응용통계연구 (2010)
23(2), 335-344

공간적 연관구조를 고려한 총범죄 자료 분석

최정순[1] · 박만식[2] · 원유복[3] · 김학열[4] · 허태영[5]

[1]Division of Biostatistics and Epidemiology, Medical University of South Carolina
[2]성신여대 통계학과, [3]서울시청 정보화기획담당관, [4]서경대학교 도시공학과
[5]한국해양대학교 데이터정보학과

(2009년 11월 접수, 2010년 1월 채택)

요 약

공간자료분석에서 공간적 상관성을 배제한 일반적인 회귀모형을 통한 모수 추정값들은 신뢰성의 문제가 지적 되어 오고 있다. 본 연구에서는 공간자료의 상관성을 고려한 모형을 구축하기 위하여 일변량 조건부자기회귀모형을 이용하였으며 베이지안 기법을 통하여 모수를 추정하고 공간상관성이 고려된 공간 가산자료모형과 고려되지 않은 일반 가산자료모형을 비교하였다. 연구 대상으로는 서울시의 25개 행정자치구별 총범죄 자료를 이용하였으며 자료분석을 통하여 도시계획과 같은 국가 정책의 수립에 참고자료로 활용될 수 있으리라 판단된다.

주요용어: 총 범죄건수, 공간적 연관성, 조건부자기회귀모형, 일반화포아송분포, 음이항분포.

1. 서론

사회의 경제적 발전과 급속한 도시화에 따라 우리나라는 범죄발생 및 유형에 있어서 갈수록 다양화, 흉폭화, 연소화 등 심각한 문제를 야기하고 있다. 전체 인구수에서 5대 강력범죄(살인, 강도, 강간, 절도, 폭력)의 총 범죄의 건수는 1998년에 330,304건에서 2005년 487,847건으로 약 67% 이상 증가 한 것으로 드러났으며 이러한 범죄의 발생은 도시 지역이 더욱 심각한 양상을 보이고 있다. 따라서 본 연구에서는 서울시 각 행정자치구별로 발생된 범죄 자료를 이용하여 범죄에 영향을 미치는 변수들을 규명하고 공간적 연관성이 존재하는지를 확인하고자 한다. 일반적으로 공간 자료는 특정 위치 또는 특정 지역에서의 관측값이나 측정값으로 구성되며 공간 자료의 특성에 따라 지리통계 자료, 격자 자료, 점 패턴 자료의 세가지 유형으로 구분된다. 본 논문에서 사용된 자료는 서울시의 각 행정자치구의 총범죄 자료이며 지역 간의 공간상관성을 고려하여 범죄자료에 대한 공간 가산자료모형(spatial count data model)을 적용하였다.

자료가 서로 독립적이지 않고 어떤 상관성을 가지는 경우 그 상관성의 정보가 배제된 상태에서 일반적인 회귀모형을 적용하면 추정된 회귀계수는 효율적이지 못한 추정량이 되어 통계적 유의성 검정 및 예측오차 등의 정확성을 신뢰할 수 없게 된다 (Griffith, 1996). 따라서 공간정보를 포함한 공간자료의 경우 공간상관성을 반영한 모형의 적용이 필요하다. 공간상관성을 고려한 범죄모형의 연구는 다음과 같다. 이성우 (2004)와 이성우와 조중구 (2006)는 공간계량모형을 이용하여 종속변인으로 인구 10만명당 총 발생 건수에 대하여 변수변환을 이용하여 정규분포 기반의 공간범죄모형을 연구하였으며, 윤성도

[5]교신저자: (606-791) 부산시 영도구 동삼동 1, 한국해양대학교 데이터정보학과, 조교수.
E-mail: heoty@hhu.ac.kr

(2004)는 이성우 (2004)와 동일한 자료를 이용하여 범죄사건의 크기에 따라 0 또는 1의 이진수의 값으로 구분하여 공간로지스틱 모형을 구축하고 각 변수들의 영향력을 파악하였다.

본 연구에서는 범죄에 미치는 영향요인으로 서울시에서 제공한 재산세, 인구밀도, 유동인구, 청소년비율, 고학력비율, 개발제한구역비율, 주택연상비율, 숙박연상비율 등을 고려하였다 (이성우, 2004). 범죄의 양상 및 행태는 범죄를 발생시키는 주요 요인들에 의하여 결정되지만 지역의 환경과 같은 공간적 특성과 밀접한 관련성을 가지고 있어 공간 영향력을 조건부자기회귀모형을 이용하여 범죄 모형에 포함시켜 총범죄수에 대하여 지역적 상관성이 존재하는가에 대한 탐색적 연구를 진행하였다.

