Application of Spatial Econometrics Analysis for Traffic Accident Prediction Models in Urban Areas

Jiyeon HONG [a], Soobeom LEE [b], Joonbeom LIM [c], Jeonghyun KIM [d]

[a] *Research Professor, Dept of Transportation Engineering, University of Seoul, Seoul, Korea;E-mail:cathy56@uos.ac.kr*
[b] *Professor, Same as the first author; E-mail: cathy56@uos.ac.kr*
[c] *Ph.D Canidate, Same as the first author; E-mail :t_safety @hanmail.net*
[d] *Principal Researcher, Korea Railroad Research Institute, Uiwang, Korea; E-mail: Kimjh@krii.re.kr*

Abstract: The previous researches on the prediction of accidents frequently assumed the independence among the error terms, from the standpoint of traditional statistics. However, as spatial data including information about geographical spaces, data collected in each traffic analysis zones are not randomly distributed in space and have the spatial correlation each other. That violates the basic assumption. To control such autocorrelation, spatial econometrics analyses need to be considered. The aim of this study is to identify the spatial correlation of traffic accidents and to develop prediction models with spatial econometrics analysis in urban areas. As a result, it was found that traffic accidents revealed high spatial correlation and spatial econometrics models showed a more enhanced explanatory power than normal linear regression model. Moreover, spatial econometrics models were more excellent than it, when verifying it through SRMSE.

Keywords: Spatial Econometrics Model, Spatial Correlation, Traffic Accidents, Traffic Accident Prediction Models

1. INTRODUCTION

Traffic accidents occur at specific time and location, but include the comprehensive interaction between spatial road environments, human factors and the other influencing factors during the process.

Previous studies on accident rate prediction were based on the traditional and statistical assumptions that the variables are random and the error terms are independent. However, in general, the traffic accident data are the spatial data which were collected by the administrative zone including the information on the geographical space. These are not distributed randomly in the space but represent the spatial autocorrelation. It is then against the traditional assumption on the independency of the error terms (Doreian, 1980 & 1981). The traditional linear regression analyses can hardly control the spatial dependence of and interactions between the socio-economic phenomena. Then, the traditional models have the limitations to consider the comprehensive spatial influences in practice.

Many researchers have recognized the influences of the geographical space, i.e. roads, and the efforts have been given to consider the influences of space to the traffic accidents. After the studies of the spatial analyses for the traffic accident data by Black(1991) and Loveday(1992), the spatial patterns of traffic accidents and the autocorrelations were reviewed by N. Levine(1995), G. Lee(2004) and Y. Lee(2007). N. Levine(1995) examines spatial patterns in motor vehicle crashes for the city. The additional studies were then followed by Benoît Flahaut(2004), Mohammed A. Quddus(2008), Cha Wang(2009), and S. Park(2010).

The spatial econometrics analysis is a statistical method analyzing the relationships between the variable related spatially considering the spatial autocorrelation.

This study is to review the spatial correlation of traffic accident frequencies collected by the administrative zone, and develop the traffic accident frequency prediction model in urban area by the spatial econometrics analysis. It can then identify the characteristics of traffic accidents by regional unit, and predict the effectiveness of the countermeasures against the traffic accidents.

2. METHODOLOGY

2.1 Diagnosis of Spatial Correlation

The frequencies of traffic accidents by the administrative spatial unit and the spatial weighted matrix are necessary as the input data to review the spatial correlation of the accidents. The spatial contiguity matrix is applied for the weighted matrix, then "0" or "1" is given by the contiguity, which is derived by the spatial statistics on the ArcMap 10.0.

The spatial correlation of accident frequencies by type (Y_i) is tested according to the *Moran's I* statistic. *Moran's I* can determine whether the spatial patterns of accidents are "clustered", "dispersed" or "random". The positive value of *Moran's I* means the entities in the space have the similar values and clustered, and the negative value represents the spatial entities have the different values and dispersed. When the spatial pattern is random, the value is "0".

$$Moran's\ I = \frac{N}{\sum_i \sum_j w_{ij}^*} \frac{\sum_i \sum_j w_{ij}^* (Y_i - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \quad (1)$$

where,

Y_i : Actual number of accidents by administrative spatial unit

$\bar{Y}$: Average number of accidents by administrative unit

N : Number of administrative spatial units

i, j : Administrative spatial units

w_{ij}^*: The standardized spatial weights matrix

$$w_{ij}^* = \frac{w_{ij}}{\sum_i w_{ij}},\ w_{ii} = 0 \quad (2)$$

w_{ij} is the spatial weights matrix, and "1" or "0" is given when the unit is neighbored or separated respectively. The Rock-Based Contiguity was applied in this study. The sum of the weighted values for the spatial units related a unit was standardized by row(w_{ij}^*), and the value of "1" is given whereas the own weighted value is "0".

