IDENTIFYING HAZARDOUS LOCATIONS BASED ON SEVERITY SCORES OF HIGHWAY CRASHES

PARK, Shin Hyoung, Seoul National University, Seoul, Korea, shpark76@snu.ac.kr

KIM, Dong-Kyu, Seoul National University, Seoul, Korea, kimdk95@snu.ac.kr

KHO, Seung-Young, Seoul National University, Seoul, Korea, sykho@snu.ac.kr

RHEE, Sungmo, Seoul National University, Seoul, Korea, rheesm@snu.ac.kr

ABSTRACT

Identifying hazardous locations is an essential step for safety improvement programs or projects since it provides decision makers with logical and scientific basis in the allocation of budgets and other resources in a cost-effective manner. There have been numerous studies conducted to develop suitable methods for identifying hazardous locations; however, the majority of them did not consider spatial interactions (e.g. spatial dependency and spatial heterogeneity) which complicatedly appeared in accident analyses. With improvements of Geographical Information Systems (GIS) technology, it is possible to use various spatial analysis tools on traffic safety studies. Of those, Geographically Weighted Regression (GWR) and Kernel Density Estimation (KDE) are applied to perform this research. GWR is used to verify the effect of spatial dependency and spatial heterogeneity on the outbreak of traffic accidents. The role of the KDE in this study is displaying crash-clustered area under the consideration of an appropriate bandwidth and kernel function which determine extents and severity levels of accidents. This paper aims to develop a method for identifying hazardous locations based on severity scores of highway crashes. The method developed in this paper is applied to real-world data of Korean expressways. The results imply the necessity of examining spatial dependency and spatial heterogeneity in accident analyses and exploring hazardous locations based on crash severity. Findings from this research will contribute to both of saving time and efforts spent on manual tasks and enhancing accuracy for identifying hazardous locations by practical use of a comprehensive method.

Keywords: bandwidth, kernel density estimation, geographically weighted regression, hazardous locations, highway crashes

INTRODUCTION

Identifying hazardous locations is an essential step for safety improvement programs or projects since it provides decision makers with logical and scientific basis in the allocation of budgets and other resources in a cost-effective manner. Failure to identification of true hazardous locations results in errors such as false positives (i.e., identifying sites for safety improvements that should not have been selected) or false negatives (i.e., not identifying sites that should have been selected) and reduces effectiveness of safety improvement projects by waste of the resources. For decades, a variety of studies has been conducted to find appropriate methods; however, most of them did not take the effect of spatial interactions among traffic accidents into account properly.

With improvements of Geographical Information Systems (GIS) technology, it is possible to use various spatial analysis tools on traffic safety studies. Of those, Geographically Weighted Regression (GWR) and Kernel Density Estimation (KDE) are adopted to perform this research.

GWR is used to verify the effect of spatial dependency and spatial heterogeneity on the outbreak of traffic accidents. The common research framework based on traditional statistical models such as Ordinary Least Squares (OLS) regression models ignores spatially autocorrelated residuals and does not provide any way of exploring spatial heterogeneity across geographical space. In this study, by using Moran's I and comparing the results from both of OLS regression and GWR, it is confirmed that spatial autocorrelation and spatial heterogeneity are important considerations of accident analyses.

Kernel Density Estimation (KDE) is used for identifying hazardous locations for highway crashes recently, which is one of the most popular methods for analyzing the first order properties of a point event distribution. Using GIS tools, it can also visualize hazardous locations with smoothly tapered surface on the map. However, the role of KDE in the previous studies has only limited to displaying crash-clustered area without any sufficient theoretical consideration of an appropriate bandwidth and kernel function which determine extents and severity levels of accidents.

In Korea, hazardous locations on expressways have been identified based on the number of crashes within a 200m-radius segment from a crash location. If the crash count is five or more in metropolitan area or three or more in the other regions, that segment is selected as a hazardous location. This method is very simple and decisive because it assumes that locations which have many crashes are obviously more risky than other locations which have respectively less crashes. However, since it gives equal weights to all types of fatal and injury crashes, it cannot take the severity of each crash into account. In order to improve this drawback, we assigned severity weights to each crash according to its severity (e.g. add 12 per fatality and 3 per injured person to crash frequency of a road segment). And then, KDE is performed to identify hazardous locations. Reducing overall system severity is more reasonable than reducing overall crash count because the former can save injury or loss of life and thus save social costs caused by traffic accidents.

The main objective of this paper is to develop a crash severity based method for identifying hazardous locations under the consideration of spatial interactions among traffic accidents. The flow of the study is as follows: (1) we firstly verify whether the crash data are spatially autocorrelated using spatial analysis methods; (2) we propose a severity-weighted KDE method for hazardous location analysis.

LITERATURE REVIEW

Spatial Dependency and Spatial Heterogeneity

Traffic accidents are basically spatial events occurring in geographic space. These are caused by various factors such as human behaviors, mechanical failure of vehicles, roadway geometries and environmental conditions. In addition, spatial interaction among these factors which is not identified on collision reports also can be a significant to crashes as latent factors. As known as 'Tobler's first law of geography' (Tobler, 1970), “Everything is related to everything else, but near things are more related than distant things.” This means that objects in geographic space do not distributed randomly but interact with each other. This phenomenon sometimes results in spatial dependency and spatial heterogeneity.

