

A Study on Risk Evaluation of Crime in the Seoul Metropolitan Area based on Poisson Regression Model

HagYeol Kim¹ · HyeKyung Yu² · Man Sik Park³ · Tae-Young Heo⁴

¹Department of Urban Planning and Engineering, Seokyeong University

²Department of Information & Statistics, Chungbuk National University

³Department of Statistics, Department of Statistics, Sung Shin Women's University

⁴Department of Information & Statistics, Chungbuk National University

(Received July 1, 2012; Revised July 6, 2012; Accepted July 18, 2012)

Abstract

In this study, we identify the variables that affect the number of crime and spatial correlation in the Seoul metropolitan area, in addition, we measure the relative risk on the incidence of crime by a Poisson regression model. We suggest a statistical methodology to make a risk map for crime based on relative risk instead of the total event of crime by region using the Geographic Information System. To demonstrate the use and advantages of this methodology, this study presents an analyses of the total crime count in 25 wards in the Seoul metropolitan area.

Keywords: Poisson regression model, crime risk, Bayesian, geographic information system.

1. 서론

사회의 경제적 발전과 급속한 도시화에 따라 우리나라의 범죄발생 및 유형은 갈수록 다양화, 흉폭화, 연소화 등의 심각한 문제를 야기하고 있으며, 주거환경에 따라 범죄의 형태도 다르게 나타나는 것으로 알려져 있다. 구체적으로 지역 내 유흥주점 비율과 인구의 불안정성이 높아지면 5대 범죄의 발생 건수도 증가하는 것으로 나타났다. 특히 2006년부터 2008년 8월까지 서울지역의 범죄발생건수는 강남구가 최대이며, 인구 1만 명당 비율로 환산하면 중구가 가장 높은 것으로 나타났다. 이는 유동인구와 유흥주점의 비율이 높다는 것이 원인으로 나타났다.

최근 범죄발생에 대한 예방을 위해 지리정보시스템 기술을 이용한 범죄에 대한 공간분석(Spatial data analysis)에 대한 연구가 활발하게 이루어지고 있다. 공간분석은 일반적으로 특정 위치 또는 특정 지역에서의 관측값이나 측정값으로 구성된 공간자료의 통계적 분석 방법을 의미한다. 공간상관성의 정보가 기존의 회귀모형에서 배제된다면 추정된 회귀계수는 효율적이지 못한 추정량이 되어 회귀계수의 통계적 유의성 검증 및 예측오차 등의 정확성을 신뢰할 수 없게 된다 (Griffith, 1996).

따라서 범죄발생에 영향을 미칠 것으로 판단되는 지역의 공간적 상관성을 반영한 모형을 통해 범죄자료에 대한 연구가 활발하게 진행되고 있으며, 공간상관성을 반영한 범죄모형의 연구는 다음과 같다. Lee

⁴Corresponding author: Assistant Professor, Department of Information & Statistics, Chungbuk National University, Cheong-Ju, Chungbuk 361-763, Korea. E-mail: theo@cbnu.ac.kr

(2004)는 다양한 공간계량모형을 이용하여 종속변인으로 인구 10만 명당 총 발생 건수 및 5대 범죄에 대하여 정규성을 가정하고 각각의 범죄유형별로 모형을 적합시켰다. Yoon (2004)는 범죄사건의 크기에 따라 0 또는 1의 이진수의 값으로 구분하여 공간로지스틱 모형(Spatial logistic model)을 구축하고 각 변수들의 영향력을 파악하였다. Park (2003)는 서울시 성북구의 범죄자료를 통해 공간적 패턴과 범죄 발생에 대한 Hot spot의 공간적 분포에 대해 연구하였으며, 범죄에 영향을 주는 요인들의 가중치를 통해 범죄예측지도를 구축하였다. Lee와 Cho (2006)는 서울시를 대상으로 범죄유발 요인을 공간적 특성을 중심으로 분석하여 범죄예방에 관한 도시계획적 함의를 제시하였다. 이 연구에서는 공간적 특성을 반영한 다양한 공간계량경제기법을 적용하였으며, 서울시 범죄자료의 경우 공간적 종속성과 공간적 이질성이 존재하는 것으로 나타났다. 특히, 유동인구수와 유흥업소수가 범죄발생과 가장 밀접한 관련성이 있다고 제시하였다. 그러나 이 연구에서는 범죄발생에 영향을 주면 변수의 확인과 함께 공간적 종속성과 이질성을 대한 연구를 주로 다루었으며 본 연구에서 제시한 지역별 범죄발생에 대한 지역별 상대위험도에 대한 연구는 제시되지 못하였다.

Kakamu 등 (2007)은 일본 내에서의 범죄발생에 대해 시간과 공간적 연관관계를 반영한 시공간 모형을 통해 베이지안(Bayesian)기법을 이용하여 추정하였다. Shin (2005)과 Jeong 등 (2009)은 도시범죄에 대해 공간적 패턴 분석을 통해 시공간 Hot spot 지역의 탐색방법론에 대해 연구하였으며, Choi 등 (2010)은 서울시의 각 자치구의 범죄자료에 대해 지역 간의 공간상관성을 고려하여 범죄발생자료에 대한 공간가산자료모형(Spatial count data model)을 제안하였다.

