5.2.

2.5.3.2 Mechanism of CART

Using the samples formed in the previous section as input, I will use CART to select key features and detect key feature patterns associated with "hot spots". Figure 2.2 illustrates a regression tree used to explore relationships between highway collision frequencies (counts of collisions de vided by registered vehicles) and census data for some of Maryland ZIPs. The collision frequency data in this illustration were artificial. All nodes with descendents are expressed by blue rectangles in the diagram while nodes without descendents are named leaves and are expressed by red rectangles. The target (dependent) variable in this case is a numeric variable, collision frequency. Percentage of residents using public transportation, population density (residents per square mile), percentage of minorities, percentage of married registered drivers, and percentage of pickups of the registered vehicles served as input variables. Conditions based upon values of these inputs decides how to split a parent node into child nodes. Each leaf node can be viewed as the final outcome following the decision path, which begins with the root node and ends in the final position in the tree decided by the regression tree algorithm.

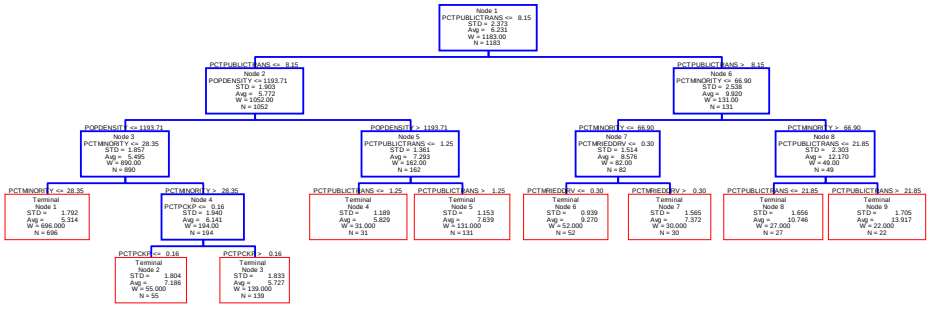


Figure 2.2 A regression tree to explore relationships between ZIP level highway collision frequencies (artificial) with Maryland 2000 census data using Salford CART

Breiman et al. (1984) laid the foundation of the mathematical theory and created algorithms for the implementation of CART. With the whole dataset as the root node in the tree structure, each parent node can be split into two child nodes following certain "splitting rules". The binary partitioning process can be applied “recursively”, so the tree building process goes on until the process is stopped.

Typically CART analysis consists of four basic steps: tree building, tree building stopping, tree pruning, and optimal tree selection. In the process of tree building, the optimal split is selected so splitter and split point are both set by splitting rules. Splitting rules are always in the form

An instance goes left if *CONDITION*, and goes right otherwise

where the CONDITION is generally expressed as "attribute $X_j \leq X_j^R$". Splitting rules depend on algorithms. Typically these algorithms work top-down on a node

by exhaustively searching over all X variables for the best split that minimizes the total impurity of its two child modes. The measure of impurity, often referred as the impurity function, depends on algorithms. When target variable is categorical, one commonly used impurity function is the Gini measure which takes the form

$$i(t) = \sum_{k \neq l} p(k|t)p(l|t) \tag{2.47}$$

where t is the index of node, k, $l = 1, \ldots, K$ are indexes of classes of the target variable, $p(l|t)$ is the conditional probability of class l given node t. For regression, CART often uses least squares (LS, sum of squared prediction error) or least absolute deviation (LAD, sum of absolute prediction errors) to measure the improvement of each split.

The tree growing process stops when: (1) there is only one instance in each of the child nodes so the recursion is impossible; (2) further splitting cannot bring any gain for prediction following splitting rules; (3) the designed stop rule ends the recursion, such as limit of number of levels in "maximal" trees, which are trees grown to a maximal size without the use of a stopping rule.

In pruning the maximal tree, a sequence of simpler and simpler trees are generated from which the final optimal tree is extracted. The CART pruning mechanism begins with a cost-complexity measure based upon training data defined in the form

$$R_\alpha(T) = R(T) + \alpha\,|T| \tag{2.48}$$

where $R(T)$ is the cost of the tree. For the maximal tree $R(T) = 0$ because there is no prediction error. α is the complexity parameter and $|T|$ is the number of leaf nodes in the tree. As α increases, the minimum cost-complexity tree will cut away the bottom splits, which improve prediction the least. The selection of optimal tree from the pruned sequence needs independent test data or cross validation to the learning data. By setting the appropriate complexity parameter α, the information in the learning dataset is fit but not overfit and the optimal tree achieves minimum cost on the test data.

In the process of analysis, CART can rank predictor importance, which is based upon the sum of the improvements in all nodes in which the predictor appears as a splitter. Variable Importance (VI) is defined under the context of algorithms. For the Random Forests algorithm, the most advanced VI is the "permutation accuracy importance". By randomly permuting the values of a predictor variable, its original association with the target is broken and thus there exists a difference in prediction accuracy before and after permuting a variable. The permutation accuracy importance is based on the average of these differences over all trees. Breiman and Cutler (2008) define it as

$$VI(X_i) = \frac{\sum_{t=1}^{n_{tree}} VI^{(t)}(X_i)}{n_{tree}}$$

where t denotes the index of a tree and n_{tree} is the total number of trees constructed. This can be rescaled to a "z-score"

$$z(X_i) = VI(X_i)/(\frac{\widehat{\sigma}}{\sqrt{n_{tree}}})$$

where $\frac{\widehat{\sigma}}{\sqrt{n_{tree}}}$ is the standard error.

2.5.3.3 Classification Tree Output and Regression Tree Validation

Let $X_1, X_2, \ldots, X_m$ be key features selected from the classification tree described in the previous section, $X_1, X_2, \ldots, X_m$ are in descending order of their importance score and thus X_1 has the highest importance level. I define hot-spot patterns first and then develop methods

The categorical response variable z defined in (2.46) denotes the source component from which the jth observation originates, and thus z has levels valued in $0, 1, \ldots, g_0$. In (2.46) the source component is already sorted by its value from the least to the greatest, and thus for the observation y_j, if $z_j \in \{h, h+1, \ldots, g_0\}$, then it is classified as highly risky.

Let A denote the key feature space defined by $X_1, X_2, \ldots, X_m$ and their corresponding domains. I partition key feature space such that $A = A_1 \cup A_2 \cup \cdots \cup A_{g_0}$ and $A_i \cap A_j = \emptyset$ where $i, j \in (1, 2, \ldots, g_0)$ and $i \neq j$, and there exists a mapping $F_c(A_i) \longrightarrow i$, where F_c denotes the classification tree prediction algorithm, thus we define the hot-spot key feature space (highly risky) as

$$A_{HR} = A_h \cup A_{h+1} \cup \cdots \cup A_{g_0} \tag{2.49}$$

We denote P_j be the path from the root node to the leaf corresponding to the observation y_j, and z_j be its predicted component label, hot-spot patterns P_{HR} are defined as

$$P_{HR} \underset{j \in \{1, 2, \ldots g_0\}}{=} \{P_j\} \tag{2.50}$$

2.5.4 Summary of Chapter 2

In this chapter, I have built methods extending EM (expectation maximization) algorithm to Poisson point processes with incomplete data structure to undercover the underlying components characterizing highway loss events. With component information obtained, I have developed methods that use classification and regression trees along with visualization procedures to identify key features influencing highway loss intensities, and detect key feature patterns of the "hot spot" loss areas.

3. Prediction of Highway Loss Incidents by the Use of Bayesian Hierarchical Spatio-Temporal Model

In Chapter 2, methods for determining subpopulations of highway loss incidents in the study area D_s at time t have been developed, and features having key influences on the highway loss intensity have been identified. Meanwhile, key feature vector patterns corresponding to "hot spots" of losses also have been defined. In this chapter, I start from clustering cells of the study area by mapping the partition of the key feature space to the partition of the geographical space. By doing so, cells in the key feature space "close" to each other are aggregated so that the homogeneity can be built and prediction of future losses on the study area can be based upon aggregated cells instead of on each single cell. Then, for each cluster, a Bayesian Hierarchical Model (BHM) is designed to predict losses at $t+1$ using the posterior of the current losses at time t, and the posterior of the most recent past losses at time $t-1$ in the Bayesian modeling. The proposed Bayesian model has an updating mechanism and thus adds adaptation to the Bayesian approach.

3.1 Key Feature Space Partition and Study Area Partition

3.1.1 *K*-means Clustering Algorithm

The *K*-means clustering algorithm was named by MacQueen (1967) and Hartigan and Wong (1979) detailed the algorithm in Fortran. It is old yet vigorous and powerful. A more recent development was by Ahmad and Dey (2007), which moved *K*-means clustering a big step forward by allowing the use of mixed numerical and categorical data.

Let $\{c_1, c_2, \ldots, c_n\}$ be the set of geographic cells, which partitioned the study region $D_{\boldsymbol{s}}$ and $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be the n m-dimensional key feature vectors corresponding to $\{c_1, c_2, \ldots, c_n\}$. Here $\boldsymbol{x}_j$ is the observed feature vectors on c_j, and the mapping between $\boldsymbol{x}_j$ and c_j is a one-to-one mapping $\boldsymbol{x}_j \leftrightarrow c_j$. Each element of the m-dimensional vector is a key feature identified by Section 2.5.3.3. I denote the key feature space $\boldsymbol{E}_{\boldsymbol{x}} \subset \mathcal{R}^m$ and $\boldsymbol{E}_{\boldsymbol{x}}$ is a m-dimensional Euclidean space. The *j*th observation of the feature vector set $\boldsymbol{x}_j$ takes the form $(x_{j1}, x_{j2}, \ldots, x_{jm})$ and $j = 1, 2, \ldots, n$.

To cluster $(c_1, c_2, \ldots, c_n)$ into K homogeneous disjoint partitions, C_1, $C_2, \ldots, C_K$ ($K \leq n$), I use the *K*-means clustering method on n vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and partition them into K sets $\boldsymbol{S} = \{S_1, S_2, \ldots, S_K\}$. Here clustering is used to partition and assign a set of objects into homogeneous clusters on the basis of measures of distance so that the objects (vectors/points) in the same cluster are more similar/closer to each other than to those in other clusters. In the Euclidean space $\boldsymbol{E}_{\boldsymbol{x}}$, the distance between two vectors, $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$, is measured by the distance function

$$\mathrm{d}(\boldsymbol{x}_u,\boldsymbol{x}_v) = \mathrm{d}(\boldsymbol{x}_v,\boldsymbol{x}_u) = \sqrt{\sum_{i=1}^{m}(x_{ui} - x_{vi})^2} \tag{3.1}$$

where x_{ji} is the ith element of the vector $\boldsymbol{x}_j$.

The whole process of the partition can be described as to cluster n vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ into K sets $\boldsymbol{S} = \{S_1,\ S_2, \ldots, S_K\}$ so as to minimize the within-cluster sum of squares (WCSS)

$$\arg\min_{\boldsymbol{S}} \sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in \mathrm{S}_i} ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2 \tag{3.2}$$

where $\boldsymbol{\mu}_i$ is the mean of points within cluster S_i and $||\boldsymbol{x}_j - \boldsymbol{\mu}_i||$ denotes the distance between $\boldsymbol{x}_j$ and $\boldsymbol{\mu}_i$ as defined in (3.1).

The nature of the *k*-means algorithm is to search for a *k*-partition with the locally optimal within-cluster sum of squares by moving points from one cluster to another. Other versions of *k*-means clustering followed the same idea and can be summarized in two steps after setting the initial *K* means, $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}$,

Assignment- step: Assign each vector to the cluster whose mean is closest to it.

$$S_i^{(t)} = \{\boldsymbol{x}_j : ||\boldsymbol{x}_j - \boldsymbol{\mu}_i|| \ \leq ||\boldsymbol{x}_j - \boldsymbol{\mu}_l|| \ \forall \quad l \neq i\}$$

where $\mathrm{S}_i^{(t)}$ is the ith cluster at tth iteration, $1 \leq l \leq K$, each point can only be assigned to one cluster

Update - step

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\boldsymbol{x}_j \in S_i^{(t)}} \boldsymbol{x}_j$$

where $|S_i^{(t)}|$ is the cardinality of $S_i^{(t)}$.

The algorithm repeats the Assignment-*step* and the Update -*step* until no further change of assignments are made and thus converged.

3.1.2 Modification of the Distance Function

The key features $X_1, X_2, \ldots, X_m$ decided in Section 2.5.3.3 were sorted in descending order by their importance level. Thus, the first element of the vector $\boldsymbol{x}_j$, x_{j1} is a realization of the feature with the highest importance level, X_1, and the last element x_{jm} is a realization of the feature with the lowest importance level, X_m. I make some modifications to the traditional Euclidean distance based upon two considerations: first, I need to take into consideration influences of scale of key features on distance functions; second, I want the features with higher importance levels to contribute more to the distance function.

I apply the robust MAD standardization (median absolute deviation from median) to X_i first. The median absolute deviation is defined as,

$$\text{MAD}_{\cdot i} = \text{median}_j(|x_{ji} - \text{median}_l(x_{li})|)$$

The MAD standardization uses median absolute deviation as scale and uses the median as location, thus

$$x'_{ji} = \frac{x_{ji} - \text{median}_l(x_{li})}{\text{MAD}_{\cdot i}} \tag{3.3}$$

where x'_{ji} is the MAD standardized x_{ji}.

Let VI_i be the importance level of X_i thus $VI_1 \geq VI_2 \geq \ldots \geq VI_m > 0$, and define the relative variable importance level as ω_i, where

$$\omega_i = \frac{VI_i}{VI_1}. \tag{3.4}$$

Thus, the distance function is defined as

$$\mathrm{d}(\boldsymbol{x}'_u, \boldsymbol{x}'_v) = \mathrm{d}(\boldsymbol{x}'_v, \boldsymbol{x}'_u) = \sqrt{\sum_{i=1}^{m} \omega_i^2 (x'_{ui} - x'_{vi})^2}. \tag{3.5}$$

3.1.3 Determination of Number of Clusters

The *K*-means algorithm needs the number of clusters, k, as input. There are no completely satisfactory methods that can be used to determine the number of population clusters for cluster analysis. Fang and Wang (2012) proposed a method that selected the number of clusters via the bootstrap method. Liang et al. (2012) developed a method determining the number of clusters using information entropy for mixed data. Wang (2010) presented a method via cross-validation. Typically the data used for testing the above methods contained no more than 10 clusters. It seems that there is no hypothesis test good enough to produce an exact k.

Different criterion may produce different results in the number of clusters selected. Wong and Shaack (1982) described the kth-nearest-neighbor density estimate. Based upon this concept when varying values of k yields a constant number of modal clusters it is strong evidence that at least that many modes are in the population. Here I use Hartigan's index defined in Hartigan (1975) to determine the number of clusters to partition the key feature space. Let *W(k)* denote the within-cluster sum of squares. Hartigan's index defined as following

$$H(k) = \gamma(k) \frac{W(k) - W(k+1)}{W(k+1)} \tag{3.6}$$

where $\gamma(k) = n - k - 1$. Hartigan (1975) showed that from k to $k+1$, $\frac{H(k+1)}{H(k)}$ is not monotone, thus by comparison of improvement, an optimal value of k can be decided. Here I developed the following procedure,

Inputs

MAD standardized matrix of $\boldsymbol{X}'_{n\times m}$

Control parameters

g_0

n

Initialization

apply k-means clustering to $\boldsymbol{X}'_{n\times m}$ assuming g_0 clusters

compute $W(g_0)$ and $\gamma(g_0)$

for $k = g_0, \ldots, n$

apply k-means clustering to $\boldsymbol{X}'_{n\times m}$ assuming $k+1$ clusters

compute $W(k+1)$, $\gamma(k+1)$, and $H(k)$

if $H(k-1)$ exists

if $H(k)/H(k-1)$ no longer monotone, exit for loop

end if

end if

$k \leftarrow k+1$

end for

Return

3.1.4 Mapping the Key Feature Space Partition to the Study Area Partition

Let $\boldsymbol{S} = \{S_1, S_2, \ldots, S_k\}$ be partitions decided by the k-means clustering procedure in Section 3.1.4 and let $\boldsymbol{\mu}_i$ be the mean of cluster S_i thus

$$S_i^{(t)} = \{\boldsymbol{x}_j : ||\boldsymbol{x}_j - \boldsymbol{\mu}_i|| \leq ||\boldsymbol{x}_j - \boldsymbol{\mu}_l|| \ \forall \quad l \neq i\}.$$

Based upon the one-to-one relationship of $\boldsymbol{x}_j \leftrightarrow c_j$, I map $S_i \rightarrow C_i$

$$C_i = \{c_j : ||\boldsymbol{x}_j - \boldsymbol{\mu}_i|| \leq ||\boldsymbol{x}_j - \boldsymbol{\mu}_l|| \ \forall \quad l \neq i\}. \tag{3.7}$$

The equation (3.7) partitions D_s into K clusters (groups) of geographic cells thus

$$D_s = \bigcup_{i=1,\ldots,k} C_i \text{ and } C_i \bigcup_{i \neq l} C_l = \emptyset.$$

3.2 Bayesian Hierarchical Model on Spatio-Temporal Process

3.2.1 Bayesian Hierarchical Model (BHM)

The whole Spatio-Temporal point process of highway losses can be thought as a temporal process of spatial point processes, let $Z(\cdot\,;\cdot)$ be the counting process as described in Section 2.1 and its conditional intensity function is defined as

$$\psi(\boldsymbol{s};t) \equiv \lim_{\substack{|d\boldsymbol{s}| \to 0 \\ d_t \to 0}} \frac{E(Z(d_{\boldsymbol{s}};d_t)\,|\mathcal{H}_t)}{\nu(d\boldsymbol{s})d_t}.$$

The BHM proposed here first separate spatial components from temporal components (always discrete) to avoid the correlation entanglement between these two kinds of components, and then at a given time point t partition the nonstationary study area into subregions so that the stationarity can be well assumed in each subregion.

The term Hierarchical Model (HM) here means the uncertainty in data. The uncertainty in the modeling has to be decomposed into two or more levels, and hence involves several levels of conditional distributions. I follow the terminologies that Berliner (1996) used to describe the levels of modeling discussed in Chapter 2.

Data model $[Z|Y,\theta]$, the top level expresses the distribution of the data given a hidden process, e.g., the observed number of events in certain areas at time point t once the underlying Poisson point process is given.

Process model $[Y|\theta]$, underneath the top level is the process model level. This level models the uncertainty of the hidden process in the above data model through a conditional probability distribution, given that all parameters prior to the hidden process are known. For example, in Section 2.4 the parameters of each component of the mixture model have been decided, a multinomial distribution "decides" the dominating Poisson process of sub regions/cells of the study area.

Parameter model $[Y|\theta]$, the bottom level models the uncertainty of parameters prior to the process model, e.g., the parameters of components in the mixture model in Section 2.4.

It should be noted that the parameter model can also be made up of submodels through sublevel conditional prior distributions. It is also possible that a Hierarchical Model does not have the process model, but has the data model and multilevel parameter models.

The Bayesian approach, which is fundamentally different from the classical frequentist approach described in Chapter 2, is applied in the Hierarchical Model

described above. The origin of the Bayesian approach comes from Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{3.8}$$

where $P(A)$ and $P(B)$ respectively denote the probabilities of event A and B while $P(A|B)$ and $P(B|A)$ denote the conditional probability of event A given event B, and the conditional probability of event B given event A, respectively.

