The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 781-790

Hierarchical Bayesian Analysis of Spatial Data with Application to Disease Mapping[1)]

Dal Ho Kim[2)], Sang Gil Kang[3)]

Abstract

In this paper, we consider estimation of cancer incidence rates for local areas. The raw estimates usually are based on small sample sizes, and hence are usually unreliable. A hierarchical Bayes generalized linear model is used which connects the local areas, thereby enabling one to 'borrow strength'. Random effects with pairwise difference priors model the spatial structure in the data. The methods are applied to cancer incidence estimation for census tracts in a certain region of the state of New York.

1. 서 론

최근 암이나 백혈병 같은 특정 질병에 대한 발병률이나 사망률의 지역적 분포를 표시한 질병지도의 작성이 역학연구에서 매우 중요한 관심사이다. 여러 지리적 지역들에 걸친 상대위험도(relative risk)에 대한 질병지도 작성이 특정 질병이 다발적으로 발생한 집락을 찾거나 특정 질병을 유발하는 환경적 결정자를 파악하는데 도움이 되기 때문이다.

상대위험도에 대한 전통적인 추정치는 최우추정법에 근거한 표준화된 사망률비(standardized mortality ratio; SMR)이다. 그러나 질병지도 작성시 연구지역(study area)이 보다 작은 국소적 지역(local areas)으로 나뉘어 지면서 단지 소수의 사람들만이 특정 질병에 걸린 것으로 조사된 지역이 많아지게 되었다. 따라서 원추정치인 SMR은 표준오차와 변동계수가 커지게 되어 신뢰성을 잃게 된다. 이는 흔하지 않은 질병(rare disease)의 경우 더욱 그러하다. 또한 SMR은 조사대상 지역의 공간적 패턴(spatial patterns)을 전혀 고려할 수 없다.

따라서 각 국소적 지역에 대한 상대위험도를 평활(smoothing)하기 위해서 이웃한 지역들(neighboring areas)로부터 정보를 "borrow strength" 할 필요가 있다. 이러한 필요를 가장 잘 충족시키는 방법이 계층적 베이즈(hierarchical Bayes; HB) 및 경험적 베이즈(empirical Bayes; EB) 절차들이다. 왜냐하면 이들 방법들이 모형을 통하여 국소적 지역들을 체계적으로 연결시켜주기 때문이다.

상대위험도를 평활하기 위한 두 베이지안 절차들의 닮은 점은 흔히 초모수(hyperparameter)라 불리우는 사전분포의 모수들을 알지 못하는데 기인한 불확실성을 인지하는 것이다. 그러나 EB 방

1) 이 논문은 1998년 한국학술진흥재단의 학술연구비에 의하여 지원되었음.
2) Assistant Professor, Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea.
3) Lecturer, Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea.

법은 관측치들의 주변분포로부터 초모수들을 추정하는 반면, HB 방법은 다음 단계에서 초모수들에 대한 사전분포를 부여한다. 이미 잘 알려진 대로 소박한(naive) EB방법으로 추정된 사후분산은 초모수의 추정에 기인하는 추가 변이를 제대로 반영하지 못한다.

공간자료의 분석에 베이지안 방법이 선호되는 또 하나의 다른 이유는 베이지안 방법을 사용하면 자료에 주어진 공간구조를 사전분포로 쉽게 모형화 할 수 있다는 것이다. 예컨대 로그 상대위험도에 대해 Clayton과 Kalder(1987)은 조건부 자기회귀(conditional autoregression; CAR) 모형을 사전분포로 고려했고, Cressie(1992)는 랜덤 마르코프 장(random Markov field) 성질을 가지는 가우스모형을 사전분포로 사용했다.

본 논문의 목적은 질병지도 작성을 위한 질병의 상대위험도 추정문제에 대해 일반화 선형모형(GLM)하에서 자료에 주어진 공간구조를 사전분포로 모형화하는 계층적 베이즈 공간모형(HB spatial model)을 도입하여 후리컨티스트(frequentist) 추론에 대응하는 완전히 베이지안적(fully Bayesian) 추론을 제안하고, HB 방법에서 수반되는 고차원 적분 문제에 대한 해결책으로서 깁스 표집기를 사용한 통계적 추론을 연구하고자 한다.

공간효과를 모형화하는 시도 가운데 우리의 관심을 끄는 것은 최근 영상분석에서의 베이지안 이론을 질병지도 문제에 접목시킨 Besag et al.(1991)과 Besag et al.(1995)이다. 그들은 각 지역 로그 상대위험도를 국소적 집락을 나타내는 구조적 이질성(structured heterogeneity)과 전혀 구조를 가지지 않은 이질성(unstructured heterogeneity)의 합으로 표현할 것을 제안했다. 이러한 제안에 공변량(covariates)의 사용에 대한 최근의 연구결과를 합하여 우리의 사전분포를 표현하고, 이 사전분포를 사용한 구체적 공간자료에 대한 계층적 베이지안 분석이 본 논문의 주된 관심사이다.