본 연구에서는 각 지역별로 발생되는 범죄 건수가 정규분포 기반의 선형회귀모형보다 가산자료 기반의 모형(예, 포아송분포, 음이항 분포, 일반화 포아송 분포 등)을 적합시키는것이 타당할 것으로 판단되어 총범죄수에 대하여 다양한 공간가산자료모형을 적합시켜 비교하였다. 본 연구의 구성은 다음과 같다. 2장에서는 공간모형 중 조건부자기회귀모형에 대하여 간략하게 설명하였으며 3장에서는 가산자료모형인 음이항모형과 일반화된 포아송모형에 대해 소개하였으며 4장에서는 서울시 자치구별 총범죄에 대한 실증예제를 다루었으며 5장에서는 결론을 제시하였다.

2. 조건부자기회귀모형

격자 자료에서는 대표적으로 공간자기회귀(Simultaneous autoregressive; SAR)모형과 조건부자기회귀(conditional autoregressive; CAR)모형 (Sain과 Cressie, 2002; Carlin과 Banerjee, 2003; Jin 등, 2005) 등을 이용하여 공간 상관성을 회귀모형 내에 포함시키게 되며 다양한 분야에서 광범위하게 사용되어지고 있다. SAR 모형은 CAR 모형으로 표현 가능하기 때문에, 본 연구에서는 조건부자기회귀모형(CAR)을 이용하였다.

2.1. 모형의 구조

조건부자기회귀모형의 간략한 소개는 다음과 같다. $\{Z(\mathbf{s}_i) : \mathbf{s}_i \in \mathbb{D}\}$이 가우시안 확률과정(Gaussian stochastic process)을 따른다고 하고 $Z(\mathbf{s}_i)$가 지역 $\mathbf{s}_i$에서의 관측값이라고 하면 $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$이 집합 $\mathbb{D}$의 격자의 형태를 가진다고 하자. 여기서, 지역 $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$은 $\mathbb{D}$를 분할하게 되고 다음의 조건을 만족하게 된다. 모든 $i = 1, \ldots, n$에 대하여,

$$\bigcup_{i=1}^{n} \mathbf{s}_i = \mathbb{D}, \quad \mathbf{s}_i \bigcap_{i \neq k} \mathbf{s}_k = \phi.$$

$\{Z(\mathbf{s}_i)\}$을 아래와 같은 조건부 자기회귀모형으로 표현할 수 있다.

$$Z(\mathbf{s}_i) | \{Z(\mathbf{s}_{-i})\} \sim N\left(\mu_i + \sum_{k=1}^{n} b_{ik} \left(Z(\mathbf{s}_k) - \mu_k\right), \tau_i^2\right), \tag{2.1}$$

여기서 $\{Z(\mathbf{s}_{-i})\} = \{Z(\mathbf{s}_k) : k \neq i\}$이고 $E(Z(\mathbf{s}_i)) = \mu_i$이다. 또한 $b_{ii} = 0$이고 τ_i^2는 조건부 분산(conditional variance)이다. $n < \infty$에 대하여 $\mathbf{z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$의 결합분포(joint distribution)는 Besag (1974)의 인수분해정리(Factorization theorem)에 의하여 다음과 같이 표현된다.

$$\mathbf{z} \sim N\left(\boldsymbol{\mu}, (\mathbf{I}_n - \mathbf{B})^{-1}\mathbf{T}\right),$$

여기서, $\mathbf{B} = \{b_{ij}\}_{i,j=1,\ldots,n}$, $\mathbf{T} = \text{Diag}\{\tau_1^2, \tau_2^2, \ldots, \tau_n^2\}$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$이며 $\mathbf{I}_n$은 n-차원 단위행렬(identity matrix)을 나타낸다. 그리고 $(\mathbf{I}_n - \mathbf{B})$이 역행렬이 존재하기 위해서 $b_{ik}\tau_k^2 = b_{ki}\tau_i^2$를 만

족하여야 한다. 행렬 $(\mathbf{I}_n - \mathbf{B})^{-1}\mathbf{T}$는 대칭이며 양정치행렬이다. 공간적 구조를 설정하기 위하여 인접성 기반의 이웃정보를 사용하기 위하여 $b_{ij} = w_{ij}/w_{i+}$로 정의하였다. 여기서, $\mathbf{s}_i$와 $\mathbf{s}_j$가 서로 이웃하면 $w_{ij} = 1$, 그렇지 않으면 $w_{ij} = 0$이며 $w_{i+} = \sum_j w_{ij}$으로 $\mathbf{s}_i$에 인접한 지역의 개수를 의미한다. 그리고 $\tau_i^2 = \sigma_c^2/w_{i+}$를 가정하였다. 여기서, σ_c^2은 공간 영향의 평활성을 나타내는 모수이다. 이 모형을 고유자기상관(Intrinsic autoregressive; IAR)모형이라 하며, CAR$(1, \sigma_c^2)$으로 표현할 수 있다.