I follow notations and definitions in Section 4.1 of Shao (2003) to introduce the Bayesian approach under the context of decision theory. Let θ be a realization of a random vector $\boldsymbol{\theta}$ whose *prior* distribution is Π on Θ. A sample Y is drawn from the conditional distribution of Y given $\boldsymbol{\theta} = \theta$, $P_\theta = P_{y|\theta}$, the observed sample $Y = y$ is then used to obtain an updated prior distribution, the *posterior* distribution, $P_{\theta|y}$, and its density function takes the form

$$p_{\theta|y} = \frac{dP_{\theta|y}}{d\lambda} = \frac{f_\theta(y)\pi(\theta)}{m(y)}, \tag{3.9}$$

where $P_{y|\theta}$ is dominated by the σ-finite measure ν and $f_\theta(y) = \frac{dP_{y|\theta}}{d\nu}$ is a Borel function on $(\mathcal{Y} \times \Theta, \sigma(\mathcal{B}_\mathcal{Y} \times \mathcal{B}_\Theta))$ and $\mathcal{Y}$ is the range of y; $\frac{d\Pi}{d\lambda} = \pi(\theta)$ is dominated by the σ-finite measure λ and $m(y) = \int_\Theta f_\theta(y)\pi(\theta)\, d\lambda$. The posterior distribution $P_{\theta|y}$, conditional on the observed $Y = y$, contains all the information needed to make statistical decisions and inference. I define an action space $\mathbb{A}$ in a decision problem and define $L(\theta, a) \geq 0$ as a loss function (e.g. squared error loss), for any $y \in \mathcal{Y}$, a Bayes action with respect to Π is any $\delta(y) \in \mathbb{A}$ such that

$$E[L(\theta, \delta(y))|\, Y = y] = \min_{a \in \mathbb{A}} E[L(\theta, a)|\, Y = y] \tag{3.10}$$

where the expectation is with respect to the posterior distribution $P_{\theta|y}$. In the context of this dissertation, the Bayes action exists and is unique.

I use $(f_\theta(y) = \frac{dP_{y|\theta}}{d\nu},\ \pi(\theta) = \frac{d\Pi}{d\lambda})$ to denote the *Bayesian statistical model*, a *Bayesian Hierarchical Model* is a Bayesian statistical model that has either three layer models (data model, process model and parameter model) or two layer models (data model and parameter model) whose parameter model is a multilevel model, where the prior distribution $\pi(\theta)$ is decomposed into conditional distributions

$$\pi_1(\theta|\theta_1),\ \pi_2(\theta_1|\theta_2), \ldots, \pi_n(\theta_{n-1}|\theta_n)$$

and a marginal distribution $\pi_{n+1}(\theta_n)$ such that

$$\pi(\theta) = \int_{\Theta_1 \times \Theta_2 \times \cdots \times \Theta_n} \pi_1(\theta|\theta_1)\, \pi_2(\theta_1|\theta_2) \cdots \pi_n(\theta_{n-1}|\theta_n)\ \pi_{n+1}(\theta_n) d_{\theta_1} \cdots d_{\theta_{n+1}} \tag{3.11}$$

where Θ_i is the parameter space for θ_i and $i = 1, 2, \ldots, n$.

3.2.2 Assumptions Made in the Bayesian Hierarchical Model

I make two assumptions for the BHM proposed in this dissertation. The first is the homogeneity assumption. At a certain time point t of study interest, let C_i be a cluster generated from Section 3.1.4 based upon k-means clustering on the key feature space, Completely Spatial Randomness (CSR) is assumed within C_i.

The second is the first-order Markov property. Denote a time series by $\{Y_t : t = 0, 1, \ldots, T\}$, and denote distribution of Y_t as $[Y_t]$, then the joint distribution of the time series is denoted as $[Y_0, Y_1, \ldots, Y_t]$.

I assume

$$[Y_0, Y_1, \ldots, Y_t] = [Y_0]\prod_{t=1}^{T}[Y_t|Y_{t-1}]. \tag{3.12}$$

It means only the most recent past of the whole past determines the conditional probabilities about the present.

3.2.3 Prior Information and Prior Distribution of the BHM

In the Bayesian interpretation a probability measures a degree of belief while in the frequentist interpretation it measures a proportion of outcomes. The prior probability distribution, denoted by $\pi(\theta)$, is a hypothesis made on the uncertainty of θ before observed evidence is obtained from the distribution $P_{y|\theta}$ dominated by θ. The posterior distribution $P_{\theta|y}$ can be viewed as the result of the correction the observed evidence made on the prior distribution. Thus, the posterior is determined by two factors, the *prior* $\pi(\theta)$, and $P_{y|\theta}$, also known as *likelihood*. Under Bayesian approach, when a prior is known, the derivation of the posterior is obtained by dividing the resulting joint distribution by its marginal distribution. Once the posterior is produced, inference, estimation, and prediction can be made based upon it. It is obvious the prior distribution is the key to Bayesian inference.

A convincing Bayesian inference to a large extent depends on making the right decision in selecting an appropriate prior. Practically there are two difficulties in the selection of the prior distribution: first, there is no prior information precise enough to provide a basis for the selection; second, prior information is enough while there are more than one distributions compatible with that prior information which makes the selection not unique. Historically critics of Bayesian paradigm have focused their criticisms on hypothesis of prior distributions. The recent

developments in Bayesian robustness analysis and the introduction of hierarchical modeling have largely quelled these criticisms.

In recent years, the objective Bayesian inference theory has made great progress. This theory allows that prior distributions used to make an inference to be least informative in a certain information-theoretic sense (Berger, 2009a, 2009b). In contrast, Williamson (2010) agrees that priors usually represent subjective judgments of opinion in practice that cannot be rigorously justified. I think in practice the selection of priors is influenced by the research interest along with the prior information. If prior information is available about θ, it should be taken into consideration in the design of $\pi(\theta)$, especially when the present data are related to previous data in a certain way, and noninformative priors can serve as a validation in the belief that they should not produce results significantly inconsistent. On the other hand, the indeterminacy in the selection of prior distribution influences the posterior distribution, even if the prior information is precise. Thus, ideal priors should be robust and the process of selection should limit the arbitrariness. Additionally the resulting posterior should take close form without adding complexity to the model.

Berger (1990) introduced the concept of classification into the robustness analysis. In this paper, Berger proposed the uncertainty about the prior distribution $\pi(\theta)$ could be represented by the assumption that $\pi(\theta)$ belonged to a class of distributions $\mathcal{P}_\theta$. Berger (1990) recommended that *conjugate* prior classes should be used when the likelihood was in an exponential family, which takes a generic form $f(y|\theta) = h(y)\, e^{\theta y - \phi(\theta)}$, then its priors can be expressed as

$$\pi(\theta|\mu,\nu) = K(\mu,\nu)\, e^{\theta\mu - \nu\phi(\theta)} \tag{3.13}$$

where $K(\mu,\nu)$ is a normalizing constant, and the corresponding posterior distribution takes form $\pi(\theta|\mu + y,\ \nu + 1)$. Here $\frac{d\Pi}{d\lambda} = \pi(\theta)$ is dominated by the σ-finite measure λ and $\nu > 0$ and $\frac{\mu}{\nu} \in \mathring{N}$ holds.

The use of conjugate priors is desirable. When conjugacy holds, the posterior is in the same family as the prior distribution, thus the evidence only corrects the parameter of the hypothesis. Conjugate priors also have intuition and rationale when showing how the evidence updates the priors. In addition, the computation and the following inference are convenient.

The use of hierarchical models adds extra submodels to model the uncertainty of parameters of priors, e.g., the first-level prior distribution can be denoted as $\Pi_{\theta|\xi}$, and if necessary, the second-level distribution can be introduced, $\Psi_{\xi|\eta}$, the hierarchical levels can increase until it meets the demand of modeling. Because misspecifying a second-level prior is much less serious than misspecifying a first-level prior (Berger, 1985, Section 4.6), the multi-level model brings more robustness, and the use of noninformative priors in the second-level prior is better justified than in the first-level prior.

3.2.4 Design of the BHM

3.2.4.1 Conjugacy

Let C_k be the kth ($k = 1, \ldots, K$) cluster determined in Section 3.1.4 resulting from the partition of study region $D_{\boldsymbol{s}}$ at time t, and t is discrete and $t-1$ be its most recent past. At time t, let c_{k_j}, the cell indexed by k_j, be the jth element of C_k where $k_j = k_1, \ldots, k_J$ and let $J_{(k)}$ denotes the dimension (number of cells) of C_k, thus $J_{(k)}$ is decided by C_k and we have $\sum_{k=1}^{K} J_{(k)} = n$, which is the total

number of cells of $D_{\boldsymbol{s}}$. Let Y_{k_j}, a discrete stochastic random variable, be the number of events observed on c_{k_j} and $|c_{k_j}|$ be the area of c_{k_j}, and let y_{k_j} be the realization of Y_{k_j}, thus $y_{k_1}, \ldots, y_{k_J}$ is a random sample of highway loss events in C_k at time t. Let Y'_{k_j} be the corresponding discrete stochastic random variable of Y_{k_j} at the most recent past $t-1$ and y'_{k_j} be its realization, and $y'_{k_1}, \ldots, y'_{k_J}$ form a random sample of highway loss events in C_k at time $t-1$.

Based upon the CSR assumption made on Section 3.2.2, let λ_k denote the intensity dominating the Poisson counting process on C_k, then the probability mass function of Y_{k_j} on c_{k_j} is

$$f(y_{k_j}|\lambda_k) = \Pr(Y_{k_j} = y_{k_j}) = \frac{(\lambda_k |c_{k_j}|)^{y_{k_j}}}{y_{k_j}!} e^{-(\lambda_k |c_{k_j}|)}. \tag{3.14}$$

Here I propose a Gamma distribution $\pi(\lambda_k)$ as the prior distribution of the likelihood $P_{y_{k_j}|\lambda_k}$ whose probability mass function is $f(y_{k_j}|\lambda_k)$, the density function of the prior is

$$\pi(\lambda_k|\alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \lambda_k^{\alpha_k - 1} e^{-\beta_k \lambda_k}.$$

The likelihood is

$$\begin{aligned} f(y_{k_1} \ldots y_{k_J}|\lambda_k) &= \prod_{k_j=k_1}^{k_J} \frac{(\lambda_k |c_{k_j}|)^{y_{k_j}}}{y_{k_j}!} e^{-(\lambda_k |c_{k_j}|)} \\ &\propto (\lambda_k)^{\sum_{k_j=k_1}^{k_J} y_{k_j}} e^{-(\sum_{k_j=k_1}^{k_J} |c_{k_j}|)\lambda_k} \end{aligned}$$

after omitting constants. The posterior

$$\pi(\lambda_k|y_{k_1} \ldots y_{k_J}) \propto f(y_{k_1} \ldots y_{k_J}|\lambda_k)\, \pi(\lambda_k|\alpha_k, \beta_k)$$

$$\propto (\lambda_k)^{\sum_{k_j=k_1}^{k_J} y_{k_j}+\alpha_k-1} e^{-(\sum_{k_j=k_1}^{k_J}|c_{k_j}|+\beta_k)\lambda_k}, \qquad (3.15)$$

and also takes the form of a Gamma distribution,

$$\pi(\lambda_k|y_{k_1}\dots y_{k_J}) \sim \text{Gamma}(\sum_{k_j=k_1}^{k_J} y_{k_j}+\alpha_k, \sum_{k_j=k_1}^{k_J}|c_{k_j}|+\beta_k),$$ it has been shown in (3.15) that the proposed prior is a conjugate prior.

3.2.4.2 Updating Mechanism

Feature space formation (also known as the key feature selection) and the key feature space partition are done at each time point t. Thus, they depend on t. Bayesian Hierarchical Model is based on loss data $y'_{k_1},\dots,y'_{k_J}$ on C_k at $t-1$ first using a Gamma-Poisson hierarchical model implemented by Gibbs sampler (described in later section)

Let $\lambda_k^{'}$ denote the intensity on C_k at the most recent past $t-1$, and propose the distribution density of $\lambda_k^{'}$ as

$$\pi(\lambda_k^{'}|\alpha_k^{'},\beta_k^{'}) = \frac{{\beta'_k}^{\alpha'_k}}{\Gamma(\alpha'_k)}{\lambda'_k}^{\alpha'_k-1}e^{-\beta'_k\lambda'_k}.$$

Its posterior distribution

$$\pi(\lambda'_k|y'_{k_1}\dots y'_{k_J},\alpha_k^{'},\beta_k^{'}) \propto f(y'_{k_1}\dots y'_{k_J}|\lambda'_k)\,\pi(\lambda'_k|\alpha'_k,\beta'_k)$$
$$\propto (\lambda'_k)^{\sum_{k_j=k_1}^{k_J} y'_{k_j}+\alpha'_k-1} e^{-(\sum_{k_j=k_1}^{k_J}|c_{k_j}|+\beta'_k)\lambda'_k}, \qquad (3.16)$$

serves as the prior distribution for present losses at t. The corresponding posterior distribution of λ_k for the present t, denotes as $f(\lambda_k|y_{k_1}\dots y_{k_J})$ in (3.15), serves as the prior distribution in modeling the intensity distribution of the nearest future $t+1$. Thus an updating mechanism is well built.

3.2.4.3 BHM Modeling at $t-1$

The completely spatial randomness (CSR) assumption made in Section 3.2.2 is based on time specific feature space formation and partition at time t. I extend this assumption to $t-1$ which is justified for the following reasons. First, changes in feature space formation and partition from $t-1$ to t are expected to be limited, thus, the compromise on CSR should not be beyond acceptance. Second, even a compromise exists, the loss of precision is limited in the prior at t (the prior is by nature a hypothesis) and the prior will have a chance to be corrected by observations (evidence) at t. Next, I propose one hierarchical model for the most generic case and two empirical alternatives.

Hierarchical model

Figure 3.1 is the graphic model for the most generic case where all hyperparameters (α', β') are unknown. Notations in Figure 3.1 follow the directed acyclic graph (DAG) for the Bayesian network originally defined in Spiegelhalter (1998). In DAGs, a node ζ, is referred to be a parent node of ξ if an arrow emanating from ζ points to ξ, and ξ is said to be a child node. Stochastic dependencies are denoted by single-edged arrows while functional dependencies are denoted by double-edged arrows. Rectangular nodes denote known constants while elliptical nodes denote deterministic relationships or stochastic quantities. Repetitive entities such as loops are denoted by overlapped plates. My interest is primarily focused on stochastic nodes. Thus, constants are ignored and deterministic relationships are collapsed in the description of probabilistic relationships between stochastic nodes,

Figure 3.1 illustrates the two-stage BHM at time $t-1$ where $\alpha_k^{'}$, $\beta_k^{'}$, $y_{k_j}^{'}$, $\lambda_k^{'}$ are all stochastic nodes having unknown parameters. $\lambda_k^{'}$ is the first-stage prior with unknown hyperparameters $\alpha_k^{'}$ and $\beta_k^{'}$ while $\alpha_k^{'}$ and $\beta_k^{'}$ are second-stage priors whose hyperparameters A_α, B_α, and B_β are tuning parameters already known in the implementation.

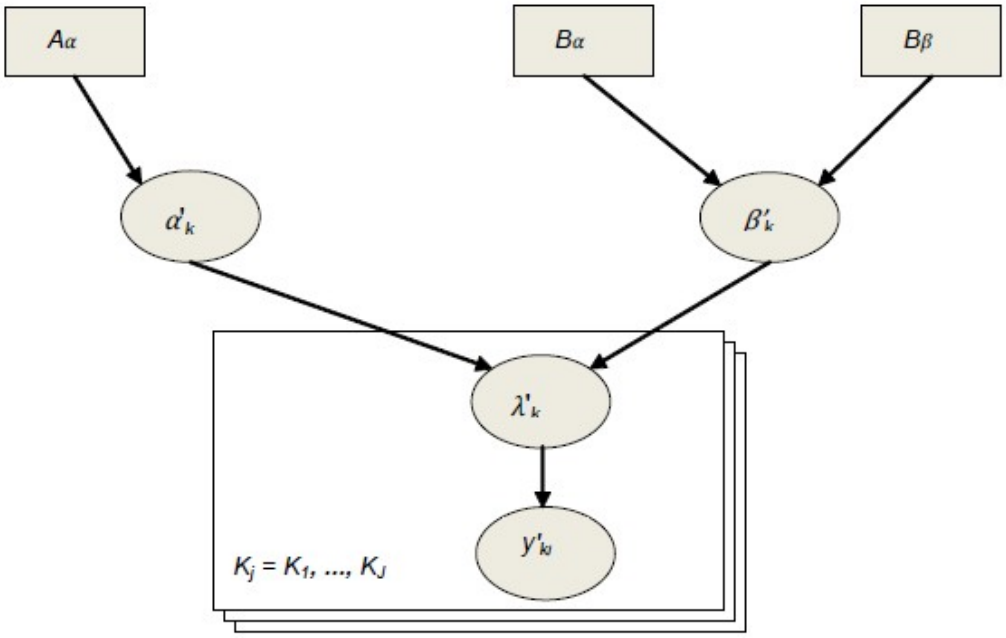


Figure 3.1 Graphic model of the Hierarchical model

$\lambda_k^{'}$ is the proposed intensity dominating the Poisson counting process on C_k=$\{c_{k_1}, \ldots, c_{k_j}, \ldots, c_{k_J}\}$. $y_{k_j}^{'}$ is the number of loss events observed, conditional on λ', and the density mass function has the same form as in (3.13)

$$f(y'_{k_j}|\lambda'_k) \ = \Pr\ (Y'_{k_j} = y'_{k_j}) \frac{(\lambda'_k\ |c_{k_j}|\ )^{\ y'_{k_j}}}{y'_{k_j}!} e^{-(\lambda'_k\ |c_{k_j}|)}. \tag{3.17}$$

λ'_k is proposed to follow a Gamma distribution conditional on $\alpha'$and $\beta'$where both hyperparameters unknown. It has the same form with (3.14)

$$\pi(\lambda'_k|\alpha'_k,\ \beta'_k) \ = \frac{\beta'^{\alpha'_k}_k}{\Gamma(\alpha'_k)} \lambda'^{\alpha'_k-1}_k e^{-\beta'_k\lambda'_k}. \tag{3.18}$$

α' is proposed to follow an exponential distribution with a known parameter A_α,

$$\pi(\alpha'_k) \ = A_\alpha e^{-A_\alpha\alpha'_k}. \tag{3.19}$$

β' is proposed to follow a Gamma distribution with known parameters B_α and B_β,

$$\pi(\beta'_k) \ = \frac{B_\beta{}^{B_\alpha}}{\Gamma(B_\alpha)} \beta'^{\ B_\alpha-1}_k e^{-B_\beta\beta'_k}. \tag{3.20}$$

Assuming independence between $\alpha'$and β', then $\pi(\alpha'_k,\ \beta'_k) = \pi(\alpha'_k)\pi(\ \beta'_k)$ and $\alpha'_k > 0$ and $\beta'_k > 0$.

To get the full conditional distribution for $\alpha'_k$and β'_k, I began with the distribution of α' and β' conditional on λ', after ignoring constants,

$$\pi(\alpha'_k,\ \beta'_k|\lambda'_k) \ \propto \frac{\beta'^{\alpha'_k}_k}{\Gamma(\alpha'_k)} \lambda'^{\ \alpha'_k-1}_k e^{-\beta'_k\lambda'_k} \ \cdot \pi(\alpha'_k,\ \beta'_k)$$

and

$$\pi(\alpha'_k|\lambda'_k,\ \beta'_k) \ \propto \frac{\beta'^{\alpha'_k}_k}{\Gamma(\alpha'_k)} \lambda'^{\ \alpha'_k-1}_k \cdot \pi(\alpha'_k) \tag{3.21}$$

and

$$\pi(\ \beta'_k|\lambda'_k, \alpha'_k) \ \propto \beta'^{\ \alpha'_k}_k e^{-\beta'_k\lambda'_k} \cdot \pi(\beta'_k). \tag{3.22}$$

It must be noted that $\lambda_k' > 0$ in (3.18), (3.21) and (3.22).

