응용통계연구 (2008)
21(5), 855–866

이웃정보시스템을 이용한 공간 소지역 추정량 비교

김정숙[1] · 황희진[2] · 신기일[3]

[1]한국외국어대학교 통계학과; [2]한국외국어대학교 통계학과; [3]한국외국어대학교 통계학과

(2008년 6월 접수, 2008년 7월 채택)

요 약

최근 격자자료(lattice data) 분석 방법을 이용한 소지역 추정(small area estimation)이 연구되고 있으며 좋은 결과를 주고 있는 것으로 알려져 있다. 소지역 추정에 주로 사용되는 격자자료(lattice data) 분석의 경우 가장 자료를 잘 설명할 수 있는 이웃정보시스템을 사용하여야 분석의 효율을 향상시킬 수 있다. 최근 이강석과 신기일 (2008)은 지리정보시스템을 이용하여 만들어진 여러 이웃정보시스템을 비교, 분석하였다. 본 논문에서는 이강석과 신기일 (2008)이 제안한 여러 이웃정보시스템이 소지역 추정에 얼마나 영향을 미치는지를 MSE, 커버리지, 캘리브레이션 그리고 회귀분석 방법 등을 이용하여 비교하였다. 2001년 경제활동인구조사의 실업자수 자료가 비교에 사용되었다.

주요용어: 공간통계, 지리정보시스템, 공간자기회귀모형, 직접추정량.

1. 서론

변수의 총계를 추정하거나 평균을 추정하기 위해 표본 조사가 일반적으로 사용된다. 표본 규모가 충분히 큰 경우, 직접추정량 또는 설계기반 추정량(Design based estimator)이 사용되며 이 경우 우리가 원하는 수준의 정도(Precision)를 얻을 수 있다. 그러나 표본 수가 작거나 소지역 추정의 경우는 다르다. Rao (2003)에 따르면 소지역은 조사 설계 당시 계획되어 있지 않아서 '우리가 원하는 정도(Precisoin)의 추정량을 얻을 수 있을 만큼 크지 않은 모든 도메인'으로 정의된다. 따라서 소지역 추정에서는 계획에 없던 지역을 추정해야 하는 어려움이 따르며 가장 일반적인 문제점은 해당 소지역에 정도를 높일 수 있을 만큼 충분한 표본 수가 할당되지 않는다는 것이다. 이를 극복하기 위한 여러 가지 방안에 대한 연구가 국내외적으로 활발히 진행되고 있다. 소지역 추정량은 크게 직접추정량, 합성추정량, 복합추정량 등의 자료기반 추정량(Data based estimator)과 관심 변수와 상관성이 높은 설명 변수가 존재할 때 사용 가능한 모형기반 추정량(Model based estimator)으로 나눌 수 있다. 모형기반 추정량은 자료기반 추정량에 비해 우수한 결과를 주는 것으로 알려져 있으며 회귀분석 방법, 경험적 베이지안 추정법(EB: Empirical Bayesian method), 계층적 베이지안 추정법(HB: Hierarchical Bayesian method) 등이 있다. 김달호와 김재광 (2004)의 연구에서도 알 수 있듯이 일반적으로 HB 추정량이 가장 우수한 결과를 주는 것으로 알려져 있다. 그런데 모형기반 추정량의 기본 가정은 '충분한 설명변수가 있을 경

이 연구는 2008년도 한국외국어대학교 교내연구비에 의해 수행되었음

[1](449-791)경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 자연과학대학 통계학과, 석사과정.
E-mail: hjs31428@naver.com

[2](449-791)경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 자연과학대학 통계학과, 박사과정.
E-mail: lshhj01@naver.com

[3]교신저자: (449-791)경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 자연과학대학 통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

우'이며 이 가정이 만족되지 않을 경우 모형기반 추정량의 사용은 제한적일 수밖에 없다. 충분한 설명변수의 존재 여부는 추정량의 정도에 결정적 영향을 미치는 매우 중요한 요인이지만 적당한 설명변수를 구하는 것이 쉽지 않다. 최근 국내에서는 추가적인 정보를 얻기 위한 여러 방법이 연구되었으며 그 중에서도 간단하면서도 효용성이 높은 방법으로 공간 통계 기법이 제안되었다. 이에 관한 논문은 김정오와 신기일 (2006), 이상은 (2006)을 참조하기 바란다.