Spatial dependency implies a certain degree of redundancy in the additional information that is provided by the nearby locations within geographic space (Goodchild et al., 1992). Spatial dependency leads to the spatial autocorrelation problem in statistics because this violates underlying assumptions of many traditional (non-spatial) statistical methods. Therefore, traditional statisticians view spatial autocorrelation as a bad thing that needs to be removed from the data. On the contrary, GIS analysts view spatial autocorrelation as an evidence of important spatial processes at work (ESRI, 2008).

Spatial heterogeneity is characterized by spatial or regional dissimilarity between locations of objects in geographic space. The results of any analysis over a limited area can be different from the results that would be obtained for the other areas. These concepts tend to affect almost any kind of spatial analysis conducted on geographic data (de Smith et al., 2008).

Previous studies on accident prediction models analyzed crash causal factors from a traditional statistical standpoint without any sufficient consideration of spatial interactions. Safety Performance Function (SPF), one of popular accident prediction models, estimates the expected number of traffic accidents per unit of time using independent variables such as traffic flow rates and geometric design features (Zhong et al., 2009). SPF assumes that: (1) the rate of traffic collisions along a highway is spatially uncorrelated; (2) the rate at which collisions occur within the segment remains constant; (3) the factors causing high collision rates reside within the segment. (Chung et al., 2009)

For the first assumption, we examined spatial autocorrelation by estimating Moran’s I (Moran, 1950) which describes the existence of spatial dependency in the model. The second and the third assumptions are not appropriate because crash risk varies on each location even in the same segment and the factors causing high collisions spread to contiguous segments.

In order to consider these spatial effects, previous studies developed models based on homogeneous road sections divided by a geometric element as well as a variety of explanatory variables such as geometric features of the site, traffic volume and other environmental features (Miaou and Lum, 1993; Shankar et al., 1995). However these models could not solve these impractical assumptions properly due to the limits of traditional statistical analyses. Since crash data should be analyzed under the understanding of spatial data, spatial analysis techniques should be also applied in accordance with the data.

Kernel Density Estimation

In general, crash maps do not exactly reflect the crash concentrations of locations having more than one crash because the symbols for each of the crashes at one location lie on top of each other and thus are not shown distinctly (Pulugurtha et al., 2007). Therefore crash density estimation using a kernel function is a very useful method in the analysis of hazardous locations.

Kernel Density calculates a magnitude per unit area from point or line features using a kernel function to fit a smoothly tapered surface to each point or line. The surface value is highest at the location of the point, diminishes away from the point, and reaches zero at the radius distance from the point (Silverman, 1986). The radius distance is referred to as bandwidth, also called the smoothing parameter or window width, which is the most important criterion for determining the most appropriate density surface (Silverman, 1986; Fotheringham et al., 2002). Generally there are two methods for choosing optimal bandwidth. One approach for approximating the optimal bandwidth is to make use of cross-validation (CV) (Fotheringham et al., 2002).

$$\mathrm{CV} = \sum_{\mathrm{i}}^{\mathrm{n}} [\mathrm{y_i} - \hat{\mathrm{y}}_{\neq \mathrm{i}}(\mathrm{h})]^2$$

where, n is the number of data points and $\hat{\mathrm{y}}_{\neq \mathrm{i}}(\mathrm{h})$ is the fitted value of $\mathrm{y_i}$ with data from point i omitted from the calibration. Lower values of CV indicate better model fits.

The other approach is minimizing the corrected Akaike Information Criterion (AIC_c) (Akaike, 1974).

$$\mathrm{AIC_c} = \mathrm{n}\ln(2\pi\hat{\sigma}^2) + \mathrm{n}\left\{\frac{\mathrm{n} + \mathrm{tr(s)}}{\mathrm{n} - 2 - \mathrm{tr(s)}}\right\}$$

where, n is the sample size, $\hat{\sigma}$ is the estimated standard deviation of the error term, and tr(S) denotes the trace of the hat matrix which is a function of the bandwidth. Optimal bandwidth is a trade-off between bias and variance. Too small a bandwidth leads to large variance in the local estimates and too large a bandwidth leads to large bias in the local estimates.

However, an appropriate choice of the bandwidth should be determined by the purpose of the estimate. Silverman (1986, Section 3.4.1) suggests a subjective choice of the bandwidth if the purpose of the estimation is to explore the data in order to propose possible statistical

models and hypotheses. Anderson (2009) also suggests that the process of deciding the bandwidth is somewhat subjective. The choice of a bandwidth depends on the purpose of the analysis. For a global vision of the risk of a given road, a global estimator may be a good starting point. A smaller bandwidth allows for a narrowing of the global description (Flahaut et al., 2003). In order to find an appropriate bandwidth for identifying hazardous locations, Xie and Yan (2008) examined the impacts of search bandwidth at local and larger spatial extents. Six search bandwidths were used, including 20, 100, 250, 500, 1000, and 2000 m. It appears that the narrow bandwidths (20, 100, and 250 m) might produce patterns suitable for presenting local effects or hotspots at smaller scales. As the search bandwidth increased, the local hot spots were gradually combined with their neighbors, and thus larger clusters appeared. The larger bandwidths (500, 1000, 2000 m) seemingly gave better sense of locations of the hot spots at larger spatial scales.