공간적 자료의 분석에 대한 통계적 모형과 후리컨티스트적 방법론에 관한 연구 결과는 최근 Cressie(1991)에 매우 잘 소개되어 있다. 최근 표본조사에서의 일반화 선형모형(GLM)을 사용한 소지역 추정(small area estimation) 문제에 대한 활발한 연구가 공간자료에 관한 연구로 연결되어 있는데, 이는 공간적 자료에서 특정 질병의 발병률에 관한 문제가 본질적으로 소지역 추정 문제에서 국소적 지역 혹은 소지역 추정에 관한 문제와 유사하기 때문이다. (Ghosh와 Rao(1994), Ghosh et al.(1998)). 그러나 소지역 추정과는 달리, 질병지도에서는 공간적인 구조와 더불어 구조화 할 수 없는 이질성도 중요한 역할을 한다.

자료가 가지는 공간적 패턴에 대한 모형화는 여러 가지로 형태로 여러 연구에서 제안되었다. Cressie와 Chan(1989) 그리고 Cressie(1992)는 랜덤 마르코프 장 성질을 가지는 가우스모형을 우도 및 사전분포에 각 각 사용하였고, Clayton과 Kaldor(1987)는 로그 상대위험도에 대해 조건부 자기회귀(CAR) 모형를 사전분포에 사용하였다. 또한 영상분석에서의 베이지안 이론과 질병지도와의 관련성에 근거하여 Besag et al.(1991)과 Besag et al.(1995)는 국소적 집락과 구조를 가지지 않은 이질성의 합으로 로그 상대위험도를 표현하여 프랑스와 영국에서의 암 사망률에 관한 세 자료를 분석하였다. 나아가 Besag et al.(1995)는 공간적 자료뿐만 아니라 여러 베이지안 응용분야에 사용할 수 있는 매우 일반적인 "짝진 차이 사전분포(pariwise difference prior)" 를 제안했다.

최근 베이지안 추론에서의 다차원 적분 문제가 MCMC(Markov Chain Monte Carlo) 방법에 의해 해결됨으로 짝진 차이 사전분포를 사용한 복잡한 모형에서도 베이지안 접근이 가능하게 되었다. 그러나 모형이 복잡한 관계로 다소 어렵지만 베이지안 추론에 쓰이는 사후분포가 진(proper)임을 밝히는 것 또한 우리의 중요한 관심사이다.(Natarajan과 McCulloch(1995), Ghosh et

al.(1998)).

본 논문의 대략적 개괄은 다음과 같다. 2절에서 일반적인 계층적 베이즈 공간적 일반화 선형모형을 도입하고, 비정보적(noninformative) 사전분포 하에서 사후분포가 진(proper)일 충분조건을 찾는다. 3절에서는 현실 자료를 사용하여 계층적 베이즈 분석을 예시하고, 후리컨티스트 추론의 결과와 비교한다.

2. 계층적 베이즈 공간적 일반화 선형모형

공간구조를 모형화 할 필요가 있는 여러 경우를 포함하는 매우 일반적인 다음과 같은 계층적 베이즈 공간적 일반화 선형모형(hierarchical Bayes spatial GLM)을 생각하자.

(I) $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)^T$가 주어졌을 때, $Y_1,\ldots,Y_m$은 서로 독립이면서 다음과 같은 밀도함수를 가진다고 하자.

$$f(y_i|\theta_i)=\exp(y_i\theta_i-\psi_i(\theta_i))h(y_i).$$

(II)
$$\theta_i=q_i+\boldsymbol{x}_i^T\boldsymbol{b}+u_i+v_i \quad (i=1,\ldots,m), \tag{2.1}$$

여기서 q_i는 알려진 상수이고, u_i와 v_i는 서로 독립이면서 $v_i \overset{iid}{\sim} N(0,\sigma_v^2)$이고 u_i는 다음과 같은 결합 밀도함수를 가진다.

$$f(\boldsymbol{u}) \propto (\sigma_u^2)^{-m/2}\exp\Big[-\sum\sum_{i\neq j}(u_i-u_j)^2w_{ij}/(2\sigma_u^2)\Big], \tag{2.2}$$

여기서 모든 $1\leq i\neq j\leq m$에 대해서 $w_{ij}\geq 0$이다.

(III) $\boldsymbol{b}$, σ_u^2 그리고 σ_v^2는 서로 독립이면서 $\boldsymbol{b}\sim$ Uniform (R^p) $(p<m)$, $(\sigma_u^2)^{-1}\sim$ Gamma $(a/2,g/2)$ 그리고 $(\sigma_v^2)^{-1}\sim$ Gamma $(c/2,d/2)$를 따른다고 하자. 여기서 Gamma (α,β)는 밀도함수가 $f(z)\propto\exp(-\alpha z)z^{\beta-1}$ 형태임을 의미한다.

여기서 우리의 관심은 $\boldsymbol{y}=(y_1,\ldots,y_m)^T$가 주어졌을 때, $\boldsymbol{\theta}$와 $\boldsymbol{b}$의 사후분포와 나아가 이들 사후분포들의 사후 평균, 분산 및 공분산을 찾는 것이다.

