The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.231
28(2), 231–239

Zero Inflated Poisson Model for Spatial Data

Junhee Han[a,1] · Changhoon Kim[b]

[a]Research And Statistical Support, Research Institute of Convergence for Biomedical Science and Technology, Pusan National University Yangsan Hospital
[b]Department of Preventive Medicine, Pusan National University School of Medicine

(Received March 16, 2015; Revised March 31, 2015; Accepted March 31, 2015)

Abstract

A Poisson model is the first choice for counts data. Quasi Poisson or negative binomial models are usually used in cases of over (or under) dispersed data. However, these models might be unsuitable if the data consist of excessive number of zeros (zero inflated data). For zero inflated counts data, Zero Inflated Poisson (ZIP) or Zero Inflated Negative Binomial (ZINB) models are recommended to address the issue. In this paper, we further considered a situation where zero inflated data are spatially correlated. A mixed effect model with random effects that account for spatial autocorrelation is used to fit the data.

Keywords: Counts data, Poisson model, zero inflated data, spatial data, CAR, WinBUGS.

1. 서론

특정 암에 의한 사망자수나 교통사고 사망자수와 같이 임의의 기간 동안 발생한 사건의 건수를 기록한 가산 자료(Counts data)는 흔히 포아송 모형으로 설명이 된다. 하지만, 실제 자료들은 평균과 분산이 같아야 한다는 포아송 분포의 기본 가정을 만족하지 못하고 분산이 평균보다 크거나(과대산포, over-dispersion) 또는 그 반대(과소산포, under-dispersion)인 경우가 많다. 이런 경우도 유사 포아송(Quasi-Poisson)이나 음이항(Negative Binomial) 모형 등을 이용해서 어느 정도 설명이 가능하다. 특히 과산포 가산 자료 중 포아송분포에서 기대되는 0의 관측빈도보다 더 많은 0이 관측이 되는 경우 위의 모형만으로는 적합도가 떨어 질 수 있다. 이렇게 0이 기대치 이상으로 많은 자료를 영과잉(Zero inflated) 자료라고 하고 Lambert (1992) 등이 이런 자료를 설명하기위해 영과잉 모형(Zero inflated models) 제안했다.

최근 들어서 많은 가산 자료들이 위치 정보를 포함한 공간자료의 형태로 제공이 되고, 특히 인구기반(population based) 자료들은 많은 경우 시군구 또는 읍면동과 같은 행정구역 단위 자료들로 주어진다. 이런 자료들을 격자자료(Lattice data)라고 부르고 본 논문에서 분석 하고자는 2004년 기준 부산시 남성의 동별 갑상선암 발생자수 자료가 그 좋은 예이다. 이렇게 가산 자료가 공간 정보까지 담

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry for Health and Welfare, Republic of Korea (0920050).
[1]Corresponding author: Research And Statistical Support, Research Institute of Convergence for Biomedical Science and Technology, Pusan National University Yangsan Hospital, Yangsan, Korea.
E-mail: pnuyh.rass@gmail.com

고 있는 경우 관측치 간의 유의한 공간상관관계가 존재 할 수 있어서 기존의 포아송 일반화선형모형(Generalized Linear Model; GLM) 만으로는 설명이 충분하지 않을 수 있다. 이 경우 공간효과를 설명하기 위해 랜덤 효과(random effects)를 포함한 혼합효과모형(mixed effects model)이 필요 할 수 있고, 베이지안적 접근법으로 모형을 적합할 수 있다. 이때 공간효과는 공간자기회귀(Spatial Autoregressive; SAR) 모형, 조건부자기회귀(Conditional Autoregressive; CAR) 모형, 또는 공간이동평균(Spatial Moving Average; SMA) 모형 등의 공간모형들을 사전함수로 이용해서 설명 될 수 있다. 주로 CAR 모형이 선호되고 본 논문에서도 CAR 모형을 공간효과에 대한 사전함수로 이용했다.