Empirical model 1

The empirical model illustrated in Figure 3.2 is a simplified version of the above hierarchical model. α' becomes a constant, β' takes the same prior distribution as in (3.20), and the full conditional distribution of λ_k' in (3.17) simplifies to

$$\pi(\lambda_k' | \beta_k') \propto \beta_k'^{\alpha_k'} \lambda_k'^{\alpha_k'-1} e^{-\beta_k' \lambda_k'}. \tag{3.23}$$

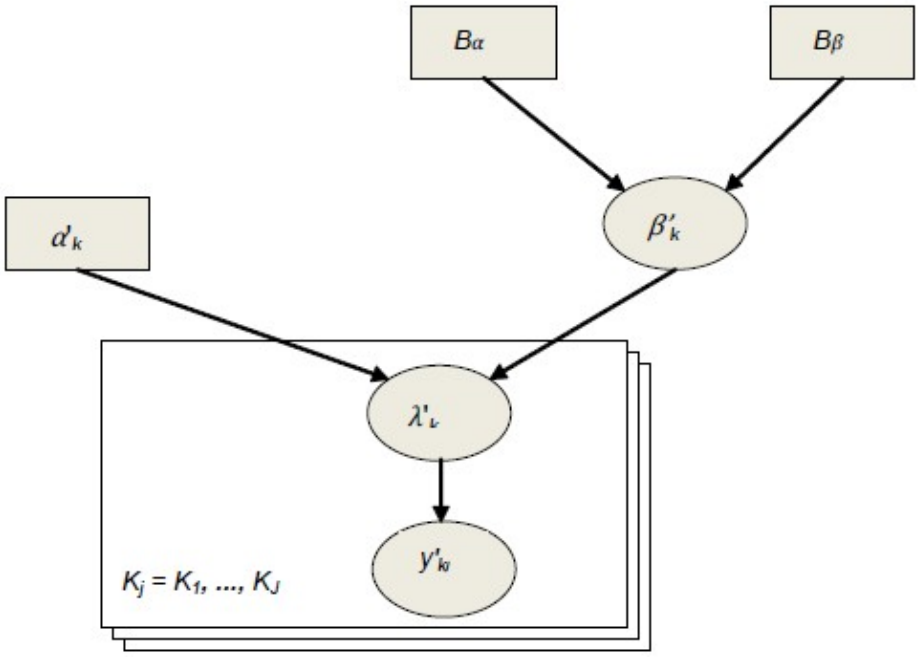


Figure 3.2 Graphic model of Empirical model 1

$$\pi(\beta_k^{'}|\lambda_k^{'},\alpha_k^{'}) \propto {\beta_k^{'}}^{\alpha_k^{'}} e^{-\beta_k^{'}\lambda_k^{'}} \cdot \pi(\beta_k^{'}). \tag{3.24}$$

In an empirical Bayes spirit, I derive $\alpha_k^{'}$ by a method-of-moments empirical argument. Let $Y'_{k_1},\ldots,Y'_{k_J}$ be random variables measuring highway loss events in C_k at time $t-1$ from the distribution $P_{\lambda'_k}$ with density mass function of $f(y'_{k_j}|\lambda'_k)$ in (3.16), then the lth sample moment, $\hat{\mu'_l}$ is an unbiased estimator of the lth moment, $\mu'_l = E(Y'_{k_j})^l$. Let η'_{k_j} be the resultant loss event intensity on c_{k_j} at $t-1$ and $\eta'_{k_j} = \frac{y'_{k_j}}{|c_{k_j}|}$, then

$$E(\eta'_{k_j}) = \frac{\alpha'_k}{\beta'_k} \approx \overline{\eta'_k} \tag{3.25}$$

$$\begin{aligned} V(\eta'_{k_j}) &= VE(\eta'_{k_j}|\lambda'_k) + EV(\eta'_{k_j}|\lambda'_k) \\ &= VE(\tfrac{y'_{k_j}}{|c_{k_j}|}|\lambda'_k)) + EV(\tfrac{y'_{k_j}}{|c_{k_j}|}|\lambda'_k)) = V(\lambda'_k) + \tfrac{1}{|c_{k_j}|}E(\lambda'_k) \\ &= \tfrac{\alpha'_k}{(\beta'_k)^2} + \tfrac{1}{|c_{k_j}|}\tfrac{\alpha'_k}{\beta'_k} \approx S^2_{\eta'_k} \end{aligned} \tag{3.26}$$

where $\overline{\eta'_k}$ and $S^2_{\eta'_k}$ are respectively the sample mean and sample variance of the resultant loss event intensity η'_{k_j}, $\overline{\eta'_k} = \frac{1}{J_{(k)}}(\sum_{k_j=k_1}^{k_J}\eta'_{k_j})$ and $S^2_{\eta'_k} = \frac{1}{J_{(k)}}(\sum_{k_j=k_1}^{k_J}(\eta'_{k_j} - \overline{\eta'_k})^2)$, $J_{(k)}$ is number of c_{k_j}s in C_k.

From (3.25) and by further averaging (3.26),

$$\alpha' \approx \frac{(\overline{\eta'_k})^2}{S^2_{\eta'_k} - \frac{\overline{\eta'_k}}{J_{(k)}}(\sum_{k_j=k_1}^{k_J}\frac{1}{|c_{k_j}|})} \tag{3.27}$$

Empirical model 2

This proposed empirical model, as illustrated in Figure 3.3, is a further simplified version of the above empirical model shown in Figure 3.2. In addition to the change of α' from a stochastic node, which is unknown, to a constant node, β' also changes from a stochastic node to a constant node.

In this model, the value of α' is determined in the same way as in (3.27), and from (3.25) and (3.27),

$$\beta'_k \approx \frac{\alpha'_k}{\overline{\eta'_k}} \approx \frac{\overline{\eta'_k}}{S^2_{\eta'_k} - \frac{\overline{\eta'_k}}{J_{(k)}}\left(\sum_{k_j=k_1}^{k_J} \frac{1}{|c_{k_j}|}\right)} \tag{3.28}$$

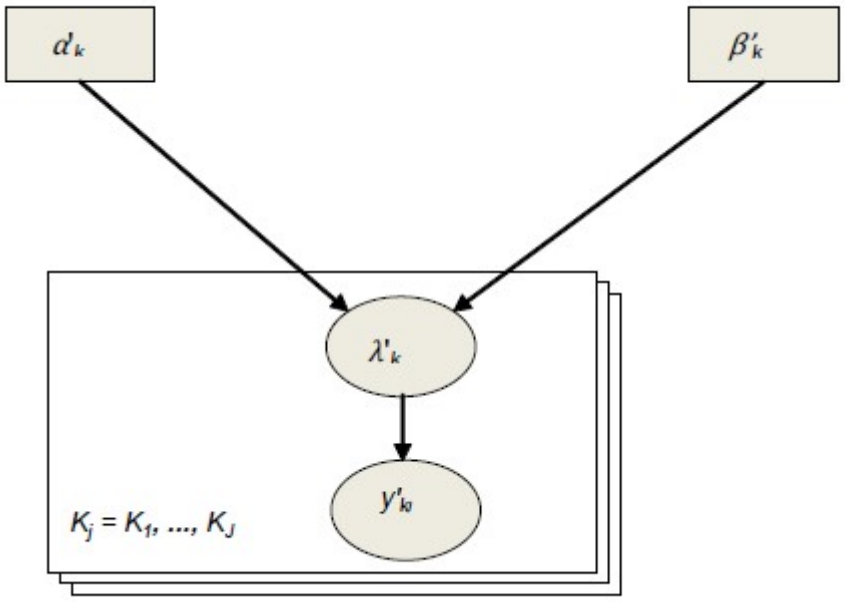


Figure 3.3 Graphic model of Empirical model 2

3.2.4.4 Gibbs Sampling

Gibbs sampling, detailed in Geman and Geman (1984), Gelfand and Smith (1990), and Bolstad (2010), is a Markov Chain Monte Carlo (MCMC) randomized algorithm used for obtaining a sequence of samples from multivariate probability distributions where direct sampling is difficult. Gibbs sampling has two advantages. First, it is simpler to sample from the distribution of one variable conditional on all other variables (full conditional distribution) than to sample from the marginal distribution of that variable by integrating over the joint distribution of all variables. Second, Gibbs sampler only requires the conditionals up to proportionality, the procedure of normalization, often the most difficult step, is not needed.

Let $\{x_1, \ldots, x_n\}$ be a sample from a joint distribution $f(x_1, \ldots, x_n)$, the full conditional distribution of x_j takes the form

$$f(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, 1, \ldots, x_n) = \frac{f(x_1,\ldots,x_n)}{f(x_1,\ldots,x_{j-1},x_{j+1,1},\ldots,x_n)}$$

$$\propto f(x_1, \ldots, x_n). \qquad (3.29)$$

Suppose I want to get I samples from the joint distribution, the Gibbs sampler proceeds as follows:

1. Begin with initial values $\{x_1^{(0)}, \ldots, x_n^{(0)}\}$ for each variable.
2. For each sample $i = \{1, \ldots, I\}$, sample each $x_j^{(i)}$ from the full conditional distribution $f(x_j^{(i)}|, x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, 1, \ldots, x_n^{(i-1)})$. The whole process uses the most current value of one variable once it's sampled.

As $i \to \infty$, $f(x_1^{(i)}, \ldots, x_n^{(i)}) \xrightarrow{d} f(x_1, \ldots, x_n)$.

3.3 BHM-Based Highway Loss Event Intensity Prediction

3.3.1 Focused Area Loss Intensity Prediction

Let A be a focused area of study interest between D_s and its elements (cells). A is composed of cells in D_s and the set of cells of A is a subset of cells of D_s. If D_s is a state and each cell is a census tract, a good example of such an area of study interest could be a city or a county. Let c_{a_j} be a cell of A indexed by a_j, $a_j = 1, \ldots, a_J$. $A = \{c_{a_1}, \ldots, c_{a_J}\}$, $c_{a_j} \in \{c_1, c_2, ..., c_n\}$, $a_J < n$, and $\{c_{a_1}, \ldots, c_{a_J}\} \subset \{c_1, c_2, ..., c_n\}$.

Denote $\widehat{\alpha_k}'$, $\widehat{\beta_k}'$ as the BHM estimators of α_k' and β_k' defined in Section 3.2.4.3 and determined in Section 3.2.4.4 by use of Gibbs sampler. The prior of loss intensity on C_k at time $t-1$ follows a Gamma distribution,

$$\pi(\lambda_k' | \widehat{\alpha_k}', \widehat{\beta_k}') = \frac{\widehat{\beta_k}'^{\widehat{\alpha_k}'}}{\Gamma(\widehat{\alpha_k}')} \lambda_k'^{\widehat{\alpha_k}'-1} e^{-\widehat{\beta_k}'\lambda_k'}, \tag{3.30}$$

and its posterior distribution takes the form

$$\pi(\lambda_k' | y_{k_1}', \ldots, y_{k_J}', \widehat{\alpha_k}', \widehat{\beta_k}') = \frac{\widehat{\beta_k}^{\widehat{\alpha_k}}}{\Gamma(\widehat{\alpha})} \lambda_k'^{\widehat{\alpha_k}-1} e^{-\widehat{\beta_k}\lambda_k'}, \tag{3.31}$$

where $\widehat{\alpha_k} = \sum_{k_j=k_1}^{k_J} y_{k_j}' + \widehat{\alpha_k}'$, and $\widehat{\beta_k} = \sum_{k_j=k_1}^{k_J} |c_{k_j}| + \widehat{\beta_k}'$.

$\pi(\lambda_k' | y_{k_1}', \ldots, y_{k_J}', \widehat{\alpha_k}', \widehat{\beta_k}')$, the posterior loss intensity distribution on C_k at time $t-1$, also serves as the prior distribution of the loss intensity λ_k at time t, and its posterior, $\pi(\lambda_k | \widehat{\alpha_k}, \widehat{\beta_k}, y_{k_1}, \ldots, y_{k_J})$, serves as the predicted intensity on C_k at the nearest future $t+1$.

$$\pi(\lambda_k|\hat{\alpha},\hat{\beta},y_{k_1},\ldots,y_{k_J}) = \frac{\widetilde{\beta}_k^{\widetilde{\alpha}}}{\Gamma(\widetilde{\alpha}_k)}\lambda_k^{\widetilde{\alpha}_k-1}e^{-\widetilde{\beta}_k\lambda_k} \tag{3.32}$$

where $\widetilde{\alpha}_k = \sum_{k_j=k_1}^{k_J} y_{k_j} + \widehat{\alpha}_k$, and $\widetilde{\beta}_k = \sum_{k_j=k_1}^{k_J} |c_{k_j}| + \widehat{\beta}_k$.

For the convenience of predicting the loss intensity on A at time $t+1$, the predicted intensity on C_k at time $t+1$ as is denoted as $\widetilde{\lambda}_k$, and the predicted intensity on A at time $t+1$ is denoted as $\widetilde{\lambda}(A)$.

Let $\omega_{a_j} = \frac{|c_{a_j}|}{\sum_{a_j=a_1}^{a_J}|c_{a_j}|}$, which is the proportion of the area of the cell c_{a_j} on A, be

$$\widetilde{\lambda}(A) = \sum_{a_j=a_1}^{a_J} \omega_{a_j} \cdot (\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_{a_j})) \tag{3.33}$$

where $I_{C_k}(c_{a_j})\begin{cases} 1 & \text{if } c_{a_j} \in \{c_{k_1},\ldots,c_{k_j},\ldots,c_{k_J}\} \\ 0 & \text{otherwise} \end{cases}$.

Thus, the predicted loss intensity at time $t+1$ takes the form of a mixture Gamma distribution, and it has an expectation of

$$E(\widetilde{\lambda}(A)) = \sum_{a_j=a_1}^{a_J} \omega_{a_j} \cdot (\sum_{k=1}^{K} \frac{\widetilde{\alpha}_k}{\widetilde{\beta}_k} \cdot I_{C_k}(c_{a_j})), \tag{3.34}$$

and a variance of

$$V(\widetilde{\lambda}(A)) = \sum_{a_j=a_1}^{a_J} \omega_{a_j}^2 \cdot (\sum_{k=1}^{K} \frac{\widetilde{\alpha}_k}{(\widetilde{\beta}_k)^2} \cdot I_{C_k}(c_{a_j})). \tag{3.35}$$

3.3.2 Loss Intensity Prediction of the Whole Study Area

D_s is the study area with jth cell $c_j, j = 1,\ldots,n$, and $D_s = \{c_1, c_2, ..., c_n\}$. Let $\omega_j = \frac{|c_j|}{\sum_{j=1}^{n}|c_j|}$ be the proportion of the area of the cell c_j on D_s, the predicted loss intensity of D_s at time $t+1$ is

$$\widetilde{\lambda}(D_s) = \sum_{j=1}^{n} \omega_j \cdot (\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_j)) \tag{3.36}$$

where $I_{C_k}(c_j)\begin{cases} 1 & \text{if } c_j \in \{c_{k_1},\ldots,c_{k_j},\ldots,c_{k_J}\} \\ 0 & \text{otherwise} \end{cases}$.

Similarly, the predicted loss intensity of D_s at time $t+1$ also takes the form of a mixture Gamma distribution, and has an expectation of

$$E(\widetilde{\lambda}(D_s)) = \sum_{j=1}^{n} \omega_j \cdot (\sum_{k=1}^{K} \frac{\widetilde{\alpha_k}}{\widetilde{\beta_k}} \cdot I_{C_k}(c_j)), \tag{3.37}$$

and a variance of

$$V(\widetilde{\lambda}(D_s)) = \sum_{j=1}^{n} \omega_j^2 \cdot (\sum_{k=1}^{K} \frac{\widetilde{\alpha_k}}{(\widetilde{\beta_k})^2} \cdot I_{C_k}(c_j)). \tag{3.38}$$

3.3.3 Predicted Loss Centroid of the Whole Study Area

Let $\boldsymbol{s}_{c_j} = (s_{1_{c_j}}, s_{2_{c_j}})$ be the geographic centroid of c_j, and let $\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_j) \cdot |c_j|$ be the predicted number of events at time $t+1$ with an abstract centroid of $(s_{1_{c_j}}, s_{2_{c_j}})$. I abstract the loss centroid of D_s at time $t+1$ to

$$s_{1_{D_s}} = \frac{\sum_{j=1}^{n}((\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_j)) \cdot |c_j| \cdot s_{1_{c_j}})}{\sum_{j=1}^{n}((\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_j)) \cdot |c_j|)}, \tag{3.39}$$

$$s_{2_{D_s}} = \frac{\sum_{j=1}^{n}((\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_j)) \cdot |c_j| \cdot s_{2_{c_j}})}{\sum_{j=1}^{n}((\sum_{k=1}^{K} \widetilde{\lambda}_k \cdot I_{C_k}(c_j)) \cdot |c_j|)}. \tag{3.40}$$

3.3.4 Summary of Chapter 3

In this chapter, I have developed methods using a *k*-means based algorithm and specially tailored distance functions to partition the key feature space into

homogeneous clusters, and map this partition to the geographical space partition. Then, I have developed theory and methods of a Bayesian hierarchical model (BHM) that uses the current time point loss information and most recent past loss information to predict the future losses for each cluster. The BHM has a good updating mechanism and adds adaptation to the Bayesian approach.

4. NHTSA FARS Data and Proposed Bayesian Hierarchical Spatio-Temporal Model

In Chapter 2 and Chapter 3, I have built the methodology framework composed of the following components: the selection of key features, "hot spots" pattern detection, key feature space and geographical space partition, and the Bayesian Hierarchical Model (BHM) with adaptation in the prediction of future losses. In this chapter, I apply these methods to 2009, 2010 and 2011 Fatality Analysis Reporting System (FARS) data published by National Highway Traffic Safety Administration (NHTSA) of U.S. Department of Transportation. The 2009-11 FARS data are the most current data available for this dissertation. It should be noted the software used in this chapter is SAS 9.3 unless otherwise stated.

4.1 FARS Data and the Poisson Point Process

4.1.1 FARS Data

FARS is a census of all crashes of motor vehicles traveling on public roadways in which a person died within 30 days of the crash (NHTSA, 2012). The deceased person can be either an occupant of a vehicle or a non-motorist. FARS was created by NHTSA and has been operational since 1975. According to NHTSA (NHTSA, 2010), FARS is the only source of U.S. real-world fatal crash data to serve the public use in "conducting basic research, identifying problem areas,

developing effective countermeasures, identifying program and rulemaking needs, developing and evaluating programs, rules, and standards..." by legislature institutes, governments, academic researchers, medical community, automotive industry, insurance industry and other traffic safety stakeholders.

FARS data originate from police-reported fatal motor vehicle traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. Data are input by FARS analysts in each state and are then transmitted to DOT for quality assurance and analysis. Data sources include: police accident reports, state vehicle registration files, state driver license files, state highway department data, vital statistics data, death certificates, etc. (NHTSA, 2010a). The collection, standardization, quality control and analysis process lead to the lag in FARS data. Typically after September, NHTSA publishes the FARS data and initial analysis for the previous year.

The content of FARS data collected includes, but not limited to: the time and location of the crash, number of people and vehicles involved, vehicle type(s), impact points, driver's license status of all drivers, demographics of all persons involved, their role in crash (driver, passenger, etc), injury severity, and seatbelt restraint use. Driver and nonoccupant blood alcohol content measures are also collected (NHTSA, 2010b).

The time and location of the crash is of greatest interest in this dissertation. The date of the crash has been included in the data since 1975 and from 1999 on the exact location of the crash was added to FARS data (NHTSA, 2013). The exact geographic location of a crash is expressed by its "Global Position" in the latitude and the longitude, and is often collected either by GPS systems on the site of the

crash or by Geographic Information Systems (GIS) after the crash. The format for the latitude is: dd mm ss.ss (Degrees/Minutes/Seconds) and the format for the longitude is: ddd mm ss.ss (Degrees/Minutes/Seconds).

Although FARS data quality has been improving over time, users still may face difficulties, e.g. missing values of variables used. In Maryland, 5 out of 515 fatal events in 2009, 3 out of 463 fatal crashes in 2010, and 5 out of 455 fatal crashes in 2011, had missing GPS coordinates and thus lost exact locations.

4.1.2 Census Data and Geocoding

Table 4.1 is a simplified sample of 2010 census tract attributes. The 2010 census defined 1406 census tracts and 12 of them are pure water area and thus are excluded from consideration in this dissertation. It should be noted that the land area was in square meters instead of square miles.