위의 모형 (I)부분에서 다음과 같은 세 경우가 특히 우리의 관심거리이다. 첫째, Y_i가 독립인 Bin (n_i,p_i)인 경우이다. 이 경우 $\theta_i=\log(p_i/(1-p_i))$ 그리고 $\psi(\theta_i)=n_i\log(1+\exp(\theta_i))$이다. 둘째, Y_i가 독립인 Poisson (λ_i)인 경우, $\theta_i=\log\lambda_i$ 그리고 $\psi(\theta_i)=\exp(\theta_i)$이다. 셋째, $Y_i\sim N(\theta_i,1)$인 경우, $\psi(\theta_i)=\theta_i^2/2$이다.

위의 모형 (II)부분은 오프셋(offset) 모수 q_i를 포함한다. 그리고 $\boldsymbol{u}$에 가정된 사전분포는 Besag et al.(1995)에서 주어진 분포의 한 특별한 경우이다. 왜냐하면 Besag et al.(1995)은 $(u_i-u_j)^2$ 대신에 대칭인 임의의 함수 ψ를 사용한 $\psi(u_i-u_j)$를 생각했기 때문이다. 이러한 u_i의 결합분포는 "짝진 차이 사전분포(pariwise difference prior)" 라 불리운다. 실제 응용에서 흔히 w_{ij}는 이웃한 지역에 대해서는 1, 그렇지 않으면 0를 사용한다.

위에서 제안된 사전분포 모형은 지리적 구조를 반영한다. 구체적으로, θ_i의 추정치들은 이웃한 지역들로부터 강하게 영향을 받지만, 모든 다른 지역들의 추정치들로는 단지 간접적으로 영향을 받는다. 그 결과 각 추정치들은 전체(global) 평균값보다는 국소적(local) 평균값을 향하여 더욱 축소(shrinking)된다. 또한 공변량의 포함으로 이웃한 지역들간의 교환가능(exchangeability)은 성립하지 않는다.

위의 모형 (III)부분은 회귀계수에게 균등(uniform) 사전분포를, 그리고 분산 성분들에게 역 감마(inverse gamma) 사전분포들을 독립적으로 주는 것으로 이는 계층적 베이즈 분석에서는 매우 자연스러운 것이라 할 수 있다.

위에서 제안된 모형 (I) - (III)하에서의 통계적 추론을 위해 $\boldsymbol{y}$가 주어졌을 때 $\boldsymbol{\theta}$의 사후분포가 진(proper)임을 이론적으로 점검할 필요가 있다. 이에 대한 충분조건으로 주어진 $\boldsymbol{y}$에 대해 $\boldsymbol{\theta}$, $\boldsymbol{y}$, σ_u^2 그리고 σ_v^2의 결합 사후분포(joint posterior)가 진임을 밝히고자 한다.

정리 1. $f(y_i|\theta_i)$가 모든 i에 대해 유계(bounded)임을 가정하자. 또한 $y_{i_1},\cdots,y_{i_n}$ $(1<i_1<\cdots<i_n\le m;\ p\le n\le m)$가 존재하여 $j=1,\cdots,n$에 대해 $\int_{-\infty}^{\infty}\exp[\theta y_{i_j}-\phi(\theta)]d\theta<\infty$이고 그에 대응하는 설명 변수 $\boldsymbol{x}_{i_1},\cdots,\boldsymbol{x}_{i_n}$로 구성된 행렬 $\boldsymbol{X}_{i*}=(\boldsymbol{x}_{i_1}-\bar{\boldsymbol{x}}_{i\cdot},\cdots,\boldsymbol{x}_{i_n}-\bar{\boldsymbol{x}}_{i\cdot})^T$ $(\bar{\boldsymbol{x}}_{i\cdot}=n^{-1}\sum_{j=1}^{n}\boldsymbol{x}_{ij})$가 완전계수(full rank) p를 가진다고 가정하자. 그러면, 만약 $a>0$, $c>0$, $m+g>0$ 그리고 $n+d>0$이면, 결합 사후분포 $p(\boldsymbol{\theta},\boldsymbol{b},\sigma_u^2,\sigma_v^2|\boldsymbol{y})$가 진(proper)이다.