국립암센터를 비롯한 각 지역별 암센터에는 지역별로 암등록 자료를 수집, 보관, 관리하고 있고 각 암종별 발생률, 사망률 등에 대한 추정을 비롯해서 이를 설명할 수 있는 사회경제적 위험요인들을 파악하는 것 등에 많은 관심을 가지고 있다. 기존의 많은 분석 방법론들은 공간자료의 가장 큰 특징인 공간상관관계를 무시한 경우가 많았다. 하지만, 공간자료에 유의한 공간상관관계가 있는 경우 분석 결과가 틀리거나 또는 과장될 가능성이 있어서 주의를 요한다. 더구나 읍면동 단위의 소지역 자료나 희귀 질환 자료는 많은 수의 0이 관측되는 소위 공간 영과잉 가산 자료(Spatial zero inflated counts data)인 경우가 많다. 이런 자료의 경우 위에서 제시된 두가지 문제를 동시에 가지게 되고 기존의 영과잉 모형에 공간효과를 설명하기 위한 공간모형까지 결합된 혼합모형이 필요하게 된다. 즉, 본 연구에서 사용된 2004년 기준 부산시 남성의 동별 갑상선암 발생자수 자료의 경우, 소지역단위에서 수집된 희귀한 암종 자료로서 영과잉 자료이자 격자자료로서 공간 영과잉 가산 자료에 해당이 된다.

암을 비롯한 많은 만성 질환들의 지역 간 차이는 그 지역들의 생활수준, 주거수준 차이와 밀접한 연관이 있다는 것은 잘 알려져 있다. 흔히 사용되는 생활·주거 수준의 지수로는 지역박탈지수(Deprivation Index)가 있다. 박탈지수란 낙후된 주거환경 비율, 노인 인구 비율, 고졸 미만 학력 인구 비율, 가구원 기준 하위 사회계급 비율, 아파트 가구 비율, 자동차 미소유 가구 비율, 독거 가구 비율, 여성 가구주 비율 등 8개 지표를 표준화해서 종합해 산출한 것으로 박탈지수가 클수록 사회적 경제적 여건이 취약하다는 의미이다 (Kim 등, 2014).

다른 서구 개발국가와 마찬가지로 한국사회의 감상선암의 발생은 증가하고 있는데 (Ferlay 등, 2010; Jung 등, 2015) 갑상선암의 발생률은 증가하는데 사망률은 일정하게 유지되고 있고, 증상이 없는 작은 결절을 동반한 유두상 갑상선이 대부분을 차지하고 있어 (Cho 등, 2013) 과다진단(overdiagnosis) 및 과다치료(overtreatment)의 가능성에 대한 관심이 증가하고 있다 (Ahn 등, 2014; Han 등, 2011; Lee와 Shin, 2014). 또한 이들로 설명되지 않은 환경적, 유전적, 생활습관과 관련된 가능성도 또한 제기 되고 있다 (Chen 등, 2009; Li 등, 2013; Londero 등, 2013). 미국 연구에서는 사회경제적 수준이 높은 지역에 발생이 높아서 검진서비스의 접근성과 관련되어 이들의 효과를 지역의 의료서비스 양상과의 관련성을 확인하기도 하였다 (Li 등, 2013). 부산지역의 갑상선암의 발생율과 관련된 연구에서는 점차 확산되고 있는 의료서비스의 진료양상과 인근 원자력발전소의 영향에 대한 우려와 함께 복합적으로 고려해야 하는 상황을 가지고 있어, 부산시의 동별박탈지수와 갑상선암 발생자수의 연관성을 보는 것 또한 큰 의미를 가질 것이다.

본 연구에서는 2004년 기준 부산시 남성의 동별 갑상선암 발생자수를 반응변수로, 지역박탈지수를 설명변수로 두고 둘 사이의 관계를 설명하기 위해 가능한 여러 가지 모형들을 베이지안적 접근법으로 적합해 보고 어떤 모형이 더 적절한지 보기 위해 DIC(Deviance Information Criteria; Spiegelhalter 등, 2002) 값들을 비교하였다. 이 모형들은 일반화선형 포아송(Generalized Linear Model; GLM) 모형, 영과잉 포아송(Zero Inflated Poisson; ZIP) 모형, 공간 영과잉 포아송(Spatial Zero Inflated Poisson; SZIP) 모형 등을 포함한다. 각 모형들은 WinBUGS (The BUGS project, 1997) 프로그램를 이용하여 적합되었다.