There have been a few efforts to identify hazardous locations through KDE based on the understanding of the spatial interaction existing between contiguous crash locations. Flahaut et al. (2003) compared two methods for identifying and delimiting black zones: one method is based on spatial autocorrelation indices, the other one on kernel estimators. The spatial autocorrelation method could allow a better adaptation to the local spatial structure for a given road by giving the risk of a black zone and its length, while the kernel method gives the risk of each distance away from a crash location within bandwidth. Pulugurtha et al. (2007) developed a methodology to study the spatial patterns of pedestrian crashes in order to identify pedestrian black zones, and evaluate methods to rank these zones. They first selected 29 black zones using KDE, and then computed ranks for these zones using crash frequency, crash frequency based on severity, crash density based on area, crash rate based on vehicular volumes, crash rate based on population, sum-of-the-ranks method, and crash score method. Erdogan et al. (2008) identified hazardous locations using two methods: repeatability analysis based on a traditional statistical model using Poisson distribution and spatial analysis based on kernel density estimation. Results from the methods were almost overlapped and indicated the same locations.

DATA DESCRIPTION

A total of 842 crash data occurred on the northbound of the Gyeongbu expressway from 2006 through 2008 were used for this study. In accordance with the results from Kwak et al. (2010) and Park et al. (2010), the roadway is divided into 524segments on the basis of horizontal curve which is more influential to crash severity between horizontal curve and vertical grade. And then, values of variables in the crash data are aggregated on each of the segments.

Table 1 – Variables description

Category	Variables	Description
Dependent Variable	Freq	Total crash count in 3-year period (2006 – 2008)
	WFreq1	Each crash is multiplied by a weight based on the crash severity. WFreq1 = (1 x Property Damage Only Crash) + (3 x Injury Crash)

		+ (12 x Fatal Crash)
	WFreq2	A sum of weights based on the number of the injured person(s) and/or fatalities WFreq2 = (Crash Count in a road segment) + (3 x the number of Injured persons) + (12 x the number of fatalities)
	Acc_Cost	Total costs caused by crashes in a road segment Acc_Cost = (total costs of property damages) + (a unit cost of a injured person x the number of Injured persons) + (a unit cost of a fatality x the number of fatalities)
Explanatory Variable	NumLane	The number of lanes in a segment
	Bridge	The number of bridge(s) in a segment
	Tunnel	The number of tunnel(s) in a segment
	Camera	The number of speed camera(s) in a segment
	Offramp	The number of off-ramp(s) in a segment
	OnRamp	The number of on-ramp(s) in a segment
	Restarea	The number of rest area(s) in a segment
	TG	The number of toll booth(s) in a segment
	EXPO	Exposure of a segment to traffic
	HR	Radius of horizontal curve

The followings are detailed description of some variables.

- WFreq1: If a Property Damages Only (PDO) crash is equivalent to 1, severity weights are assigned 12 to a fatal crash and 3 to an injury crash. For example, if one fatal crash, 2 injury crashes, and 3 PDOs were occurred on a roadway segment, $\mathrm{WFreq1} = 1 \times 12 + 2 \times 3 + 3 \times 1 = 21$
- WFreq2: Severity weights are assigned 12 per fatality and 3 per injured person and total weights are added up to crash frequency on a segment. In case of a segment having 3 crashes which involve 3 fatalities and 7 injuries, $\mathrm{WFreq2} = 3 \times 12 + 7 \times 3 + 3 \times 1 = 60$.
- Acc_Cost: Accident cost of each crash is calculated based on the social crash costs (see table 2) obtained from Traffic Accident Analysis System of Road Traffic Authority in Korea (http://taas.rota.or.kr/index.jsp).

Table 2 - People Damages Costs

	2006	2007	2008
Fatality Costs	2,368,091	2,342,364	2,392,364
Persons	6,327	6,166	5,870
Unit cost per person	374	380	408
Injury Costs	1,167,000	1,208,636	1,180,000
Injuries	340,229	335,906	338,962
Unit cost per person	3.43	3.60	3.48
Total Costs	3,535,091	3,551,000	3,572,364

Notation 1) Unit: People or Thousand Dollars, 1 USD ≒ 1,100 Won

- Of those severity weighted variables, WFreq1 does not take the actual number of fatalities and injuries into account whereas Acc_Cost assigns too much weight to fatalities or fatal crashes. Therefore WFreq2 is considered as the most applicable variable reflecting the magnitude and type of crashes.
- EXPO: Since the length of the road section and the traffic volume (AADT) are the most closely related to the frequency of traffic accidents, an exposure variable is introduced to apply traffic conditions.

$$\mathrm{EXPO} = \frac{\mathrm{L} \times \mathrm{Y} \times 365 \times \mathrm{AADT}}{10^6} = \frac{\mathrm{L} \times 365}{10^6} \times \left(\sum_{\mathrm{i}=2006}^{2008} \left(\mathrm{AADT_i} \times \frac{\mathrm{Freq_i}}{\sum_{\mathrm{i}=2006}^{2008} \mathrm{Freq_i}} \right) \right)$$

where, L is the length of a road segment and Y is a period (3 years)
If there is no crash occurred for three years, AADT is a average of 3-year AADT.

$$\mathrm{EXPO} = \frac{\mathrm{L} \times \mathrm{Y} \times 365}{10^6} \times \frac{\sum_{\mathrm{i}=2006}^{2008} \mathrm{AADT_i}}{3}$$

METHODS

Spatial Autocorrelation

Spatial autocorrelation means correlation of a variable with itself through space. If there is any systematic pattern in the spatial distribution of a variable, it is said to be spatially autocorrelated. This is important because most statistics are based on the assumption that the values of observations in each sample are independent of one another. In geographic space, however, spatial autocorrelation is important phenomenon for spatial data analysis.