<**증명**> 일대일 변환 $z_i=u_i-u_m$ $(i=1,\cdots,m-1)$를 사용하여 $\boldsymbol{z}=(z_1,\cdots,z_{m-1})^T$으로 나타내면, 주어진 $\boldsymbol{y}$에 대한 $\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{z},u_m,\gamma_u,\gamma_v$의 결합 사후분포는 다음과 같다.

$$\begin{aligned}p(\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{z},u_m,\gamma_u,\gamma_v|\boldsymbol{y})\propto\ &\exp\Big[\sum_{i=1}^{m}\{y_i\theta_i-\phi(\theta_i)\}\Big]\\&\times\gamma_v^{(1/2)m}\exp\Big[-\frac{1}{2}\gamma_v\sum_{i=1}^{m}(\theta_i-q_i-\boldsymbol{x}_i^T\boldsymbol{b}-z_i-u_m)^2\Big]\\&\times\gamma_u^{(1/2)m}\exp\Big[-\frac{1}{2}\gamma_u\sum\sum_{1\le i\ne j\le m}(z_i-z_j)^2w_{ij}\Big]\\&\times\exp\Big(-\frac{1}{2}a\gamma_u\Big)\gamma_u^{(1/2)g-1}\exp\Big(-\frac{1}{2}c\gamma_v\Big)\gamma_v^{(1/2)d-1}.\end{aligned}$$

여기서 $z_m=0$이다. 한편, 일반성을 잃지 않고 $i_j=j$ $(j=1,\cdots,n)$라 가정한다. 또한 $\boldsymbol{\theta}_*=(\theta_1,\cdots,\theta_n)^T$라 하고 $\boldsymbol{X}_*=\boldsymbol{X}_{i*}$로 나타내자. 여기서 $\theta_{n+1},\cdots,\theta_m$에 관해 적분하면, 주어진 $\boldsymbol{y}$에 대한 $\boldsymbol{\theta}_*,\boldsymbol{b},\boldsymbol{z},u_m,\gamma_u,\gamma_v$의 결합 사후분포는 다음과 같다.

$$p(\boldsymbol{\theta}_*, \boldsymbol{b}, \boldsymbol{z}, u_m, \gamma_u, \gamma_v|\boldsymbol{y}) \propto \exp\Big[\sum_{i=1}^{n}\{y_i\theta_i-\psi(\theta_i)\}\Big] \times \gamma_v^{(1/2)n}\exp\Big[-\frac{1}{2}\gamma_v\sum_{i=1}^{n}(\theta_i-q_i-\boldsymbol{x}_i^T\boldsymbol{b}-z_i-u_m)^2\Big] \times \gamma_u^{(1/2)m}\exp\Big[-\frac{1}{2}\gamma_u\sum\sum_{1\le i\ne j\le m}(z_i-z_j)^2w_{ij}\Big] \times \exp\Big(-\frac{1}{2}a\gamma_u\Big)\gamma_u^{(1/2)g-1}\exp\Big(-\frac{1}{2}c\gamma_v\Big)\gamma_v^{(1/2)d-1}$$

여기서 $\bar{\theta}=n^{-1}\sum_{i=1}^{n}\theta_i$, $\bar{z}=n^{-1}\sum_{i=1}^{n}z_i$, $\bar{\boldsymbol{x}}=n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i$, $\bar{q}=n^{-1}\sum_{i=1}^{n}q_i$로 나타내고 u_m에 관해 적분하면, 주어진 $\boldsymbol{y}$에 대한 $\boldsymbol{\theta}_*, \boldsymbol{b}, \boldsymbol{z}, u_m, \gamma_u, \gamma_v$의 결합 사후분포는 다음과 같다.

$$p(\boldsymbol{\theta}_*, \boldsymbol{b}, \boldsymbol{z}, \gamma_u, \gamma_v|\boldsymbol{y}) \propto \exp\Big[\sum_{i=1}^{n}\{y_i\theta_i-\psi(\theta_i)\}\Big] \times \gamma_v^{1/2(n+d)-1}\exp\Big[-\frac{1}{2}\gamma_v\Big\{c+\sum_{i=1}^{n}((\theta_i-\bar{\theta})-(q_i-\bar{q}) (\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\boldsymbol{b}-(z_i-\bar{z}))\Big\}^2\Big] \times \gamma_u^{(1/2)m}\exp\Big[-\frac{1}{2}\gamma_u\sum\sum_{1\le i\ne j\le m}(z_i-z_j)^2w_{ij}\Big].$$

따라서 $\boldsymbol{b}$에 관해 적분하면

$$p(\boldsymbol{\theta}_*, \boldsymbol{z}, \gamma_u, \gamma_v|\boldsymbol{y}) \propto \exp\Big[\sum_{i=1}^{n}\{y_i\theta_i-\psi(\theta_i)\}\Big] \times \gamma_v^{1/2(n+d)-1}\exp\Big[-\frac{1}{2}\gamma_v\Big\{c+\sum_{i=1}^{n}(\theta_i-\bar{\theta}-q_i-\bar{q}+z_i-\bar{z})^2 - \boldsymbol{v}^T(\boldsymbol{X}_*^T\boldsymbol{X}_*)^{-1}\boldsymbol{v}\Big\}\Big] \times \gamma_u^{1/2(m+g)-1}\exp\Big[-\frac{1}{2}\gamma_u\Big\{a+\sum\sum_{1\le i\ne j\le m}(z_i-z_j)^2w_{ij}\Big\}\Big]$$

이다. 여기서 $\boldsymbol{v}=\sum_{[i=1}(\theta_i-\bar{\theta}-q_i+\bar{q}-z_i+\bar{z})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$이다. 따라서 γ_u와 γ_v에 관해 적분하면 다음과 같다.