Table 4.1 Partial attributes of 2010 Maryland census tracts

STATEFP10	COUNTYFP10	TRACTCE10	GEOID10	NAME10	ALAND10	AWATER10	INTPTLAT10	INTPTLON10
24	029	950100	24029950100	9501	170,616,683	5,541,577	+39.3006660	-075.8425009
24	029	950200	24029950200	9502	279,186,393	29,289,092	+39.3109363	-076.0379468
24	017	850600	24017850600	8506	60,285,332	17,307	+38.5266341	-077.0933062
24	017	850600	24017850600	8506	60,285,332	17,307	+38.5266341	-077.0933062
24	017	851002	24017851002	8510.02	42,640,094	225,588	+38.5370101	-076.9541185
24	017	850300	24017850300	8503	34,111,856	3,466,222	+38.5530041	-077.1569782
24	011	955600	24011955600	9556	88,677,690	427,648	+38.7068793	-075.7661753
24	011	955600	24011955600	9556	88,677,690	427,648	+38.7068793	-075.7661753
24	011	955600	24011955600	9556	88,677,690	427,648	+38.7068793	-075.7661753

Field Name	**Definition**
STATEFP10	State code (Maryland = 24)
COUNTYFP10	County code, three characters
TRACT10	2010 Tract code, with leading zeroes and two implied decimal places (e.g. Tract "000302" = Tract 3.02)
GEOID10	Unique geographic ID (concatenated State + County + Tract codes)
NAME10	2010 Tract code, formatted for labeling
BLKGRP10	2010 block group
BLOCK10	2010 block
LOGRECNO	Logical record number
ALAND10	2010 Census land area in square meters
AWATER10	2010 Census water area in square meters
INTPTLAT10	Latitude in degrees of a point within the tract
INTPTLON10	Longitude in degrees of a point within the tract

The path to link the census tract dataset to FARS dataset is via geocoding. Geocoding is the process of assigning a location, in my case, the latitude and the longitude, to an address by comparing the descriptive location elements in the address to those present in the reference material. The address here has a variety of forms. It could be a narrow term such as a normal postal address, or a general term which could be a postal zone or a census tract. In this dissertation geocoding specifically means assigning the exact location described by the latitude and the longitude of a fatal crash to the 2010 census tracts.

In this dissertation, Esri ArcMap 10.0 is used for geocoding and other GIS practice such as mapping. The following graph is the geocoding results of MD 2010 fatal crash locations.

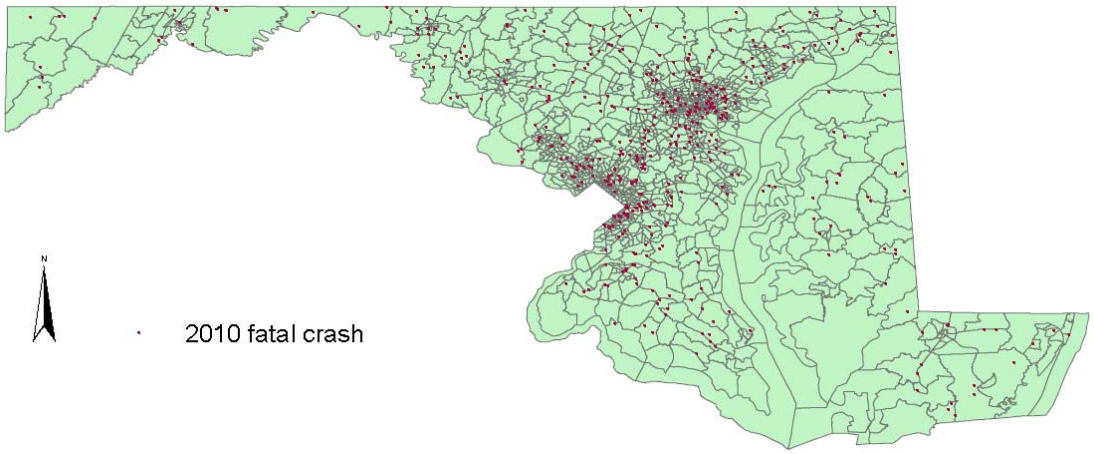


Figure 4.1 Locations of MD 2010 fatal crashes, map is created using ArcMap 10.0

Figure 4.1 shows that MD 2010 fatal crashes concentrate in outskirt census tracts of Washington D.C. and Baltimore. The source of the location coordinates was from the FARS data and the census tract definitions were from 2010 census tiger/line shapefiles published by the U.S. Census Bureau.

Point to polygon spatial join operations were conducted to join attributes of the fatal crashes to attributes of the census tracts. According to ESRI definitions, spatial join operation is used to combine two or more datasets with respect to a

spatial predicate. The predicate can be a combination of directional, distance, and topological spatial relations. The topological predicate here is if a point (the crash location) falls inside of a polygon (the census tract), attributes of that crash is appended to that census tract.

4.1.3 Fatal Crash Intensity and Poisson Point Process

Following the first-order intensity function of the Poisson point process $Z(\cdot)$ defined in Chapter 2 here the unit for the time t, is year, thus $\lambda_{c_j,t}$, the fatal crash intensity at t for the jth census tract c_j ($j = 1 \ldots 1{,}394$), which is the most basic spatial element in this dissertation, is defined as the number of fatal crashes occurring in c_j at t divided by $|c_j|$, the area (measured in square miles) of c_j assuming completely spatial randomness (CSR) of c_j.

$$\lambda_{c_j,t} = \frac{Z(c_j;t)}{|c_j|}. \tag{4.1}$$

Following the definition in (2.41) the observed intensity $\lambda_j^{'}$ is calculated as

$$\lambda_{j;t}^{'} = \frac{y_{j;t}}{|c_j|}$$

where $y_{j;t}$ is the observed number of fatal crashes at t for the census tract c_j.

It should be noted that the number of fatal crashes is different from the number of deaths since a fatal crash could result in more than 1 death. For example in 2010, Maryland had 493 fatalities in comparison to 463 fatal crashes.

Figure 4.2 illustrates the distribution of the counts of fatal crashes by MD 2010 census tracts. The distribution is highly skewed, 1,039 of the 1,394 Maryland census tracts did not have fatal crashes in 2010.

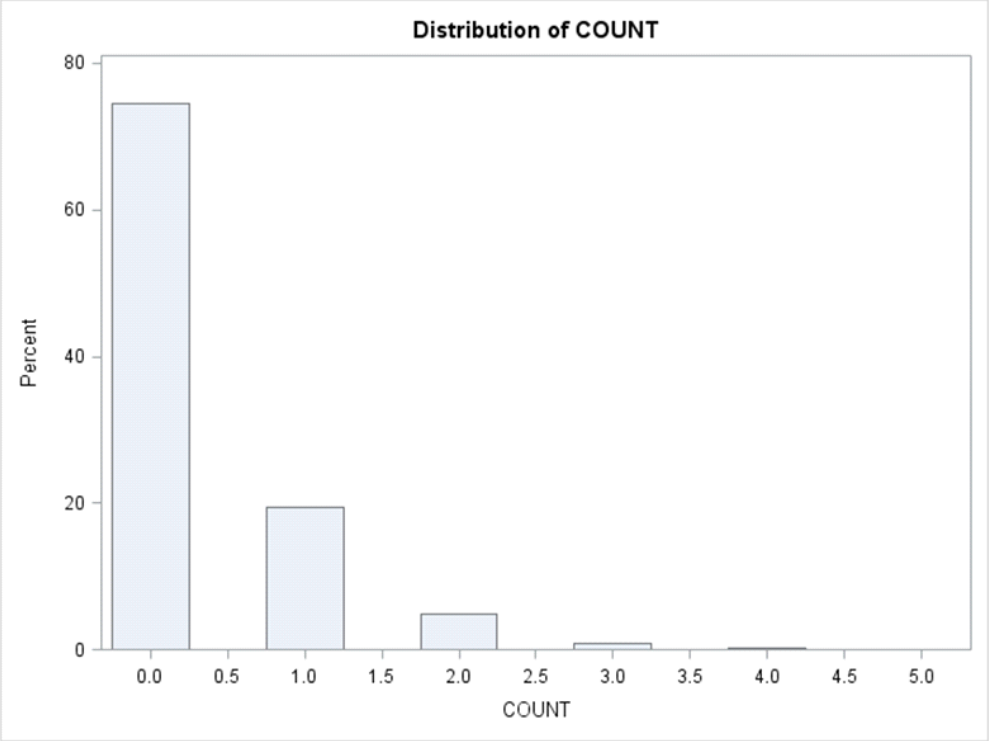


Figure 4.2 The distribution of counts of fatal crashes of MD 2010 census tracts

A Poisson regression was fitted to 2010 Maryland census tract level number of fatal crashes. The number of fatal crashes served as the event and the land area in square miles served as the trials. Maximum likelihood estimator was computed using logarithm as the link function. The regression estimated that the intensity of 2010 MD fatal crashes had a 95% confidence interval of $(0.030, 0.037)$. Roughly

a fatal crash occurred every 30 square miles. Detailed regression results are presented in the following table.

Table 4.2 Maximum likelihood estimate of 2010 Maryland fatal crash intensity

Maximum likelihood parameter estimates of MD 2010 crash intensity							
Parameter	**DF**	**Estimate**	**Standard Error**	**Likelihood Ratio 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-3.4015	0.0558	-3.5112	-3.296	3714.14	<.0001
Exponent		0.0333		0.0298	0.0370		

4.1.4 2010 Maryland Fatal Crash Intensity by Census Tract

Figure 4.3 describes the distribution of the observed intensity of the 2010 Maryland fatal crashes at the census tract level. The underlying dataset had 1,394 observations corresponding to the 1,394 census tracts. The distribution of the observed intensity was highly skewed. One thousand and thirty nine of the 1,394 Maryland census tracts had 0 intensity while the Census Tract 1901 had an intensity as high as 13.5 fatal crashes per square mile in 2010.

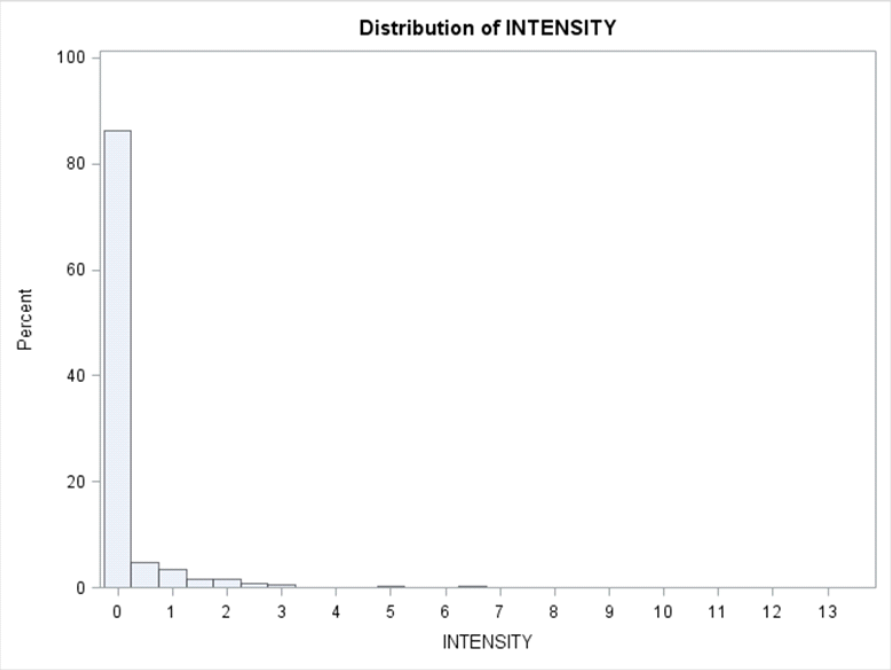


Figure 4.3 2010 Maryland fatal crash intensity distribution by census tract

To make it more intuitive, the observed intensity is further illustrated on a thematic map on which the darker color means higher value. Although census tracts close to a metropolitan center had higher intensity, there were exceptions as indicated by the highlighted area.

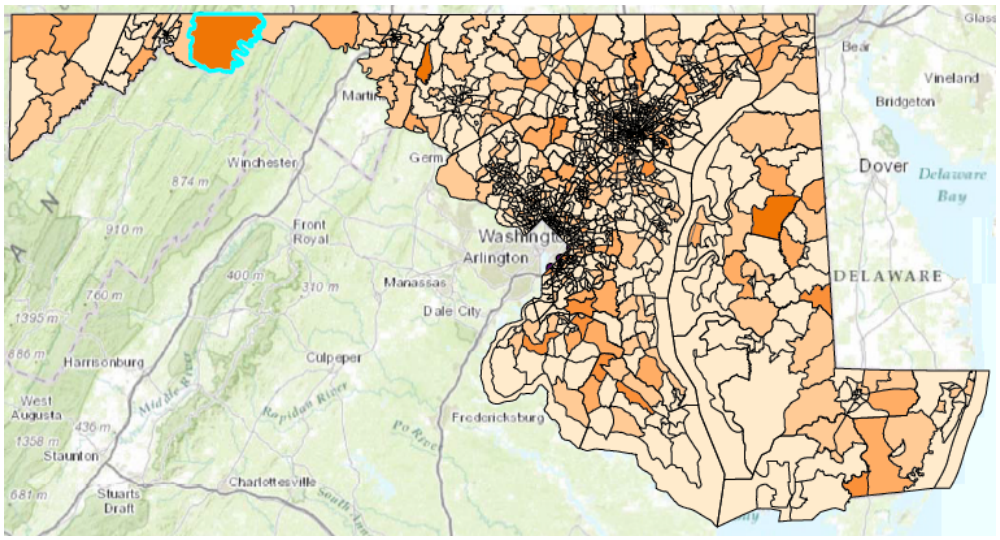


Figure 4.4 A thematic map on 2010 Maryland fatal crash intensity, created using ArcMap 10.0

4.2 Finite Mixture Model on FARS Data

4.2.1 Kernel Density Estimation of the Fatal Crash Intensity

In this section, I applied the methodology proposed in Chapter 2 to identify the sub populations (components) which dominated the 2010 Maryland fatal crashes. Before using kernel density estimation (KDE) to investigate the multimodality of the MD 2010 observed fatal crashes, I first explored the percentile of the observed intensity.

Table 4.3 Percentiles of the observed MD 2010 fatal crash intensity by census tract

MD 2010 fatality intensity quantile estimates											
Quantile	**100% Max**	**99%**	**95%**	**90%**	**75% Q3**	**50% Median**	**25% Q1**	**10%**	**5%**	**1%**	**0% Min**
Estimate	13.5203	4.7845	1.5135	0.6181	0.0141	0	0	0	0	0	0

More than half census tracts had 0 intensity which makes the KDE very hard to choose the right Sheather-Jones plug-in bandwidth. First I demonstrated the intensity distribution for all observations in the following figure.

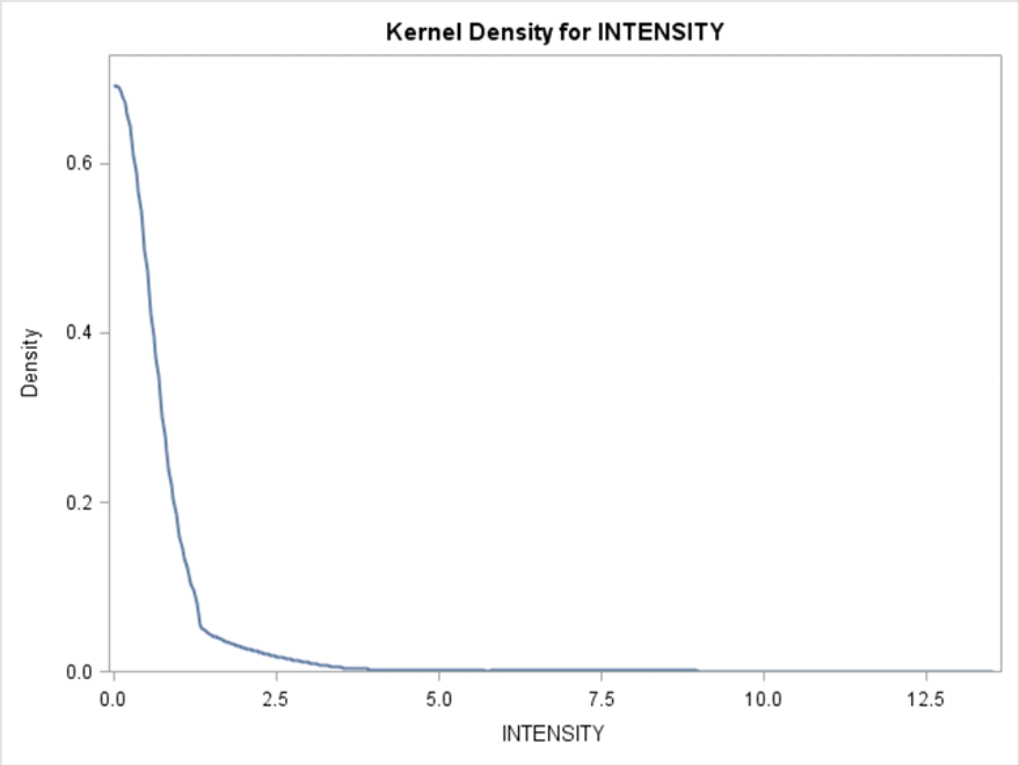


Figure 4.5 Distribution and kernel density for observed intensity of all census tracts

Here the bandwidth is 0.8 and a unimodal distribution is shown which is very close to 0. In the application of KDE the choice of bandwidth had a great influence on the number of modes detected. In this case, the mode close to 0 masked all other potential modes. Thus a second KDE was conducted over the census tracts having positive intensities.

In Figure 4.6 two obvious modes can be identified plus another potential candidate. The first one is very close to 0 from the right side, another one is larger

than 1.0 but smaller than 1.5. I lack the confidence to claim the third mode existing between 2.0 and 3.0, just like I did for the first two modes.

As for the multimodality, it's safe to say there might be three or four modes that can be obviously identified by Kernel density estimate visualization. This finding provides preliminary estimate to tune the algorithm to identify the exact number of components in the finite mixture model.