Kim과 Park (2009)은 도시공간정보 기반의 범죄발생 확률 모형 및 위험도에 대한 확률지도를 생성하는 연구를 진행하여 도시공간정보에 가중치를 적용하여 위험도를 계산하고 위험도에 대한 확률지도를 구축하였으나 지역적 환경 요인을 반영하지 못하였다는 아쉬움을 가지고 있다.

따라서 본 연구에서는 범죄발생에 밀접한 연관이 있을 것으로 판단되는 요인으로 재산세, 인구밀도, 유동인구, 청소년비율, 고학력비율, 개발제한구역비율, 주택연상비율, 숙박연상비율을 설정하였으며, 서울시 행정구별 범죄발생건수 자료와 같은 공간가산자료는 공간적 연속성과 연계성을 가지고 있을 것으로 판단되어 범죄를 발생시키는 주요 요인들뿐만 아니라 지역적으로 밀접한 행정구역에 따른 공간 영향력을 조건부자기회귀모형(Conditional Autoregressive Model; CAR)을 이용하여 범죄모형에 포함시켜 총 범죄건수에 대하여 지역적 연관성을 반영하였다. 또한 각 지역별 범죄발생에 대한 상대위험도(Relative risk)를 제시하기 위해 범죄발생 건수가 정규분포를 기반으로 하는 회귀모형이 아닌 가산자료를 기반으로 하는 포아송회귀모형을 베이지안기법을 통해 각 지역별 상대위험도를 추정하고 지리정보 시스템을 통해 표출하였다.

본 논문의 구성은 다음과 같이 2장에서는 가산자료모형 중 포아송모형에 대하여 간략하게 설명하였으며, 3장에서는 공간상관성을 반영한 포아송모형, 조건부 자기회귀모형 그리고 상대위험도에 대해 간략히 소개하였으며, 4장에서는 서울시 행정구별 총 범죄에 대한 실증예제 및 분석결과를 제시하고, 5장에서는 결론으로 맺음 하였다.

2. 포아송모형

본 연구에서 사용된 반응변수는 서울시의 각 행정구별 범죄건수이며, 반응변수인 범죄건수에 영향을 미치는 요인들로 간주되는 설명변수로는 각 행정구의 환경 및 인구변수가 사용되었다.

본 연구에서 사용된 범죄발생건수는 항상 양의 상수이므로 포아송(Poisson) 분포를 따른다는 점을 이용하여 일반화선형모형 중에서 포아송모형(Poisson model)을 이용하였으며, 본 연구에서 고려한 포아송

모형의 일반식은 식 (2.1)과 같다.

$$P(Y_i = y_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}. \quad (2.1)$$

여기서 $P(Y_i = y_i)$ 는 지역 i 에서의 범죄발생건수(Y_i)가 y_i 번 발생할 확률을 의미하며, λ_i 는 지역 i 에서 평균 범죄발생건수를 의미한다. 식 (2.1)에서의 λ_i 는 식 (2.2)와 같이 표현할 수 있다.

$$\begin{aligned} \log[E(Y_i|x_i)] &= \log(\lambda_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \\ E(Y_i|x_i) &= \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = \lambda_i. \end{aligned} \quad (2.2)$$

여기서 Y_i 와 $E(Y_i|x_i)$ 는 각각 지역 i 에서 관측된 범죄발생건수와, 지역 i 에서 기대되는 평균 범죄발생건수를 나타내고, x_i 는 범죄발생건수를 결정하는 지역 i ($i = 1, 2, \dots, n$)구간에서의 지역의 공간적 속성변수 및 환경변수를 의미하며, β_i 는 각 설명변수의 회귀계수를 의미한다.

3. 공간연관성을 고려한 포아송회귀모형

3.1. 공간 포아송회귀모형

d -차원의 유클리드 공간(R^d) 상의 특정 공간 D 내에서 특정한 자료(\mathbf{y})를 관측한다고 가정하자. 그리고 이 자료들은 D 내의 여러 측정지역에서 관측되게 되는데 이 지점들을 $\{s_1, s_2, \dots, s_n\}$ 라고 표현한다면 다음과 같이 표현할 수 있다.

$$\{Y(s_i) : s_i \in D \subset R^d, i = 1, 2, \dots, n\}.$$

여기서, $\mathbf{y} = (Y(s_1), Y(s_2), \dots, Y(s_n))'$ 를 다변량 확률장(multivariate random field)라고 정의한다. 그리고 집합 D 의 속성에 따라 $\{Y(s_i) \in D\}$ 는 지리통계자료(geostatistical data), 지역자료(lattice data), 공간 점 패턴자료(spatial point pattern data)로 구분되며 이 논문에서는 지역자료를 이용한 공간범죄모형을 고려하고자 한다. 지역자료가 되기 위한 기본적인 조건들은 모든 $i \neq j$ 에 대하여 다음과 같다.

$$\bigcup_{i=1}^n s_i = D, \quad s_i \cap s_j = \phi.$$

지역자료에서 측정된 관측값을 일반화선형모형에 적용한다면 다음과 같은 모형식으로 표현할 수 있다.

$$g(\mathbf{y}(s_i)) = \mathbf{x}'(s_i)\beta.$$

여기서 $\mathbf{x}'(s_i)$ 는 지역 s_i 에서 관측된, $(1 \times p)$ 크기의 독립변수이고, β 는 성공의 확률에 대한 독립변수들의 효과의 크기를 의미하며 $(p \times 1)$ 크기의 열벡터(column vector)이다. 또한 $g(\cdot)$ 는 랜덤성분(random component)인 \mathbf{y} 와 독립변수들의 선형결합인 체계적 성분(systematic component)과 선형의 관계를 형성하는 함수이다.