$$p(\boldsymbol{\theta}_*, \boldsymbol{z}|\boldsymbol{y}) \le K\exp\Big[\sum_{i=1}^{n}\{y_i\theta_i-\psi(\theta_i)\}\Big]\Big[a+\sum\sum_{1\le i\ne j\le m}(z_i-z_j)^2w_{ij}\Big]^{-1/2(m+g)}.$$

여기서 $K(>0)$는 $\boldsymbol{\theta}_*$와 $\boldsymbol{z}$에 전혀 의존하지 않는 상수이다. 마지막으로, $z_m=0$이고 $\sum\sum w_{ij}(z_i-z_j)^2$는 $m-1$개의 변수 $z_1,\cdots,z_{m-1}$만 관계된다. 따라서 $\boldsymbol{z}$에 관해 적분하면, 다변량 t-분포의 구조를 이용하여 다음을 얻는다.

$$p(\boldsymbol{\theta}_*|\boldsymbol{y}) \le K\exp\Big[\sum_{i=1}^{n}\{y_i\theta_i-\psi(\theta_i)\}\Big].$$

따라서 가정에 의해서 $\int p(\boldsymbol{\theta}_*|\boldsymbol{y})d\boldsymbol{\theta}_* < \infty$이다.

정리 1은 모든 $y_1,\cdots,y_n$에 대해 적분이 유한하다는 가정이 필요한 Ghosh et al.(1998)의 결과

를 일반화한다. 예컨대, 이항분포의 경우, 위의 정리 1은 m 부분지수 중 적어도 p개에 대해 $1 \le y_i \le n_i - 1$인 조건이 필요하나, Ghosh et al.(1998)의 정리 1은 모든 $i = 1, \cdots, m$에 대해 $1 \le y_i \le n_i - 1$인 조건이 필요하다. 한편 포아송 경우 위의 정리 1은 m 부분지수 중 적어도 p개에 대해 $y_i \ge 1$인 조건이 필요하나, Ghosh et al.(1998)의 정리 1은 모든 $i = 1, \cdots, m$에 대해 $y_i \ge 1$인 조건이 필요하다.

위의 계층적 베이즈 공간 일반화 선형모형에서의 통계적 추론은 다차원 적분문제의 어려움이 있지만 최근에 베이지안 추론에서 널리 쓰이는 깁스 표집에 의해 해결이 가능하다. $\gamma_u = \sigma_u^{-2}$ 그리고 $\gamma_v = \sigma_v^{-2}$라 나타내면, 깁스 표집을 수행하기 위한 조건부 밀도함수를 다음과 같이 구할 수 있다.

(i) $\gamma_u \mid \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{u}, \gamma_v, \boldsymbol{y} \sim \text{Gamma}(\{a + \sum\sum_{i \neq j}(u_i - u_j)^2 w_{ij}\}/2, \ (m+g)/2)$;

(ii) $\gamma_v \mid \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{u}, \gamma_u, \boldsymbol{y} \sim \text{Gamma}(\{c + \sum_i(\theta_i - \boldsymbol{x}^T \boldsymbol{b} - u_i)^2\}/2, \ (m+d)/2)$;

(iii) $u_i \mid \boldsymbol{\theta}, \boldsymbol{b}, u_j(j \neq i), \gamma_u, \gamma_v, \boldsymbol{y}$

$\sim \text{N}[(\gamma_v + k_i)^{-1}\{\gamma_v(\theta_i - \boldsymbol{x}_i^T \boldsymbol{b})^2 + \gamma_u \sum_{j(\neq i)} u_j w_{ij}\}, \ (\gamma_v + k_i)^{-1}], \ (k_i = \sum_j w_{ij})$;

(iv) $\boldsymbol{b} \mid \boldsymbol{\theta}, \boldsymbol{u}, \gamma_u, \gamma_v, \boldsymbol{y} \sim \text{N}[(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T(\boldsymbol{\theta} - \boldsymbol{u}), \ \gamma_v^{-1}(\boldsymbol{X}^T \boldsymbol{X})^{-1}]$;

(v) $p(\theta_i \mid \theta_j(j \neq i), \boldsymbol{b}, \boldsymbol{u}, \gamma_u, \gamma_v, \boldsymbol{y}) \propto \exp(y_i \theta_i - \psi(\theta_i))\exp[-\gamma_v(\theta_i - q_i - \boldsymbol{x}_i^T \boldsymbol{b} - u_i)/2]$.