김정오와 신기일 (2006)에서 사용한 공간 통계 추정 방법은 SAR 모형(Spatial autoregressive model)을 이용한 방법이며 이를 사용하기 위해서는 이웃정보시스템(Neighborhood information system)이 필요하다. 대부분 "경계를 공유하는 경우 이웃"을 이용한 이웃정보시스템을 사용하였으나 이강석과 신기일 (2008)은 GIS를 이용한 여러 이웃정보시스템을 제안하였고 Moran's I를 이용하여 이들을 비교하였다. 일반적으로 SAR 모형에서는 이웃정보시스템이 매우 중요한 역할을 하는 것으로 알려져 있다. 따라서 이웃정보시스템이 공간 소지역 추정에 미치는 영향을 비교, 분석하는 것은 매우 중요하다고 하겠다.

본 논문에서는 먼저 2장에서 이강석과 신기일 (2008)이 제안한 이웃정보시스템을 구하는 다양한 방법을 간단히 살펴보고, 3장에서는 소지역 추정에 사용되는 직접추정량과 여러 가지 이웃정보로 얻어진 공간추정량 그리고 이들을 선형결합한 선형결합추정량들을 살펴보았다. 최근 여러 소지역 추정량을 비교하기 위한 방법이 제안되었으며 4장에 제안된 비교 통계량을 소개하였다. 5장에서는 비교 통계량을 이용하여 소지역 추정량을 비교하였으며 이웃정보시스템에 따라 공간통계 추정량이 어떠한 영향을 받는지 살펴보았다.

2. 이웃정보 구하는 방법

공간 추정법에서 이웃을 결정하는 방법에는 크게 각 조사 지역의 공간적 위치와 거리를 기준으로 결정하는 방법과 각 지역의 상호 교류 내역을 기준으로 정하는 방법이 있다. 본 논문에서는 이강석과 신기일 (2008)에서 제안한 위치와 거리를 기준으로 이웃을 정하는 방법을 살펴보았다.

2.1. 경계기준

가장 간단하게 이웃을 정하는 방법은 지도상에서 같은 경계를 갖는 지역을 모두 이웃이라 정하는 것이다. 따라서 지역마다 이웃의 개수는 모두 다르다. 일반적으로 경계를 어느 정도 같이 공유하고 있는지 그 길이에 따라 가중치를 주는 방법도 있으나 이강석과 신기일 (2008)의 분석 결과, 가중치가 "1"인 경우 공간상관관계가 가장 높게 나타났으므로 본 논문에서는 경계를 공유하면 경계의 길이에 상관없이 가중치 "1"을 갖도록 하였다.

2.2. 거리기준

최근 GIS의 발달로 지역 간 거리를 측정하는 것이 수월해졌다. 이를 이용하면 미리 특정 거리를 정해놓고 정해진 거리 안에 있는 모든 지역을 이웃(Neighbor)으로 정하는 이웃정보시스템을 만들 수 있다. 이 경우 이웃의 개수는 조사 지역에 따라, 거리기준에 따라 달라지게 된다. 예를 들어 전라북도 정읍시의 경우 경계를 같이 하는 시군구는 김제시, 완주군, 임실군, 순창군, 고창군, 부안군 6개이다. 그런데 각각 시군청간의 거리를 조사해 보면 김제시 26,159m, 완주군 40,458m, 임실군 38,949m, 순창군 33,548m, 고창군 20,390m, 부안군 21,069m이다. 만약 이웃이 되는 거리의 제한을 30km로 정한다면 3개가 이웃이 되고 40km로 변경하면 5개가 이웃이 된다. 여기서는 이강석과 신기일 (2008)에서와 같이 30km를 이웃의 경계로 정하였다. 이 경우도 거리가 멀어지면 상관관계가 약화되도록 가중치를 줄 수도 있으나