Moran's I (Moran, 1950) is one of the oldest indicators of global spatial autocorrelation and is still used for determining spatial autocorrelation. It compares the value of the variable at any one location with the value at all other locations.

$$I = \frac{N \sum_i \sum_j W_{i,j} (X_i - \overline{X})(X_j - \overline{X})}{(\sum_i \sum_j W_{i,j}) \sum_i (X_i - \overline{X})^2}$$

where *N* is the number of cases, X_i is the variable value at a particular location, X_j is the variable value at another location, X is the mean of the variable, and W_{ij} is a weight applied to the comparison between location *i* and location *j.* Wij is a distance-based weight matrix which is the inverse distance between locations i and j (1/dij).

Given a set of features and an associated attribute, Global Moran's I evaluates whether the pattern expressed is clustered, dispersed, or random. When the Z score indicates statistical significance, a Moran's I value near +1.0 indicates clustering while a value near –1.0 indicates dispersion. To test whether or not we can reject the null hypothesis, a Z score value is calculated as

$$Z(I) = \frac{I - E(I)}{S_{E(I)}}$$

where, E(I) is the expected value of Moran's I and $S_{E(I)}$ is an estimate of the theoretical standard deviation.

To determine if the Z score is statistically significant, it is compared to the range of values for a particular confidence level. For example, at a significance level of 0.05, a z score would have to be less than –1.96 or greater than 1.96 to be statistically significant (ESRI, 2008).

Geographically Weighted Regression

In an OLR model, model parameters are estimated globally. Once estimated, they apply universally, although the influence of some independent variables on the dependent variable may vary across space. This inability to take into consideration the spatial variability of the influence of independent variables may result in large model errors, thus weakening the explanatory power of a model (Zhao and Park, 2004). The underlying idea of GWR is that parameters may be estimated anywhere in the study area given a dependent variable and a set of one or more independent variables which have been measured at places whose location is known. We might expect that if we wish to estimates parameters for a model at some location then observations which are nearer that location should have a greater weight in the estimation than observations which are further away (Charlton and Fotheringham, 2009).

When the Ordinary Least Regression (OLR) model is written as

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i,$$

the GWR model can be written as

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \epsilon_i$$

where
y_i = dependent variable at location i (i = 1, 2, . . . , n, where n is the number of observations),
x_{ik} = independent variable of the k^{th} parameter at location i,
β_{ik} = estimated k^{th} parameter at location i for the GWR model,
β_k = estimated k^{th} parameter for the OLR model,
ϵ_i = error term at location i, and p = number of parameters.

The OLR estimator takes the form:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where $\hat{\beta}$ is the vector of estimated parameters, X is the design matrix which contains the values of the independent variables and a column of 1s, *y* is the vector of observed values, and $(X^T X)^{-1}$ is the inverse of the variance-covariance matrix.

In GWR, to estimate local parameters $\hat{\beta}(i)(i = 0, 1, \ldots, n)$ for any given location or data point *i,* the weights are conditioned on the location relative to the other observations in the dataset and hence change for each location. The estimator takes the form:

$$\hat{\beta}(i) = \left(X^T W(i) X\right)^{-1} X^T W(i) y$$

W(i) is an $n \times n$ matrix of weights relative to the position of i in the study area; $X^T W(u) X$ is the geographically weighted variance-covariance matrix (the estimation requires its inverse to be obtained), and *y* is the vector of the values of the dependent variable.

The W(i) matrix contains the geographical weights in its leading diagonal and 0 in its off-diagonal elements.

$$W(i) = \begin{bmatrix} w_1(i) & 0 & \cdots & 0 \\ 0 & w_2(i) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n(i) \end{bmatrix}$$

The weights themselves are computed from a weighting scheme that is also known as a *kernel*. A number of kernels are possible: a typical one has a Gaussian shape:

$$W_k(i) = e^{-0.5\left(d_k(i)/h\right)^2}$$

where $w_k(i)$ is the geographical weight of the k^{th} observation in the dataset relative to the location i, $d_k(i)$ is some measure of the distance between the k^{th} observation and the location i, and *h* is a quantity known as the bandwidth. The distances are generally Euclidean distances when Cartesian coordinates are used and Great Circle distances when spherical coordinates are used. The bandwidth in the kernel is expressed in the same units as the coordinates used in the dataset. As the bandwidth gets larger the weights approach unity and the local GWR model approaches the global OLR model.

Hotspot Identification

In this study, a Kernel Density Estimation (KDE) is suggested as a method for identifying hazardous locations considering crash severity and spatial interactions. To apply this method, we need to give attention to two parameters which determine the shape of kernel density: bandwidth and the function K, called a “kernel”. While bandwidth means a horizontal range of accident likelihood, kernel means the magnitude of risk at each location in the range. That is to say, bandwidth and kernel can be regarded as a horizontal element and a vertical element, respectively.