그러나 공간자료인 경우, 인접한 지역들(이웃)은 서로 유사한 양상을 보일 수 있고, 이러한 연관성은 오차항의 전통적인 가정 중 하나인 독립성을 위배하게 된다. 따라서 이 연구에서 분석하고자 하는 서울시 25개 구에서 측정된 자료의 경우, 이웃정보의 상호연관성을 모형에 포함시키는 것이 필요하며 이를 위해 오차에 대한 추가적인 모형화가 요구된다. 이러한 오차에 대한 추가적인 공간적 연관성을 포함한 모형식은 다음과 같이 표현된다.

$$g(\mathbf{y}(s_i)) = \mathbf{x}'(s_i)\beta + \epsilon(s_i). \quad (3.1)$$

여기서 $\epsilon(s_i)$ 는 지역 s_i 에서의 오차(error)를 의미하며, 오차항은 동일한 분포를 따르나 다른 지역들과의 공간적 상호연관성을 내포할 수 있으므로 일반적인 오차항의 독립성을 담보할 수 없다. 따라서 오차 벡터 $\epsilon = (\epsilon(s_1), \epsilon(s_2), \dots, \epsilon(s_n))' \sim (\mathbf{0}, \Sigma)$ 이고 이 때 분산-공분산행렬 Σ 는 공간상관성의 모수들을 포함하게 된다. 여기서 $\mathbf{0} = (0, 0, \dots, 0)'$ 이다. 식 (3.1)을 추정하기 위해 조건부자기회귀모형을 통해 확률효과를 소개하고자 한다. 공간적 상호연관성을 오차의 속성에 포함시키기 위해 일반적으로 사용되는 모형은 조건부자기회귀모형이다. 특정한 공간 D 내의 s_1, s_2, \dots, s_n 에서 얻게 되는 자료 $\{\epsilon(s_i)\}$ 가 정규 확률과정(Gaussian stochastic process)을 따른다고 가정하자. 이러한 가정 하에서 $\{\epsilon(s_i)\}$ 을 모든 s_i 에 대하여 다음 식 (3.2)와 같은 조건부자기회귀모형으로 나타낼 수 있다.

$$\epsilon(s_i) | \{\epsilon(s_{-i})\} \sim N \left(\sum_{k=1}^n \omega_{ik} \epsilon(s_k), \tau_{ii}^2 \right). \quad (3.2)$$

여기서 $\{\epsilon(s_{-i})\} = \{\epsilon(s_k); \forall k \neq i\}$ 이고, τ_{ii}^2 는 s_i 지역(구)에서의 조건부분산(conditional variance)을 나타내며, 이웃정보를 포함하는 ω_{ik} 는 0 또는 1의 값을 갖는다. 즉, s_i 와 s_k 가 이웃하게 되면 $\omega_{ik} = 1$, 그렇지 않으면 $\omega_{ik} = 0$ 이다. $\omega_{i+} = \sum_{k=1}^n \omega_{ik}$ 라고 하면 조건부 분산은 $\tau_{ii}^2 = \sigma_c^2 / \omega_{i+}$ 로 표현되며, 이 때 σ_c^2 은 공통분산을 의미한다. $(n \times n)$ 의 크기인 이웃행렬, $\Omega = \{\omega_{ij}\}$ 와 조건부 분산을 대각원소로 갖는 $(n \times n)$ 의 크기인 분산행렬 $T = \{\tau_{ii}^2\}$ 을 이용하여 식 (3.2)를 다음과 같이 오차벡터 ϵ 에 대한 결합확률분포(joint probability distribution)로 표현할 수 있다.

$$\epsilon \sim N_n(\mathbf{0}, (I_n - \Omega)^{-1}T). \quad (3.3)$$

여기서 I_n 은 $(n \times n)$ 크기인 단위행렬(identity matrix)이며 N_n 은 n 차원 다변량 정규분포를 의미한다(Besag, 1974). 식 (3.3)에서 제시된 모형을 확장하여 공간적 상호연관성을 나타내는 모수를 포함한 조건부자기회귀모형은 아래 식 (3.4)와 같이 표현할 수 있다.

$$\epsilon \sim N_n(\mathbf{0}, (I_n - \rho\Omega)^{-1}T). \quad (3.4)$$

여기서 ρ 는 공간적 연관성을 의미하는 모수이며, 식 (3.4)를 $\epsilon(s_i) \sim \text{CAR}(\rho)$ 로 표현한다. 또한 식 (3.4)에서 $\rho = 1$ 이면, 식 (3.3)과 같다.

3.2. 상대위험도

지역별 발생하는 범죄발생에 대한 관심사로 범죄모형을 통해 종속변수인 범죄발생 건수에 주요한 영향력을 미치는 설명변수의 계수 추정도 중요한 의미를 갖지만, 각 행정구별 범죄발생에 대한 범죄 위험도를 파악하는 것도 매우 중요한 것으로 판단된다. 본 연구에서는 각 행정구별 범죄의 위험도를 나타내기 위해 범죄발생 건수에 영향을 주는 설명변수를 파악하고, 이를 이용하여 설명변수의 영향력을 고려한 각 행정구별 범죄발생의 위험의 정도를 상대위험도로 나타내고자 한다.