공간 상관성을 나타내는 모수 ρ를 고려한 조건부 자기회귀모형은 아래와 같이 정의할 수 있다.

$$Z(\mathbf{s}_i)|\{Z(\mathbf{s}_{-i})\} \sim N\left(\mu_i + \rho\sum_{k=1}^{n} b_{ik}\left(Z(\mathbf{s}_k) - \mu_k\right), \tau_i^2\right). \tag{2.2}$$

인수분해 정리에 의하여 다음과 같이 표현될 수 있다.

$$\mathbf{z} \sim N\left(\boldsymbol{\mu}, (\mathbf{I}_n - \rho\mathbf{B})^{-1}\mathbf{T}\right),$$

여기서, $\rho = 1$이면 식 (2.1)에서 제시한 공간 모형과 동일하다 (Heo와 Hughes-Oliver, 2009). 식 (2.2)의 모형은 CAR(ρ, σ_c^2)으로 표현할 수 있다.

3. 음이항모형과 일반화 포아송모형

일반적으로 가산자료 분석에 가장 많이 활용되는 통계모형은 포아송모형(Poisson model)이지만 평균과 분산이 동일하다는 엄격한 가정을 가지고 있다. 따라서 본 연구에서는 기본적인 가산모형인 포아송모형 외에 본 연구에서 이용된 범죄자료와 같이 자료의 평균이 분산보다 더 작은 과대산포를 고려한 모형을 사용하는 것이 더 적합하여 과대산포(over-dispersion)시 대표적으로 많이 활용되는 음이항(negative binomial; NB)모형과 일반화 포아송(generalized Poisson; GP)모형을 이용하였으며 간략한 소개는 다음과 같다.

3.1. 음이항모형

두 모수 $r > 0$과 $\lambda > 0$을 가지는 확률변수 Y의 음이항모형의 밀도함수는 다음과 같이 나타낼 수 있다.

$$P(Y = y|r, \lambda) = \frac{\Gamma(y + r)}{\Gamma(r)y!}\left(\frac{r}{\lambda + r}\right)^r\left(\frac{\lambda}{\lambda + r}\right)^y, \quad y \geq 0.$$

음이항모형은 NB(r, λ)로 표현할 수 있으며, $E(Y) = \lambda$이고, $\mathrm{Var}(Y) = \lambda(1 + \lambda/r)$이다. 따라서 분산이 평균보다 크기 때문에, 과대산포 자료는 음이항분포로 모형화할 수 있다. 만약 $r \to \infty$이면, 음이항모형은 포아송모형이 된다 (Johnson 등, 1993).

3.2. 일반화 포아송모형

일반화 포아송모형은 Consul과 Jain (1973)에 의해 처음으로 소개되었고, Consul (1989)이 자세히 연구하였다. 확률변수 Y가 일반화 포아송분포를 따른다고 하면 $Y \sim$ GPoi(λ, γ)로 표현되며, 밀도함수는 다음과 같이 정의할 수 있다.

$$P(Y = y|\alpha, \lambda) = \lambda\{\lambda + (\alpha - 1)y\}^{y-1}\frac{\alpha^{-y}}{y!}\exp\left\{-\frac{\lambda + (\alpha - 1)y}{\alpha}\right\}, \tag{3.1}$$

여기서 $\gamma = 1 - 1/\alpha$는 산포 모수이며, $\lambda \geq 0$이 평균 모수이다. 확률변수 Y의 평균과 분산은 각각 $E(Y) = \lambda$와 $\mathrm{Var}(Y) = \lambda(1-\gamma)^{-2}$이다. 만약 $\gamma > 0$이면 본 모형은 과대산포를 표현하며, 만약 $\gamma < 0$이

표 **4.1.** 모형에 사용된 자료 및 변수 설명

변수		변수 설명(단위)	자료출처
종속변수	총범죄건수	상주인구 100,000명당 범죄건수	서울지방경찰청
설명변수	가구주 1인당 재산세	재산세/가구주 인구수	서울시 기본통계(통계연보)
	인구밀도	상주인구수/행정구역면적(km^2)	통계청 인구주택 총조사
	유동인구	명	통계청 인구주택 총조사
	청소년비율	15~24세	통계청 인구주택 총조사
	고학력비율	4년제 이상 대졸	통계청 인구주택 총조사
	개발제한구역비율	km^2	서울시 기본통계(통계연보)
	주택연상비율	주택연상면적(m^2)/총건물연상면적(m^2)	서울시 건축주택통계분석시스템
	숙박연상비율	숙박연상면적(m^2)/총건물연상면적(m^2)	서울시 건축주택통계분석시스템

면 분산이 평균보다 작은 과소산포(under-dispersion)를 표현한다. $\gamma = 0$이면, 일반화 포아송모형은 표준 포아송모형이 된다.