Conceptually, the kernel function divides study area into small cells defined by users and then, calculate cell values which are highest at the locations of the point, diminish away from the points, and finally reach zero at the radius distance from the points. The value of a cell located within bandwidths of two or more points are computed as a sum of each cell value obtained from each crash location. For example, if a cell named Q locates within bandwidths

of crash points a, b, and c, and Q(a), Q(b), and Q(c) denote the individual cell values, the value of Q is computed as follows:

$$Q_{Final} = Q(a) + Q(b) + Q(c)$$

This means that the spatial characteristics of the location Q affect the crashes a, b, and c. If severity weights, W applied to each crash, this equation can be written as:

$$Q_{Final} = W(a)Q(a) + W(b)Q(b) + W(c)Q(c)$$

Then we can generalize the risk at location Q as follows:

$$Q_{Final} = \sum_{i=1}^{n} W(i)Q(i)$$

As mentioned earlier, WFreq2 is used as severity weights. Resultingly, a location having a high cell value can be identified as a hazardous location and thus we can obtain different results compared to the results based on crash frequency.

The other parameter, bandwidth means the influential range to outbreak of a crash. In this article, a bi-level method is suggested, which identifies hazardous areas at wide scale and then determine exact locations or segments at narrow scale.

RESULTS

Verification of Spatial Autocorrelation

Global Regression

For the four dependent variables defined in Data Description section, ordinary least regression has conducted and the results of diagnostic statistics are summarized in table 3.

Table 3 - A summary of statistics resulted from OLR

Diag_Name	Diag_Value			
	Freq	WFreq1	Wfreq2	Acc_Cost
R2	0.519	0.477	0.623	0.560
AdjR2	0.510	0.467	0.615	0.551
F-Stat	55.457	46.824	84.648	65.206
F-Prob	0.000000*	0.000000*	0.000000*	0.000000*
Wald	248.136	185.009	40.929	76.028
Wald-Prob	0.000000*	0.000000*	0.000012*	0.000000*
K(BP)	34.331	52.358	320.105	38.216
K(BP)-Prob	0.000162*	0.000000*	0.000000*	0.000035*

R-Squared and Adjusted R-Squared are both statistics derived from the regression equation to quantify model performance. These statistics indicate how well the model's predicted values explain the variation in the observed dependent variable values. The value of R-squared ranges from 0 to 1. If the model fits the observed dependent variable values perfectly, R-squared is 1.0. The Adjusted R-Squared value is always a bit lower than the Multiple R-Squared value because it reflects model complexity (the number of variables).

Both the Joint F-Statistic and Joint Wald Statistic are measures of overall model statistical significance. The Joint F-Statistic is trustworthy only when the K (BP) statistic (see below) is not statistically significant. If the K (BP) statistic is significant, the Joint Wald Statistic should be consulted to determine overall model significance. The null hypothesis for both of these tests is that the explanatory variables in the model are not effective. For a 95% confidence level, a p-value smaller than 0.05 indicates a statistically significant model. All of four models above are significant.

The K (BP) Statistic (Koenker's studentized Bruesch-Pagan statistic) is a test to determine if the explanatory variables in the model have a consistent relationship to the dependent variable both in geographic space and in data space. When the model is consistent in geographic space, the spatial processes represented by the explanatory variables behave the same everywhere in the study area (the processes are stationary). When the model is consistent in data space, the variation in the relationship between predicted values and each explanatory variable does not change with changes in explanatory variable magnitudes (there is no heteroscedasticity in the model). The null hypothesis for this test is that the model is stationary. For a 95% confidence level, a p-value (probability) smaller than 0.05 indicates statistically significant heteroscedasticity and/or non-stationarity. In this analysis, all models are statistically significant non-stationarity and thus, are especially good candidates for GWR analysis.

Finally, a message which recommends testing spatial autocorrelation of residuals came up for all the models and thus, Moran's I is examined to verify that.

Morans' I test

The result of Moran's I test is summarized in table 1.

Table 4 - The result of Moran's I test

Name	Freq	WFreq1	WFreq2	Acc_Cost
Default neighborhood search threshold	5269.49	5269.49	5269.49	5269.49
Moran's Index	0.045	0.043	0.052	0.013
Z Score	2.12	2.012	2.47	0.72
p-value	0.034	0.043	0.014	0.47

This result shows that when WFreq2 is set to the dependent variable, the index is the nearest to 1. Except for the Acc_Cost variable, the z-scores are greater than 1.96 and p-

values are smaller than 0.05, which indicates statistically significant at a 5% significance level meaning that there exists spatial autocorrelation. The following figure also confirms the result of spatial autocorrelation test graphically.