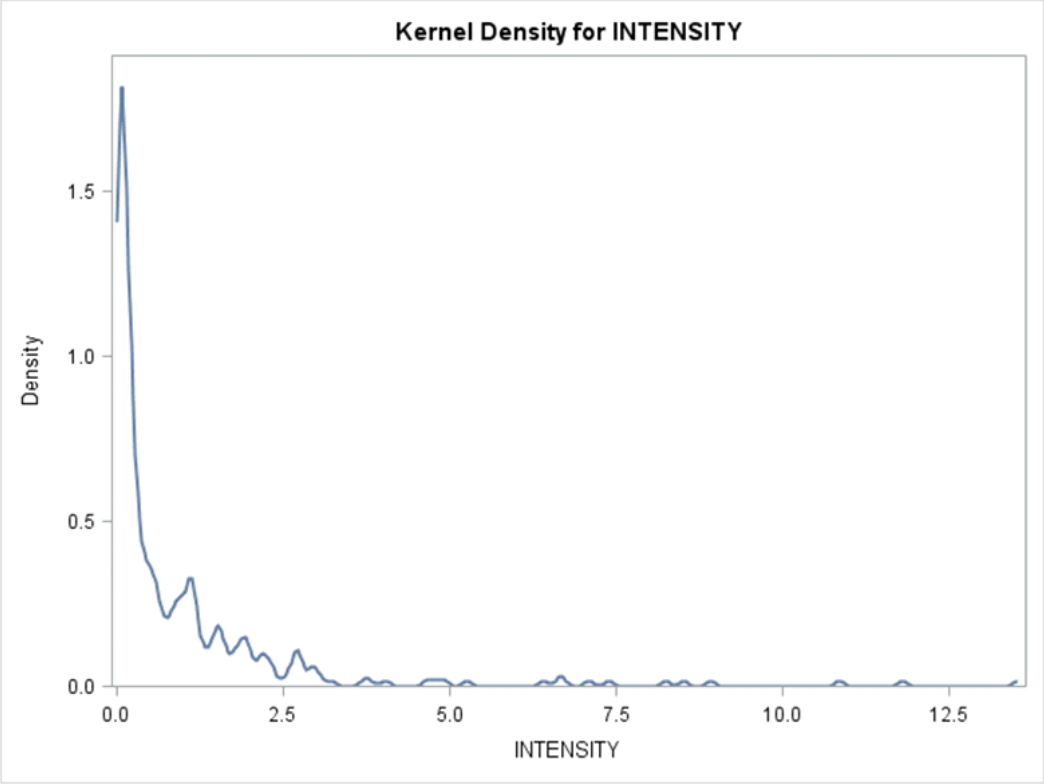


Figure 4.6 Distribution and kernel density for observed intensities > 0

Table 4.4 Model fitting statistics by number of components

Component Evaluation for Mixture Models										
Model ID	Number of Components		Number of Parameters		-2 Log L	AIC	AICC	BIC	Pearson	Max Gradient
	Total	Eff.	Total	Eff.						
1	1	1	1	1	2501.17	2503.17	2503.17	2508.41	9100.65	5.12E-06
2	2	2	3	3	2255.58	2261.58	2261.6	2277.3	2394.17	0.0036
3	3	3	5	5	2209.89	2219.89	2219.93	2246.09	1981.5	0.0002
4	4	4	7	7	2208.13	2222.13	2222.21	2258.81	1976.7	0.0008
5	5	5	9	9	2208.13	2226.13	2226.26	2273.29	1976.69	0.0008
6	6	6	11	11	2208.13	2230.13	2230.32	2287.77	1976.9	0.0067
7	7	7	13	13	2208.13	2234.13	2234.39	2302.25	1976.69	0.0022
8	8	8	15	15	2208.13	2238.13	2238.48	2316.73	1976.84	0.0088
9	9	9	17	17	2208.13	2242.13	2242.57	2331.21	1976.68	0.0029
10	10	10	19	19	2208.13	2246.13	2246.68	2345.69	1976.59	0.0057
11	11	11	21	21	2208.13	2250.13	2250.8	2360.17	1976.72	0.0027
12	12	12	23	23	2208.13	2254.13	2254.94	2374.65	1976.68	0.0025
13	13	13	25	25	2208.13	2258.13	2259.08	2389.13	1976.73	0.0032
14	14	14	27	27	2208.13	2262.13	2263.24	2403.61	1976.69	0.0032
15	15	15	29	29	2208.13	2266.13	2267.41	2418.09	1976.69	0.0180
16	16	16	31	31	2208.13	2270.13	2271.59	2432.57	1976.93	0.0220
17	17	17	33	33	2208.13	2274.13	2275.78	2447.05	1976.95	0.0067
18	18	18	35	35	2208.13	2278.13	2279.98	2461.53	1976.67	0.0024
19	19	19	37	37	2208.13	2282.13	2284.2	2476.01	1976.7	0.0020
20	20	20	39	39	2208.13	2286.13	2288.43	2490.49	1976.79	0.0046
21	21	21	41	41	2208.13	2290.13	2292.68	2504.97	1976.26	0.0150
22	22	22	43	43	2208.13	2294.13	2296.93	2519.45	1976.79	0.0042
23	23	23	45	45	2208.13	2298.13	2301.2	2533.93	1976.68	0.0017
24	24	24	47	47	2208.13	2302.13	2305.48	2548.41	1976.51	0.0110
25	25	25	49	49	2208.13	2306.13	2309.78	2562.89	1976.45	0.0130
26	26	26	51	51	2208.13	2310.13	2314.08	2577.37	1976.78	0.0049
27	27	27	53	53	2208.13	2314.13	2318.4	2591.85	1976.68	0.0059
28	28	28	55	55	2208.13	2318.13	2322.73	2606.33	1976.69	0.0027
29	29	29	57	57	2208.13	2322.13	2327.08	2620.81	1976.73	0.0023
30	30	30	59	59	2208.13	2326.13	2331.44	2635.29	1977	0.0120
31	31	31	61	61	2208.13	2330.13	2335.81	2649.77	1976.08	0.0370
32	32	32	63	63	2208.13	2334.13	2340.19	2664.25	1976.84	0.0089
33	33	33	65	65	2208.13	2338.13	2344.59	2678.72	1976.47	0.0085
34	34	34	67	67	2208.13	2342.13	2349	2693.2	1976.78	0.0061
35	35	35	69	69	2208.13	2346.13	2353.43	2707.68	1976.61	0.0037
36	36	36	71	71	2208.13	2350.13	2357.86	2722.16	1976.72	0.0010
37	37	37	73	73	2208.13	2354.13	2362.31	2736.64	1976.66	0.0014

4.2.2 Decision on the Number of Components, g

The 2010 Maryland FARS data had a sample size of 1,394 and the above multimodality examination suggested 3 or 4 modes. Thus according to (2.43) the lower bound of the number of components, g_L, was set to be 2 and the upper bound, g_U, was set to be 37 (square root of sample size n). Finite mixture models were run 37 times for each g. Each model fit the variable COUNT to a Poisson distribution in which the number of fatal crashes served as events and the land area, ALAND, measured in square miles, served as trials. Modeling fitting statistics for 37 runs were compared to decide the exact number of components of the underlying Poisson point process.

Table 4.4 lists the results for 37 runs. The AIC (Akaike's Information Criterion), AICC (AIC with a correction) and BIC (Bayesian information criterion) all suggests a clear cut of 3 components.

4.2.3 Component Estimates and the Mixing Probabilities

Table 4.5 are the results of the finite mixture model with 3 components. The model was implemented assuming Poisson distribution using logarithm as the link function, and the intensity estimates were listed in the last column.

Table 4.5 Estimates of identified components

Parameter Estimates for 'Poisson' Model						
Component	Parameter	Estimate	Standard Error	z Value	Pr > |z|	Inverse Linked Estimate
1	Intercept	0.1376	0.1837	0.75	0.454	1.1475
2	Intercept	-2.1868	0.2087	-10.48	<.0001	0.1123
3	Intercept	-4.1285	0.1785	-23.13	<.0001	0.0161

The three components (intensity of underlying Poisson point processes) were respectively estimated to be 1.1475, 0.1123 and 0.0161.

Table 4.6 Estimates of mixing probabilities

Parameter Estimates for Mixing Probabilities						
Component	Parameter	Linked Scale				Probability
		Estimate	Standard Error	z Value	Pr > |z|	
1	Probability	-1.9063	0.272	-7.01	<.0001	0.0855
2	Probability	-0.5287	0.3133	-1.69	0.0915	0.3391

The finite mixture model also estimated the mixing probability for each component. The mixing probability was 0.0855 for Component 1, 0.3391 for Component 2 and 0.5754 for Component 3. Based upon Table 4.5 and Table 4.6, the underlying Poisson point process dominating the 2010 Maryland census tracts took the form,

$$f(y_j;\boldsymbol{\Psi}) \;=\; \sum_{i=1}^{g_0}\pi_i f(y_j;\,\lambda_i) = 0.0855 * Poi(1.1475) \;+$$

$$0.3391 * Poi(0.1123) \;+ 0.5784 * Poi(0.0161) \qquad (4.2)$$

In addition the model also calculated the posterior probability, $\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}) = \pi_i f_i(\boldsymbol{y}_j;\,\boldsymbol{\theta}_i\,)\;\; /\sum_{l=1}^{g}\pi_l f_l(\boldsymbol{y}_j;\boldsymbol{\theta}_l)$ which is the probability the jth observation originated from the ith component which was defined in (2.14) of Chapter 2.

Table 4.7 lists observed intensity and the posterior probabilities of components from which this observation likely arose. Census Tract 1505 had an observed intensity of 2.7275 and the likelihood it originated from the first component was 0.5838, thus it was most likely arose from the first component.

According to posterior probability 101 census tracts were most likely from the first component which was estimated having the highest fatal crash intensity, 180 census tracts were most likely from the second component, and the rest 1,113 census tracts were most likely from the third component which had the lowest estimate of the fatal crash intensity.

Table 4.7 Example of posterior probability an observation arose from a component

NAMELSAD10	COUNT	ALAND	INTENSITY	POST_1	POST_2	POST_3	COMPONENT ID
Census Tract 8505	0	37.912	0	0	0.0152	0.9849	3
Census Tract 8506	3	23.276	0.1289	0	0.9551	0.0449	2
Census Tract 8510.02	0	16.463	0	0	0.1079	0.8921	3
Census Tract 8511	0	27.666	0	0	0.0396	0.9604	3
Census Tract 8503	1	13.171	0.0759	0	0.5365	0.4635	2
Census Tract 8504	0	68.331	0	0	0.0008	0.9992	3
Census Tract 805	0	0.135	0	0.07465	0.3403	0.5850	3
Census Tract 1505	1	0.367	2.7275	0.5848	0.3316	0.0836	1

4.3 Key Feature Selection and Feature Space Formation

4.3.1 Data Source of Features

All independent variables used in this chapter were from the U.S. Census Bureau. The Census 2010 is the most recent national census of the United States. Different from the 2000 decennial census, for which some homes received a "long form" (U.S. Census Bureau, 2000) questionnaire and most homes received a "short form" questionnaire, the 2010 Census only sent "short form" questionnaire that should take about ten minutes to complete. The questions included: name, age, sex, date of birth, Hispanic origin, race, ethnicity, relationship (to the first name listed on the form), and housing tenure (whether a family owns or rents their home).

The "long form" was replaced by the American Community Survey (ACS), which samples approximately 3 million housing unit addresses across the country on a

regular basis to obtain important data on demographic, economic, social and housing information. ACS datasets are combined to produce 12 months, 36 months or 60 months of data. The following table describes the availability of ACS period estimates for geographic areas by population size (US Census Bureau, 2009).

Availability of ACS data.

Data pooled to produce	Data published for areas with
1-year data sets	populations of 65,000 or more
3-year data sets	populations of 20,000 or more
5-year data sets	populations of almost any size*

The 2010 decennial census provides demographic and household data at the census tract and the lower geography level while the census tract level economic, social and housing data can only be obtained from the 5-year ACS data because most of the census tracts have populations less than 20,000. In this dissertation all demographic and household data were from the 2010 decennial census and all other economic, social and housing data were from the 2006-10 ACS data.

4.3.2 Initial Screen by Visualization

First, only variables at least to some extent relevant to this research were kept and all other variables obviously irrelevant were filtered out from the demographic,

economic, social and housing data sets. As a result, 47 out of 372 demographic variables, 9 out of 597 social variables, 62 variables out of 549 economic variables, and 47 out of 565 housing variables, besides 3 identity and label variables, were selected as candidates for the initial screening.

Let F be the initial feature set before the visualization screen procedure which had 174 variables including the above selected variables and 9 variables in Table 4.1 which was also selected from the geography definition table from 2010 census data. The initial screen is defined by the triplet (F, v_1, F_1), where v_1 is the visualization screen procedure, and F_1 is the subset of F which is composed of the elements chosen from F in the procedure. Data in F come with the form $(x_{j1}, x_{j2}, \ldots, x_{jp})$ and data in F_1 have the form of $(x_{j1}, x_{j2}, \ldots, x_{jl})$ where $p > 0,\ l > 0$, and $p > l$. Parallel coordinate plot served as the main tool of this visualization procedure and only variables showed association with loss intensities were chosen. Here I omitted the details of the visualization procedure and gave a summary parallel coordinate plot shown in Figure 4.7.

The 31 variables in F_1 were listed in Table 4.8 in which every variable has a detailed description.

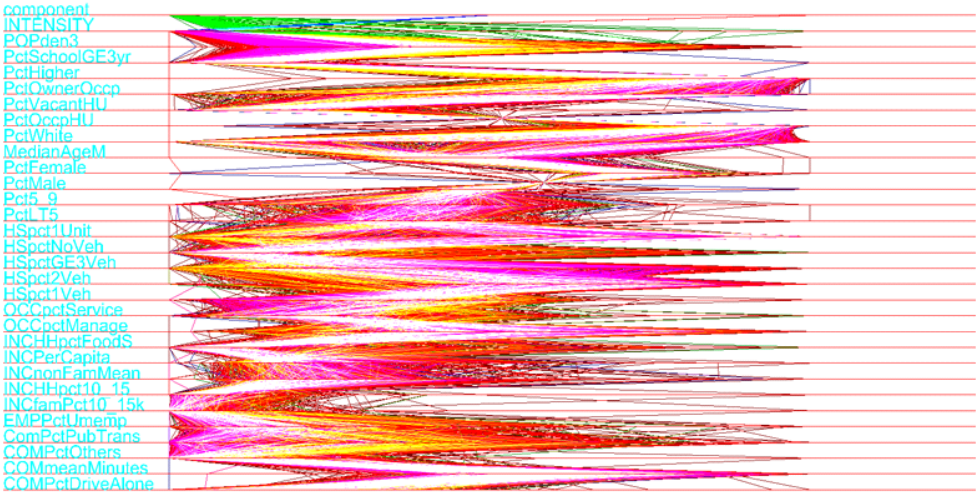


Figure 4.7 Parallel coordinate plot of the variables selected by the initial screen

Table 4.8 Variables selected by the initial screen

Variable	Variable description
Economic	
COMPCTDRIVEALONE	Percent; COMMUTING TO WORK - Car, truck, or van - drive alone
COMMEANMINUTES	Estimate; COMMUTING TO WORK - Mean travel time to work (minutes)
COMPCTOTHERS	Percent; COMMUTING TO WORK - Other means
COMPCTPUBTRANS	Percent; COMMUTING TO WORK - Public transportation (excluding taxicab)
EMPPCTUMEMP	Percent; EMPLOYMENT STATUS - Percent Unemployed
INCFAMPCT10_15K	Percent; FAMILY INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - $10,000 to $14,999
INCHHPCT10_15	Percent; HOUSEHOLD INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - $10,000 to $14,999
INCNONFAMMEAN	Estimate; INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - Mean nonfamily income (dollars)
INCPERCAPITA	Estimate; INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - Per capita income (dollars)
INCHHPCTFOODS	Percent; INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - With Food Stamp/SNAP benefits in the past 12 months
OCCPCTMANAGE	Percent; OCCUPATION - Management, business, science, and arts occupations
OCCPCTSERVICE	Percent; OCCUPATION - Service occupations
Housing	
HSPCT1VEH	Percent; VEHICLES AVAILABLE - 1 vehicle available
HSPCT2VEH	Percent; VEHICLES AVAILABLE - 2 vehicles available
HSPCTGE3VEH	Percent; VEHICLES AVAILABLE - 3 or more vehicles available
HSPCTNOVEH	Percent; VEHICLES AVAILABLE - No vehicles available
HSPCT1UNIT	
Demographic and household	
PCTLT5	Percent; SEX AND AGE - Total population - Under 5 years
PCT5_9	Percent; SEX AND AGE - Total population - 5 to 9 years
PCTMALE	Percent; SEX AND AGE - Male population
PCTFEMALE	Percent; SEX AND AGE - Female population
MEDIANAGEM	Number; SEX AND AGE - Male population - Median age (years)
PCTWHITE	Percent; RACE - Total population - One Race - White
PCTOCCPHU	Percent; HOUSING OCCUPANCY - Total housing units - Occupied housing units
PCTVACANTHU	Percent; HOUSING OCCUPANCY - Total housing units - Vacant housing units
PCTOWNEROCCP	
Social	
PCTHIGHER	Percent; EDUCATIONAL ATTAINMENT - Percent high school graduate or higher
PCTSCHOOLGE3YR	Percent; SCHOOL ENROLLMENT - Population 3 years and over enrolled in school
Derived	
POPDEN3	TRANSFORMED POPULATION DENSITY (cubic root of population density (population per square mile))
Dependent variables	
COMPONENT	SUB POPULATION (COMPONENT) a census tract was assigned by posterior probability
INTENSITY	OBSERVED FATAL CRASH INTENSITY

Economic	
COMPCTDRIVEALONE	Percent; COMMUTING TO WORK - Car, truck, or van - drive alone
COMMEANMINUTES	Estimate; COMMUTING TO WORK - Mean travel time to work (minutes)
COMPCTOTHERS	Percent; COMMUTING TO WORK - Other means
COMPCTPUBTRANS	Percent; COMMUTING TO WORK - Public transportation (excluding taxicab)
EMPPCTUMEMP	Percent; EMPLOYMENT STATUS - Percent Unemployed
INCFAMPCT10_15K	Percent; FAMILY INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - $10,000 to $14,999
INCHHPCT10_15	Percent; HOUSEHOLD INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - $10,000 to $14,999
INCNONFAMMEAN	Estimate; INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - Mean nonfamily income (dollars)
INCPERCAPITA	Estimate; INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - Per capita income (dollars)
INCHHPCTFOODS	Percent; INCOME AND BENEFITS (IN 2010 INFLATION-ADJUSTED DOLLARS) - With Food Stamp/SNAP benefits in the past 12 months
OCCPCTMANAGE	Percent; OCCUPATION - Management, business, science, and arts occupations
OCCPCTSERVICE	Percent; OCCUPATION - Service occupations
Housing	
HSPCT1VEH	Percent; VEHICLES AVAILABLE - 1 vehicle available
HSPCT2VEH	Percent; VEHICLES AVAILABLE - 2 vehicles available
HSPCTGE3VEH	Percent; VEHICLES AVAILABLE - 3 or more vehicles available
HSPCTNOVEH	Percent; VEHICLES AVAILABLE - No vehicles available
HSPCT1UNIT	
Demographic and household	
PCTLT5	Percent; SEX AND AGE - Total population - Under 5 years
PCT5_9	Percent; SEX AND AGE - Total population - 5 to 9 years
PCTMALE	Percent; SEX AND AGE - Male population
PCTFEMALE	Percent; SEX AND AGE - Female population
MEDIANAGEM	Number; SEX AND AGE - Male population - Median age (years)
PCTWHITE	Percent; RACE - Total population - One Race - White
PCTOCCPHU	Percent; HOUSING OCCUPANCY - Total housing units - Occupied housing units
PCTVACANTHU	Percent; HOUSING OCCUPANCY - Total housing units - Vacant housing units
PCTOWNEROCCP	
Social	
PCTHIGHER	Percent; EDUCATIONAL ATTAINMENT - Percent high school graduate or higher
PCTSCHOOLGE3YR	Percent; SCHOOL ENROLLMENT - Population 3 years and over enrolled in school
Derived	
POPDEN3	TRANSFORMED POPULATION DENSITY (cubic root of population density (population per square mile))
Dependent variables	
COMPONENT	SUB POPULATION (COMPONENT) a census tract was assigned by posterior probability
INTENSITY	OBSERVED FATAL CRASH INTENSITY

4.3.3 Population Density and the Observed Fatal Crash Intensity

The thematic map of the 2010 Maryland fatal crash intensity in Figure 4.4, along with previous research, suggests population density, population per square mile, had sizeable influence on fatal crash intensity. This relationship was examined between two variables and their transformations. Two transformations take the following form,

$$\text{INT2} = log\,(1 + sqrt(\text{INTENSITY}))$$

$$\text{POPden3} = \sqrt[3]{\text{POPden}}$$

The relationships were visualized in Figure 4.8a and Figure 4.8b.