2. 공간 영과잉 모형(Spatial Zero Inflated Models)

이 장에서는 포아송 분포를 가정해서 계산된 0의 빈도수보다 더 많은 0이 관측이 되는 영과잉 자료를 적합하는 영과잉 모형, 특히 영과잉 포아송(ZIP) 모형을 살펴보고 더 나아가 공간상관관계까지 고려를 해야 되는 공간 가산 자료에 대해 공간 영과잉 포아송(SZIP) 모형을 적합하는 방법에 대해서 알아본다.

2.1. 영과잉 포아송 모형(Zero Inflated Poisson Model; ZIP)

가산자료의 경우 포아송 분포를 통한 일반선형모형(Generalized Linear Model; GLM)을 이용하여 적합 될 수 있다. 즉, 확률변수 Y의 확률질량함수, $f(y;\mu)$는 아래와 같고 로그함수를 연결함수(link function)로 이용해서 모수와 설명변수와의 관계는 아래의 선형모형으로 나타내어진다.

$$\begin{cases} f(y;\mu) = \dfrac{e^{\mu}\mu^{y}}{y!} \sim \text{Poisson}(\mu), \quad y = 0, 1, \ldots, \\ \log(\mu) = X\beta, \end{cases} \tag{2.1}$$

여기서 X는 설명변수 행렬이고 β는 회귀계수 벡터를 나타낸다. 이때 본 논문에서 사용된 암발생자수 자료와 같은 인구기반 자료는 각 관측치의 노출(exposure)이 다르고 이를 오프셋(offset)으로 간주한 아래와 같은 선형모형을 사용해야 한다.

$$\log(\mu) = \log(N) + X\beta, \tag{2.2}$$

여기서 N은 해당지역의 인구수가 되고 로그값이 오프셋으로 모형에 포함이 되어진다.

영과잉 포아송(ZIP) 분포를 따르는 확률변수, Y의 확률질량함수, $f_{ZIP}(y;\mu)$는 다음과 같이 구조적인 0(structured zeros)이 관측되는 부분과 포아송분포에 의해 설명되는 부분의 혼합분포 형태를 가진다.

$$\begin{cases} f_{ZIP}(y;\mu) = p \cdot I(y=0) + (1-p) \cdot f_{count}(y;\mu), \\ f_{count}(y;\mu) = \dfrac{e^{\mu}\mu^{y}}{y!} \sim \text{Poisson}(\mu), \quad y = 0, 1, \ldots, \end{cases} \tag{2.3}$$

여기서 $I(y=0)$은 y가 0인지 아닌지를 나타내는 지시함수(Indicator function)이고 p는 평균 모수 μ를 가지는 포아송분포에서 발생되는 0 이외에 더 있는 구조적인 0(structured zeros)의 비율을 설명하고 흔히 베르누이 변수로 적합 된다. 확률변수 Y의 평균과 분산은 아래와 같이 주어지고 분산이 평균보다 큰 관계를 이용해서 과산포 문제를 해결할 수 있다.

$$\begin{aligned} E(Y) &= (1-p)\mu, \\ \text{Var}(Y) &= (1-p)\mu + p(1-p)\mu^2. \end{aligned}$$

영과잉 포아송 분포에 기반한 여러 회귀모형이 제안되었고 그 중 Lambert (1992)는 일반선형모형에서와 같이 포아송 분포, 베르누이 분포에 대한 모수들에 대해 각각 로그함수와 로짓함수를 연결함수로 이용해서 설명변수와의 관계를 아래와 같이 제시했다.

$$\begin{aligned} \log(\mu) &= X\beta, \\ \text{logit}(p) &= \log\left(\frac{p}{1-p}\right) = Z\gamma. \end{aligned} \tag{2.4}$$

여기서 X와 Z는 설명변수 행렬이고 β와 γ는 각각에 대한 회귀계수 벡터를 나타낸다. 이 때, 행렬 X와 Z는 같은 설명변수들을 나타낼 수도 있고 아닐 수도 있다.