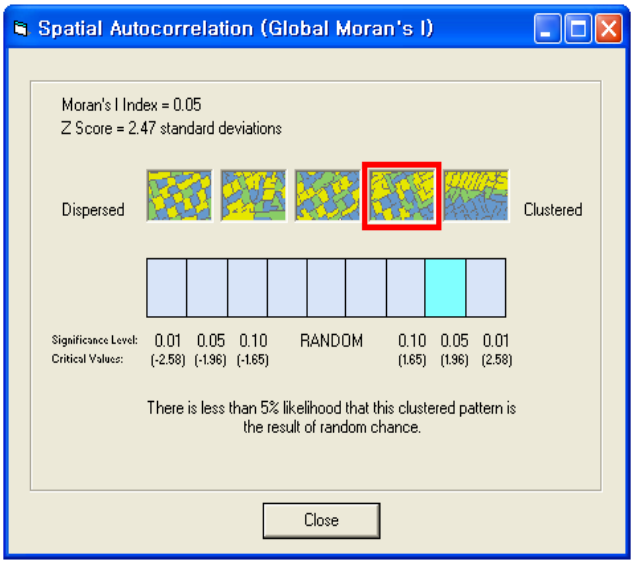


Figure 1 - The result of Moran's I test

A Comparison of results between OLR and GWR

As results from global regression and Moran's I test, the spatial interaction in crash data arose an important problem to be considered in accident analyses. Since a global regression model estimates parameters globally, it cannot be allowed for the spatial variability of the influence of explanatory variables. For this reason, GWR is performed in this study, which is able to describe different relationships between the dependent and explanatory variables at different locations in geographic space and thus is more appropriate to exploring hazardous locations.

The results are summarized in Table 5.

Table 5 - A comparison of diagnostic statistics between OLR and GWR

Name	OLR				GWR			
	Freq	Wfreq1	Wfreq2	Acc_Cost	Freq	Wfreq1	Wfreq2	Acc_Cost
AICc	2066.07	3,390.38	3,928.31	19,931.50	2,040.16	3,390.46	3,870.34	19,888.26
R^2	0.519	0.477	0.623	0.560	0.586	0.520	0.694	0.633
Adjusted R^2	0.510	0.467	0.615	0.551	0.551	0.484	0.668	0.602

Akaike Information Criterion (AIC) is a relative measure of performance used to compare models; the smaller AIC indicates the superior model. For Freq, WFreq2, and Acc_Cost variables, AICc values from GWR are smaller than those from OLR. Both R^2 and Adjusted R^2 also show that GWR model is superior to OLR model for the all dependent variables. These

results indicate that GWR model is better than OLR model, respectively and furthermore, testify that spatial analyses are essential for accident analyses.

Identification of Hazardous Locations

First of all, candidates for hazardous locations are identified at a route-wide level. After putting WFreq2 into severity weights and 5270 into a bandwidth parameter a density map is created by the 'Kernel Density Estimation' tool of ArcGIS 9.3. The value, 5270 is automatically determined through the Moran's I test, which means the analytical search radius for the examination of spatial autocorrelation and thus, proper bandwidth for the wide-level analysis.

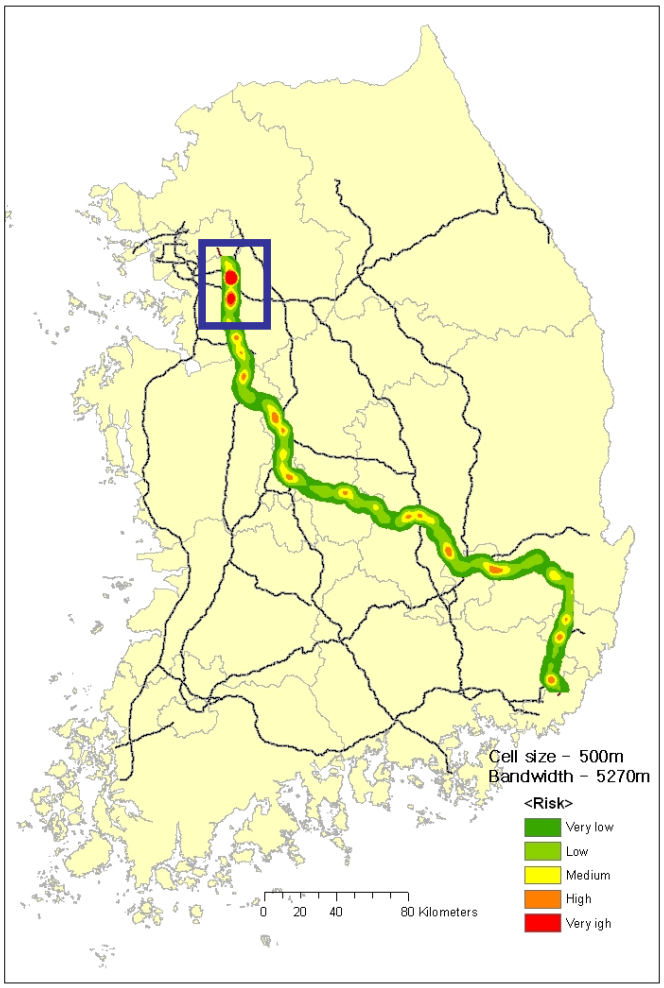


Figure 2 - a wide-level density map

Figure 2 shows a wide-level density. The color changes from green to red as the risk of locations increases. In this map, as the density is the highest in the area within a blue box, further study is performed for the area at a narrow level.

With a large bandwidth, some segments which have low crash risk can be included in hazardous area because local variations tend to be removed. This results in unnecessary dissipation of the resources such as budgets and time by including low risk segments in detailed engineering study sites for the road safety improvement projects. Therefore, a detailed investigation is conducted at a narrow level.