본 장에서 소개하는 상대 위험도는 포아송모형을 통해 계산될 수 있으며, 지역별 상대위험도를 계산함으로써 범죄발생에 대한 위험지역을 평가할 수 있다. 이러한 통계적 모형을 통한 범죄발생에 대한 위험지역의 평가는 범죄예방을 위해 유용하게 사용될 수 있다.

본 연구에서는 포아송회귀모형을 이용하여 각 행정구별 상대범죄위험도를 계산하고 그 패턴을 확인한다. i 번째 행정구에서 범죄발생건수를 Y_i 라 하고, 기대 범죄건수를 E_i 라 하면, 다음 식 (3.5)와 같은 포아송모형을 가정할 수 있다.

$$Y_i \sim \text{Poi}(\theta_i E_i). \quad (3.5)$$

여기서 θ_i 는 i 번째 행정구에서 범죄가 발생할 위험의 정도로 정의할 수 있으며, 다음과 같이 설명변수의 함수로서 표현할 수 있다.

$$\text{Model 1(without spatial effects)} : \log(\theta_i) = \beta_0 \sum_{j=1}^k \beta_k x_{ij}, \quad (3.6)$$

$$\text{Model 2(with spatial effects)} : \log(\theta_i) = \beta_0 \sum_{j=1}^k \beta_k x_{ij} + \epsilon_i. \quad (3.7)$$

식 (3.6)은 기존의 공간상관성을 반영하지 않은 포아송모형이며, 식 (3.7)은 공간상관성을 반영한 포아송모형으로 ϵ_i 가 3.2절의 공간상관성을 나타내는 조건부자기회귀모형을 따르는 확률효과를 의미한다.

또한 기대범죄발생건수 E_i 는 노출(Exposure)의 값에 따라 보정이 필요하며 다음과 같은 식을 이용하여 보정하여 기대범죄발생건수를 계산할 수 있으며, 이 값이 offset으로 활용될 수 있다.

$$E_i = \frac{\ln(P_i)}{\sum_{j=1}^n \ln(P_j)} \times C. \quad (3.8)$$

여기서 E_i 는 i 번째 행정구의 기대범죄발생건수, C 는 관측된 전체 행정구의 범죄발생건수, P_i 는 i 번째 행정구의 인구수를 나타내며 자연로그를 취하였다. 따라서 위와 같이 각 지역별로 인구수가 다를 경우 인구수 별로 범죄발생건수가 보정이 되고, 각 행정구별 범죄발생에 대한 상대위험도를 계산할 수 있다는 장점을 가지고 있다. 인구수와 같은 변인을 노출계수라고 하며 범죄발생건수 보정을 위해 해당 행정구의 인구 수 뿐만 아니라 면적 등을 이용할 수 있다.

또한 평균적으로 예측된 범죄발생건수가 기대범죄발생건수 보다 크다면 상대위험도가 증가하는 양상을 보이고, 평균적으로 예측된 범죄발생건수가 기대범죄발생건수 보다 작다면 상대위험도가 감소하는 양상을 보인다. 따라서 각 행정구별로 상대위험도를 파악할 수 있는 모형을 개발하는 것이 타당하며, 상대위험도가 높은 행정구에 대해 범죄발생 유발요인들을 역으로 파악하면 유발요인들에 대한 범죄발생에 미치는 영향을 파악할 수 있다는 장점을 가지고 있다.

4. 자료 및 분석결과

4.1. 자료

본 연구에서는 서울시의 자치구별 범죄발생과 관련된 설명변수와의 인과관계를 모형화 하고 분석하기 위하여 서울시에서 제공한 범죄, 사회 경제학적 변수 및 도시계획 변수 자료를 통합하여 분석에 필요한 단일 자료를 구축하였다. 서울시의 자치구별 범죄발생에 대한 자료는 2007년 서울시의 자치구별 5대 범죄(살인, 강도, 강간, 폭력, 절도)에 대한 총 범죄건수를 이용하여 공간상관성을 반영한 포아송회귀모형과 공간상관성을 반영하지 않은 포아송회귀모형을 제안하고 그 결과를 비교하였다. 본 연구 사용된 범죄발생과 관련하여 선택된 설명변수는 서울시의 25개 행정구별 재산세, 인구밀도, 유동인구, 청소년비율, 고학력비율, 개발제한구역비율, 주택연상비율, 숙박연상비율이며, 이를 구별 자료로 재구성하여 단일 자료를 구축하였다(서울시, 통계청). 종속변수인 범죄발생건수는 2007년도 발생한 5대 강력범죄(살인, 강도, 강간, 절도, 폭행)건수로 서울시에서 제공하고 있는 자료를 사용하였다.

4.2. 베이지안 추정방법

본 연구에서 범죄모형을 구축하여 서울시 행정구별 범죄발생에 대한 지역별 상대위험도를 계산하기 위하여 포아송회귀모형을 제안했으며, 포아송회귀모형의 모수 추정을 위해 베이지안 방법을 이용하였다.