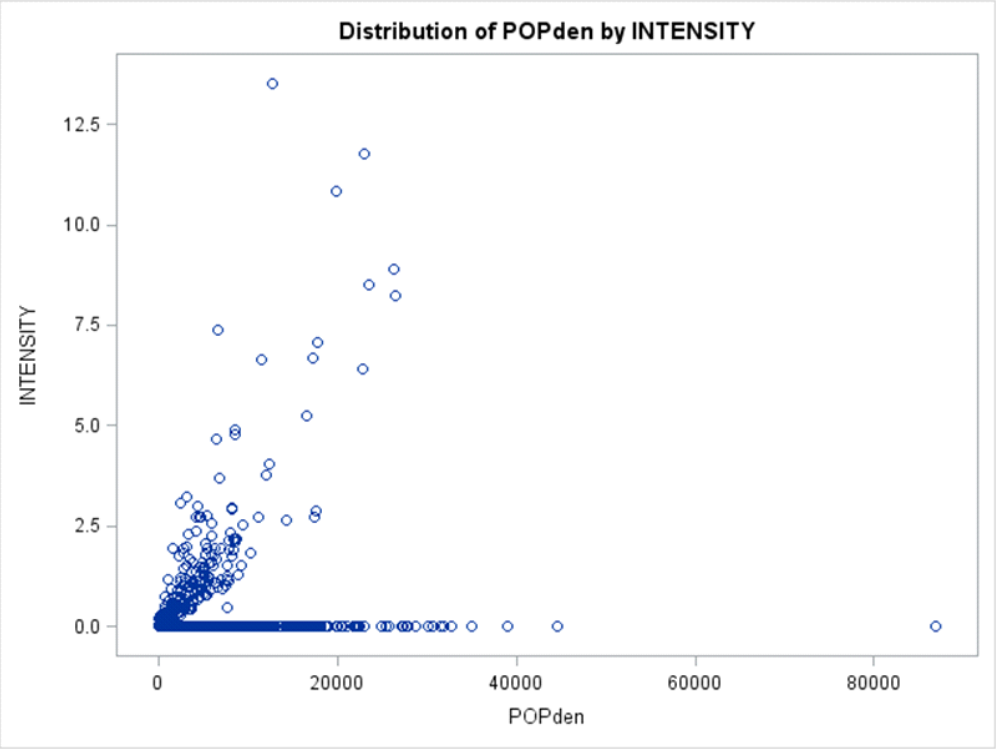


Figure 4.8a Scatter plot of fatal crash intensity by population density

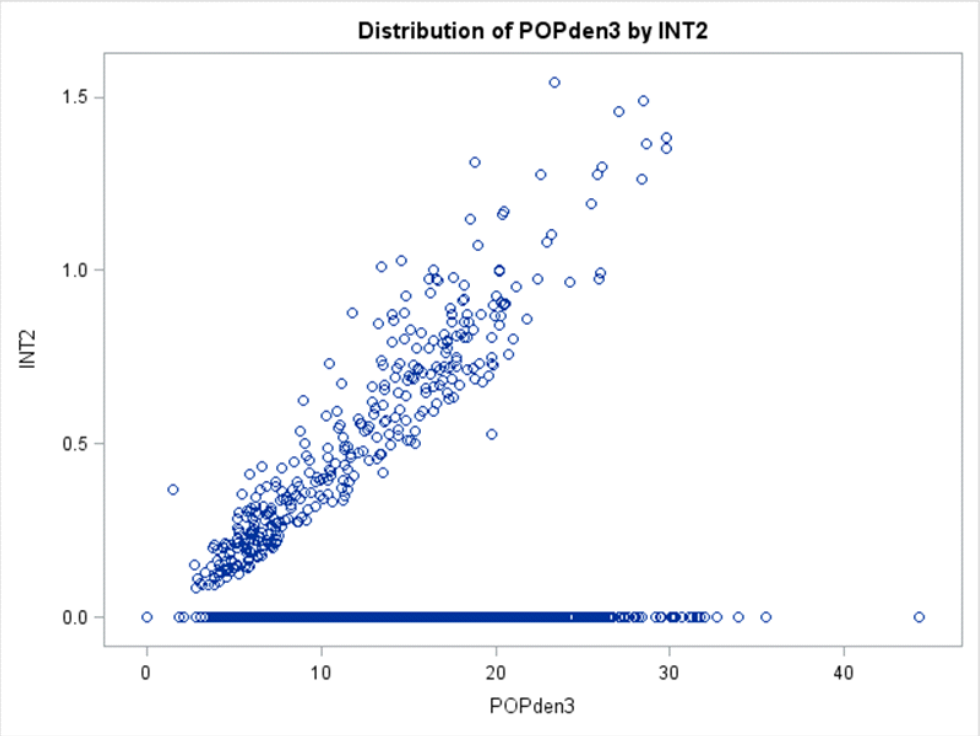


Figure 4.8b Scatter plot of fatal crash intensity by population density, both transformed.

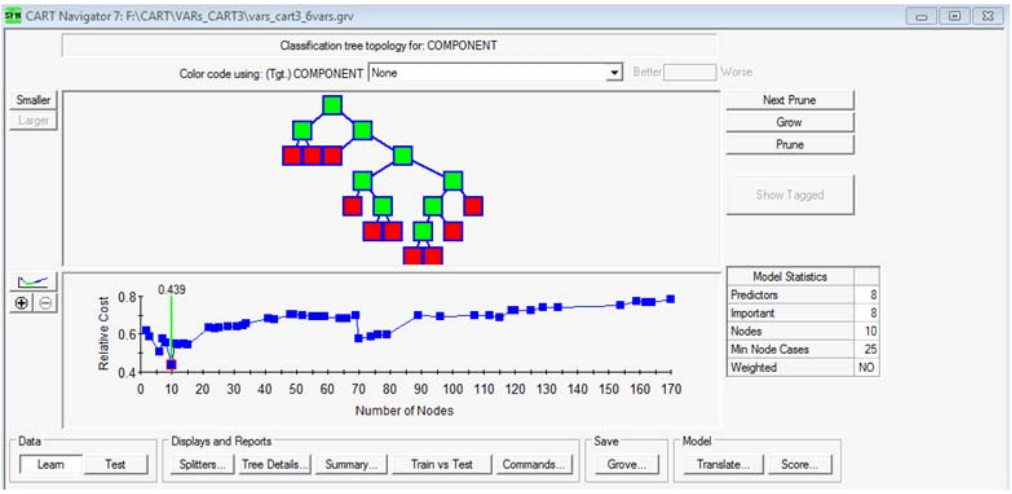
In Figure 4.8b a linear relationship between two transformed variables is clear for some census tracts while for others intensity remains constant at 0 however population density varies.

4.3.4 Classification and Regression Tree and Phase 2 Feature Selection

4.3.4.1 Classification Tree

The Salford SPM software was used to conduct classification and regression trees analysis for key feature selection and key feature space formation. In the classification tree analysis, the variable COMPONENT, a categorical variable which had been decided in the finite mixture model, was used as the target variable. COMPONENT had three values, 1, 2 and 3, respectively representing the first component (with the highest intensity estimate), the second component and the third component (with the lowest intensity estimate). All other variables except another target variable INTENSITY and its transformation were used as predictors. Gini measure served as the impurity function in splitting the nodes.

The classification tree analysis aimed at finding the feature pattern(s) of "hot spot" and thus focused on correctly predicting the first component, and 90% of the sample was used for training and 10% of the sample was used for testing in this case. This ratio was purposely set lower than usual case which is 75-80% vs. 20-25%.



Actual Class	Total Class	Percent Correct	1 N = 46	2 N = 74	3 N = 23
1	4.00	100.00%	4.00	0.00	0.00
2	25.00	92.00%	2.00	23.00	0.00
3	114.00	20.18%	40.00	51.00	23.00
Total:	143.00				
Average:		70.73%			
Overall % Correct:		34.97%			

Figure 4.9 Classification tree analysis model results, produced using Salford SPM 7.0

Figure 4.9 shows the classification tree analysis that produced the best tree with 10 nodes. In the training sample of 97 census tracts, it predicted 89 correctly with a success rate of 91.75%. It predicted correctly for the whole testing sample which contained 4 randomly selected observations. It also had high accuracy to predict the testing sample for the second component.

Table 4.9 Classification tree identified key features and importance scores, produced using Salford SPM 7.0

Variable	Score	
POPDEN3	100.00	
HSPCTGE3VEH	46.49	
COMPCTPUBTRANS	36.67	
HSPCT1VEH	35.13	
HSPCT2VEH	33.71	
PCTWHITE	31.20	
PCTOCCPHU	23.97	
PCTHIGHER	13.52	

The model also identified 8 variables from the 29 candidates and it also produced importance score for each variable for which the most important variable, POPen3, was set to have an importance score of 100. The list of variable and importance score was in Table 4.9. The detailed best classification tree was shown in Figure 4.10 to help some readers to understand the concept of a classification tree.

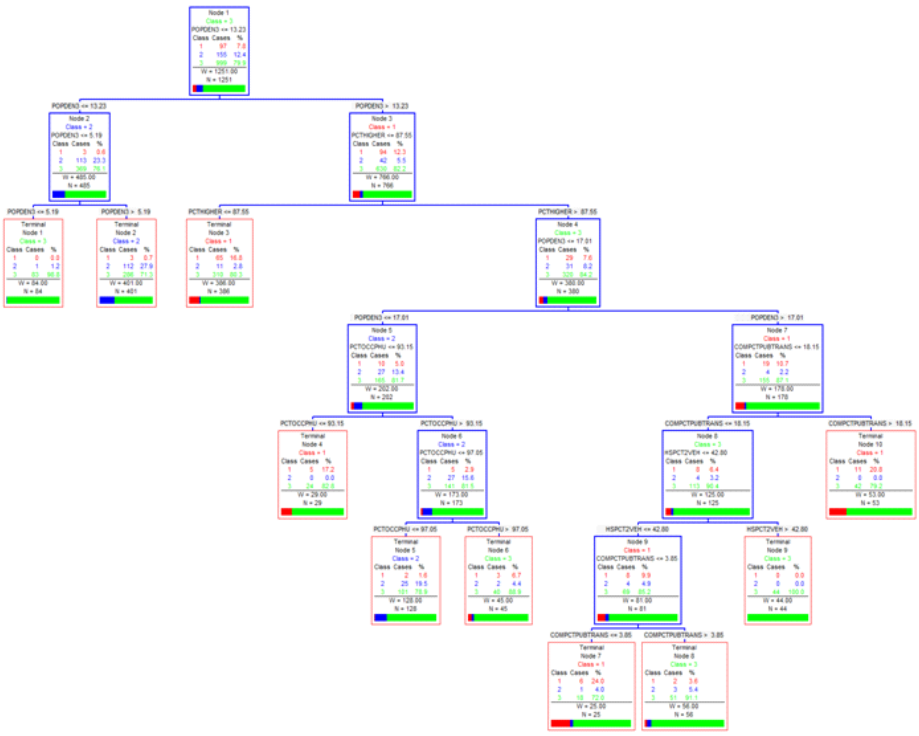


Figure 4.10 Detailed classification tree, produced using Salford SPM 7.0

4.3.4.2 Regression Tree

In the regression tree analysis, the variable INT2, the transformed fatal crash intensity, was used as the target variable. All other variables except the INTENSITY and the COMPONENT were used as predictors. Least squares served as the impurity function in splitting decisions. The regression tree model statistics and the model summary are shown in Figure 4.11. The resulted best tree

had 182 nodes and the root mean square of error (RMSE) and the mean square error of the prediction can also be found in that figure.

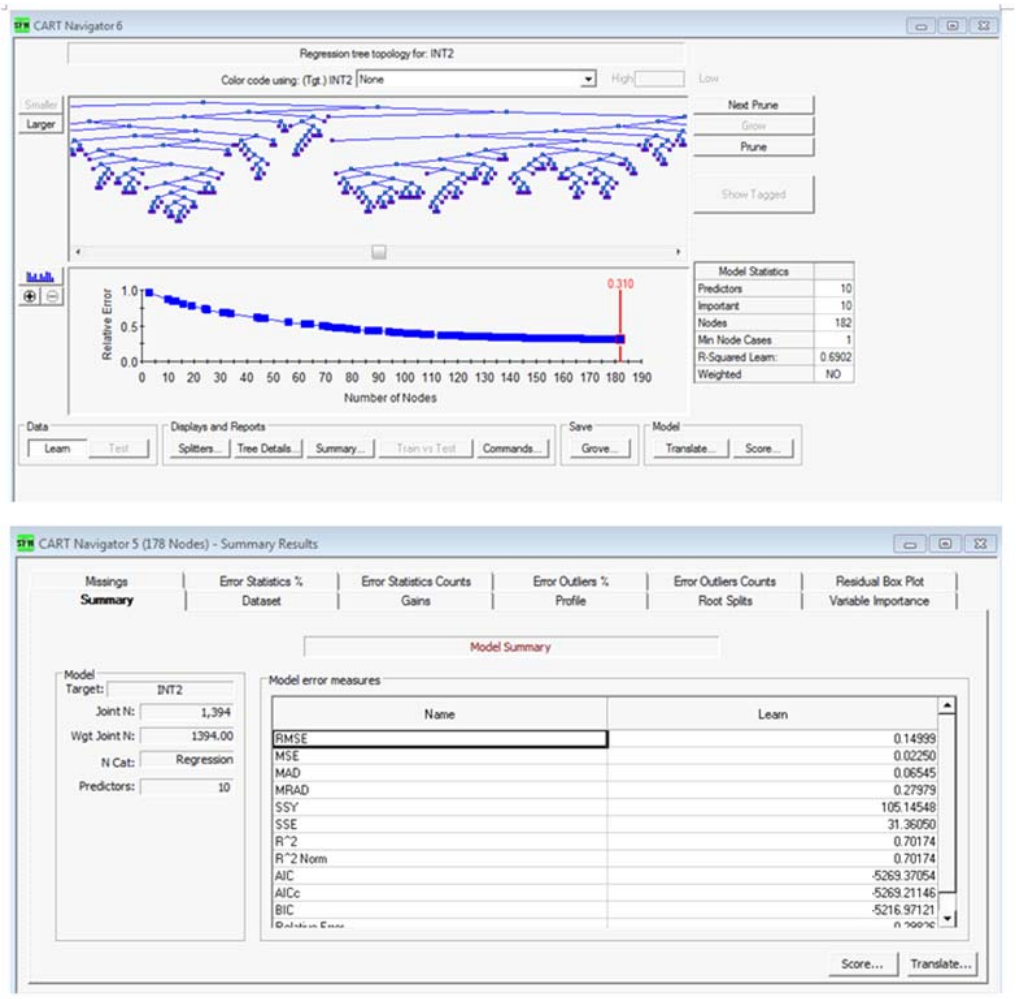


Figure 4.11 Regression tree analysis model results, produced using Salford SPM 7.0

The regression tree analysis model also identified 10 variables from the 29 candidates. The most important variable, POPden3 again, was set to have an importance score of 100. The selected variables and their importance scores are presented in Table 4.10.

Table 4.10 Regression tree identified key features and importance scores

Variable	Score	
POPDEN3	100.00	
PCTHIGHER	75.63	
EMPPCTUMEMP	73.59	
HSPCTGE3VEH	68.87	
INCNONFAMMEAN	60.61	
INCHHPCTFOODS	52.56	
PCTWHITE	51.04	
HSPCT2VEH	50.71	
COMMEANMINUTES	48.95	
INCHHPCT10_15	40.35	

Although the best regression tree took the same form as in Figure 4.10, it had 182 nodes, which were too many to be presented in a figure. It should be noted that the regression tree here served for key feature selection instead of for real

prediction. A recent successful use of regression tree can be found in Falcone and Wong (2012).

4.3.4.3 "Hot Spot" Feature Patterns

Classification tree analysis was conducted mainly aiming at the detection of the "hot spots" key feature patterns. However, the resulted detailed tree was too big to examine "hot spots" feature patterns by visualization. Here I used the three components identified in finite mixture model to represent the level of the fatal crash intensity, with 1 for high risk, 2 for middle risk and 3 for low risk. Means and medians of key features identified by the classification tree are computed by these 3 levels for comparison in Table 4.11. In addition, the medians of these features by fatal crash intensity level are illustrated in Figure 4.12. The means of these features by the fatal crash intensity levels showed approximately the same pattern and hence are omitted.

The population densities, the percent of households owning 3 or more vehicles, the percent of workers commuting to work by public transportation, and the percent of white exhibited different patterns as the intensity levels varied. More expertise and further explorations are needed to give in depth interpretation on what these features really mean and how these features affect the traffic safety.

Table 4.11 Median and Mean of key features by fatal crash intensity level

Label	Median			Mean		
COMPONENT	1	2	3	1	2	3
INTENSITY	1.92	0.26	0.00	2.74	0.38	0.00
POPDEN3	18.16	10.35	15.09	18.80	10.66	14.79
HSPCTGE3VEH	11.40	28.50	19.20	13.17	28.66	20.91
COMPCTPUBTRANS	15.10	2.80	5.40	16.35	5.16	9.83
HSPCT1VEH	41.10	26.90	32.25	40.71	27.11	32.14
HSPCT2VEH	30.00	38.80	37.90	29.30	39.22	36.47
PCTWHITE	37.60	79.00	67.10	39.28	66.57	57.36
PCTOCCPHU	92.90	94.95	93.70	90.17	93.55	91.11
PCTHIGHER	84.10	90.80	89.60	81.08	90.47	86.78

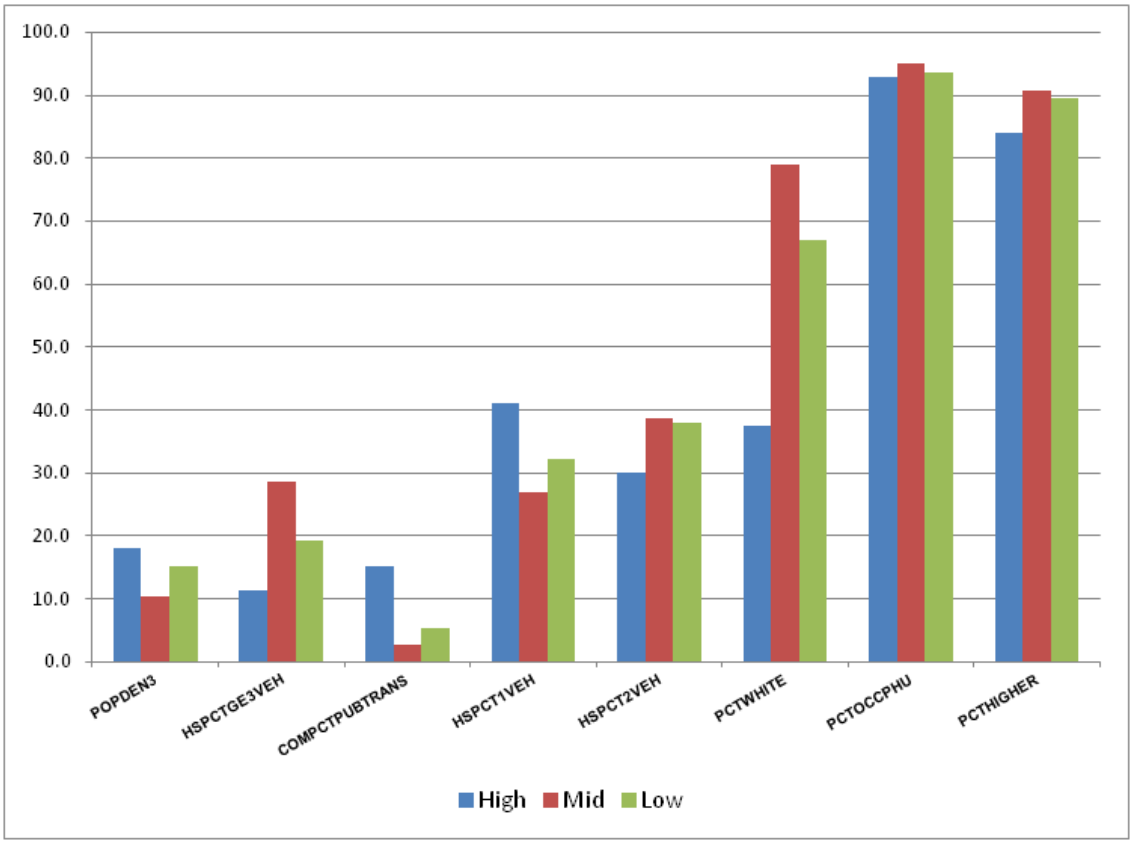


Figure 4.12 Key feature medians by fatal crash intensity level

4.3.4.4 Key Feature Space Formation

The fatal crash intensity to be predicted for time $t+1$ will be in numeric form, and the 10 variables identified by regression tree will be used to form a 10-D key

feature space for the prediction. The domain of each variable is listed in Table 4.12.