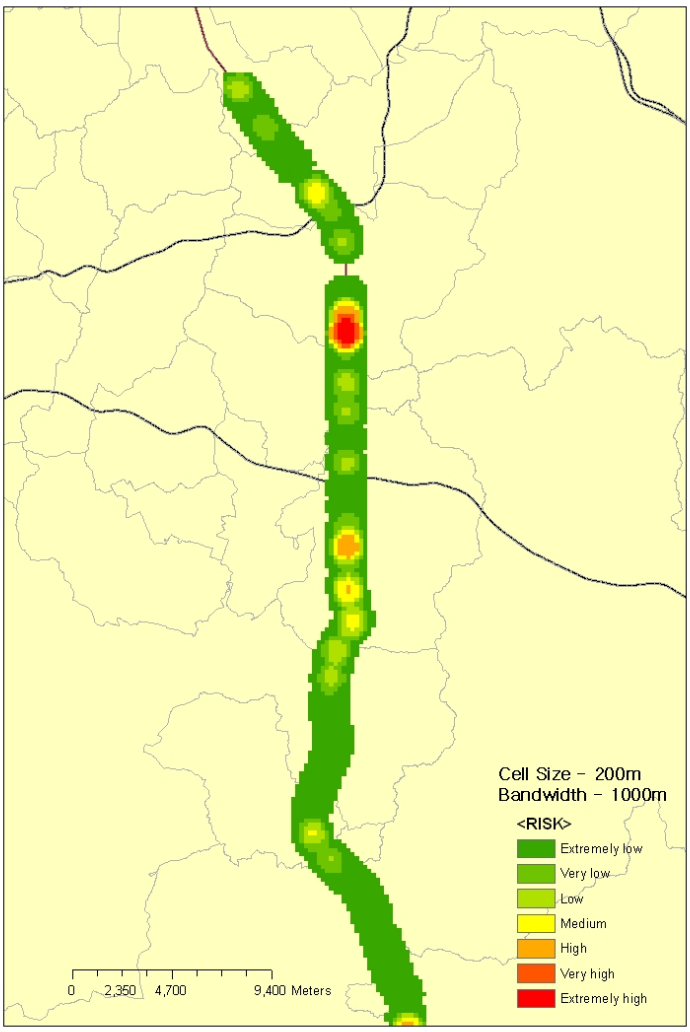
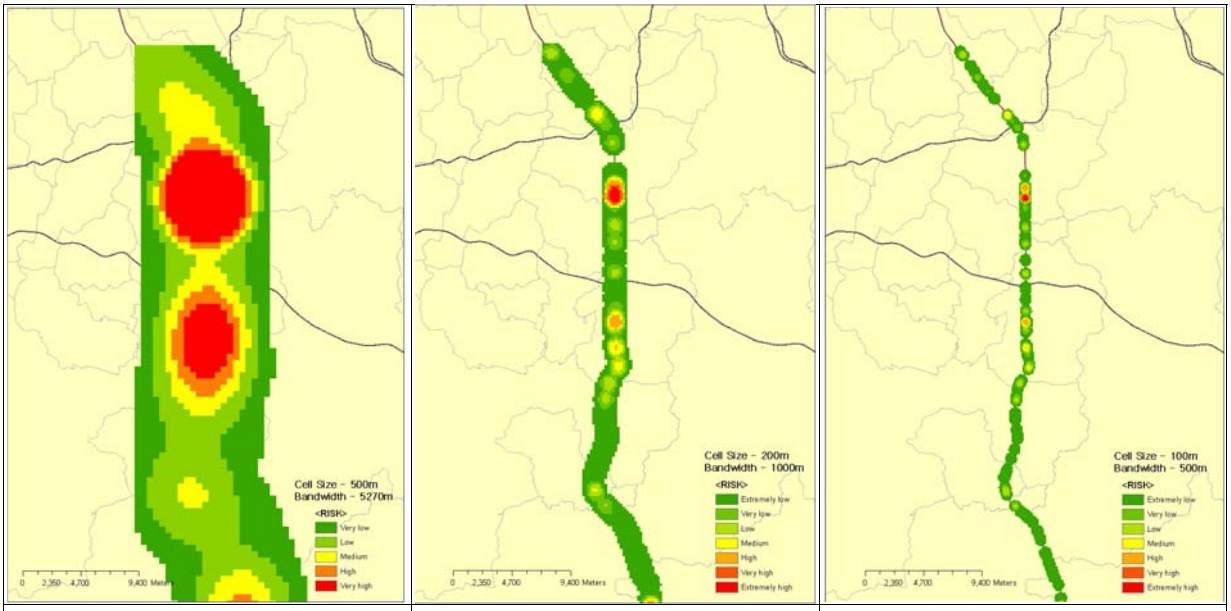


Figure 3 - a narrow-level density map

With a 200m cell size and a 1000m bandwidth, hazardous locations are appeared more clearly at this scale. While the length of high risk area at wide level is approximately 3500m, the length on this map is approximately 1200m.

At the next step, three bandwidths are compared at the same scale, including 500m, 1000m and 5270m with 100m, 200m and 500m cell sizes, respectively.



Bandwidth = 5270m	Bandwidth = 1000m	Bandwidth = 500m

Figure 4 - A comparison of the results from three bandwidths

The result for 500m bandwidth shows more distinctive hazardous locations which is shortened from 1200m to 400m. When a road improvement project is implemented on expressway, the length of a target segment usually varies from 200m to 400m. Therefore, 500m bandwidth can be the practically best option in the field.