일반선형 포아송 모형과 마찬가지로, 본 연구에서는 해당지역의 인구수, N의 로그값을 오프셋으로 아래와 같이 선형모형에 포함시켰다.

$$\log(\mu) = \log(N) + X\beta,$$
$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(N) + Z\gamma. \tag{2.5}$$

2.2. 공간상관관계(Spatial Autocorrelation)

일반적으로 자료간의 독립성을 가정하는 경우와는 달리 공간 자료는 그 특성상 가까운 관측치일 수록 더 유사한 값을 가지고 멀어질 수로 그렇지 않을 것이라는 가정이 타당하다. 이로 인해 자료간의 독립성을 기반으로 한 여러 분석들은 더 이상 유효하지 않게 되고 이 공간 상관성을 고려한 모형이 요구된다. 물론, 유의한 공간상관관계가 존재하는지를 검정하는 일이 우선되어야하고 실제로 유의한 공간상관관계가 존재한다면 이를 적절히 설명할 수 있는 공간모형이 적합 되어야 한다. 공간상관관계를 측정하는 통계량 중 Geary's c와 Moran's I가 가장 널리 이용되고 본 연구의 자료는 Moran's I 값을 구해서 검정을 했다 (Moran, 1950). 지역 i; $i = 1, \ldots, n$에서의 관측값을 Y_i라고 하면 Moran's I 값은 아래와 같이 구해진다.

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2},$$

여기서 $\bar{Y}$는 Y_i들의 평균이고, w_{ij}는 지역 i와 j의 이웃관계를 나타내는 가중치로 일반적으로 서로 이웃이면 1, 그렇지 않으면 0으로 주어진다. 두 지역이 이웃인지 아닌지에 대한 기준은 보통 공통의 경계를 가지면 이웃으로 정의되고 필요에 따라서는 다르게 정의되기도 한다.

공간상관관계가 없다는 가정 하에 표준화 시킨 통계량 I 값은 근사하게 정규분포를 따른다고 알려져 있고 이를 이용해서 공간상관관계의 유의성을 검정하게 된다.

2.3. 공간 영과잉 포아송 모형(Spatial Zero Inflated Poisson Model; SZIP)

공간 영과잉 가산 자료는 설명변수를 포함한 고정효과(fixed effects)와 공간효과를 포함한 랜덤효과(random effects)를 함께 고려하는 혼합효과모형(mixed effects models)을 이용하여 적합 될 수 있다.

즉, 식 (2.5)의 영과잉 포아송 모형에 공간효과를 설명하는 변수 s, t를 더 추가한 아래와 같은 선형모형이 사용될 수 있다.

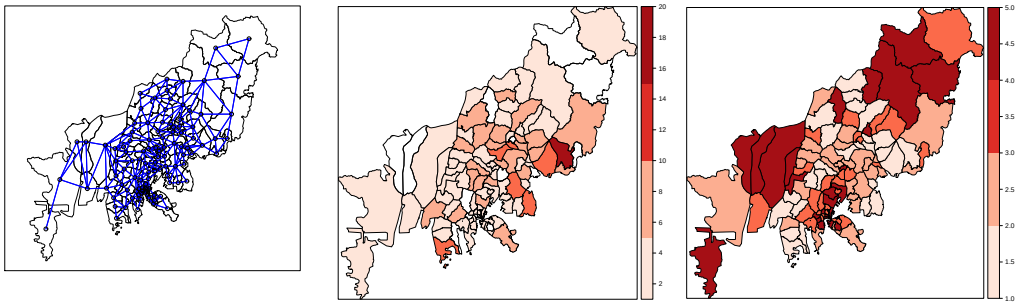
$$\log(\mu) = \log(N) + X\beta + s,$$
$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(N) + Z\gamma + t, \tag{2.6}$$

여기서 s, t는 공간효과를 나타내는 변수로, 본 연구에서는 CAR(Conditional Autoregressive) 모형을 사용하여 적합되었다. 가령 s_i를 지역 $i = 1, \ldots, n$에서의 공간효과라고 한다면, CAR 모형에서 지역 간 공간효과의 관계는 아래의 식들로 정의된다.

$$E(s_i | s_j; j \neq i) = \rho \sum_{j=1}^{n} c_{ij} s_j,$$
$$\text{Var}(s_i | s_j; j \neq i) = \frac{1}{\tau_i}, \tag{2.7}$$

Table 3.1. Summary of data.

