응용통계연구 제 19권 3호, 2006년, pp. 421–430

공간통계량을 활용한 베이지안 자기 포아송 모형을 이용한 소지역 통계*

이상은[1)]

요 약

표본조사에서는 일반적으로 지형학적 범위가 넓거나 흔히 우리가 알고 있는 지형적 범위 즉 시 또는 도 단위로 표본설계가 이루어진다. 그러므로 지형학적 범위가 작은 소지역은 충분한 표본의 확보가 불가능하며 따라서 정확한 소지역 통계를 얻는 것은 매우 어렵다. 이러한 문제로 정확한 소지역 통계를 얻기 위한 연구가 활발히 진행되고 있다. 최근 신기일과 이상은(2003)은 공간통계 모형을 이용한 소지역 추정을 연구하였다. 본 논문은 신기일과 이상은(2003)의 공간자기회귀(Spatial Autoregressive: SAR) 모형을 확장한 모형인 베이지안 자기 포아송 모형(Bayesian Auto-Poisson Model: BAPM)을 이용한 소지역 추정에 관하여 연구하였다. 분석에 사용된 자료는 호주의 1998년 장애인 통계(Survey of Disability, Aging and Cares:SDAC)이며 MSE, MB 그리고 회귀분석을 이용한 편의 분석기법이 비교에 사용되었다.

주요용어: 베이지안 자기 포아송 모형, 공간 자기회귀 모형, 공간상관관계

1. 서론

소지역 통계는 자료에 기반을 둔 자료기반(data based) 방법을 시작으로 모형에 기반을 둔 모형기반(model based) 기법에 이르기까지 많은 연구가 활발하게 진행되고 있다. 모형기반 방법은 회귀분석을 선두하여 시계열 분석 등 많은 모형들이 적용되고 있다. 그러나 모형기반 기법에서는 모형설정에 필요한 설명변수를 얻어야 하는 어려움이 있다. 이러한 어려움을 극복하기 위한 방법으로 신기일과 이상은(2003)은 공간통계 모형을 적용하였고 모형설정에 필요한 설명변수를 생성하여 분석의 효율성을 높일 수 있음을 보였다. 또한 김재두 등(2004)에서는 공간시계열 모형을 소지역 통계에 적용할 경우 효율을 높일 수 있음을 역시 보였다. 이러한 기존의 분석에서는 자료가 정규분포를 따른다는 가정 하에서 분석이 이루어졌으며 실업자 수나 장애자 수와 같은 자료는 변환을 한 후 분석을 하였다. 그러나 이러한 자료는 포아송 분포를 따르는 것으로 알려져 있기 때문에 포아송 분포를 이용한 분석이 더욱 타당할 것이다. 이에 본 논문에서는 포아송 분포를 가정한 공간통계모형을 소지

* 이 연구는 경기대학교 해외 파견 연구비에 의해 이루어짐

1) (442-720) 경기도 수원시 영통구 이의동 94-6,경기대학교 응용정보통계학, 부교수
E-mail : sanglee62@kyonggi.ac.kr