4. 실증예제

서울시의 자치구별 범죄발생과 다양한 설명 변인과의 인과 관계를 모형화하고 분석하기 위하여 서울시에서 제공한 범죄, 사회 경제학적 변수 그리고 도시계획 변수 자료를 통합하여 분석에 필요한 단일 자료를 구축하였다. 서울시의 자치구별 범죄발생에 대한 자료는 2007년 서울시의 자치구별 총범죄수를 이용하였고 가산자료모형에 대하여 공간상관성을 반영한 모형과 반영하지 않은 모형을 적용하여 그 결과를 비교하였다. 본 연구 사용된 변수들에 대한 간략한 설명은 아래와 같다. 서울시의 25개 행정자치구별로 집계한 재산세, 인구밀도, 유동인구, 청소년비율, 고학력비율, 개발제한구역비율, 주택연상비율, 숙박연상비율을 설명변수로 고려하였다(서울시, 통계청). 종속변수인 2007년도 총범죄수의 자료는 서울시에서 제공하고 있는 자료를 사용하였다. 모형에 사용된 자료 및 자세한 설명변인에 관한 설명은 표 4.1에 나타내었다. 그리고 그림 4.1은 서울시의 25개 자치구에 대한 지도이다.

4.1. 모형의 적합

모형을 구축하기 위하여 범죄에 영향을 미치는 변수들과 공간연관성을 동시에 고려한 공간 가산자료모형을 시도하였으며, 사용된 모형들은 아래와 같다.

$$Y(\mathbf{s}_i) \sim \text{Poisson}\,(\lambda(\mathbf{s}_i)) \quad \text{or} \quad \text{NB}\,(\lambda(\mathbf{s}_i), r) \quad \text{or} \quad \text{GP}\,(\lambda(\mathbf{s}_i), \gamma)\,,$$

여기서, $\log(\lambda(\mathbf{s}_i)) = \mu(\mathbf{s}_i) + Z(\mathbf{s}_i)$, $\mu(\mathbf{s}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_8 x_{i8}$는 평균함수이고 설명변수들은 로그변환을 하였다. $Z(\mathbf{s}_i)$는 공간상관성을 반영한 공간모형으로 표현될 수 있다. 본 연구에서는 조건부 자기회귀모형의 일반형인 CAR(ρ, σ_c^2)를 이용하여 공간 상관의 강도를 나타내는 모수인 ρ를 추정하여 공간영향을 추정할 수 있도록 하였다.

포아송모형, 음이항모형 그리고 일반화 포아송모형을 고려하였으며, 각각의 모형에서 공간상관성을 고려한 경우와 고려하지 않은 경우의 전체 6가지 모형을 제시하였으며, 각 모형별로 모수를 추정하였다. 모수를 추정하기 위해 베이지안 방법을 이용하였으며, 이를 위해 아래와 같이 모수들에 대한 사전분포(prior distribution)를 정의하였다 (Heo와 Hughes-Oliver, 2009).

$$\beta_p \sim N(0, 10^6), \qquad p = 0, 1, \ldots, 8$$
$$\sigma_c^2 \sim \text{Inverse Gamma}(0.5, 0.0005)$$