동의 개수	발생자수 평균	발생자수 분산	포아송분포 가정하의 기대 0 빈도수	실제 관측된 0 빈도수
105	3.5	10.3	3.2	16



(a) Neighborhood structure (b) Cancer Incidence (c) Quartiles of Deprivation Index

Figure 3.1. Thyroid cancer data for male in 2004 Busan

여기서 ρ는 이웃 간의 상관관계의 강도를 나타내는 모수이고 c_{ij}는 지역 i와 j의 이웃관계를 나타내는 가중치로 일반적으로 서로 이웃이면 1, 그렇지 않으면 0으로 주어진다. 그리고 $1/\tau_i$를 통해 지역 i에 대한 조건부 분산이 정의된다.

3. 자료분석

3.1. 2004년 부산시 남성의 갑상선 발생자수 자료와 박탈지수

본 연구에서 사용된 2004년 기준 부산시 남성의 동별 갑상선암 발생자수 자료는 희귀 암종으로 분류되어 실제로는 2003, 2004, 그리고 2005년의 3년간의 자료를 합한 자료이다. 보통의 가산자료와 마찬가지로 우선 포아송분포를 적합하기 위한 가정이 적절한지 보기 위해 자료의 평균과 분산을 계산해보면 각각 3.5와 10.3으로 분산이 평균의 거의 3배에 가까운 과대산포 자료임을 알 수 있다. 또한 포아송분포를 가정해서 기대되는 0의 빈도수 (3.2개 동) 보다 훨씬 많은 수 (16개 동)의 0이 관측이 되어 영과잉 가산자료임을 확인 할 수 있다 (Table 3.1). 그리고 분석에서 설명변수로 이용된 동별 박탈지수는 발생자수 자료의 마지막 연도 (2005년) 기준 계산값으로 수준에 따른 더 적절한 비교를 위해 원래 박탈지수값들을 각각의 4분위수(Quartiles)로 변환해서 이용하였다. 즉, 1분위에서 4분위로 분위수가 높아질수록 해당 동의 주거수준이 더 열악하다는 의미이다.

3.2. 공간상관관계 확인

동별 암발생자수 자료는 가까운 동 간의 유사성을 고려해야 되는 공간자료이기도 해서 먼저 자료 간에 공간상관관계가 있는지 확인해야 한다. 이를 위해 우선 부산시 동별 이웃관계를 정의해야 하는데 기본적으로 공통의 지리적 경계를 가지는 동들은 서로 이웃으로 정의가 되고 영도구와 중구에 속한 몇몇 동의 경우 교각을 통하여 서로 연결된 동들 또한 이웃으로 본다 (Figure 3.1(a)). 반응변수인 암 발생자수와 설명변수인 동별 박탈지수(4분위수)는 Figure 3.1(b)와 (c)의 색지도(choropleth) 통해서 확인이 할 수 있듯이 비슷한 색의 동들이 서로 모여 있는 것처럼 보인다. 이는 유의한 공간상관관계의 가능성을 내

Table 3.2. Summary of results

모형	$\hat{\beta}$ (95% C.I.*)	$\hat{\gamma}$	DIC
GLM	−8.51 (−8.58, −8.43)	해당없음	5502.8
ZIP	−8.50 (−8.58, −8.43)	−19.24 (−29.29, −14.25)	5505.0
SZIP	−4.49 (−4.64, −4.35)	−16.40 (−32.71, −4.56)	882.3

∗: 95% MCMC Credible Interval

포하는 것으로 실제로 Section 2.2에서 설명된 Moran's I 통계량을 암 발생자수와 동별 박탈지수에 대해서 계산해보면, 각각 0.19 (p-value = 0.0006), 0.28 (p-value << 0.0001)로 두 변수 모두 공간상관관계가 아주 유의함을 알 수 있다.