2.2 Modeling

2.2.1 Multiple Linear Regression Analysis

A multiple linear regression model is developed in advance of the proposed spatial econometrics models. The multi-linear regression model is the most popular to predict the traffic accident rate by region. In the process of parameter estimations, the explanatory

variables for the proposed models are determined. The linear model can be a reference to verify the performance of the proposed models.

2.2.2 Spatial Econometrics Analysis

Two kinds of the spatial econometrics models are applied to develop the traffic accident frequency prediction model in urban area such as the Spatial Autoregressive Model (SAR) and the Spatial Errors Model (SEM) both of which are a kind of global spatial regression models. The SAR includes the spatial correlation in the dependent variable as shown in the equation (3), and the SEM does the spatial correlation in the error term as in the equation (4). Each model was developed by the "Maximum Likelihood Estimation" with the "OpenGeoDa 1.2.0".

$$y = \rho W_y + X\beta + \epsilon \quad (3)$$

$$y = X\beta + \mu, \mu = \lambda W_\mu + \epsilon \quad (4)$$

where,
y : A dependent variable
W_y: A spatially lagged dependent variable for spatial weights matrix W
ρ : The scalar for spatial lag coefficient
β : The parameters to be estimated
X : The matrix of exogenous explanatory variables
μ : The error term expressing spatial dependence
λ : The spatial autoregressive coefficient

2.3 Comparison of Models

The performance of the proposed model was verified by comparing the performance of the multi-linear regression model in term of goodness of fit, spatial correlation and measuring efficiency.

At first, the goodness of fit could be compared with the "R^2". The second verification may be the test for the spatial correlation of the residuals. The spatial correlations of the residuals are tested for both of the multi-linear and proposed models. Then, the implications of the results from the models should be explained.

The comparison of the measuring efficiencies can be performed with the "Standardized Root Mean Square Error (SRMSE)" which is recommended to compare the performances of different models with the same data. It is regarded that the model is desirable when the SRMSE is less than 0.08 (JUNG W. H., 2011).

$$\mathrm{SRMSE} = \frac{\sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2}{N}}}{\frac{\sum_i Y_i}{N}} \quad (5)$$

where,
Y_i : actual number of accidents by administrative spatial unit
$\hat{Y}_i$: estimated number of accidents by administrative unit
N : number of administrative spatial units
i, j : Administrative spatial units

3. DATA DESCRIPTION

The scope of this study was limited in City of Seoul in 2010 and the spatial analysis unit was based on the administrative units. It is because all the data is collected by the administrative unit each year, and the data in 2010 was the most recent available.

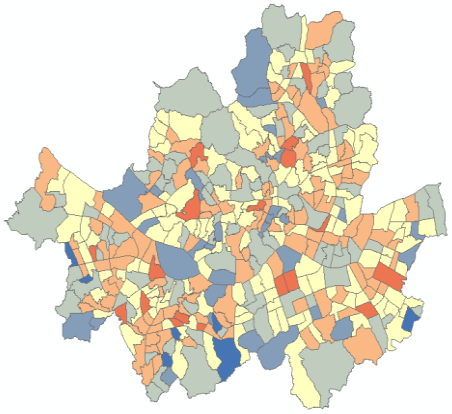

Figure 1. Administrative unit in city of Seoul

Figure 2. Points of Traffic Accident in Seoul

The input data were acquired from the police department (traffic accident statistics, number of police enforcements), the city statics (numbers of residents and households), the Seoul GIS Potal (numbers of crosswalks, speed humps and crosswalk acoustic signs, and statistics on exclusive bus lanes), Center for Seoul Metropolitan Transportation (OD by modes), and the Road Name and Address Management System (building floor area by use). The data were collected by the administrative unit and the dataset based on the GIS were established as following Table 1.