역 추정에 이용하였다. 연구에 이용된 자료는 호주 통계청(ABS)에서 1998년 전국 조사로 행해진 호주 장애인 통계조사이며, 전국 기준 조사를 이용하여 각 주(state) 단위의 장애인 수를 추정하고 그 추정치를 각 주 단위의 복지예산 편성의 기초자료로 이용하려고 하였다. 또한 주를 소지역으로 하는 모형 기반 방법을 사용하기 위해 사용 가능한 설명변수를 알아본 결과 각 소지역 단위의 기본정보(남녀 인구수 등)만이 가능했다. 그러므로 회귀 모형 등과 같은 설명변수의 절대적인 모형의 응용에는 한계가 있음을 알 수 있다. 또한 김재두 등(2004)에서와 같이 공간시계열 모형을 적용하기에는 5년에 한 번씩 이루어지는 자료로 단지 두 번의 조사가 이루어진 상태여서 공간시계열 분석의 적용은 적절치 않았다. 그러므로 충분한 설명변수가 존재하지 않고, 조사 결과인 장애인 수는 각 소지역 별로 매우 적은 수로 얻어졌기 때문에, 분석은 장애인 수가 포아송 분포를 따른다는 가정 하에서 공간 통계분석을 이용하는 것이 타당할 것이다. 따라서 본 논문에서는 각 소지역의 장애인수를 지역당 흔치 않은 사건의 수에 적합한 분포인 포아송 모형에 적용하였으며 이 때 각 소지역에서 얻어질 수 있는 표본오차를 랜덤 효과로 적용한 일반선형모형을 포아송 분포의 모수에 사전함수로 적용하였다. 즉 사전함수를 단순 분포함수가 아닌 일반선형모형으로 적용하였다. 이와 같은 모형을 본 논문에서는 베이지안 자기 포아송 모형(BAPM)이라 정의했으며 이는 2장에 설명하였다. 또한 신기일과 이상은(2003)이 제안한 공간통계모형인 SAR 방법과 BAPM 방법을 MSE와 MB로 비교하였으며 또한 Brown 등(2001)이 제안한 편의진단(bias diagnostics)법을 이용 비교하였다. 본 논문은 2장에 베이지안 자기 포아송 모형과 공간통계 SAR 모형을 소개하였고, 3장에서는 자료 분석으로 자료의 소개와 모형 적용 결과를 보였으며 4장은 3장에서 얻은 결과 값을 이용하여 두 방법을 비교하였으며 또한 이 분야의 장래 연구방향을 제시하였다.

2. 베이지안 자기 포아송 모형(Bayesian Auto Poisson Model : BAPM)

공간통계에서의 이산형 Lattice 자료의 모형에서 자료가 발생 횟수로 얻어지는 경우 일반적으로 자기 포아송 모형(Auto-Poisson Model : Cressie 1993)을 고려하게 되며 이 때의 모형은 다음과 같다.

$$P(Y_i = y_i | Y_j = y_j, i \neq j) = \frac{e^{-\mu_i(y_i, i \neq j)} \mu_i(y_i, i \neq j)^{y_i}}{y_i!}, y_i = 0, 1, 2, \cdots \infty \qquad (2.1)$$

여기서 Y_i는 지역 i에서의 사건 발생 건수가 되며 $\mu_i(y_i, i \neq j)$는 y_i의 함수인 $E(Y_i) = \mu_i$이다. 따라서 분석에 사용되는 모형은 다음과 같다.

$$E(Y_i) = \mu_i(y\{s_i, i \neq j\}) = exp(\alpha_i + \sum_{j=1}^{n} \rho_{ij} y_j) \qquad (2.2)$$

여기서 $y(s_i)$는 i 지역의 y 값이 공간통계변수 s_i 의 함수형태로 표현되며

$$\rho_{ij} = \begin{cases} 1, & \text{지역 } i,\ j\text{가 이웃하는 경우} \\ 0, & \text{그렇지 않은 경우} \end{cases}$$

이다. 본 논문에서는 위의 모형을 바로 적용하지 않고 베이지안적 접근법인 베이지안 자기 포아송 모형(Bayesian Auto Poisson Model : BAPM)을 이용하였다. 즉 BAP 모형은 SAR 모형을 포아송 분포에서 정의된 평균 모수인 μ_i의 사전함수로 이용하였다.
그러므로 BAP 모형은 다음과 정의할 수 있다. 즉 포아송 분포를 따르는 확률변수, Y_i는 다음과 같다.

$$(Y_i|\mu_i) \sim Poi(\mu_i),\ i = 1, \cdots, I \tag{2.3}$$

$$\mu_i = N_i\theta_i \tag{2.4}$$

여기서 N_i는 i지역의 모집단 크기이고 θ_i는 i지역 특성치의 비율이다. 이때 모형에 있는 θ_i에 대한 사전분포함수는 다음과 같이 SAR 모형의 형태를 갖는 일반 선형모형을 이용한다. 즉

$$\log(\theta_i) = X_i\beta + S_i\gamma_i + \nu_i \tag{2.5}$$

여기서 X_i : i 지역의 설명변수 백터, β : 설명변수의 계수, S_i : i 지역의 랜덤효과를 고려한 공간통계 변수, 그리고 γ_i: 랜덤효과 변수의 계수이다. 또한 $\nu_i \sim N(0, \sigma_\nu^2)$이라 가정하고 γ_i와 ν_i는 서로 독립이라 가정한다. 다음으로 $\tau_\nu = 1/\sigma_\nu^2$와 γ_i와 β의 모호 사전분포를 다음과 같이 정한다.