CONCLUSIONS

In this study, Geographically Weighted Regression (GWR) and Kernel Density Estimation (KDE) are applied to perform this research. GWR is used to verify the effect of spatial dependency and spatial heterogeneity on the outbreak of traffic accidents. The role of the KDE is displaying crash-clustered area under the consideration of an appropriate bandwidth and kernel function which determine extents and severity levels of accidents. This paper aims to develop a method for identifying hazardous locations based on severity scores of highway crashes. The method developed in this paper is applied to real-world data of Korean expressways. The results imply the necessity of examining spatial dependency and spatial heterogeneity in accident analyses and exploring hazardous locations based on crash severity. Findings from this research will contribute to both of saving time and efforts spent on manual tasks and enhancing accuracy for identifying hazardous locations by practical use of a comprehensive method.

The next step will be developing a GWR based accident prediction model. Studies on estimating severity indices and determining an optimal bandwidth will also be valuable topics for road safety improvement programs.

ACKNOWLEDGEMENTS

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (No. 2009-0076811). The generous support from the Engineering Research Institute of Seoul National University and Korea Expressway Corporation are also gratefully acknowledged. The opinions expressed in the paper, however, are solely of the authors and do not necessarily reflect the opinions of the respective agencies.

REFERENCES

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

Anderson, T. K. (2009). Kernel Density Estimation and K-means Clustering to Profile Road Accident Hotspots. Accident Analysis and Prevention, 41, 359-364

Anselin, L. (1998). Exploratory Spatial Data Analysis in a Geocomputational Environment. in Longley, P. A. and Brooks, S. M. (eds.), Geocomputation: A Primer, John Wiley & Sons, Chichester, 77-94.

Charlton M. and A. S. Fotheringham (2009). Geographically Weighted Regression: a White Paper. From http://ncg.nuim.ie/ncg/GWR/GWR_WhitePaper.pdf

Chung, K., D. R. Ragland, S. Madanat, and S. M. Oh (2009). The Continuous Risk Profile Approach for the Identification of High Collision Concentration Locations on Congested Highways. Transportation and Traffic Theory 2009: GoldenJubilee, Papers selected for presentation at ISTTT 18, a peer reviewed series since 1959, William H. K. Lam, S. C. Wong and Hong K. Lo, eds., Springer, August 2009, pp. 463-480

de Smith, M. J., M. F. Goodchild, and P. A. Longley (2008). Geospatial Analysis - a Comprehensive Guide to Principles, Techniques and Software Tools, 3rd edition, Matador, Leicester

Doreian, P. (1981). Estimating Linear Models with Spatially Distributed Data. Sociological Methodology, 12, 359–388, John Wiley & Sons, Chichester.

Erdogan, S., I. Yilmaz, T. Baybura, and M. Gullu (2008) Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accident Analysis and Prevention, 40, 174-181

ESRI (2008). ArcGIS version 9.3. *Environmental Systems Research Institute (ESRI).* Redlands, California, U.S.A.

Federal Highway Administration (2002). Safety Analyst: Software Tools for Safety Management of Specific Highway Sites. White paper.

Flahaut, B., M. Mouchart, E. S. Martin, and I. Thomas (2003). The Local Spatial Autocorrelation and the Kernel Method for Identifying Black Zones: A comparative approach. Accident Analysis and Prevention, 35, 991-1004

Fotheringham, A. S., C. Brunsdon and M. Charlton, (2002) Geographically Weighted Regression- the analysis of spatially varying relationships. John Wiley & Sons Ltd, Chichester

Goodchild, M. F., R. Haining, S. Wise and 12 others (1992). Integrating GIS and spatial data analysis: problems and possibilities, International Journal of Geographical Information Science, 6(5), 407-423

Kwak, H., D. K. Kim, S. H. Park, and S. Y. Kho (2010). Development of a Safety Performance Function for Korean Expressways. submitted for the 12th World Conference on Transport Research, Lisbon, Portugal

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

Park, S., S. H. Park, D. K. Kim, and K. S. Chon. (2010). Factors that Influence the Level of Accident Severity in Vehicle Crashes: A Case Study of Accidents on Korean Expressways. submitted for the 12th World Conference on Transport Research, Lisbon, Portugal

Pulugurtha, S. S., V. K. Krishnakumar, and S. S. Nambisan (2007). New Methods to Identify and Rank High Pedestrian Crash Zones: An Illustration. Accident Analysis and Prevention, 39, 800-811

Shankar, V., F. Mannering, and W. Barfield. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention, 27(3), 371-389

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society, Ser. B 13: 238–241.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, 46(2), 234-240

Xie, Z. and J. Yan (2008). Kernel Density Estimation of Traffic Accidents in a Network Space. Computers, Environment and Urban Systems, 32, 396-406

Zhao, F. and N. Park (2004). Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic. Transportation Research Record: Journal of the Transportation Research Board, 1879, 99–107.

Zhong, L. D., X. D. Sun, Y. L. He, X. M. Zhong, and Y. S. Chen (2009). Safety Performance Function for Freeway in China. Presented at 88th Annual Meeting of the Transportation Research Board, Washington D. C.