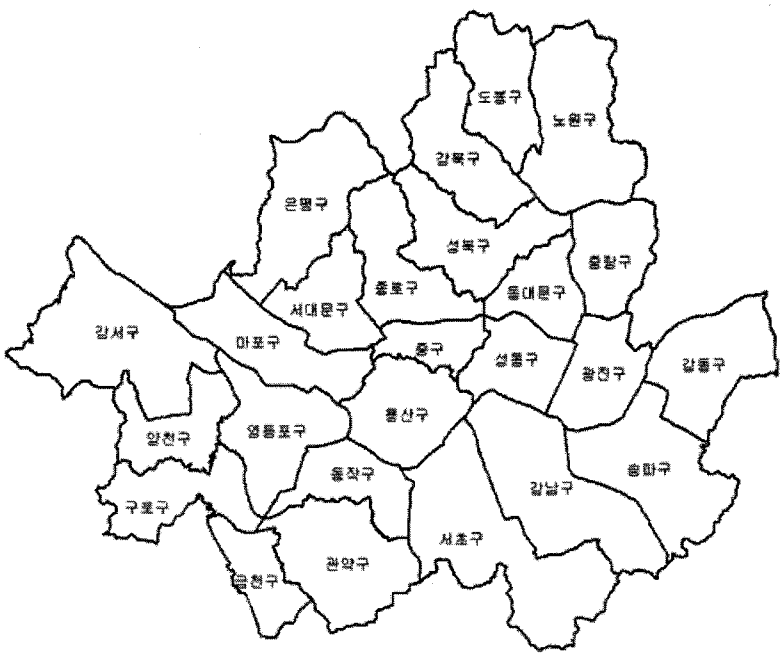


그림 **4.1.** 서울특별시 25개 행정자치구 경계지도

$$\rho \sim \text{Uniform}(-1, 1)$$
$$r \sim \text{Inverse Gamma}(1, 0.0005)$$
$$\gamma \sim \text{Uniform}(0, 1).$$

각 모수 $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_8)^T$, σ_c^2, ρ, r(or γ) 간에는 상호 독립임을 가정하며, 공간상관성을 고려한 모형의 경우, 모수벡터 $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_c^2, \rho, \gamma)^T$에 대한 사후분포(posterior distribution)는 다음과 같이 정의된다.

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

모수벡터 $\boldsymbol{\theta}$에 대한 추론은 사후분포를 통해서 이루어지며, 제시된 모형의 모수를 추정하기 위하여 마코프 연쇄 몬테카를로(Markov Chain Monte Carlo; MCMC) 방법을 사용하여 베이지안 통계 패키지인 WinBUGS(`http://www.mrc-bsu.cam.ac.uk/bugs/`)를 사용하였다. 모수추정의 결과는 세 개의 초기값을 활용하여 20,000번 반복 후 처음 10,000번까지 제거한 나머지 결과값이며 구체적인 모수 추정값은 표 4.2에 제시하였다.

표 4.2에서 알 수 있듯이 고려한 3개의 모형에서 모두 공간상관계수 ρ가 통계적으로 유의하지 않기에 범죄자료의 경우 공간상관이 존재하고 있지 않음을 알 수 있다. 음이항 모형과 일반화 포아송 모형의 산포모수의 추정치가 모두 양수이기 때문에, 자료의 과대산포를 고려한 모형이 적합함을 알 수 있다.

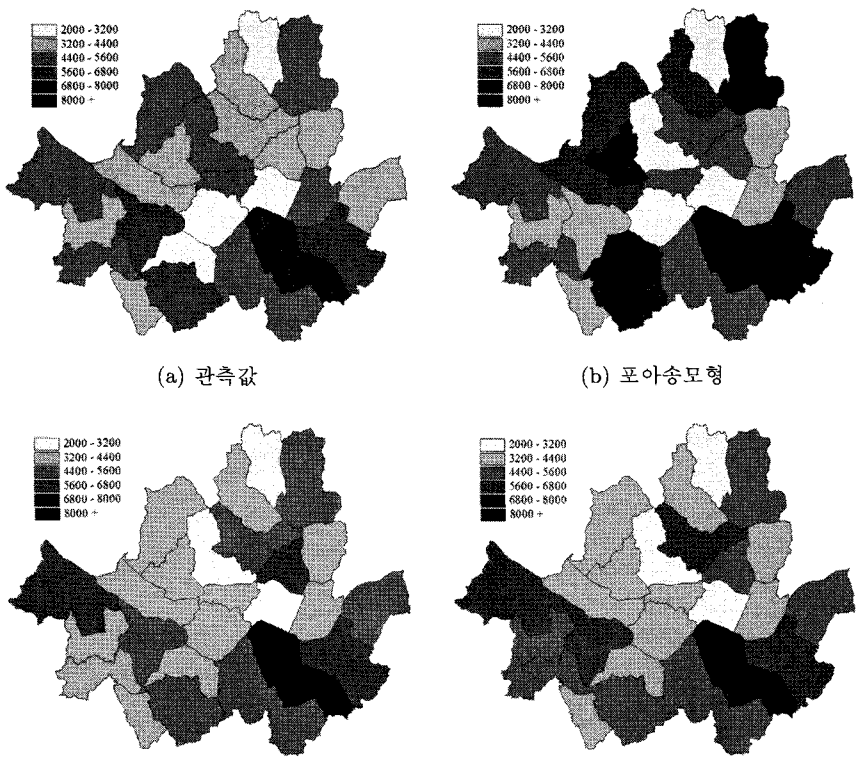
공간상관성을 반영하지 않고 과대산포 영향을 고려하지 않은 포아송모형의 경우, 유동인구 변수를 제외한 모든 변수에서 통계적으로 유의하게 나타난 반면, 공간 상관성을 반영하지 않은 음이항모형과 일반화 포아송모형은 주택연상비율 만이 통계적으로 유의하였다. 공간상관성을 반영한 모형의 경우 포아송