$$\begin{aligned} \tau_\nu &\sim Gam(p_0,\ q_0) \\ \beta &\sim N(0,\ \sigma_{\beta_0}) \\ \gamma_i &\sim N(0,\ \sigma_{\gamma_0}) \end{aligned}$$

여기서 p_0, q_0, σ_{γ_0}, σ_{β_0}들은 알려진 초 모수 값들이다. 위의 가정을 이용하게 되면 μ_i의 추정량, $\hat{\mu}_i$,는 N_i 와 $\hat{\theta}_i$ 의 곱으로 구하게 되며 $\hat{\theta}_i$은 베이즈 추정량으로 이의 계산을 위한 θ_i의 사후분포 $h(\theta_i|y_i)$는 다음과 같다.

$$h(\theta_i|y_i) \ \propto\ f(y_i|\theta_i)\ g(\theta_i) \tag{2.6}$$

여기서 $f(*)$는 우도함수, $g(\theta_i)$는 사전함수로 θ_i의 베이지안 예측확률함수의 형태로 사용되며 $g(\theta_i|X_i, S_i)$ 로 표현될 수 있다. 이는 다음과 같다.

$$g(\theta_i|X_i, S_i) = \int_\beta \int_\gamma \int_{\tau_\nu} l(\theta_i|X_i, S_i, \beta, \gamma, \tau_\nu) h(\beta, \gamma, \tau_\nu|X_i, S_i) \partial\beta\partial\gamma\partial\tau_\nu \tag{2.7}$$

여기서

$$l(\theta_i|X, S) \propto \tau_\nu exp[-\tau_\nu(log(\theta_i) - X_i\beta - S_i\gamma_i)^2]$$

$$h(\beta, \gamma, \tau_\nu) = exp[-\beta^2]exp[-\gamma_i^2]\tau_\nu^{(p_0-1)}exp[\frac{-\tau_\nu}{p_0}]$$

이제, μ_i의 추정량, $\hat{\mu}_i$ 은

$$\hat{\mu}_i = N_i\hat{\theta}_i$$

이고 여기서 $\hat{\theta}_i$ 는 다음과 같다.

$$\hat{\theta}_i = E(\theta_i|X_i,\ S_i) \tag{2.8}$$

참고로 베이즈 추정량은 소지역 i의 추정량이 되며 본 논문에서는 베이지안 통계 패키지인 윈벅(Winbugs)에서 MCMC를 이용하여 구하였다.
다음으로 신기일과 이상은(2003)에서 사용된 SAR 모형을 간단히 소개하기로 한다. 이 논문에서는 Freeman-Tukey 변환이 사용되었다. 먼저 Y_i를 Freeman-Tukey 변환된 반응변수라 하면 사용된 SAR 모형은 다음과 같다.

$$Y_i = X_i\beta + \rho_i S_i + \nu_i \tag{2.9}$$

여기서 X_i : i 지역의 설명변수, β : 설명변수의 계수, S_i: i지역의 랜덤효과를 고려한 공간통계 변수, 그리고 ρ_i : 랜덤효과 변수의 계수이다. 그리고 $\nu_i \sim N(0,\ \sigma_\nu^2)$라 가정하고 ρ_i와 ν_i는 서로 독립이라 가정한다. 소지역 추정량은 주어진 자료에 의해 계수들을 추정한 후 이에 따른 예측값을 각 소지역 i의 추정량으로 하였다. 전술한 것처럼 신기일과 이상은(2003)의 SAR 모형은 공간상관관계가 존재하는 경우 공간통계량을 설명변수로 모형에 활용하는 방법이다.