포아송 회귀모형의 베이시안 방법을 통한 모수추정을 위해 다음과 같이 각 모수에 대한 사전분포(prior distribution)를 정의하였다 (Hughes-Oliver 등, 2009).

$$\begin{aligned}\beta_j &\sim N(0, 10^6), \quad j = 0, 1, 2, \dots, k, \\ \rho &\sim \text{Unif}(-1, 1), \\ \sigma_c^2 &\sim \text{IG}(0.5, 0.0005).\end{aligned}$$

여기서 IG는 역감마(Inverse Gamma distribution)분포를 의미한다. 각 모수 $\beta = (\beta_0, \beta_1, \dots, \beta_j)'$, ρ 와 σ_c^2 간에는 상호 독립임을 가정한 경우 모수 벡터 $\delta = (\beta, \rho, \sigma_c^2)$ 에 대한 사후분포(posterior distribution)는 다음과 같이 정의 된다.

$$f(\delta|\mathbf{Y}) \propto f(\mathbf{Y}|\delta) \times f(\delta).$$

모수벡터 δ 에 대한 추론은 사후분포를 통해 이루어지며, 포아송모형의 모수를 추정하기 위해 마코프 연쇄 몬테카를로(Markov Chain Monte Carlo; MCMC)방법을 사용하였다. 모수추정결과는 세 개의 초기값(initial value)을 활용하여 20000번 반복 후 처음 3000번까지 제거한 나머지 결과값으로 구체적인 모수 추정값은 결과에 제시하였다 (Gelfand와 Smith, 1990; Gelman과 Rubin, 1992; Tanner와 Wong, 1987).

4.3. 모형의 평가

지역별로 발생한 범죄발생의 상대위험도 계산을 위해 제시된 공간상관성을 반영한 포아송회귀모형의 성능평가를 위해 공간상관성을 반영하지 않은 포아송회귀모형의 결과와 비교하였다. 범죄에 대한 지역별 상대위험도 계산을 위해 제시된 모형의 적합성을 판단하기 위한 기준으로 베이시안적 모형 비교 기준인 DIC(Deviance Information Criterion; Spiegelhalter 등, 2002)를 이용하였다. 베이시안적 모형 비교 기준인 DIC 값이 작을수록 좋은 모형이라고 판단한다. DIC는 자료에 대한 모형의 적합도(goodness of fit)와 모형의 복잡도(complexity)를 동시에 반영한 측도로서 일반적인 적합도 기준인 AIC(Akaike Information Criteria)와 비슷한 측도이다. DIC는 이탈도(deviance)함수의 사후분포를 기반으로 계산되며, 이탈도는 다음과 같이 정의된다.

$$D(\delta) = -2\log[L(y|\delta)].$$

여기서 $L(y|\delta)$ 는 모수 δ 가 주어졌을 때 자료의 우도함수이다. DIC는 자료의 적합도를 나타내는 이탈도 $D(\delta)$ 와 모형의 복잡도를 나타내는 p_D 로 구성된다.

$$\text{DIC} = \bar{D} + p_D.$$

위의 식에서 이탈도의 사후평균으로 정의되는 \bar{D} 와 모형에서의 유효한 모수의 수를 의미하는 p_D 는 각각 다음과 같이 정의된다.

$$\text{DIC} = E_{\delta|y}[D(\delta)], \quad p_D = E_{\delta|y}[D(\delta)] - D[E_{\delta|y}(\delta)].$$

Table 4.1은 베이시안 관점에서의 모형 비교의 측도를 나타내고 있다.

Table 4.1에서 DIC 값으로 종합하여 모형을 비교한 결과 공간상관성을 반영한 포아송회귀모형(Model 2)이 공간상관성을 반영하지 않은 모형(Model 1)보다 적합함을 알 수 있다. 이는 자료의 공간적 연관성을 반영할 수 있는 모형이 기존의 포아송회귀모형보다 지역적 연관성을 가진 범죄발생건수 자료를 더 잘 설명하고 있음을 알 수 있다.

Table 4.1. Model comparison using Bayesian criteria

Model	\bar{D}	p_D	DIC
Model 1	2296.44	8.90	2305.34
Model 2	279.51	24.93	304.44

Model 1: Poisson regression model without spatial effects

Model 2: Poisson regression model with spatial effects

Table 4.2. Parameter estimation of Poisson regression model without spatial effects

Parameter	Estimate	Standard Deviation	2.5%	Median	97.5%
Intercept	0.0278	0.0031	0.0218	0.0278	0.0339
Property taxes	0.1440	0.0087	0.1270	0.1440	0.1610
Population density	-0.0919	0.0066	-0.1048	-0.0919	-0.0789
Floating population	-0.0205	0.0063	-0.0329	-0.0205	-0.0082
Proportion of youth	0.0361	0.0045	0.0273	0.0362	0.0450
PHE*	-0.1684	0.0079	-0.1839	-0.1684	-0.1527
PDRA**	-0.0534	0.0052	-0.0635	-0.0534	-0.0432
Proportion of housing	-0.1426	0.0064	-0.1551	-0.1426	-0.1302
Proportion of accommodation	0.0468	0.0071	0.0328	0.0467	0.0607

*PHE: Proportion of highly educated

**PDRA: Proportion of development restricted area

4.4. 결과

서울시 지역별 범죄발생에 대한 상대 위험지역을 확인하기 위해 식 (3.6)과 식 (3.7)에서 제시한 공간상관성을 고려하지 않은 포아송모형과 공간상관성을 고려한 포아송모형을 통해 지역별 상대위험도를 계산하였다. 다음의 Table 4.2는 식 (3.6)의 공간상관성을 고려하지 않은 포아송회귀모형의 결과로 총 범죄수와 관련이 있다고 판단되는 9가지 변수 중에 재산세(Property taxes), 청소년비율(Proportion of youth), 숙박연상비율(Proportion of accommodation)은 총 범죄수와 정(+)의 관계가 있다고 확인되었으며, 나머지 변수는 부(-)의 관계가 있다고 판단되었으며, 자세한 결과값은 Table 4.2에 제시되어 있다.