여기서 (i) - (iv)에 주어진 조건부 밀도함수로부터는 쉽게 표본를 추출할 수 있지만, (v)의 경우는 다르다. 그러나 $p(\theta_i \mid \theta_j(j \neq i), \boldsymbol{b}, \boldsymbol{u}, \gamma_u, \gamma_v, \boldsymbol{y})$의 로그-오목성(log-concavity)을 밝히면 Gilks와 Wild(1992)의 ARS(adaptive rejection sampling) 알고리즘을 사용할 수 있다. 여기서 로그-오목성은 다음의 계산으로 어렵지 않게 밝힐 수 있다. 왜냐하면

$$\frac{\partial^2 \log p(\theta_i \mid \cdot)}{\partial \theta_i^2} = -n_i \psi''(\theta_i) - \gamma_v < 0$$

이기 때문이다. 여기서 $\gamma_v > 0$ 그리고 $\psi''(\theta_i) = V(Y_i \mid \theta_i) > 0$이다.

위 모형의 질병지도 작성에의 응용으로 Clayton과 Kaldor(1987) 그리고 Besag et al.(1991)은 $Y_i \mid \phi_i \overset{iid}{\sim} \text{Poisson}(E_i \phi_i)$를 고려하였다. 여기서 $\mu_i (= \log \phi_i) = \boldsymbol{x}_i^T \boldsymbol{b} + u_i + v_i$이고 E_i는 나이나 성별 같은 공통 위험 요인들에 관한 어떤 '표준적' 모집단에 비교되는, 지역 i에서 '기대되는' 사례의 수를 나타낸다. 흔히 μ_i는 로그-상대위험도로 불리운다. 따라서 $\theta_i = \log \lambda_i = \log E_i + \log \phi_i = \log E_i + \boldsymbol{x}_i^T \boldsymbol{b} + u_i + v_i$이다. 이 모형이 다음 절에서의 공간적 자료의 분석에 구체적으로 사용될 모형이다.

3. 예 제

이 절에서는 앞 절에서 제안한 계층적 베이즈 모형을 현실 자료(real data)로 이 분야에서 많이 인용되는 뉴욕주의 백혈병 자료(New York leukemia data)에 적용해 보고자 한다. (Turnbull et al.(1990), Kulldorff와 Nagarwalla(1995), Waller et al.(1992, 1994), Waller와 McMaster(1997)). 여기서 우리의 관심은 백혈병 발병률뿐만 아니라 연구지역 내 11개의 지하수 오염원으로 알려진 화학물질인 트리크롤로에틸렌(trichloroethylene; TCE)을 포함한 비활성 유독 산업 폐기물의 매립장소와의 근접 거주가 백혈병 발병률에 어떤 영향을 미치는지를 알아보고자 한다.

연구 지역은 281개의 센서스 조사구(census tracts)에 포함된 1,057,673명을 조사한 뉴욕주의 북부지방 8개 군(county) 지역이다. 뉴욕주 보건당국에 의하면 1978년과 1982년 사이에 발생한 백혈병 환자의 수는 597명이고, 각 환자는 거주지에 따라 어느 센서스 조사구에 포함될지를 결정했다.

TCE에 대한 노출정도를 측정한 값이 구체적으로 없으므로 노출에 대한 대리값(surrogate)으로 거리의 역수를 사용한다. TCE 지점에서의 거리의 역수의 효과를 수량화하기 위해서 2절에서 제안된 HB 방법을 적용하여 뉴욕주 백혈병 발병률을 다음과 같이 모형화 한다. $\boldsymbol{\theta}$가 주어졌을 때, $Y_1, \cdots, Y_{281}$은 서로 독립인 포아송 $(E_i\psi_i)$, $i=1,\cdots,281$ 분포를 따른다. 여기서 θ_i를 다음과 같이 모형화 한다.

$$\theta_i = \log E_i + \beta x_i + u_i + v_i \quad (i=1,\cdots,281) \tag{3.1}$$

여기서 x_i는 가장 가까운 TCE를 포함한 유독 폐기물의 매립지와 i번째 센서스 조사구의 중심과의 거리의 역수이며 u_i와 v_i는 모집단에서의 초과 이질성을 각각 나타낸다. 비교를 위해서 Waller et al.(1994)에 따라 E_i를 281개의 모든 조사구에 걸친 상수 발병률(0.00054)에 대한 기대 발병수로 정의한다. 한편 모형 II 단계에서 i와 j가 이웃한 조사구이면 $w_{ij}=1$, 그렇지 않으면 $w_{ij}=0$을 취한다. 또한 모형 III 단계에서 $p=1$, $a=g=0.001$ 그리고 $b=d=1$을 사용한다.

구체적으로 모형을 설정함에 있어 중요한 것은 이웃한 지역(neighbors)의 정의이다. 여기서 이웃한 지역은 주어진 센서스 조사구의 발병률과의 상관(correlation)이 있는 센서스 조사구들을 의미한다. 이웃한 지역에 대한 전통적 정의는 주어진 조사구에 지리적으로 인접한 모든 조사구들을 의미한다. 이는 모든 조사구가 동일한 지리적 면적과 인구 크기를 가진 경우에 적절하다. 그러나 우리의 경우는 양 측면에서 매우 이질적이다. 연구 지역이 대부분 농촌 지역이나 두 도시 지역(Cyracuse, Binghamton) 및 세 개의 큰 도시(Auburn, Cortland, Ithaca)를 포함하고 있다. 원래 센서스 조사구는 약 3,000-4,000명 정도 포함하도록 설계되었으나, 우리의 연구 조사구들은 인구 크기에 있어서 9명에서 13,014명까지 걸쳐있다. 이러한 이질적 요인으로 인하여 우리의 연구에서는 이웃한 지역을 각 조사구에 대해 조사구 중심에서 30km 반경 내의 조사구 중심을 가진 모든 조사구로 정의한다. 따라서 전통적인 지리적으로 인접한 조사구를 이웃으로 사용 할 때 보다 상당히 더 많은 조사구를 이웃한 지역으로 포함하게 된다. 그림 1이 TCE 지점, 센서스 조사구 중심, 발병률이 없는 조사구 표시 및 이웃을 정의하는 30km 반경 등을 나타낸다.