3. 자료분석

3.1. 자료 설명

이 논문에서는 호주에서 1998년에 이루어진 장애인 통계(Survey of Disability, Aging and Cares : SDAC)에서 장애인 수를 반응변수로 이용하였다. 전국에 걸쳐 43,000명을 대상으로 조사가 이루어졌으며 장애인은 5개의 변수로 분류된다. 1: Physical(PHYS), 2: Sensory (SENS), 3: Intellectual(INTEL), 4: Psychological(PSYC), 5: Head (HEAD). 그러나 분석에 사용된 장애인 수는 전체의 장애인 총수인 Disabled(DIS) 변수가 사용되었다. 즉

$$DIS_i = PHYS_i + SENS_i + PSYC_i + INTEL_i + HEAD_i \tag{3.1}$$

설명변수로는 다양한 행정통계에서 장애인 통계를 설명할 수 있는 변수들로 연령별 인구수가 얻어졌다. 또한 도시와 소지역이 얼마나 떨어져 있는 가를 표시한 remonteness 변수가 있다. 다음 표 3.1과 표 3.2는 반응변수와 설명변수들을 설명한 것이다.

표 3.1: 장애인 형태에 따른 지역별 장애인수 비율

		장애인 형태					
주(state)	ssd	PHYS	SENS	INTEL	PSYC	HEAD	DISABLED
1	10505	0.10243	0.0226	0.0037	0.0094	0.0229	0.1612
1	10510	0.05936	0.0224	0	0.0056	0.0058	0.0933
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	80540	0.1084	0.0649	0	0	0.0181	0.1915
8	81005	0.0917	0.0148	0	0.0047	0.0306	0.1419

표 3.2: 지역별 보조변수 보조

		보조 변수				
주(state)	ssd	PF00	PF01	···	Intel_dsp	sens_dsp
1	10505	24665	59181	···	981	269
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	80540	2971	3423	···	448	23
8	81005	16008	15548	···	1487	11

이 자료에서의 소지역은 주(state)단위이며 지역 자리수의 첫 자리 숫자가 된다. 지역 자리수의 처음 5 자리의 수 즉,10505, 10510.. 등은 ssd 지역단위를 나타낸다. ssd는 주 안의 있는 작은규모 즉 "구" 단위와 같은 규모의 지역 단위이다. 자료는 총 8개의 소지역 자료로 이루어져 있으며 각각의 소지역에는 ssd가 각각 48개, 45개, 30개, 19개, 28개, 8개, 10개, 8개로 구성되어있다. 이 때 사용가능한 자료는 ssd레벨의 장애인수의 비율과 인구수 그리고 보조변수들의 자료는 행정자료에서 얻은 ssd레벨의 총 수이며 이를 구체적인 자료 설명은 생략하기로 한다. 일반적으로 i번째 ssd의 총 장애인 비율은 10% 에서 20%미만으로 장애인 발생 건수는 매우 드물게 나타난다. 이는 현 우리나라의 실업률 통계와 매우 흡사 한 특성을 나타낸다.

전술한 데로 본 논문에서 사용된 반응변수는 총 장애인 수, 즉 각 소지역의 5개의 범주에 속하는 모든 장애인 수의 합으로 하였다. 소지역의 장애인수를 추정하는 데 있으며 사용 가능한 설명변수는 표 3.2의 행정자료 외에 공간통계가 사용될 수 있다. 따라서 공간통계의 활용을 가능한가를 살펴 보기 위해 공간상관관계를 보기로 하자.

3.2. 공간 상관관계(Spatial correlation : Moran's I)

공간통계를 설명변수로 사용하기 위해 우선 공간통계의 사용이 적절한가를 보기로 한다. 일반적으로 공간상관관계를 측정하는 여러 가지 통계 중에서 Moran's I가 주로 사용되며 이 장에서도 Moran's I를 이용하여 공간 통계 활용의 타당성을 보도록 하자. Cressie(1993)의 Moran's Index는 다음과 같으며 총 장애인수(DIS)인 반응변수의 Moran's I 값은 S-PLUS를

이용하여 구해질 수 있으며 그 결과는 표 3.3과 같다.

$$\text{Moran's Index} = \text{I} = \frac{n\sum_i\sum_j \delta_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{2T\sum_i (Y_i - \bar{Y})^2}$$

여기서 δ_{ij}은 지역 i와 j의 이웃 관계를 나타내는 지시함수(Indicator function)이다.