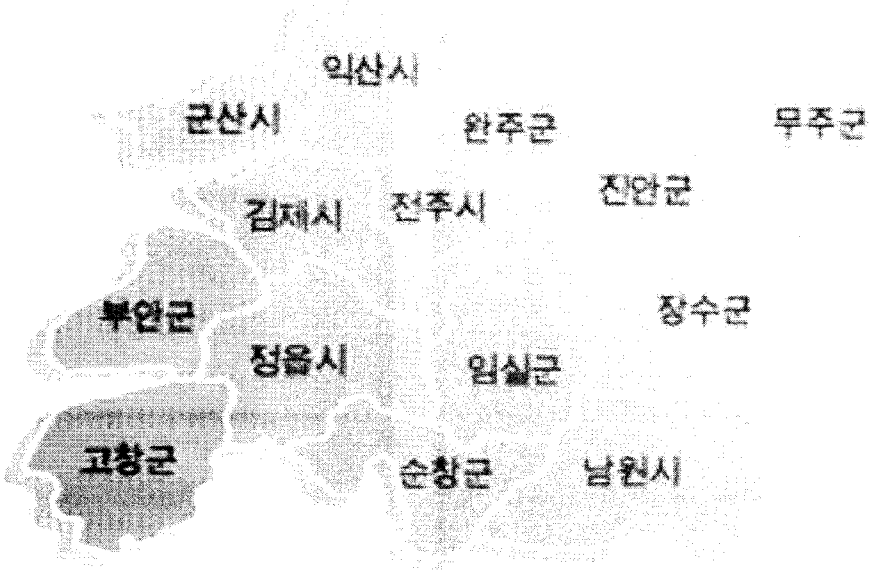

그림 **2.1.** 전라북도 지역 시군구 지도

경계기준에서와 마찬가지로 가중치가 "1"인 경우 공간상관관계가 가장 높았으므로 가중치는 모두 "1"로 같게 하였다.

2.3. 최근거리 기준

다음으로 많이 사용하는 것이 최근점 이웃(Nearest neighbor)이다. 먼저 특정한 k개를 이웃으로 하기로 정한 후 거리기준으로 가장 가까운 k개 지역을 이웃으로 정한다. 예를 들어 $k=3$을 정했다면 정읍시의 경우 고창군, 부안군, 김제시가 이웃이 된다. 이강석과 신기일 (2008)의 결과에 의하면 $k=3$인 경우가 가장 높은 Moran's I를 주기때문에 본 논문에서는 $k=3$을 사용하였다.

2.4. 가중치를 사용한 경우

위의 방법은 모두 거리 또는 경계선의 길이에 상관없이 이웃 지역이 같은 영향을 준다는 가정 하에서 얻어진 이웃정보이다. 하지만 거리가 멀어지면 영향력도 그만큼 약화될 가능성이 높으며 이러한 현상을 가중치에 적용하여 이웃정보시스템을 만들 수 있다. 다음은 이강석과 신기일 (2008)에서 사용한 가중치를 구하는 방법이다.

$$w_{ij} = \begin{cases} w_{ij}, & \text{만약 } i\text{와 } j\text{가 이웃이면,} \\ 0, & \text{그렇지 않으면,} \\ 0, & \text{만약 } i{=}j\text{이면,} \end{cases}$$

여기서 w_{ij}는 다음의 여러 가지 방법으로 정의 될 수 있으며 $\sum_j w_{ij} = 1$을 만족한다.

1) $w_{ij} = d_{ij}^{-\gamma}$, $\gamma > 0$, d_{ij}는 조사 지역 i와 j의 거리로 본 논문에서는 각 시군청간의 거리를 나타낸다.