3.3. 모형의 적합과 비교

영과잉 문제를 고려하지 않은 일반선형 포아송(GLM) 모형과 영과잉을 문제를 설명하기 위한 영과잉 포아송(ZIP)모형, 또한 공간효과까지 고려한 공간 영과잉 포아송(SZIP) 모형에 대한 각각의 선형모형은 식 (2.2), (2.5), (2.6)와 같이 주어지고, X와 Z는 유일한 설명변수인 동별 박탈지수이고 각각의 회귀계수인 β와 γ 그리고 공간효과에 대한 사전함수에서 사용되는 모수, ρ와 τ (식 (2.7) 참고, 지역첨자 i 생략) 대한 사전분포(prior distribution)는 아래와 같다.

$$\begin{aligned}\beta, \gamma &\sim N(0.0, 0.01),\\ \rho &\sim \text{Uniform}(0, 1),\\ \tau &\sim \text{Gamma}(0.01, 0.01),\end{aligned}$$

여기서 β와 γ의 초모수(hyper parameters) 값 0.01은 WinBUGS에서의 표기법과의 일치성을 위해서 정도(precision)로 표현된 것으로 분산 100에 해당한다. 선택된 각 초모수값들에 대해서는 2개의 MCMC를 생성해봄으로써 민감도(sensitivity) 테스트를 거쳤음을 밝혀둔다. 각 모형에 대해 베이지안적 접근법을 이용하여 구해진 MCMC(Monte Carlo Markov Chain)는 WinBUGS의 기본 알고리즘인 깁스 샘플러(Gibbs sampler)와 메트로폴리스 헤스팅스(Metropolis Hastings) 알고리즘을 이용하여 생성되었고 이 샘플들을 이용하여 주관심인 모수값에 대한 추정량과 해당 95% C.I.(Credible Interval)을 계산하였고 각 회귀계수들에 대한 유의성 또한 검정했다 (The BUGS project, 2007).

Table 3.2에서 볼 수 있듯이 모든 모형에서 주 관심인 동별 박탈지수에 대한 회귀계수(β) 값은 모두 음수이고 유의하게 나타났다. 영과잉 포아송모형이나 공간 영과잉 포아송 모형에서는 추가적으로 구조적인 '0'이 나타나는 확률 부분에 대한 회귀계수(γ)도 구할 수 있고 이들 또한 유의한 음수값들이 나왔다.

이는 동별 박탈지수의 분위수가 높을수록, 즉, 주거 수준이 열악해 질수록, 갑상선암 발생자가 줄어들었다는 의미이다. 일반선형 모형이나 ZIP 모형의 경우 각각 −8.51과 −8.50으로 거의 같은 값을 보인 반면 공간효과를 고려한 SZIP의 경우 −4.49로 거의 절반에 가까운 값을 보여 공간효과를 고려하지 않는 모형에서는 동별 박탈지수의 영향이 과장이 되었을 가능성을 보여준다.

각 모형들은 DIC(Deviance Information Criterion) 값들을 이용해서 비교 되었고 DIC 값이 작을수록 더 좋은 모형임을 의미한다 (Spiegelhalter 등, 2002). 일반선형 포아송모형과 영과잉 포아송모형의 경우는 회귀계수의 결과와 마찬가지로 거의 비슷한 DIC 값을 보여서 모형의 적합성에 큰 차이가 없음을 보인다. 반면, 공간상관관계를 고려한 공간 영과잉 포아송모형은 훨씬 작은 DIC 값을 보여 공간상관관계를 고려하지 않은 두 모형에 비해서 적합성이 월등히 뛰어남을 보여준다.

4. 결론 및 향후 연구

부산시 남성의 동별 갑상선암 발생자수 자료를 이용하여 2003–2005년의 갑상선암 발생율과 개인의 사회경제적 지위와 지역의 사회경제적 수준을 나타내는 박탈지수와의 관련성을 적합한 통계 모형을 이용하여 확인하고자 하였다. ZIP에 비해 SZIP에서 회귀계수의 절대값이 감소하여 과장된 관련성을 다소 보정하였음에도 박탈지수와 암발생의 회귀계수가 음의 값으로 나타나고 있어 사회경제적 수준이 높을수록 발생이 높은 것을 보여주고 있다. 이는 과다진단 및 과다치료의 논란에서 제기 되었던 것처럼 사회경제적 수준과 관련된 검진서비스의 비용적, 거리적 접근성과 관련되었을 가능성을 시사하고 있어, 국내의 다른 연구결과의 맥락과 일치하는 결과를 보여준다 (Cho 등, 2013; Kweon 등, 2013; Lee 등, 2012).