표 4.2. 모형별 모수 추정값과 베이지안 신뢰구간

	Without Spatial Effects					With Spatial Effects				
	Mean†	Std‡	2.5%	50.0%	97.5%	Mean†	Std‡	2.5%	50%	97.5%
Poisson model										
β_0	11.320	0.288	10.760	11.320	11.880	10.650	3.255	3.968	10.720	16.820
β_1	0.043	0.011	0.020	0.043	0.065	0.062	0.122	−0.177	0.063	0.291
β_2	−0.167	0.016	−0.197	−0.167	−0.137	−0.143	0.159	−0.458	−0.137	0.163
β_3	−0.018	0.020	−0.058	−0.018	0.022	−0.049	0.223	−0.476	-0.051	0.392
β_4	0.454	0.070	0.318	0.454	0.591	0.716	0.715	−0.707	0.733	2.103
β_5	−0.358	0.027	−0.410	−0.358	−0.305	−0.401	0.275	−0.949	−0.390	0.119
β_6	−0.002	0.001	−0.003	−0.002	−0.001	0.000	0.006	−0.012	0.000	0.012
β_7	−0.781	0.025	−0.828	−0.781	−0.731	−0.777	0.246	−1.244	−0.789	−0.270
β_8	0.043	0.004	0.034	0.043	0.051	0.050	0.050	−0.052	0.051	0.148
ρ						−0.013	0.500	−0.912	−0.001	0.871
σ_c^2						0.121	0.045	0.061	0.112	0.232
Negative Binomial model										
β_0	10.550	3.416	3.640	10.570	17.340	10.410	3.518	3.567	10.460	17.200
β_1	0.048	0.130	−0.214	0.048	0.302	0.057	0.135	−0.207	0.055	0.329
β_2	−0.212	0.167	−0.543	−0.212	0.120	−0.197	0.176	−0.535	−0.201	0.153
β_3	0.036	0.238	−0.440	0.040	0.499	0.026	0.248	−0.475	0.032	0.511
β_4	0.679	0.775	−0.888	0.686	2.196	0.647	0.787	−0.889	0.653	2.248
β_5	−0.460	0.295	−1.031	−0.460	0.134	−0.451	0.299	−1.030	−0.454	0.160
β_6	−0.002	0.006	−0.015	−0.002	0.011	−0.002	0.007	−0.015	-0.002	0.011
β_7	−0.735	0.261	−1.267	−0.733	−0.228	−0.722	0.272	−1.263	−0.721	−0.180
β_8	0.032	0.045	−0.057	0.032	0.120	0.037	0.046	−0.056	0.037	0.129
r	37.880	12.790	17.150	36.700	66.770	63.650	81.260	17.830	42.000	304.600
ρ						−0.008	0.567	−0.948	−0.003	0.936
σ_c^2						0.020	0.037	0.000	0.003	0.135
Generalized Poisson model										
β_0	11.050	3.259	4.627	11.040	17.350	11.150	3.420	3.977	11.220	17.860
β_1	0.059	0.130	−0.196	0.057	0.322	0.060	0.135	−0.212	0.056	0.341
β_2	−0.127	0.190	−0.495	−0.129	0.256	−0.116	0.183	−0.475	−0.116	0.238
β_3	−0.017	0.222	−0.440	−0.021	0.435	−0.057	0.254	−0.558	−0.053	0.438
β_4	0.270	0.826	−1.436	0.282	1.901	0.409	0.807	−1.212	0.414	1.988
β_5	−0.335	0.320	−0.977	−0.328	0.293	−0.322	0.324	−0.995	−0.317	0.306
β_6	−0.001	0.006	−0.014	−0.002	0.012	−0.001	0.006	−0.013	−0.001	0.012
β_7	−0.761	0.287	−1.323	−0.765	−0.176	−0.787	0.295	−1.379	−0.787	−0.162
β_8	0.042	0.052	−0.063	0.043	0.148	0.048	0.049	−0.053	0.047	0.147
γ	0.908	0.016	0.875	0.908	0.938	0.773	0.243	0.081	0.893	0.934
ρ						0.002	0.559	−0.939	0.010	0.950
σ_c^2						0.041	0.056	0.000	0.008	0.181

Notes. †: posterior mean, ‡: standard deviation.