Table 1. Summary statistics

Classification		Total	Average	Max	Min
# of Accidents	Frequency	40,046	95.58	578	8
	ln(Frequency)		4.319		
Road	Total Length(㎞)	8,142	19.62	78.28	2.10
	# of Intersections (#)	5,959	14.36	92.00	-
	Intersection Density (#/km)	0.73	-	5.55	0
	Total Length of Exclusive Bus Lane (m)	190,577	459	3,713	0
	Ratio of Exclusive Bus Lane (%)	0.02	-	0.36	-
	Total Length of One-way(m)	590,245	1,422	10,885	0
	Ratio of One-way(%)	0.07	-	0.92	-
Land Use	Area(㎢)	605.22	1.46	12..69	0.23
	Developed Area(㎢)	367.55	0.89	2.87	0.23
	Total Building Floor Area (㎢)	476.06	1.15	7.78	0.0002

Table 2. Summary statistics(continue)

Classification		Total	Average	Max	Min
Socio-Economic Factors	# of Residents (#)	10,575,447	25,483	92,527	1,658
	# of Employees among Residents (#)	5,051,673	12,173	42,855	465
	# of Employees in Area(#)	4,138,416	9,972	143,055	914
	% of Registered Cars (#)	2,695,142	6,494	22,221	304
	Traffic Volume (trip)	26,756,769	64,474	465,679	8,733
Traffic Safety Facilities & Policy	# of Crosswalks (#)	425,440	1,025	15,175	17
	Ratio of Advance Crosswalk Warning Signs (%)	69,521	168	1,277	3
	Ratio of Crosswalk Acoustic Signals (%)	10,292	27	205	0
	Ratio of Remaining Time Signs on Crosswalks (%)	10,010	26	134	0
	# of School Zonee (#)	1,721	4	19	0
	# of Speed Humps (#)	27,705	67	574	0
	# of Cat's Eye Systems (#)	44,225	107	1,217	2

4. RESULTS

4.1 Analysis for Spatial Correlation of Traffic Accidents

The global spatial correlation was tested with the null hypothesis and the alternative hypothesis as shown below. It was then concluded that there is the spatial correlation in the traffic accident frequency by the administrative spatial unit (ln(acc. Frequency)), and the null hypothesis could be rejected.

H_0 : traffic accident frequency by spatial unit (ln(# of acc.)) is not spatially correlated.
H_1 : traffic accident frequency by spatial unit (ln(# of acc.)) is spatially correlated.

Table 2. Result of spatial correlation (ln(# of acc.))

Moran's I	Z-score	p-value
0.199385	6.659904	0.000000

4.2 Prediction of Traffic Accident Frequency in Urban Area

The coefficient of determination for the multi-linear regression model was 0.572, and it could be regarded to be explanatory. For the convenience of comparison between the models, the variables in the multi-linear regression model were kept in the spatial econometrics

models.

The explanatory variables could also be significant as the variables in the spatial econometrics analysis, and the ranges of the regression coefficients were similar.

According to the results from the SAR model, the elements of the spatial lag, ρ was 1.20, and the t-statistic was 4.2306 which could be significant at 1% level of significance. The spatial correlation was also verified with the likelihood ratio of 16.64279 (p=0.00005). The result from the SEM model represented the λ value of 0.2659 and the t-statistic of 4.044, which could be regarded to be significant at the significance level of 1%. The likelihood ratio was 13.3156 (p=0.00026), so the spatial correlation was verified.

Thirty two factors were reviewed for the explanatory variables such as nine socio-economic, nine road and land use, and 14 traffic safety facilitiy and policy ones. By the analyses of the correlations and the multicollinearities between the variables, five variables on road and land use and three on traffic safet facility and policy were selected. The variables on road and land use were the total length of roads, the total building floor area, the ratio of exclusive bus lane, and the numbers of intersections and crosswalks. The selected variables on traffic safety facility and policy were the ratio of advance corsswalk warning signs and the numbers of speed humps and the police enforcements.

The variables on road and land use are the exogenous variables which cannot be controlled for the traffic safety, even though those influence the traffic accident exposure directly related to the accidents positively(+). The variables on traffic safety facility and policy have the negative(-) relationship with the accident frequency, and can be controlled in terms of traffic safety improvement.