표 3.3: DISABLED 변수의 Moran's I

Statistics	Estimates	Std. error	p-value
Morans' I	0.178	0.05065	2.998e-4

표3.3의 Moran's I를 살펴보면 공간상관관계를 존재함을 확인할 수 있으며 따라서 이는 공간통계 분석 기법을 이용할 수 있다는 근거를 주고 있다.

3.3. 베이지안 자기 포아송 모형(Bayesian Auto Poisson Model : BAPM)

앞선 2장에서 소개한 베이지안 자기 포아송 모형을 고려하자. 먼저 DIS_i를 소지역 i의 총 장애인수라 하면

$$DIS_i \sim Poi(\mu_i),\ i = 1, \cdots, I \tag{3.2}$$

이 된다. 이때

$$E(DIS_i) = \mu_i = N_i\theta_i \tag{3.3}$$

이다. 여기서 N_i는 소지역 i의 총 인구수이고 θ_i는 소지역 i의 장애인 비율이다. 다음으로 θ_i의 사전함수로 SAR 모형을 적용하기로 한다. 이는 공간상관관계의 유의성에 따라 일반 선형모형에 공간상관변수를 활용한 SAR 모형을 선택했기 때문이다. 또한 모형설정을 위해 가능한 행정자료인 각 연령별 성별인구수 등을 최대한 활용하기 위해 step-wise로 모형에 사용할 수 있는 5개의 보조변수가 선택되었다. 이 논문에서 설명변수의 선택에는 큰 비중을 두지 않았다. 우선은 가능한 행정자료에서 반응변수와 상관관계성에 유의성을 가진 변수를 찾기 어려웠으므로 공간통계량외의 설명변수는 가능한 변수들 중 유의성이 높은 변수들을 선택하여 모형식에 추가하였으며 이에 따라 5개의 설명변수가 선택되었다. REMOTE, PF00, PF02, PM00, PM03. 각 변수들의 의미는 모형식에서 설명되었다. 참고로 설명변수의 선택은 전국 자료를 기준으로 하였으며 이 변수들이 소지역 모형에 적용되었다. 또한 결과의 비교를 위해 모형을 각 소지역별로 고정 하였다.
이제 SAR를 모수 θ_i에 적용하면 다음과 같다.

$$\log(\theta_i) = \beta_1 Remote_i + \beta_2 F00_i + \beta_3 F02_i + \beta_4 M00_i + \beta_5 M03_i + \rho_i^* S^*_{DIS_i} + \nu_i \tag{3.4}$$

여기서

$Remote_i$: 소지역 i가 행정 중심가와 어느 정도 떨어져있는가를 나타내는 remoteness 변수,
$F00_i$: 소지역 i의 20세 미만의 여자 인구 비율,
$F02_i$: 소지역 i의 20세 이상 40세 미만의 여자 인구 비율,
$M00_i$: 소지역 i의 20세 미만의 남자 인구 비율,
$M03_i$: 소지역 i의40세 이상 65세 미만의 남자 인구 비율,
$S^*_{DIS_i} = \sum_k DIS_k/n_k$: 소지역$i$ 주변에 있는 모든 소지역의 모든 타입의 장애인수 (DISABLED)의 평균이며 이때 랜덤 효과의 계수는 알려진 행렬이다. 또한 $\beta_j, j = 1, \cdots, 5$은 회귀계수이고, ρ_i^*은 랜덤효과 변수로 $S_{DIS_i^*}$의 계수이며 $\nu_i \sim N(0, \sigma_\nu^2)$라 가정한다. 그리고 $\tau_\nu = 1/\sigma_\nu^2$라 할 때 주어진 모호 사전함수는 다음과 같다.

$$\tau_\nu \sim Gam(0.001, 0.001) \tag{3.5}$$

$$\beta_j \sim N(0.0, 0.01),\ j = 1, \cdots, 5 \tag{3.6}$$

$$\rho_i^* \sim N(0.0, 0.01) \tag{3.7}$$

윈벅(Winbugs)의 MCMC에 의해 총 장애인수의 베이즈 추정량, $\hat{\mu}_i = N_i\hat{\theta}_i$는 $E(\theta_i|X_i, S_i) = \hat{\theta}_i$이 계산되어지며 이는 소지역 i의 추정량이 된다.