Table 4.12 Ranges for key features selected by regression tree

Variable	Min	Max
POPden3	0.0	44.3
PctHigher	19.3	100.0
EMPPctUmemp	0.0	40.8
HSpctGE3Veh	0.0	62.0
INCnonFamMean	9,694.0	251,267.0
INCHHpctFoodS	0.0	54.1
PctWhite	0.5	98.7
HSpct2Veh	0.0	62.9
COMmeanMinutes	10.5	50.2
INCHHpct10_15	0.0	63.6

4.4 Key Feature Space Partition

4.4.1 Transformation and Imputation

The standardization method described in section 3.3.1, robust MAD standardization, is applied to the key feature vector. The MAD standardization uses the median absolute deviation as scale and uses the median as location. The

transformation of ith feature of the jth observation is done by taking $x'_{ji} = \frac{x_{ji} - \text{median}_l(x_{li})}{\text{MAD}_{\cdot i}}$.

It should be noted that all features except POPden3 had missing values. An imputation procedure is applied here to set the missing value of a feature to its median. Table 4.13 lists the number of missing values for each feature.

Table 4.13 Number of missing values for key features selected by regression tree

Variable	N	Number of missing values
POPden3	1394	0
PctHigher	1389	5
EMPPctUmemp	1385	9
HSpctGE3Veh	1386	8
INCnonFamMean	1383	11
INCHHpctFoodS	1386	8
PctWhite	1390	4
HSpct2Veh	1386	8
COMmeanMinutes	1384	10
INCHHpct10_15	1386	8

4.4.2 Decision on the Number of Clusters, k_0

In this research the distance function between two points (already standardized) $\boldsymbol{x}'_u, \boldsymbol{x}_v$ is defined as $\text{d}(\boldsymbol{x}'_u, \boldsymbol{x}'_v) = \text{d}(\boldsymbol{x}'_v, \boldsymbol{x}'_u) = \sqrt{\sum_{i=1}^{m} \omega_i^2 (x'_{ui} - x'_{vi})^2}$ where ω_i is the relevant variable importance score for which the most important feature

POPden3 has the highest ω_i of 1.0 and the least important feature in Table 4.10, INCHHPCT10_15, has the lowest ω_i of 0.40.

A distance matrix is computed for each pair of the 1,394 census tracts. Using the median (mean) of this distance as radius, the cluster modality analysis suggest number of clusters should not be smaller than 7.

Following the algorithm based upon Hartigan's index described in section 3.1.3, WCSS (within-cluster sum of squares) of distances by each increase of number of clusters was produced and compared. The procedure repeated 31 times with the minimum cluster set to be 7 and the maximum number of cluster set to be 37 (square root of N)

.Table 4.14 Within-cluster sum of squares change for 1 cluster increase

k1	WCSS k1	k2	WCSS k2	WCSS change
7	9,469.72	8	8,998.07	-471.64
8	8,998.07	9	8,172.11	-825.96
9	8,172.11	10	8,172.02	-0.09
10	8,172.02	11	7,957.79	-214.23
11	7,957.79	12	7,780.77	-177.02
12	7,780.77	13	7,229.25	-551.52
13	7,229.25	14	6,857.56	-371.69
14	6,857.56	15	6,796.65	-60.91
15	6,796.65	16	7,019.34	222.69
16	7,019.34	17	6,360.84	-658.50
17	6,360.84	18	6,319.46	-41.39
18	6,319.46	19	6,197.69	-121.77
19	6,197.69	20	5,900.30	-297.39
20	5,900.30	21	6,103.40	203.10
21	6,103.40	22	5,757.99	-345.41
22	5,757.99	23	5,812.12	54.12
23	5,812.12	24	5,730.47	-81.65
24	5,730.47	25	5,602.40	-128.07
25	5,602.40	26	5,402.17	-200.23
26	5,402.17	27	5,325.53	-76.64
27	5,325.53	28	5,524.50	198.97
28	5,524.50	29	5,305.91	-218.59
29	5,305.91	30	5,301.28	-4.63
30	5,301.28	31	5,259.50	-41.78
31	5,259.50	32	4,926.87	-332.63
32	4,926.87	33	5,008.44	81.57
33	5,008.44	34	4,922.44	-86.00
34	4,922.44	35	4,764.47	-157.97
35	4,764.47	36	4,872.86	108.39
36	4,872.86	37	4,687.27	-185.59

Typically WCSS decreases with the increase of k, the number of clusters, and when k reaches the sample size n, WCSS becomes 0. When WCSS increased as k increased to $k+1$, it suggests k might be the number of clusters sought. Table

4.14 gives 4 candidates of k_0, the optimal number of k, respectively at 15, 20, 22, 27, 32, 35. After compared the observed intensities by clusters from the suggested k_0 with the component estimates resulted from the finite mixture model and studied within cluster standard deviations of observed intensities, I set k_0 to 15.

4.4.3 Study Area Partition

Following the method described in section 3.1.4 and the one-to-one relationship of $\boldsymbol{x}_j \leftrightarrow c_j$ defined in (3.7), the key feature space key partition was mapped to the study area partition. Figure 4.13 illustrates this partition. Each color in Figure 4.13 represents a specific cluster. The 12 census tracts in white are pure-water area and were not included in the clustering. They are kept only to ensure the integrity of the map. The statistics for the observed intensities are listed in Table 4.15.

Table 4.15 Statistics of census tract level observed intensity by clusters

Cluster	Number of census tracts	Mean	Std Dev
1	**164**	0.43	1.08
2	**35**	0.75	2.59
3	**404**	0.06	0.17
4	**125**	0.44	1.49
5	**4**	0.00	0.00
6	**56**	0.31	0.93
7	**1**	0.00	.
8	**31**	0.60	2.25
9	**2**	0.00	0.00
10	**4**	0.00	0.00
11	**8**	2.83	4.05
12	**16**	0.01	0.04
13	**1**	0.00	.
14	**330**	0.28	0.70
15	**213**	0.09	0.22

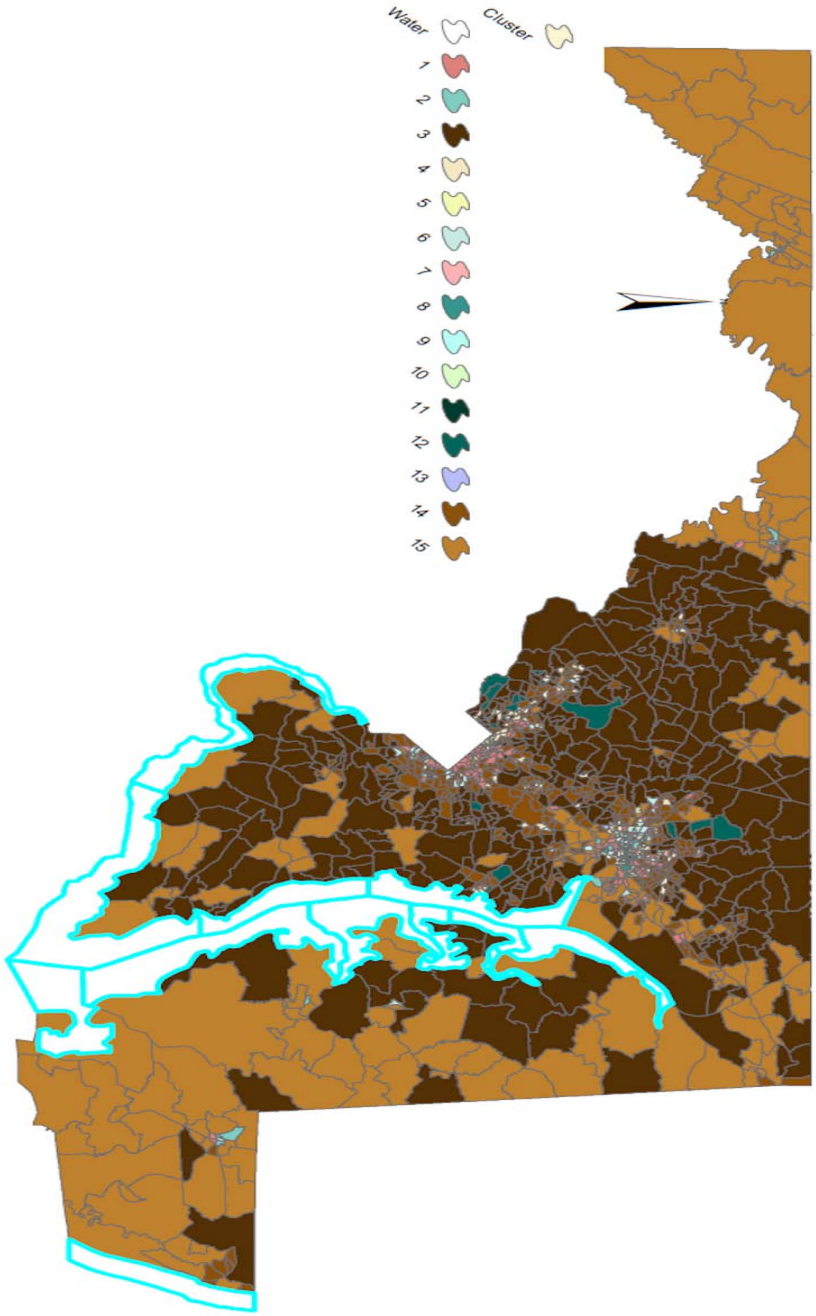


Figure 4.13 Partition of study area, map produced by use of ArcMap 10.0

4.5 Prediction of 2011 Maryland Fatal Crash Intensities

4.5.1 Settings of Priors

In this section, the three approaches of BHM described in section 3.2.4 were applied using 2009-10 Maryland FARS data as basis to predict the 2011 FARS crash intensities. The actual 2011 FARS data were used to check the prediction accuracy. The BHM modeling at $t-1$ for the kth cluster was implemented either by a hierarchical model in which λ_k' (prior distribution of fatal crash intensity of $t-1$) depended on the unknown $\alpha_k^{'}$, $\beta_k^{'}$, or by a half empirical model in which $\alpha_k^{'}$ became constant while $\beta_k^{'}$ remained stochastic, or by a pure empirical model in which both $\alpha_k^{'}$ and $\beta_k^{'}$ became constant.

For hierarchical model, the hyperparameter α_k followed an exponential distribution with a known parameter A_α, here A_α was tuned to be 2, thus

$$\pi(\alpha_k^{'}) = 2e^{-2\alpha_k^{'}}.$$

The hyperparameter $\beta^{'}$ followed a Gamma distribution with known parameters B_α and B_β, B_α and B_β were set to be 0.5 and 2.5 respectively, thus

$$\pi(\beta_k^{'}) = \frac{2.5^{0.5}}{\Gamma(0.5)}\beta_k^{'\,-0.5}e^{-2.5\,\beta_k'}.$$

The prior distribution followed a Gamma distribution conditioned on $\alpha_k^{'}$, $\beta_k^{'}$,

$$\pi(\lambda_k^{'}|\alpha_k^{'},\, \beta_k^{'}) = \frac{\beta_k'^{\alpha_k'}}{\Gamma(\alpha_k')}\lambda_k'^{\alpha_k'-1}e^{-\beta_k'\lambda_k'}.$$

The approach tuning of B_α and B_β was a technique similar to equation (3.27) and (3.28), and the same setting was applied to all clusters.

For the half empirical model, the setting of $\alpha_k^{'}$ was different. In this case, $\alpha_k^{'}$ became a constant parameter for a given cluster. The calculation of $\alpha_k^{'}$ was already shown in equation (3.27). Hence the value of $\alpha_k^{'}$ varied with the clusters but the setting of $\beta_k^{'}$ remained the same across clusters.

For the empirical model, both $\alpha_k^{'}$, $\beta_k^{'}$ became constant. The setting of $\alpha_k^{'}$ was exactly the same with the above half empirical model. $\beta_k^{'}$ was computed by equation (3.28). In the case of the empirical model, both $\alpha_k^{'}$and $\beta_k^{'}$ depended on the clusters. Details of these settings were listed in Table 4.16.

Table 4.16 Settings of prior parameters for half empirical model and empirical model

Cluster	Number of census tracts	Observed 2009 fatal intensities		Half empirical	Empirical	
		Mean	Standard deviation	α prior 2009	α prior 2009	β prior 2009
1	164	0.4276	1.2383	0.5425	0.5425	1.2688
2	35	0.6250	1.7265	0.7726	0.7726	1.2362
3	404	0.0751	0.1870	0.6049	0.6049	8.0544
4	125	0.4452	1.8171	0.1098	0.1098	0.2467
5	4	1.6353	3.2706	0.0935	0.0935	0.0572
6	56	0.4554	1.2398	0.5495	0.5495	1.2066
7		0.0000	.			0.0000
8	31	0.9763	2.1879	0.2089	0.2089	0.2140
9	2	2.4755	3.5009	0.0816	0.0816	0.0330
10	4	0.0000	0.0000			
11	8	0.0000	0.0000			
12	16	0.2617	0.5356	0.5270	0.5270	2.0137
13	1	0.0000	.			
14	330	0.2893	0.6542	0.7936	0.7936	2.7435
15	213	0.1464	0.4140	0.1779	0.1779	1.2153

It should be noted that the methods described in equations (3.27) and (3.28) did not apply to cluster 7, 10, 11 and 13 because these clusters had the 2009 intensity mean (or standard deviation) valued at 0, which made (3.27) and (3.28) mathematically invalid.

4.5.2 Posterior Results of Three Models / Comparison of the Three Models

Following the updating mechanism described in section 3.2.4, the posterior estimates of 2009 parameters became the 2010 prior parameters and the 2010 posterior parameters served as the 2011 prior parameters. Based upon the 2010 posterior parameter estimates, the predicted cluster level fatal crash intensity means and standard deviations are listed in Table 4.17. The third column of Table 4.17 are the observed results for comparison.

Table 4.17 Predictions on 2011 fatal crash intensity based upon 2010 posterior results

Cluster	n	Actual	Hierarchical	Half empirical	Empirical
3	404	0.0419	0.0362	0.0362	0.0393
14	330	0.1861	0.2209	0.2222	0.2187
15	213	0.0233	0.0244	0.0244	0.0273
1	164	0.3427	0.5058	0.5068	0.4454
4	125	0.2438	0.3296	0.3257	0.2847
6	56	0.488	0.2542	0.2579	0.2927
2	35	0.186	0.2951	0.3087	0.3564
8	31	0.2841	0.7544	0.7188	0.8541
12	16	0.024	0.0278	0.0301	0.0564
11	8	0.6711	1.7574		
5	4	0	0.7732	0.1428	1.0344
10	4	5.4054	0.5242		
9	2	3.2258	1.1418	0.1838	1.6564
13	1	0	0.2717		
7	1	0	0.2785		

In Table 4.17, predictions for cluster 15, 3, 12, and 14 were closer to actual results than other clusters. Number of census tracts of these 4 clusters accounted 69% of all census tracts. For all three models, clusters with lower fatal crash intensities and greater numbers of census tracts turned out to be ones with more accurate predictions.

Further comparison of the three models found that the hierarchical model did not have an obvious advantage over the half empirical and the empirical model models except that it could be applied to clusters to which the other two models were inapplicable.

4.5.3 Predicted Fatal Crash Centroid Shift

Based upon the hierarchical model results and following equation (3.37) and (3.38), the predicted fatal crash intensity for Maryland, all census tracts in 2011, was estimated to be 0.0478 with a standard deviation of 0.0059. The abstract intensity centroids for 2010, 2011 actual fatal crashes and the 2011 predicted fatal crashes were obtained by applying equation (3.39) and (3.40). The abstract fatal crash intensity centroid shift was presented in Figure 4.14.

In this chapter, I have applied the methods developed in previous two chapters to FARS Maryland 2009-11 data.

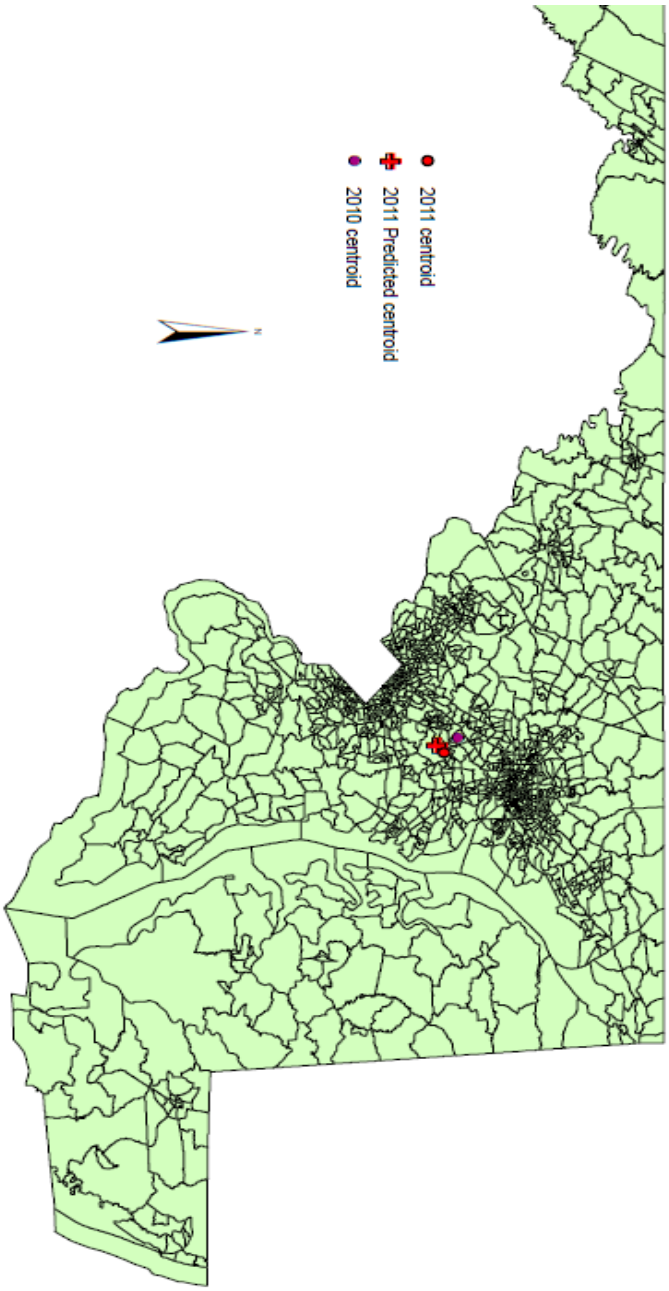


Figure 4.14 Maryland fatal crash intensity abstract centroid shift from 2010 to 2011, map produced by ArcMap 10.0

5. Conclusion, Summary and Future Work

In Chapter 2 and Chapter 3, I have developed the methods of the Bayesian hierarchical model for spatial-temporal processes whose spatial process is characterized by a Poisson point process. In Chapter 4, I applied these methods on Maryland 2009-11 fatal crash data over 1,394 census tracts. In this chapter, I draw conclusions, summarize the research of this dissertation, discuss the limitations of the methods and the application, and propose directions for future work.

5.1 Conclusion

In the proposal, I set three primary goals of this dissertation, to build a spatial-temporal model on highway loss point processes that can: identify key features associated with highway loss event intensities; detect the key feature patterns related to "hot spot" losses; and predict future losses based upon past highway loss events. Based upon the application in Chapter 4 which implemented the methods developed in Chapter 2 and 3, this dissertation has successfully accomplished these goals, though methods built still have great potential for improvement.

In recent years, spatial-temporal model theory in statistics has been gaining momentum and to some extent become a cutting edge direction. To develop methods to deal with the spatial-temporal loss data of highway safety will not

only benefit the highway safety community but also the whole society. In this dissertation, I have tried to focus my contribution efforts in two aspects.

First, I have made progress in interdisciplinary area of statistics and spatial point processes. Although the application of GIS is limited in this dissertation, it has shown great potential for future work.

Second, I find solutions from perspectives different from traditional statistical models. Traditional spatial-temporal models either use a separable model to separate spatial process from temporal process, or use nonseparable models through the introduction of an correlation function, and often requires the process to be stationary.