Table 4.3에 제시한 상대위험도를 살펴보면 2007년 총 범죄자료를 기준으로 중구(Junggu), 종로구(Jongrogu), 금천구(Keumchengu), 용산구(Yongsangu), 강남구(Kangnamgu) 등이 위험지역으로 나타난 반면 도봉구(Dobonggu), 양천구(Yangcheongu) 등은 상대적으로 덜 위험한 지역으로 나타났다. 여기서 ○는 관측된 위험도가 기대위험도보다 통계적으로 높은 지역, △는 관측된 위험도와 기대위험도가 통계적으로 유사한 지역, ×는 관측된 위험도가 기대위험도보다 통계적으로 낮은 지역을 의미한다. 따라서 ○ 지역은 총 범죄발생에 있어 매우 높은 위험도를 지닌 행정구라 판단할 수 있으며, 향후 총 범죄수를 억제하기 위해 ○ 지역에 대한 다양한 정책이 필요하다.

Table 4.4에서는 공간상관성을 반영한 모형을 통한 총 범죄에 영향을 주는 변수들을 확인하였다. 확인 결과 공간상관성을 반영하지 않은 모형과 마찬가지로 재산세(Property taxes), 청소년비율(Proportion of youth), 숙박연상비율(Proportion of accommodation)은 총 범죄수에 정의 관계이며, 나머지 변수는 부의 관계가 있음을 확인하였다. 또한 공간상관성을 나타내는 모수는 음의 값을 가지고 있지만 통계적으로 유의하지 않아 공간상관성이 나타나지 않음을 알 수 있다. 따라서 2007년의 서울시 각 행정구별 범죄발생에 대한 공간적 패턴을 존재하지 않음을 알 수 있다. 하지만 이러한 공간적 패턴은 해당 년도의

Table 4.3. Relative risk estimation of Poisson regression model without spatial effects

District	Risk Estimate	Standard Deviation	2.5%	Median	97.5%	Type
Jongrogu	1.9050	0.0250	1.8560	1.9050	1.9550	○
Junggu	3.2800	0.0441	3.1930	3.2790	3.3670	○
Yongsangu	1.3090	0.0155	1.2780	1.3080	1.3390	○
Seongdonggu	0.8137	0.0092	0.7957	0.8136	0.8320	×
Kwangjingu	0.9611	0.0083	0.9449	0.9611	0.9774	×
Dongdaemungu	1.1200	0.0100	1.1000	1.1200	1.1390	○
Jungryangu	0.8733	0.0075	0.8586	0.8732	0.8881	×
Seongbukgu	0.9843	0.0067	0.9711	0.9843	0.9975	×
Kangbukgu	1.0230	0.0091	1.0060	1.0230	1.0410	○
Dobonggu	0.6731	0.0067	0.6601	0.6731	0.6862	×
Nowongu	0.7656	0.0046	0.7567	0.7656	0.7747	×
Eunpyeungu	0.8019	0.0068	0.7886	0.8019	0.8152	×
Seodaemungu	0.9900	0.0075	0.9752	0.9900	1.0050	△
Mapogu	1.0050	0.0078	0.9893	1.0050	1.0200	△
Yangcheongu	0.7426	0.0074	0.7283	0.7425	0.7574	×
Kangseogu	1.0270	0.0069	1.0140	1.0270	1.0410	○
Gurogu	0.9525	0.0065	0.9398	0.9526	0.9651	×
Keumchengu	1.3460	0.0107	1.3250	1.3460	1.3670	○
Youngdeungpogu	1.1250	0.0107	1.1040	1.1240	1.1450	○
Dongjakgu	0.7398	0.0059	0.7286	0.7398	0.7515	×
Kwanakgu	0.8584	0.0066	0.8454	0.8584	0.8712	×
Seochogu	0.9825	0.0129	0.9574	0.9824	1.0080	△
Kangnamgu	1.2460	0.0134	1.2200	1.2460	1.2720	○
Songpagu	0.9760	0.0062	0.9638	0.9760	0.9883	×
Kangdonggu	0.9297	0.0064	0.9172	0.9297	0.9423	×

Table 4.4. Parameter estimation of Poisson regression model with spatial effects

Parameter	Estimate	Standard Deviation	2.5%	Median	97.5%
Intercept	0.0365	0.0452	-0.0447	0.0337	0.1496
Property taxes	0.1276	0.0958	-0.0565	0.1284	0.3106
Population density	-0.0704	0.0660	-0.2021	-0.0725	0.0631
Floating population	-0.0124	0.0624	-0.1342	-0.0148	0.1072
Proportion of youth	0.0576	0.0492	-0.0427	0.0600	0.1499
PHE*	-0.1791	0.0735	-0.3264	-0.1821	-0.0375
PDRA**	-0.0274	0.0524	-0.1282	-0.0277	0.0787
Proportion of housing	-0.1761	0.0673	-0.3027	-0.1783	-0.0313
Proportion of accommodation	0.0522	0.0764	-0.0850	0.0461	0.2168
Spatial correlation	-0.2158	0.5860	-1.3340	-0.2061	0.8418

특성이기에 시간과 공간을 결합한 시공간모형(Spatial-temporal model) 또는 시공간패널모형 등과 다양한 접근의 모형을 통해 범죄의 공간적 패턴을 확인할 수 있다.