3.4. 공간통계의 SAR 모형

신기일과 이상은(2003)에서 소개한 모형에서 SAR 모형은 다음과 같다.

$$DIS_{Z_i} = \beta_1 Remote_i + \beta_2 F00_{zi} + \beta_3 F02_{zi} + \beta_4 M00_{zi} + \beta_5 M03_{zi} + \rho_i^* S^*_{DIS_{zi}} + \nu_i$$

여기서

DIS_{zi} : Freeman-Tuckey 변환한 지역 i의 총 장애인수,
$Remote_i$: 소지역 i가 행정 중심가와 어느 정도 떨어져있는가를 나타내는 remoteness 변수,
$F00_{zi}$: Freeman-Tuckey 변환한 소지역 i 의 20세 미만의 여자 인구수,
$F02_{zi}$: Freeman-Tuckey 변환한소지역 i의 20세 이상 40세 미만의 여자 인구수,
$M00_{zi}$: Freeman-Tuckey 변환한 소지역 i의 20세 미만의 남자 인구수,
$M03_{zi}$: Freeman-Tuckey 변환한 소지역 i의40세이상 65세 미만의 남자 인구수,
$S^*_{DIS_{zi}} = \sum_k DIS_{zk}/n_k$: 소지역 i주변에 있는 모든 소지역의 모든 타입의 장애인수 (DISABLED)의 평균이며 이때 랜덤 효과의 계수는 알려진 행렬이다. 또한 $\beta_j, j = 1, \cdots, 5$은 회귀계수이고, ρ_i^*은 랜덤효과 변수로 $S_{DIS_i^*}$의 계수이다. 참고로 여기에서 사용한 설명변수 역시 BAPM에서 선택한 변수들과

동일하게 하였다. 이는 두 모형의 비교를 위해 사용되는 설명변수를 동일하게 하였다.

4. 결론과 향후 연구

이 장에서는 우선 BAP 모형과 SAR 모형에서 얻어진 8개의 소지역의 추정량을 MSE와 MB를 이용하여 비교하였다. 여기서 사용된 통계량은 다음과 같다.

$$MSE^{1/2} = [\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2]^{1/2} \tag{4.1}$$

$$MB = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) \tag{4.2}$$

표 4.1: SAR와 BAPM의 MSE, MB 값

		SAR		BAPM	
state	# of ssd	$MSE^{1/2}$	MB	$MSE^{1/2}$	MB
1	49	22411.14	2737.99	6006.93	-3685.17
2	45	10211.48	-606.43	6252.30	-3927.34
3	30	14983.78	8128.80	3612.30	-743.09
4	19	16323.69	5818.20	2793.11	854.47
5	28	6176.77	2077.95	3620.73	-1711.43
6	8	7561.42	6641.27	2313.34	1308.06
7	10	1353.55	-532.89	1947.46	-1702.75
8	8	1353.55	-1179.87	2837.76	-2636.57
전국		14719.60	2832.06	4564.41	-2228.11

표 4.1을 살펴보면 MSE를 기준으로 하였을 때 BAP 모형이 매우 효율적임을 알 수 있다. 그러나 MB의 경우는 다양하게 나타난다. 이는 BAP 모형이 소지역 추정의 분산을 줄이는 데는 효율적이나 편의를 줄이는 데는 크게 효율적이지 않다는 것을 나타내고 있다. 여기서 Brown 등(2001)이 제안한 편의 진단(bias diagnostic)을 보기로 하자. Brown 등(2001)에서 사용된 기본 개념은 직접 추정량이 불편성을 만족하므로 모형기반 추정량의 추정값과 직접추정값이 거의 같게 추정되어야 한다는 것이다. 따라서 모형기반 추정값과 직접 추정값을 그래프로 그린 후 회귀직선을 적합하면 기울기가 "1"이고 절편이 "0"인 직선이 얻어 지는지 판단을 하면 되는 것이다. 다음의 표 4.2는 Y_i와 $\hat{Y}_i$의 절편이 없는 회귀식에 적합시켜 얻은 결과이다.