모형, 음이항모형, 일반화 포아송모형 모두 주택연상비율 만이 통계적으로 유의하였다. 따라서 공간영향력이 그다지 높지 않더라도 모수의 추정값에 영향을 미치고 있는 것으로 나타나서 공간정보를 포함하고 있는 공간 가산자료모형의 구축시 공간상관성과 과대산포의 영향을 고려한 모형을 선택하는 것이 올바르다고 판단된다.



(a) 관측값 (b) 포아송모형

(c) 음이항모형 (d) 일반화 포아송모형

그림 **4.2.** 실제 관측지도와 공간상관성을 고려한 모형들의 예측지도

4.2. 모형의 평가

본 연구에서는 모형의 선택을 위하여 베이지안 적합도 기준인 DIC(Deviance information criterion) (Spiegehalter 등, 2002)를 이용하였다. 모형의 설명 능력은 우도함수의 값로써 측정될 수 있으나 복잡한 모형보다는 간단한 모형이 선호되어야 한다는 원리에 의하여 추정할 모수의 수가 많은 것에 대해 벌칙(penalty)을 부가함으로써 모형의 절약성을 고려할 수 있다. DIC는 일반적인 적합도 지수인 AIC (Akaike, 1973)나 BIC (Schwarz, 1978)와 같이 모형의 복잡성을 고려한 적합도 기준이다. 모형 적합도는 $D(\theta) = -2\log f(y|\theta) + 2\log[h(y)]$의 사후분포 평균인 $\bar{D} = \mathrm{E}_{\theta|y}[D(\theta), \theta = (\beta_0, \beta_1, \ldots, \beta_8, \rho, \sigma_c^2, \gamma)]$이며 $h(y)$는 모형에 의하여 영향을 받지 않는다. 모형의 복잡성은 모형에 영향을 주는 모수의 수인 $p_D = \bar{D} - D(\bar{\theta})$으로 측정되어질 수 있다. 여기서 $D(\bar{\theta})$는 모수의 사후평균, $\bar{\theta}$에서 평가된 편이(deviance)값이다. 결과적으로 베이지안 적합도 지수인 DIC는 $\mathrm{DIC} = \bar{D} + p_D$로 정의되며 작은 값일수록 모형적합이 잘 되었다고 할 수 있다.

공간상관성을 고려한 세 가지 모형에 대하여 최적의 모형을 선택하기 위하여 DIC와 다음의 RMSPE(Root Mean Squared Prediction Error)를 고려하였다.

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Y(\mathbf{s}_i) - \hat{Y}(\mathbf{s}_i)\right]^2}.$$

표 4.3. 모형의 적합도 비교

Model	Without Spatial Effects			With Spatial Effects		
	DIC	p_D	RMSPE	DIC	p_D	RMSPE
P	2230.470	9.053	870.465	304.282	24.856	1617.253
NB	410.829	10.410	916.409	407.680	13.196	903.043
GP	410.785	10.132	869.266	381.285	3.876	944.966

표 4.3에서 알 수 있듯이 공간상관성을 배제한 모형들의 경우 DIC 기준으로 과대산포 영향을 고려한 음이항모형과 일반화 포아송모형의 적합이 포아송 모형보다 상대적으로 우월함을 알 수 있고 RMSPE 기준에서는 일반화 포아송모형이 음이항모형보다 상대적으로 낮은 RMSPE를 가진다. 공간상관성을 반영한 모형의 경우 DIC 관점에서는 포아송모형이 음이항모형이나 일반화 포아송모형보다 모형 적합이 잘 이루어졌음을 알 수 있으나, RMSPE 관점에서는 음이항모형이 가장 좋은 것으로 나타났다. 포아송 모형의 경우 공간상관성이 과대산포 문제에 영향을 주어 DIC가 가장 작게 측정된 것으로 판단된다. 즉, 범죄자료의 경우 지역적 연계성이 낮고 지역별로 범죄의 양상이 서로 달라 공간영향력이 과대산포의 효과를 증가시킨 것으로 판단된다. 결과적으로 공간상관성을 고려하지 않은 포아송모형은 부적합한 모형이며 음이항모형과 일반화 포아송모형은 유사한 모형적합의 결과를 보이고 있다.