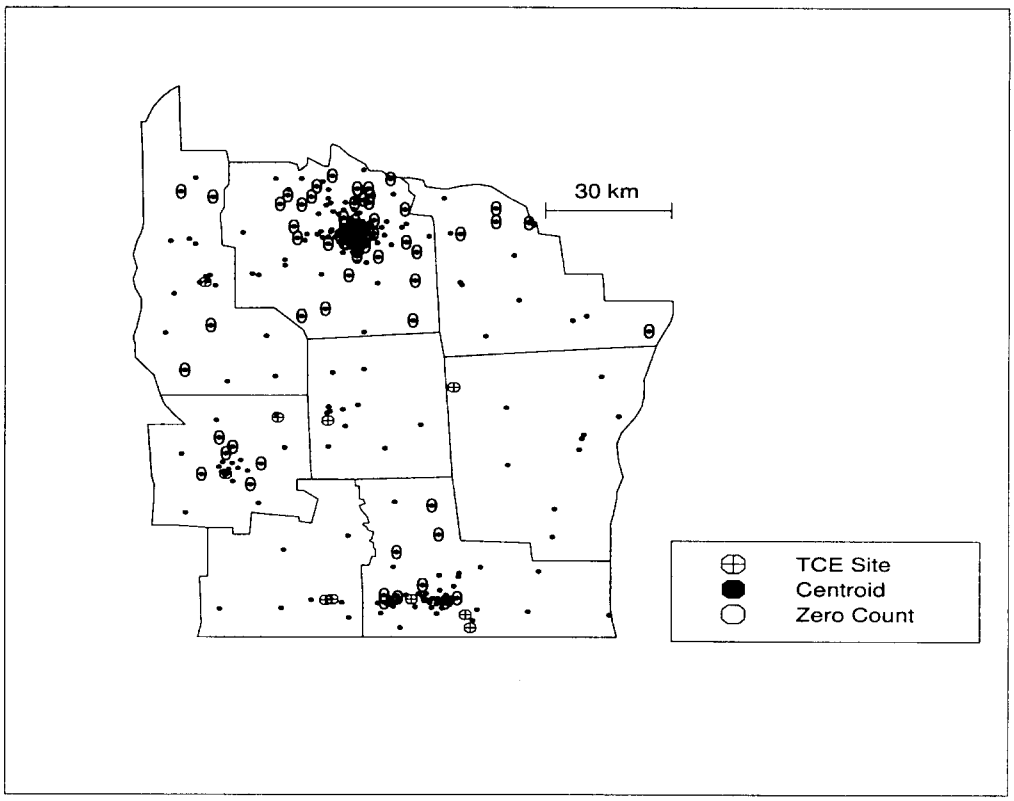

그림 1. 뉴욕주의 북부지방 8개 군 지역에 대한 센서스 조사구 중심

그림 2는 백혈병 발병률에 대한 원추정치(crude rates)와 모형을 사용한 사후 중위수 추정치(posterior median rates)를 등고선 지도로 비교한 것이다. 두 등고선 지도에서 전체 형태는 비슷함을 볼 수 있고, 연구 지역의 중심 부분에서 높은 발병률이 있음을 알 수 있다. 차이점은 기대한 바와 같이 모형을 사용하여 평활 HB 추정치가 모든 영 아닌 원추정치보다 비교적 작음을 알 수 있다. 이는 영(zero) 원추정치가 연구 지역 전체에 퍼져 있어서 대부분 이웃한 지역이 적어도 하나 이상의 영 원추정치를 포함함으로 이웃한 지역들로부터 "borrow strength" 한 평활 추정치를 작게 만든 것이다. 즉, 영 원추정치가 (더 높은) 이웃의 국소적 평균 발병률을 향하여 '축소(shrinking)' 되었다.

폐기물 장소 효과 b에 대한 95% 신용구간(credible set)은 (-0.112, 1.596)이다. 따라서 우리의 계층적 베이즈 분석에 의하면 백혈병 발병률이 폐기물 장소에 대한 근접성에 양의 효과(positive effect)가 있다는 약간의 근거(some evidence)는 있지만, 그러한 효과가 없을 사후 확률이 양수이다. 그러나 원추정치에 분모에서의 이질성을 수정하면, TCE 지점에 대한 거리의 역수가 평활된 발병률(smoothed rates)에 거의 영향이 없는 것으로 보인다. 이는 백혈병 발병률에 폐기물에의 근접성의 효과가 약간은 있을 수 있다는 Waller et al.(1994)와 Waller와 McMaster(1997)의 연구결과와 약간은 일치한다.