본 연구에서 사용된 자료와 같이 영과잉 자료이면서 공간상관관계까지 유의한 자료의 경우 일반선형 포아송모형이나 영과잉 포아송모형 보다는 공간효과까지 고려할 수 있는 공간 영과잉 포아송모형이 더 적절하다. 구조적인 0이 기대 빈도 이상으로 관측되더라도 그 정도가 심각하지 않은 경우는 영과잉 모형을 사용하는 효과가 미미할 수도 있다는 점도 확인되었다. 본 연구에 사용된 자료의 경우 전체 105개 동 중에서 0이 관측된 동은 16개로 기대 빈도 이상이긴 하지만, 전체 자료에서 차지하는 비율은 15% 정도로 아주 심각한 영과잉 문제가 있다고 볼 수는 없다. 통상적으로 전체 자료의 30%가 넘는 0이 있을 경우는 영과잉 모형을 적합 하는 효과가 뚜렷해지는 걸로 알려져 있지만, 이 또한 객관적인 연구에 기반한 수치라 보기 힘들다. 해서 가능한 경우 일반선형 포아송모형과 영과잉 포아송모형을 모두 적합 한 후 두 모형을 비교를 하는 것이 효과적일 것이다. 하지만, 어느 경우이든 유의하게 나타난 공간효과를 고려하지 않는 경우는 본 연구에서 보인 바와 같이 큰 차이를 보이는거나 잘못 된 결론을 내기도 하므로 공간자료에 대해서 공간상관관계를 검정한 뒤 필요한 경우 반드시 공간효과를 고려한 모형을 사용하는 것이 더 신뢰할 수 있는 결과를 얻을 수 있다.

이와 아울러, 더 적합도가 높은 통계모형과 공간통계분석 방법을 함께 활용한 연구는 갑상선 발생과 같이 의료서비스의 진료양상, 환경적, 생활습관요인 등 주요한 요인을 고려한 후에도 이들로 설명되지 않는 특별히 높은 지역을 확인하여 이들의 원인을 역학조사 등의 방법을 통하여 심층 조사가 필요한 연구에 활용될 수 있는 도구적 가치가 매우 높다. 향후 부산지역의 갑상선암 발생 연구뿐만 아니라 다른 연구에서도 효과적인 수단으로 활용될 수 있을 것이다.

References

Kim, D. J., Ki, M., Kim, M. H., Kim, Y. M., Yoon, T. H., Jang, S. R., JangChoi, K. H., Kang, A. R., Chae, H. R. and Choi, J. H. (2014). *Developing Health Inequalities Indicators and Monitoring the Status of Health Inequalities in Korea*, 2014-04, Korea Institute for Health and Social Affairs.

Ahn, H. S., Kim, H. J. and Welch, H. G. (2014). Korea's thyroid-cancer "epidemic"–screening and overdiagnosis, *The New England Journal of Medicine*, **371**, 1765–1767.

Chen, A. Y., Jemal, A. and Ward, E. M. (2009). Increasing incidence of differentiated thyroid cancer in the United States, 1988–2005, *Cancer*, **115**, 3801–3807.

Cho, B. Y., Choi, H. S., Park, Y. J., Lim, J. A., Ahn, H. Y., Lee, E. K., Kim, K. W., Yi, K. H., Chung, J. K., Youn, Y. K., Cho, N. H., Park do, J. and Koh, C. S. (2013). Changes in the clinicopathological characteristics and outcomes of thyroid cancer in Korea over the past four decades. *Thyroid : official journal of the American Thyroid Association*, **23**, 797–804.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C. and Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008, *International Journal of Cancer. Journal International du Cancer*, **127**, 2893–2917.

Han, M. A., Choi, K. S., Lee, H. Y., Kim, Y., Jun, J. K. and Park, E. C. (2011). Current status of thyroid cancer screening in Korea: results from a nationwide interview survey, *Asian Pacific Journal of Cancer Prevention: APJCP*, **12**, 1657–1663.