2) $w_{ij} = (l_{ij}/L_i)^\tau$, $\tau > 0$, 여기서 $L_i = \sum_j l_{ij}$이고 l_{ij}는 i번째 지역과 j번째 지역의 공통 경계선의 길이이다.

3) $w_{ij} = (l_{ij}/L_i)^\tau / d_{ij}^\gamma$. 두 조사 지역간의 거리와 공통 경계선의 길이를 함께 사용한 가중치이다.

이강석과 신기일 (2008)의 2001년 4월 실업자 수 자료를 이용한 분석결과를 보면 가중치가 "1"인 경우에 Moran's I가 가장 높게 나왔다. 이에 본 논문에서는 가중치가 모두 "1"인 경우만을 고려하였다. 보다 자세한 내용은 Cressie (1993) 또는 이강석과 신기일 (2008)을 살펴보기 바란다.

3. 소지역 추정량

본 논문에서 비교하려는 소지역 추정량들은 직접추정량과 이웃정보시스템을 이용한 공간추정량 그리고 이들을 선형결합한 선형결합추정량이다. 이들 소지역 추정량은 설계기반 추정량이라 할 수 있으며 보조변수의 영향을 받지 않는다. 따라서 순수하게 이웃정보시스템의 영향력을 비교할 수 있게 된다.

3.1. 직접추정량(Direct estimator: $\hat{Y}_{DE}$)

소지역 추정량에서 가장 기본이 되는 추정량은 직접추정량이다. 이 추정량은 해당 소지역에 배정된 자료를 직접 이용하여 구해지며 불편성은 만족하나 소지역 추정에서는 일반적으로 표본수가 적고, 표본이 불균형적으로 추출되기 때문에 분산이 매우 큰 것으로 알려져 있다. 그럼에도 불구하고 모형기반 추정량에 대한 비교 기준으로 사용되고 있다. 사용된 추정량은 다음과 같다.

$$\hat{Y}_{DE} = \hat{Y}_i^{DE} = \sum_j w_{ij}^* y_{ij}, \tag{3.1}$$

여기서 $\hat{Y}_i^{DE}$는 i번째 소지역 값을 의미하고 w_{ij}^*는 추출 가중치를 그리고 y_{ij}는 표본에서 얻어진 실업자 수를 나타낸다.

3.2. 공간추정량(Spatial estimator: $\hat{Y}_{SP}$)

공간 추정법은 모형기반의 소지역 추정량에서 적절한 설명변수를 구하기 어려운 경우, 관심 변수들의 공간상관관계를 활용하여 추정하는 방법이다. 일반적으로 공간상관관계가 존재할 경우 직접추정량보다 좋은 결과를 얻을 수 있다. 사용된 추정량은 다음과 같다.

$$\hat{Y}_{SP} = \hat{Y}_i^{SP} = \bar{Y}^{DE} + \hat{\rho} S_i, \tag{3.2}$$

여기서 $\hat{Y}_i^{SP}$는 i번째 소지역 추정값이며 $\hat{\rho}$는 추정된 계수이고 S_i는 i번째 소지역의 이웃을 모두 더하여 얻어진 이웃정보 변수이다. 일반적으로 $S_i = \sum_{j \in N_i} w_{ij}(Y_j^{DE} - \bar{Y}^{DE})$이나 본 논문에서는 $w_{ij} = 1$인 경우만을 살펴보았기 때문에 $S_i = \sum_{j \in N_i}(Y_j^{DE} - \bar{Y}^{DE})$가 이웃정보 변수로 사용되었다. 여기서 N_i는 이웃의 개수이고 w_{ij}는 2.4절의 가중치를 의미한다. 다른 이웃정보시스템은 다른 이웃과 다른 N_i가 사용되기 때문에 얻어진 공간추정량은 서로 다르다.