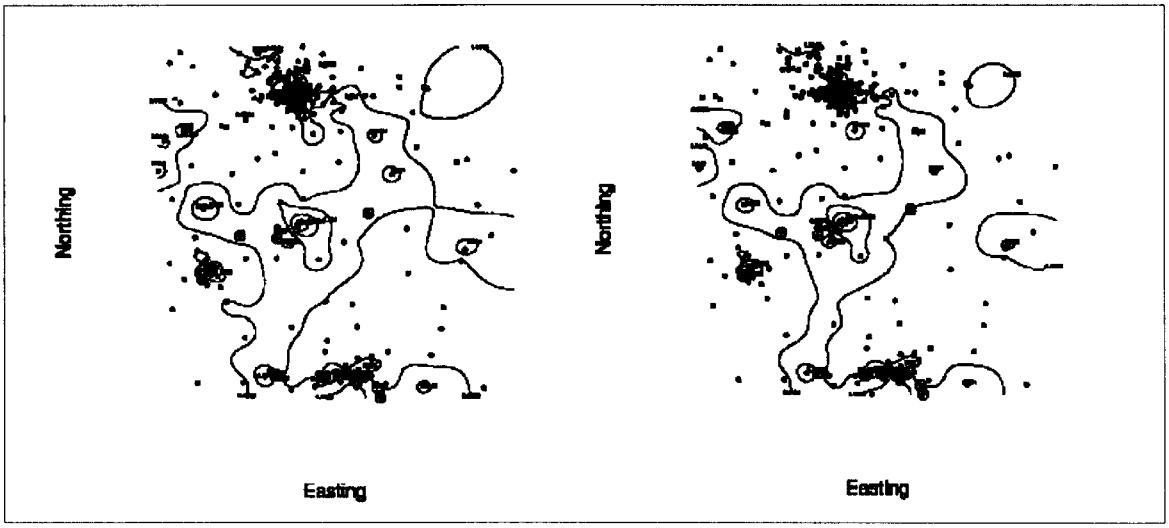

그림 2. 1978-1982년에 조사된 센서스 조사구에 의한 백혈병 발병률의 등고선 지도
(왼쪽 - 원추정치, 오른쪽 - HB 평활추정치)

관심있는 후속 연구는 좀더 정밀한 모형을 세우고, 다른 구조의 이웃(neighborhood)의 정의를 사용함으로 위의 연구와의 결과를 비교하고, 그 민감성 살피는 것이다. 또한 나이나 성별 같은 중첩요인(confounding factor)을 자료에 포함시켜 분석하는 것이다. 위의 모형에서 우리는 가장 가까운 폐기물 장소의 효과만 고려하여 질병의 발병에 대한 폐기물 장소의 효과를 동일하게 취급하였으나, 실제로는 폐기물 장소의 효과는 오염 정도, 근접성 등에 따라 공중보건에 미치는 위협이 다를 가능성이 크다. 그 결과 어떤 폐기물 장소에 대한 근접성의 총 효과(aggregate effect)가 특정한 폐기물 장소에 대한 근접성의 효과를 줄였을 수도 있다.

참 고 문 헌

[1] Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 10, 3-66.

[2] Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

[3] Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimators of age-standardised relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

[4] Cressie N. (1991). *Statistics for Spatial Data*. John Wiley and Sons, New York.

[5] Cressie N. (1992). Smoothing regional maps using empirical Bayes predictors. *Geographical Analysis*, 24. 75-95.

[6] Cressie, N. and Chan, N.H. (1989). Spatial modeling of regional variables. *Journal of the*

American Statistical Association, 84, 393-401.

[7] Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association,* 93, 273-282.

[8] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 65-93.

[9] Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling. *Journal of Royal Statistical Society, Ser B*, 41, 337-348.

[10] Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, 11, 799-810.

[11] Natarajan, R. and McCulloch, C.E. (1995). A note on the existence of posterior distribution for a class of mixed models for binomial responses. *Biometrika*, 82, 639-643.

[12] Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L., and Clark, L.C. (1990). Monitoring for clustering of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology*, 132, S136-S143.

[13] Waller, L.A. and Turnbull, B.W., Clark, L.C. and Nasca, P. (1992). Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. *Environmetrics*, 3, 281-300.

[14] Waller, L.A., Turnbull, B.W., Clark, L.C., and Nasca, P. (1994). Spatial pattern analyses to detect rare disease clusters. In *Case Studies in Biometry*, N. Lange, Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, eds. John Wiley and Sons, New York, pp.3-23.

[15] Waller, L.A. and McMaster, R.B. (1997). Incorporating indirect standardization in tests for disease clustering in a GIS environment. *Geographical Systems.*