Table 3. Traffic accident prediction model by spatial econometrics analysis

Variable	Multi-Linear Regression Model		Spatial Econometrics Models			
			SAR Model		SEM Model	
	coefficient of regression	t-statistic	coefficient of regression	t-statistic	coefficient of regression	t-statistic
Constant	3.505	42.470	2.958	19.619	3.510	40.739
Total Length of Roads	0.015	5.593	0.016	6.200	0.014	5.415
Ratio of Exclusive Bus Lane (%)	4.878	8.051	4.819	8.214	4.798	7.985
Total Building Floor Area (㎢)	1.6E-7	4.746	1.4E-7	4.273	1.2E-7	3.348
# of Intersections (#)	0.030	8.353	0.030	8.689	0.031	8.789
# of Crosswalks (#)	7.7E-5	3.063	6.9E-5	2.828	7.8E-5	3.208
Ratio of Advance Crosswalk Warning Signs (%)	-0.540	-2.428	-0.411	-1.901	-0.402	-1.836
# of Speed Humps (#)	-0.002	-4.056	-0.001	-3.714	-0.001	-3.830
# of Police Enforcements	-0.153	-3.054	-0.150	-3.089	-0172	-2.979
ρ 1)			0.120	4.230		
λ 2)					0.266	4.044
R^2	0.572		0.590		0.591	

1) Spatial Autoregressive Correlation Coefficient in SAR 2) Spatial Error Coefficient in SEM

4.3 Comparison of Models

The explanatory performance of the spatial econometrics model could be higher than that of the multi-linear regression model for the urban traffic accident frequency prediction.

When the spatial correlations of the residuals by the model estimations were investigated, it could be found on the multi-linear regression model. The Moran's I of the residuals by the spatial econometrics models represented the value close to "0", which implies that the residuals are dispersed randomly on the space and there is no spatial correlation. It means that the spatial econometrics models can include the influence of the spatial correlation for the urban traffic accident frequency prediction, which is excluded in the multi-linear regression model.

The standardized rood mean square error (SRMSE) represents the average error ration between the actual(surveyed) values and the estimations from a model. The SRMSE's of the SAR, the SEM and the multi-linear regression model were in the similar range, but that of the SAR is slightly closer to "0" than those of the other models.

Table 4. Comparative analysis between models

Classification	R2	Moran's I	z-score	p-value	SRMSE
Multi-Linear Regression Model	0.574	0.127088	4.274094	0.000019	0.1110
SAR Model	0.590	0.013814	0.535484	0.592315	0.1103
SEM Model	0.591	-0.016786	-0.474289	0.635294	0.1113

5. CONCLUSIONS

The multiple linear regression analysis is the most popular methodology to predict the traffic accident frequency in urban area, but it is hard to consider the spatial characteristics, and results in undesirable goodness of fit in the term. The accident frequency prediction models have been developed to reduce the traffic accident frequency by identifying the characteristics of accidents. This study applied a spatial econometrical method which can improve the goodness of fit of the model by considering the spatial characteristics of traffic accidents.

In advance of the model development, the spatial correlation of the traffic accidents was verified by analyzing the accident data by the administrative spatial unit. The proposed model based on the theory considering the spatial correlation showed the improved goodness of fit and prediction performances.

Conclusively, this study may suggest a new methodology in traffic accident frequency prediction, so the spatial correlation analysis could provide various future works in the field of traffic safety.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2041296).

REFERENCES

Doreian, P. (1981), Estimating Linear Models with Spatially Distributed Data, Sociological Methods & Research, August 1980, 29-60.

Black, W. R. (1992), Highway accidents: A Spatial and Temporal Analysis, Transportation Research Record, Vol.1318, 75–82.

Ned Lebine (1995), Spatial Analysis of Honolulu Motor Vehicle Crashes, Accident Analysis and Prevention, Vol. 27 No. 5, 663-674

Lee, K. H. (2004), A Study on Spatial Patterns of Traffic Accidents using GIS and Spatial Data Mining Methods: A Case Study of Kangnam-gu, Seoul, Journal of the Korean Geographical Society, Vol.39 No.3, 457-572.

Lee, Y. W. (2007), Spatiotemporal Hotspot Detection Using G Statitics – A Case of Traffic Accidents in East Japan, Seoul Studies, Vol. 8 No.3, 71-83.

Park, S. H. (2010), A Spatial Analysis Method for Identifying Hazardous Locations on Expressway, Seoul National University, Ph.D. dissertation.

Flahaut, B. (2004), Impact of Infrastructure and Local Environment on Road Un-safety Logistic, Accident Analysis and Prevention, Vol.36 No.6, 1055-1066

Quddus, M. A. (2008), Modeling Area-Wide Count Outcomes with Spatial Correlation and Heterogeneity: An analysis of London Crash Data, Accident Analysis & Prevention, Vol.40 No.4, 1486–1497.