The first approach often simplified the problem at the risk of losing useful information. The second approach often needs complicated equations to describe the correlations and in implementation requires the integral of these functions. A good example is the fatal crashes that occurred in Maryland. These crashes concentrate around Washington, D.C., and Baltimore; thus the distribution follows a two-mode pattern for which it is very hard to claim stationarity.

This dissertation differs in its approach by not defining the intensity function $\lambda(\boldsymbol{s})$ on the whole study area $D_{\boldsymbol{s}}$; rather I partitioned the study area into small cells, c_j s, and I quantified the $\lambda(c_j)$s. Instead of exploring the correlation between the geographic space and temporal space, I explored the relationship between the feature space and the temporal process.

Another advantage is that methods developed in this dissertation incorporate the Bayesian approach. The BHM model has an updating mechanism that can make

use of prior information of past events. More important, the development in Bayesian computation has made the application of these methods practical. The Bayesian approach also brings more flexibility so that models can be tailored to meet the specific requirements of applications.

Methods developed in this dissertation have good scalability. For example the application in Chapter 4 can be easily expanded. It can also be done in a small area, say, a county, so long as cells can be formed and feature information can be collected.

5.2 Summary

I summarize the contributions in this dissertation by chapter.

Chapter 2

The frame of Chapter 2 had two parts. In the first part, a mixture model was proposed to model highway loss incidents in the study area D_s. The theory of extending the expectation maximization (EM) algorithm to highway loss Poisson point process was first built for the scenario that the number of components was known. Then, for the case the number of components was unknown, I gave the criterion in determining number of components and described how to estimate the range of number of components via prior information, and in the end of this part, an algorithm was developed to determine the number of components.

In the second part of this chapter, I developed methods which can identify key features first via a visualization procedure, then by using classification and regression trees and the random forest algorithm to finalize the key feature vector

whose elements had great influences on distribution of highway loss incidents. In this part, feature patterns were defined and "hot spots" were characterized. Methods of detect the key feature patterns corresponding to "hot spots" were developed theoretically. It also introduced the concept and definition of variable importance, and described how it was quantified thus could be used in Chapter 3.

Chapter 3

Similar to Chapter 2, Chapter 3 also has two parts. The first part is about clustering, the second part is on the Bayesian model. In the first part, the mapping between the study area partition (spatial partition) and the key feature space partition was defined first. Then, a new distance function was defined for variables selected based upon their relative importance scores. Methods were developed to determine the optimal number of clusters according to changes of within cluster sum of squares. After clustering and corresponding partition of study area, cells in the key feature space "close" to each other are aggregated into clusters so that future losses can be measured and predicted in homogenous clusters (partitions) instead of on each single cells.

The second part of Chapter 3 designed a Bayesian Hierarchical Model (BHM) which can predict cluster level losses at $t+1$ using the posterior distribution of current losses at time t. The posterior of the most recent past losses at time $t-1$ was used to provide prior information for losses at time t. The Poisson-Gamma design had two advantages: the proposed prior is a conjugate prior, thus the posterior of the gamma distribution also took the form of gamma distribution; the proposed Bayesian model has updating mechanism thus adds adaptation to the Bayesian approach. In addition to the hierarchical model, a half empirical model

and a full empirical model were given as alternatives. Methods for estimating the parameters of two empirical models were specified.

Chapter 4

In Chapter 4 methods developed in Chapter 2 and 3 were applied on 2009, 2010 and 2011 FARS data for 1,394 Maryland census tracts. FARS crash data which contained the exact location (latitude and longitude) and time were first joined to census tracts shape files via geocoding. The distribution of the 2010 (current time t) observed fatal crash intensity was examined. Then finite mixture models were applied to the 2010 sample which identified three underlying subpopulations. A classification tree was used to decide variables which determine intensity levels (categorical). A regression tree was then implemented to identify variables that can predict the intensities in a quantitative way. Based upon the variables and their importance levels decided in the regression tree, the key feature space was partitioned using the clustering mechanism developed. Clusters formed in key feature space were mapped to study area and all census tracts were grouped into 15 clusters. Three Bayesian approaches were applied to these clusters, using the 2009 posteriors as 2010 priors, and then use 2010 posteriors as estimates of 2011 losses, results of three approaches were evaluated for a brief comparison. At the end of Chapter 4, the 2011 hierarchical model results were visualized.

5.3 Limitations

In this section I discuss the limitation of the application in Chapter 4 first, then give details of limitations in methods.

Limitations of the application

In Chapter 4 the prediction of 2011 fatal crash intensities had two limitations. First, from 2008 on, fatal crashes in Maryland have been declining, and it is consistent with all states trend. In the application, it was predicted that 2011 Maryland had 464 fatal crashes (stdev 21.5) which was higher than the actual results, 455, thus, the model did not catch the decline trend. The inaccuracy was largely caused by the lack of timeliness of key feature data used in the model.

The trend of fatal crash intensity was mainly influenced by two factors, in long term by crash avoidance technology development, and in the mid term by economic growth. Historically, three crash avoidance equipments effectively cut the fatal crashes, seat belt, ABS (Anti-lock braking system), and ESC (electronic stability control). In recent years the effort was shifted to active crash avoidance technologies like automatic crash warning and braking system. In mid term, e.g. 5 years, which is of interest of this dissertation, the fluctuation of fatal crash intensities to a large extent depends on economy performance. However, the best economy growth indicator ready for use of this dissertation, unemployment rate, was only available for 5-year period of 2006-10. This was also the case of all other key features in Chapter 4, all of them took 2006-10 ACS values meaning they were constant between 2006 to 2010. The lack of timeliness of predicting variables directly led to the inaccuracy of prediction of 2011 fatal crash intensities.

Limitations of methods

In this dissertation cells in topology are polygons instead of points, and loss intensities of cells are represented on abstract points instead of real points, thus area of cells should be small enough. In other words, methods developed in this dissertation must be put in the appropriate context. On the other hand, in the real world the research interests in most cases focus on "hot spot" area instead of exact point locations.

After key feature space and study area partition, within each cluster I assumed completely spatial randomness at time t, and in the BHM model I further extended this assumption to time $t-1$ for the same cluster. This assumption in theory might be too strict. A random term might be needed so that the loss intensities of cells in a cluster could have some random variation centered around the cluster loss intensity.

In the application, the finite mixture model identified 3 subpopulations, and the quantities of the estimated fatal crash intensities were in 3 different scales from high to low, and the higher risk level subpopulation had a intensity about 10 times that of next level. The proposed Poisson-Gamma BHM approach worked better for low level fatal crash intensity clusters than for clusters having higher fatal crash intensity. This is an indication that there might be better BHM updating mechanism than the Poisson-Gamma setting for high risk clusters since the Poisson model typically works better for rare events.

5.4 Future work

I have planned three types of work for future improvement and development: immediate work to extend and improve the application in Chapter 4; refinement of current methods; and new development of current methods.

Immediate future work

The application in Chapter 4 has findings that were heretofore unknown by the highway safety community: it identified three subpopulations characterized by three underlying Poisson processes; it selected key features and detected "hot spot" patterns; it partitioned the study area, and it predicted the future loss based upon current and past losses and showed accuracy. The highway safety community would have interest to know above findings for all states, thus the extension of the application to all states upon the most current FARS data would be of interest of the highway safety community.

Refinement of current methods

There are ways to improve the methods developed in Chapter 2 and 3. First, examine the correlation structure of variables selected in CART to exclude "redundant" variables. If two variables have a correlation higher than a preset threshold only one is kept. Second, add a within cluster random term to BHM model so that the loss intensity of a cells can be expressed as a sum of two components, cluster density and the random term. Third, find a more appropriate prior-posterior setting for the BHM when clusters formed were classified as high risk clusters. Fourth, more work can be done to find out better solutions that can decide the exact number of clusters in the partition of key feature space. Besides,

for applications losses with past loss information at $t-2$ or even earlier, new way of prior parameter estimates should be constructed that can make full use of all past information by assigning more current past information higher weight.

New development in methods

The following new developments will benefit highway safety and will be welcomed.

First, the constraint of transportation network will be incorporated into the model. The transportation network is a one-dimensional space that is only a subset of the two-dimensional space and almost every all highway loss event occurs on the transportation network (Yamada and Thill (2007)). Under this frame, the point process and corresponding intensity functions will be redefined, and traffic specific factors such as speed limit will be included as predictors.

Second, add seasonality to current models. The highway safety community would be interested to know how the loss intensity vary with season to answer questions like "Did the Maryland 2011 fatal crash intensity vary by month?".

Third, in addition to the current function that can predict the nearest future, I also hope the model can answer the following question such as how did the intensity vary within a relatively long study time period and what's the mechanism behind this variation.

REFERENCES

Ahmad, A. and Dey, L. (2007), "A *K*-mean clustering algorithm for mixed numeric and categorical data", *Data & Knowledge Engineering,* 63 (2007), 503-527.

Akaike, H (1974) "A new look at the statistical model identification", *IEEE Transactions on Automatic Control* 19 (6): 716–723.

Anslin, L. (2003), " An Introduction to Point Pattern Analysis using CrimeStat", Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems", *Journal of the Royal Statistical Society: Series B*, 36, 192-236

Besag, J., York, J. C., and Mollie, A. (1991), "Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1-59

Besag, J. and Kooperberg, C. (1995). "On conditional and intrinsic autoregressions", *Biometrika*, 82, 733-746.

Berger, J. (1990), "Robust Bayesian analysis: sensitivity to the prior" *Journal of Statistical Planning Inference,* 25, 303-328.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer-Verlad, New York.

Berger, J., Dernardo, J. and Sun, D. (2009a), " Natural induction: an objective Bayesian approach", *Rev Acad Sci Madrid*, A, 103, 125-159

Berger, J., Dernardo, J. and Sun, D. (2009b), "The formal definition of reference priors", *Annals of Statistics* 37, 2, 905-938.

Berliner, L. M. (1996), "Hierarchical Bayesian time-series models" *Maximum Entropy and Bayesian Methods*, 15-22, Kluwer Academic Publishers, Dordrecht, NL, 1996.

Bilonick, R.A. (1983), "Risk qualified maps of hydrogen ion concentration for the New York State area for 1966-1978", *Atmospheric Environment*, 17, 2513-2524.

Bilonick, R.A. (1985), "The space-time distribution of sulfate deposition in the Northeastern United States", *Atmospheric Environment*, 19, 1829-1845.

Bilonick, R.A. (1988), "Monthly hydrogen ion deposition maps for the Northeastern U.S. from July 1982 to September 1984", *Atmospheric Environment*, 22, 1909-1924.

Bilonick, R.A. and Nichols, D.G. (1983), "Temporal variations in acid precipitation over New York State – What the 1965-1979 USGS data reveal", *Atmospheric Environment*, 17, 1063-1072.

Block, C. (1995), "STAC hot-spot areas: A statistical tool for law enforcement decisions. In Block, C. R., Dabdoub, M., and Fregly, S. (Eds.)", Crime analysis

through computer mapping,Washington, DC: Police Executive Research Forum, p. 20036.

Bolstad, W. (2010), *Understanding Computational Bayesian Statistics*, Wiley, New York.

Breiman, L. (2001), "Random forests", *Machine Learning* 45: 5–32.

Breiman, L. and Cutler, A. (2008), *Random Forests -Classification Manual*.

Breiman, L., Friedman, J. H., Olshen, R.A., and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA. Republished by CRC Press.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003) "Efficient construction of reversible jump Markov Chain Monte Carlo proposal distributions", *Journal of the Royal Statistical Society, B, 65, 1, 3-55.*

Cane, M. A., Kaplan, A., Miller R. N., Tang, B., Hackert, E. C. and Busalacchi (1996), "Mapping tropical Pacific sea level: data assimilation via a reduced state space Kalman filter", *Journal of Geophysical Research.* 101, 22,599-22,617.

Carreira-Perpiñan, M. A. (1999) "Mode-finding for mixtures of Gaussian distributions". Technical Report CS-99-03, Department of Computer Science, University of Sheffield, UK

Cressie, N. (1993), *Statistics for Spatial Data, Revised Edition.* John Wiley & Sons, New York, p.619.

Cressie, N. (2011), *Statistics for Spatio - Temporal Data*, John Wiley & Sons, New York, p.206.

Cressie, N. (2011), *Statistics for Spatio - Temporal Data*, John Wiley & Sons, New York, p.349.

Dempster, Laird, and Rubin (1977) "Maximum likelihood from incomplete data via the EM Algorithm (with discussion)", *Journal of the Royal Statistical Society* B39, 1-38.Carlin, B. P., Banerjee, S., "Hierarchical multivariate CAR models for spatio-temporally correlated survival data", *Bayesian Statistics*, 7 (J. M. Bernardo et al., eds.), 45-64

Cressie, N. (1993), *Statistics for Spatial Data* (Second edition), John Wiley, New York.

Cressie, N. and Huang, H.-C. (1999), "Classes of nonseparable, spatio-temporal stationary covariance functions", *Journal of American Statistical Association*, 94, 1330-1340.

Diggle, P. J. (1983), *The Statistical Analysis of Spatial Point Patterns*, London: Academic Press.

Diggle, P.J. (2003), *Statistical Analysis of Spatial Point Patterns* (Second edition), Edward Arnold, London.

d'Ocagne (1885), *Coordonnėes parallèles et axiales: Mėthode de transformation gėomėtrique et procėdė nouveau de calcul graphique d´eduits de la considėration des coordonnėes parallèlles* Paris: Gauthier-Villars.

Egbert, G.D. and Lettenmaier, D.P. (1986), "Stochastic modeling of the space-time structure of atmospheric chemical deposition", *Water Resources Research*, 22, 165-179.

Everitt, B. S. (1984), *An Introduction to Latent Variable Models*, Chapman and Hall, London.

Falcone, J. and Wong, D. (2012), "Mapping urban land used in the United States by census zone using nationally available data", *Journal of Land Use Science*, 8, 4, (2013), 466-488

Fang, Y. and Wang, J. (2012), "Selection of the number of clusters via the bootstrap method", *Computational Statistics and Data Analysis*, 56 (2012), 468-477.

Fiksel, T. (1984), "Simple spatial-temporal models for sequences of geological events", *Elektronische Informationsverarbeitung und Kybernetik*, 20, 480-487.

Gelfand, A. and Smith, A. (1990), "Sampling-based approach to calculating marginal densities", *Journal of the American Statistical Association*, 85, 410, 398-409.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 6.

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley, New York.

Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136: a *K*-means clustering algorithm", *Journal of the Royal Statistical Society*, C, 28, 1, 100-108.

Hastie, T. J. and Tibshirani, R. J. (1986), "Generalized additive models (with discussion)", *Statistical Science*, 1(2): 297-318.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Insurance Information Institute, http://www.iii.org/media/hottopics/insurance/nofault/.

Insurance Institue of Highway Safety, http://www.iihs.org/research/hldi/fact_sheets /default. html.

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley, New York.

Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136: a *K*-means clustering algorithm", *Journal of the Royal Statistical Society*, C, 28, 1, 100-108.

Highway Loss Data Institute (April 2008), "Theft Losses by County", Insurance Spherical Report, A-75, 1-10.

Highway Loss Data Institute (April 2008), "Losses due to Animal Strikes", *Bulletin* 26 (5), 1-6.

Hurvich, C. M. and Tsai, C.-L. (1989), "Regression and time series model selection in small samples", *Biometrika* 76: 297–307

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, 91, 401–40

Laried, N. M. & Ware, J. H. (1982), "Random-effects models for longitudinal data" *Biometrika*, 81(3), 624-629

Levine, N. (1998), " *Hot Sport* Analysis using *CrimeStat* kernel density interpolation", Presentation of Annual Meeting of the Academy of Criminal Justice Sciences, Albuquerque, NM, March 10-14, 1998

Liang J., Zhao, X., Li, D., Cao, F. Dang C. (2012), "Determining the number of clusters using information entropy for mixed data", *Pattern Recognition*, 45 (2012), 2251-2265.

Liang, K.Y. and Zeger, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika*, 73, 13-22.

Liu, H. and Brown, D. E., "C riminal incident prediction using a point-pattern-based density model", *International Journal of Forecasting*, 19 (2003) 603-622.

MacQueen, J. B. (1967), "Some methods for classification and analysis of multivariate observations", *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*, University of California Press, 281-297.

Marchette, D. J., Priebe, C. E., Rogers, G. W., and Solka, J. L. "Filtered kernel density estimation", *Computational Statistics*, 11, 95-112.

McLachlan, G. and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker.

McLachlan, G. and Krishnan, T. (1997) *The EM Algorithm and Extensions*, Wiley, New York.

McLachlan, G. and Peel, D. (2001) , *Finite Mixture Models*, Wiley, New York.

Moustafa, R. (2011), "Parallel coordinate and parallel coordinate density plots" *WIREs Computational Statistics* 2011 3,134–148

Moustafa, R. (2011), "Andrews curves" *WIREs Computational Statistics* 2011. 3, 373–382 Mojena, R. (1977), "Hierarchical grouping methods and stopping rules: An evaluation", *Computer Journal*, 20, 359-363.

Nelder, J and Wedderburn R (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society. Series A* (General) 135 (3): 370-384

NHTSA, (2012), "2011 Motor vehicle crashes: overview", *Traffic Safety Facts,* DOT HS 811 701.

NHTSA, (2010a), "Report to congress-NHTSA's crash data collection programs", DOT HS 811 337.

NHTSA, (2010b), "FARS analytic reference guide1975 to 2009", DOT HS 811 352.

NHTSA, (2013), "Fatality analysis reporting system(FARS)- analytic users manual 1975 to 2011", DOT HS 811 693 (2011).

Pearson, K. (1894) "Contributions to the theory of mathematical evolution, II: Skew variation", *Philosophical Transactions of the Royal Society of London* A 185, 71-110.

Robert C. and Casella, G. (2011) "A short history of Markov Chain Monte

Carlo: subjective recollections from incomplete data", *Statistical Science,* 2011 Vol 0, No. 00, 1-14.

Shao, J. (2003), *Mathematical Statistic*s, Second Edition, Springer, New York.

Sheather, S. J. (2004), "Density estimation," *Statistical Science*, 2004. Vol.19 No.4, 588-597

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Spiegelhalter, J. (1998), "Bayesian graphic modelling: A case-study in monitoring health outcomes", *Applied Statistics,* 47*,* 115-133.

Stein, M.L. (1986), "A simple model for spatial-temporal processes", *Water Resources Research,* 22, 2107-2110.

US Census Bureau, (2009), "A compass for understanding and using American Community Survey data", May 2009

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E., "Hierarchical spatio-temporal mapping of disease rates", *Journal of the American Statistical Association*, 92 (1997), 607-617

Wang, J. (2010), "Consistent selection of the number of clusters via cross validation", *Biometrika,* 97, 4, 893-904.

Wegman, E. (1990), "Hyperdimensional data analysis using parallel coordinates" *Journal of the American Statistical Association,* 85, 664-675

Wegman, E. (2003), "Visual data mining" *Statistics in Medicine*, 22, 1383-1397

White, E (2003), *Tort Law in America: An Intellectual History*, Oxford University Press.

Williamson, J. (2010), "Review of Bruno di Finetti. philosophical lectures on probability", *Philosophia Mathematica* 18, 1, 130-135.

Wong, M. A. and Schaack, C. (1982), "Using the kth nearest neighbor clustering procedure to determine the number of subpopulations", *American Statistical Association* 1982 Proceedings of the Statistical Computing Section, 40-48.

Yamada, I. and Thill, J. (2007), "Local indicators of network-constrained clusters in spatial point patterns", *Geographical Analysis*, 39, *268-292*

CURRICULUM VITAE

Yongping Yan graduated from The First High School, Qian County, Shaanxi, China, in 1987. He then received his Bachelor of Engineering from TianJin University, TianJin, China, in 1991. He then went on to receive his Master of Engineering from Shanghai JiaoTong University, Shanghai, China, in 1993. In 2002, he received his Master of Science from George Mason University.