Table 4.5는 공간상관성을 반영하였을 시 상대위험도를 살펴보면 2007년 총 범죄자료를 기준으로 중

Table 4.5. Relative risk estimation of Poisson regression model without spatial effects

District	Risk Estimate	Standard Deviation	2.5%	Median	97.5%	Type
Jongrogu	2.5130	0.0375	2.4390	2.5120	2.5860	○
Junggu	3.2340	0.0479	3.1410	3.2340	3.3280	○
Yongsangu	1.0030	0.0199	0.9648	1.0030	1.0430	△
Seongdonggu	0.8581	0.0155	0.8279	0.8581	0.8887	×
Kwangjingu	1.1810	0.0172	1.1480	1.1810	1.2150	○
Dongdaemungu	1.0700	0.0163	1.0380	1.0700	1.1020	○
Jungryanggu	0.9782	0.0148	0.9494	0.9781	1.0070	×
Seongbukgu	0.7644	0.0124	0.7403	0.7643	0.7889	×
Kangbukgu	0.8944	0.0157	0.8637	0.8944	0.9253	×
Dobonggu	0.7562	0.0139	0.7292	0.7561	0.7837	×
Nowongu	0.7443	0.0107	0.7236	0.7443	0.7656	×
Eunpyeunggu	0.9163	0.0138	0.8894	0.9162	0.9435	×
Seodaemungu	1.1360	0.0175	1.1020	1.1360	1.1710	○
Mapogu	0.9723	0.0154	0.9423	0.9722	1.0030	△
Yangcheongu	0.6707	0.0112	0.6488	0.6707	0.6929	×
Kangseogu	0.8244	0.0119	0.8012	0.8244	0.8479	×
Gurogu	1.0410	0.0151	1.0110	1.0410	1.0710	○
Keumchengu	1.2660	0.0215	1.2240	1.2660	1.3090	○
Youngdeungpogu	1.2610	0.0169	1.2280	1.2610	1.2950	○
Dongjakgu	0.7254	0.0130	0.7001	0.7253	0.7511	×
Kwanakgu	1.0010	0.0133	0.9753	1.0010	1.0280	△
Seochogu	1.0310	0.0155	1.0010	1.0310	1.0610	○
Kangnamgu	1.1900	0.0143	1.1630	1.1900	1.2190	○
Songpagu	0.9354	0.0121	0.9119	0.9354	0.9594	×
Kangdonggu	0.7739	0.0126	0.7493	0.7739	0.7988	×

구(Junggu), 종로구(Jongrogu), 금천구(Keumchengu), 영등포구(Youngdeungpogu) 등이 위험지역으로 나타난 반면 양천구(Yangcheongu), 동작구(Dongjakgu) 등은 상대적으로 덜 위험한 지역으로 나타났다. Table 4.3과 Table 4.5를 살펴보면 용산구(Yongsangu)의 경우 Table 4.3에서는 총 범죄발생 건수에 있어 위험한 행정구로 분류되는 반면에 Table 4.5에서는 유사한 위험도를 가지는 행정구로 판단된다. 광진구(Kwangjingu)의 경우 공간상관성을 반영하지 않은 모형(Model 1)의 경우 위험도가 낮은 지역으로 판단하였으나 공간상관성을 반영한 모형(Model 2)의 경우 위험도가 높은 지역으로 판단하였다. 중랑구(Jungryanggu)의 경우는 공간상관성을 반영하지 않은 경우 위험도가 낮은 지역으로 판단하였으나 공간상관성을 반영한 경우 위험도가 유사한 지역으로 판단하였다. 강북구(Kangbukgu)의 경우 광진구와는 반대로 공간상관성을 반영하지 않은 모형의 경우 위험도가 높은 지역으로 판단하였으나 공간상관성을 반영한 모형의 경우 위험도가 낮은 지역으로 판단하였다. 이와 같이 공간상관성의 반영 여부에 따라 모형을 통한 각 행정구별 위험도가 달라짐에 따라 연구지역에서 자료들이 공간상관성이 존재함이 확실하다면 공간상관성을 반영한 모형을 통해 다양한 통계적 추론을 하는 것이 바람직함을 알 수 있다.

Figure 4.1은 Table 4.3과 Table 4.5에 대해 각 행정구별 위험도 추정값에 대해 시각적으로 편하게 확인하기 위해 지리정보시스템을 통해 구현한 것이다. Figure 4.1에서 보는 것과 같이 중구(Junggu), 종로구(Jongrogu)는 매우 위험도가 높은 지역임을 확연하게 알 수 있다.