그림 4.2는 실제 관측값과 공간모형을 고려한 3개의 모형(포아송모형, 음이항모형, 일반화 포아송모형)을 통해 나온 예측값을 나타낸 지도이다. 표 4.3의 RMSPE값에서도 알 수 있듯이, 포아송 모형의 예측값은 관측값과는 큰 차이가 있음을 지도에서 알 수 있다. 반면, 음이항모형과 일반화 포아송모형은 관측값과 큰 차이가 없음을 알 수 있다. 따라서, 예측력 측면에서는 음이항모형과 일반화 포아송모형이 포아송모형보다 더 좋은 결과를 보임을 알 수 있다.

5. 결론

본 연구에서는 서울시의 각 자치구별로 발생된 총범죄 자료에 대하여 포아송 회귀모형을 적합시키고, 각 자료에 대한 공간상관성의 여부를 판단하였다. 연구결과 서울시의 자치구에 따른 공간 상관성이 존재하지 않는 것으로 나타나, 현재와 같이 자치구 경계로 취합된 공간단위의 범죄 자료의 경우 공간적 연속성 또는 연계성이 유의하지 않은 것으로 분석되었다. 따라서 향후 지구단위, 블록단위, 또는 필지단위 등과 같이, 보다 세밀한 공간단위의 범죄 발생 정보가 획득되고 이와 유사한 연구가 적용된다면 본 연구 결과와 비교분석이 가능한 의미있는 연구가 될 것으로 예상된다. 또한 총범죄수 뿐만 아니라 5대범죄(살인, 강도, 강간, 폭력, 절도)의 개별 모형의 구축이 필요하며 각 범죄 간의 연관성을 고려한 다변량 범죄모형의 구축도 필요할 것으로 판단되어 연구 중에 있으며, 년도별 범죄자료를 확보하여 공간패널분석 등의 다양한 연구를 준비 중에 있다.

참고문헌

윤성도 (2004). 이산종속변인의 분석을 위한 공간계량경제모형: 베이지안 접근방법과 깁스 표본추출 방법을 응용하여, 대학원생우수논문집. 통계개발원.

이성우 (2004). <서울시 범죄발생의 도시계획적 함의>, 서울시정개발연구원.

이성우, 조중구 (2006). 공간적, 환경적 요인이 범죄피해에 미치는 영향, <서울도시연구>, **7**, 57–76.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In B. Petrox and F. Caski (Eds.), *Second International Symposium on Information Theory*, 267–281.

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.

Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics 7, eds.* J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West. Oxford Univeristy Press, Oxford.

Consul, P. (1989). *Generalized Poisson Distributions. Properties and Applications,* Marcel Dekker, Inc., New York.

Consul, P. and Jain, G. (1973). A generalization of the poisson distributions, *Technometrics*, **15**, 791–799.

Griffith, D. (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying georeferenced data, *The Canadian Geographer*, **40**, 351–367.

Heo, T. Y. and Hughes-Oliver, J. (2009). Uncertainty adjustments to determine atmospheric dispersion models, *International Journal of Environmental Pollution*, In press.

Jin, X., Carlin, B. P. and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data, *Biometrics*, **61**, 950–961.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1993). *Univariate Discrete Distributions.* 2nd ed., Wiley, New York.

Sain, S. R. and Cressie, N. (2002). Multivariate lattice models for spatial environmental data, In *ASA Proceedings of the Joint Statistical Meetings*, 2820–2825, American Statistical Association, Alexandria, VA.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

Analysis of Total Crime Count Data Based on Spatial Association Structure

Jungsoon Choi[1] · Man Sik Park[2] · Yu-Bok Won [3] · Hag-Yeol Kim[4] · Tae-Young Heo[5]

[1]Division of Biostatistics and Epidemiology, Medical University of South Carolina
[2]Department of Statistics, Sungshin Women's University
[3]Department of Information System Planning, Seoul Metropolitan Government
[4]Department of Urban Engineering, Seokyeong University
[5]Department of Data Information, Korea Maritime University

(Received November 2009; accepted January 2010)

Abstract

Reliability of the estimation is usually damaged in the situation where a linear regression model without spatial dependencies is employed to the spatial data analysis. In this study, we considered the conditional autoregressive model in order to construct spatial association structures and estimate the parameters via the Bayesian approaches. Finally, we compared the performances of the models with spatial effects and the ones without spatial effects. We analyzed the yearly total crime count data measured from each of 25 districts in Seoul, South Korea in 2007.

Keywords: Crime counts, spatial association, conditional autoregressive model, generalized Poisson distribution, negative binomial distribution.

[5]Corresponding author: Assistant professor, Department of Data Information, Korea Maritime University, 1 Dongsan-Dong, Yeongdo-Gu, Pusan 606-791, Korea. E-mail: heoty@hhu.ac.kr