표 4.2: Y_i와 $\hat{Y}_i$의 회귀식에서의 회귀계수와 결정계수

		SAR		BAPM	
state	♯ of ssd	slop(std)	R^2	slop(std)	R^2
1	49	1.012(0.045)	0.93	0.841(0.027)	0.95
2	45	0.862(0.045)	0.89	0.736(0.031)	0.92
3	30	1.038(0.028)	0.97	0.889(0.020)	0.98
4	19	1.086(0.038)	0.97	0.983(0.030)	0.98
5	28	0.987(0.022)	0.98	0.832(0.024)	0.97
6	8	1.064(0.034)	0.99	0.989(0.079)	0.95
7	10	0.939(0.059)	0.96	0.586(0.146)	0.64
8	8	0.914(0.092)	0.93	0.698(0.090)	0.89
전국		0.993(0.018)	0.93	0.840(0.013)	0.95

표 4.2에서 SAR의 경우 '1'에 가까운 기울기(slop)를 가졌으므로 편의 진단에서는 SAR 모형이 BAP 모형에 비해 효율적임을 알 수 있다. 그러나 R^2값을 살펴보면 BAP 모형이 SAR 모형을 이용한 결과보다 대부분 큰 값을 갖고 있어 BAP 모형이 더 효율적임을 알 수 있다. 결론적으로 표 4.1과 표 4.2를 동시에 고려하면 BAP 모형을 사용하게 되면 추정량의 분산을 줄일 수 있으나 편의는 크게 효과를 볼 수 없음을 알 수 있다. 따라서 BAP 모형의 편의를 줄이는 문제를 향후 연구과제로 두기로 한다.

참고문헌

김재두, 이상은, 신기일 (2004). Small Area Estimation Using STAR model, *2004년 춘계 학술발표회논문집* , 285-289.

신기일, 이상은 (2003). Model -Data Based Small Area Estimation, *The Korean Communications in Statistics*, **10**. No, 3. 37-645.

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001). Evaluation of Small Area Estimation Methods-Application to Unemployment estimations form UK LFS, *Proceedings of Statistics Canada Symposium 2001.*

Cressie, N. (1993). Statistics for Spatial Data, *John Wiley and Sons, INC.*

Freeman, M. F and Tukey, J. W. (1950). Transformation Related to the Angular and the Square Root, *Annals of Mathematical Statistics,* **21**. 607-611.

Australia Bureau of Statistics (2003). Small Area Estimation Models for Disability Methodology Advisory Committee 21, November 2003.

[2006년 2월 접수, 2006년 8월 채택]

Small Area Estimation Using Bayesian Auto Poisson Model with Spatial Statistics*

Sang Eun Lee[1)]

ABSTRACT

In sample survey sample designs are performed by geographically-based domain such as countries, states and metropolitan areas. However mostly statistics of interests are smaller domain than sample designed domain. Then sample sizes are typically small or even zero within the domain of interest. Shin and Lee(2003) mentioned Spatial Autoregressive(SAR) model in small area estimation model-based method and show the effectiveness by MSE. In this study, Bayesian Auto-Poisson Model is applied in model-based small area estimation method and compare the results with SAR model using MSE ME and bias check diagnosis using regression line. In this paper Survey of Disability, Aging and Cares(SDAC) data are used for simulation studies.

Keywords: Bayesian Auto Poisson Model, Spatial Autoregressive model, Spatial Correlations.

* This research was supported by Kyonggi University Aboard Studies Funds

1) Associate Professor, Dept. of Applied and Information Statistics, Kyonggi University, Suwon Si, Kyonngi Do, 442-720, Korea.
E-mail : sanglee62@kyonggi.ac.kr