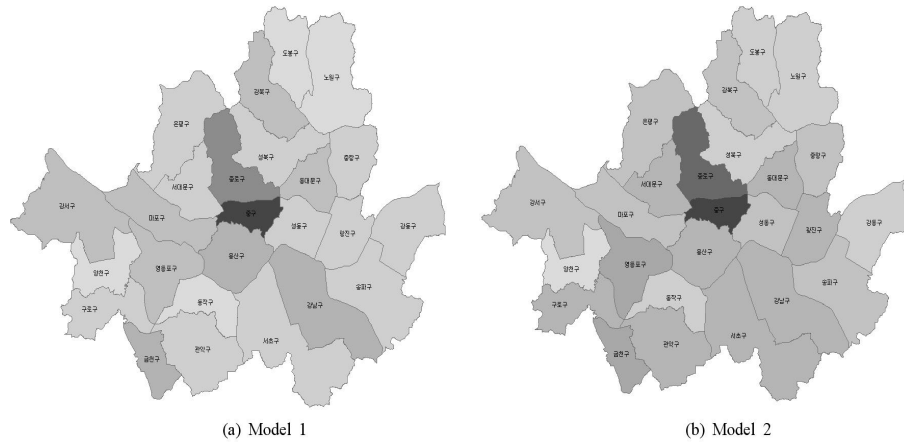


Figure 4.1. Visualization of relative risks using Geographic Information System(GIS)

5. 결론

본 연구에서는 다양한 설명변수들이 범죄발생 건수에 미치는 영향을 분석하기 위한 모형으로 포아송회귀모형을 제안하였으며, 제안된 모형을 통해 서울특별시의 각 행정구별 상대범죄위험도(relative crime risk)를 계산하고 각 행정구별로 상대범죄위험도 별로 순위를 설정할 수 있다. 따라서 본 연구에서는 기존의 범죄지도 구축 시 각 지역별 범죄사건 발생건수 기반(Event based)보다는 범죄사건 발생에 대한 상대위험도 기반의 범죄발생에 대한 범죄위험지도를 구축할 수 있는 통계학적 방법론 기반의 각 지역별 위험도를 제시하였다. 특히, 범죄발생 자료가 꾸준히 축적되면 서울시 행정구별이 아닌 행정동별 상대위험도에 대한 계산이 가능하며, 이는 향후 범죄발생 위험도가 높은 행정동에 대한 적절한 범죄예방 대책을 수립할 수 있을 것으로 예상된다. 또한 범죄발생지점에 대한 정확한 좌표에 대한 자료의 획득이 가능하면 공간점패턴 분석을 통해 범죄발생에 대한 Hot spot 분석도 가능하다.

본 연구는 2007년에 발생한 총 범죄자료를 분석하여 시기적으로 현재의 지역적 범죄의 위험도를 제시해 주고 있지 못하다는 한계가 있으나 본 연구는 지역적 범죄의 위험도를 제시하는 통계적 방법론을 제시하여 범죄발생자료가 갱신될 때마다 해당 년도의 범죄발생에 대한 지역적 위험도를 계산할 수 있다는 의의를 가지고 있다. 향후 공간적 연관성과 시간적 연관성을 반영한 공간패널자료분석 등의 연구를 통해 보다 정확한 범죄발생에 대한 시간적 공간적 연관관계를 파악할 수 있으며 향후 치안대책에 따른 범죄예방에 대한 시간적 공간적 패턴을 파악할 수 있다.

References

- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems(with Discussions), *Journal of the Royal Statistical Society, Series B*, **36**, 192-236.
- Choi, J. S., Park, M. S., Won, Y. B., Kim, H. G. and Heo, T.-Y. (2010). Analysis of total crime count data based on spatial association structure, *The Korean Journal of Applied Statistics*, **23**, 335-344.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequence (with Discussion), *Statistical Science*, **7**, 457-511.

- Griffith, D. A. (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data, *The Canadian Geographer*, **40**, 351–367.
- Hughes-Oliver, J. M., Heo, T. Y. and Ghosh, S. (2009). An Autoregressive point source model for spatial processes, *accepted in Environmetrics*.
- Jeong, K. S., Moon, T. H., Jeong, J. H. and Heo, S. Y. (2009). Analysis of Spatio-temporal pattern of urban crime and its influencing factors, *Journal of the Korean Association of Geographic Information Studies*, **12**, 12–25.
- Kakamu, K., Polasek, W. and Wago, H. (2007). Model choice for panel spatial models: Crime modeling in Japan, *Advances in Data Analysis, Springer*, 237–244.
- Kim, D. H. and Park, K. R. (2009). A probability modeling of the crime occurrence and risk probability map generation based on the urban spatial information, *Journal of the Korea Society of Computer and Information*, **14**, 207–215.
- Lee, S. W. (2004). *Study Implication of the Urban Plan for Incidence Crime on Seoul*, Technical report, Seoul Development Institute.
- Lee, S. W. and Cho, J. K. (2006). The effects of spatial and environmental factors on crime victimization, *Seoul Studies*, **7**, 57–76.
- Park, M. G. (2003). *The Implementation of Crime Probability Map Using GIS Analysis- focused on Sungbook-gu, Seoul*, Kyunghee University, Master thesis.
- Shin, W.-R. (2005). *The Analysis of Spatio-temporal Pattern for Urban Crime: Using GIS Methodology*, Kyunghee University, Master thesis.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal Royal Statistical Society, Series B*, **64**, 583–639.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association*, **82**, 528–550.
- Yoon, S. D. (2004). *Spatial Econometrics for Discrete Dependent Variable: Bayesian Approach and Gibbs Sampling Application*, Statistical Research Institute.