Jung, K. W., Won, Y. J., Kong, H. J., Oh, C. M., Cho, H., Lee, D. H. and Lee, K. H. (2015). Cancer statistics in Korea: Incidence, mortality, survival, and prevalence in 2012. *Cancer Research and Treatment : Official Journal of Korean Cancer Association.*

Kim, D. J., Ki, M., Kim, M. H., Kim, Y. M., Yoon, T. H., Jang, S. R., JangChoi, K. H., Kang, A. R., Chae, H. R. and Choi, J. H. (2014). *Developing Health Inequalities Indicators and Monitoring the Status of Health Inequalities in Korea*, 2014-04, Korea Institute for Health and Social Affairs.

Kweon, S. S., Shin, M. H., Chung, I. J., Kim, Y. J. and Choi, J. S. (2013). Thyroid cancer is the most common cancer in women, based on the data from population-based cancer registries, South Korea. *Japanese Journal of Clinical Oncology*, **43**, 1039–1046.

Lambert, D. (1992). Zero-inflated Poisson Regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Lee, J. H. and Shin, S. W. (2014). Overdiagnosis and screening for thyroid cancer in Korea, *Lancet*, **384**, 1848.

Lee, T. J., Kim, S., Cho, H. J. and Lee, J. H. (2012). The incidence of thyroid cancer is affected by the characteristics of a healthcare system. *Journal of Korean Medical Science*, **27**, 1491–1498.

Li, N., Du, X. L., Reitzel, L. R., Xu, L. and Sturgis, E. M. (2013). Impact of enhanced detection on the increase in thyroid cancer incidence in the United States: review of incidence trends by socioeconomic status within the surveillance, epidemiology, and end results registry, 1980-2008. *Thyroid : Official Journal of the American Thyroid Association*, **23**, 103–110.

Londero, S. C., Krogdahl, A., Bastholt, L., Overgaard, J., Pedersen, H. B., Frisch, T., Bentzen, J., Pedersen, P. U., Christiansen, P. and Godballe, C. (2013). Papillary thyroid carcinoma in Denmark 1996-2008: An investigation of changes in incidence, *Cancer Epidemiology*, **37**, e1–6.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena, *Biometrika*, **37**, 17–23.

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

The BUGS project | MRC Biostatistics Unit, WinBUGS 1.4.3, (2007). www.mrc-bsu.cam.ac.uk/software/bugs/.

영과잉 공간자료의 분석

한준희[a,1] · 김창훈[b]

[a]양산부산대학교병원 연구통계지원실, [b]부산대학교 의학전문대학원

(2015년 3월 16일 접수, 2015년 3월 31일 수정, 2015년 3월 31일 채택)

요 약

가산자료(counts data)를 적합 하는 경우 보통 포아송 모형이 가장 먼저 고려된다. 과산포 문제가 있을 경우도 유사 포아송(quasi Poisson) 모형이나 음이항(Negative binomial) 모형으로 대부분 설명이 가능하다. 하지만, 가산자료 중에는 포아송분포를 가정한 기대 빈도 이상으로 많은 0이 관측되는 자료가 있고 이를 영과잉(Zero inflated) 가산자료라고 부른다. 영과잉 가산자료를 설명하기 위해 영과잉 포아송(ZIP) 모형이나 영과잉 음이항(ZINB) 모형을 이용할 수 있다. 더 나아가 영과잉 가산자료가 공간상관관계까지 있을 경우 영과잉 문제뿐만 아니라 유의할 수 있는 공간효과까지 고려해야하고 이를 위해 혼합효과모형(mixed effects model)이 고려 될 수 있다. 본 연구에서 사용된 2004년 기준 부산시 남성동별 갑상선암 발생자수 자료를 이용하여, 일반선형 포아송모형, 영과잉 포아송모형, 공간 영과잉 포아송모형을 적합하여 비교해보았다.

주요용어: 가산자료, 포아송모형, 영과잉자료, 영과잉 포아송모형, 공간 영과잉 포아송모형, WinBUGS.

본 연구는 보건복지부 암정복추진연구개발사업 지원으로 이루어진 것임 (0920050).

[1]교신저자: (626-770) 경남 양산시 물금읍 금오로 20, 양산부산대학교병원 연구통계지원실.
E-mail: pnuyh.rass@gmail.com