3.3. 선형결합추정량(Linear combination estimator: $\hat{Y}_{DESP}$)

선형결합 소지역 추정량은 일반적으로 불편추정량인 직접추정량과 분산을 줄일 수 있는 모형기반 추정량의 선형결합으로 이루어진다. 예를 들면 직접추정량과 공간추정량을 선형결합함으로써 직접추정량의

변동성과 공간추정량의 편향을 동시에 줄일 수 있게 된다. 또한 직접추정량은 추정하려는 시군구에서 얻어진 자료를 이용하지만 공간추정량은 이웃에서 얻어진 정보를 이용하여 추정하기 때문에 이 두 추정량을 선형결합하여 얻어진 선형결합추정량은 의미가 있다고 하겠다. 직접추정량과 공간추정량의 선형결합추정량은 다음과 같다.

$$\hat{Y}_{DESP} = \alpha_{SP}\hat{Y}_{DE} + (1-\alpha_{SP})\hat{Y}_{SP}, \tag{3.3}$$

여기서 가중치 α_{SP}는 $\hat{Y}_{SP}$와 $\hat{Y}_{DE}$의 MSE를 사용하여 구해야 하나 많은 경우 MSE를 사용하는 대신 각 추정량의 분산을 이용하여 구한다. 즉 본 논문에서는

$$\alpha_{SP} = \frac{\text{Var}\left(\hat{Y}_{SP}\right)}{\text{Var}\left(\hat{Y}_{DE}\right) + \text{Var}\left(\hat{Y}_{SP}\right)} \tag{3.4}$$

을 이용하였다. 여기서 각 추정량의 분산은 붓스트랩 방법을 이용하여 구하였다. 즉 1,128개 자료를 복원추출하여 붓스트랩 샘플을 만들고 이를 5,000번 반복하여 분산을 구하였다. 그리고 본 논문에서 사용한 자료처럼 모든 자료가 "0"인 지역의 경우, 붓스트랩 방법을 사용하게 되면 $\hat{\text{Var}}(\hat{Y}_{DE}) = 0$이 되어 추정된 $\hat{\alpha}_{SP} = 1$이 된다. 따라서 이 지역의 선형결합추정량은 직접추정량과 같게 된다. 이 경우 본 논문에서는 $\hat{\alpha}_{SP} = 1/2$을 사용하였다. 이에 관한 자세한 설명은 Rao (2003)를 참조하고 α_{SP} 추정에 관한 내용은 Falosi 등 (1994)을 참조하기 바란다. 본 논문에서는 선형결합추정량으로 $\hat{Y}_{DE}$와 세 가지 이웃정보시스템을 이용하여 만들어진 공간추정량 3개를 선형결합하여 또 다른 3개의 선형결합추정량을 만들었다.

4. 비교 통계량

소지역 추정은 소규모의 자료에 의존하여 통계치를 생산하기 때문에 좋은 소지역 추정량이 되기 위해서는 몇 가지 조건이 필요하다. 먼저 모형에서 얻어진 소지역 추정량의 기대값이 의미가 있어야한다. 다음으로 직접추정량에서 얻어진 소지역 추정치와 모형기반 추정량에서 얻어진 소지역 추정치는 어느 정도 일치성을 보여야한다. 그리고 모형기반 추정량의 MSE는 직접추정량의 MSE보다 작아야 한다. 또한 모형기반 추정량은 직접추정량 보다 시간의 변화에 민감하지 않아야한다. 마지막으로는 소지역 추정량 결과가 각 소지역 거주 주민들 혹은 자료를 이용하는 사람들에게 상식적으로 받아들여지는 값이어야 한다는 것이다. 본 논문에서는 이러한 조건들이 통계적 모형으로 만들어져 최근에 제안된 모형검진 통계량 또는 비교 통계량을 통하여 이웃정보시스템의 영향력을 살펴보았다. 본 논문에서 사용된 비교 통계량은 Brown 등 (2001)과 McEwin와 Elazar (2006)에서 연구된 단순회귀식의 기울기와 R^2, Coverage, Calibration 그리고 MSE를 이용한 방법 등이다. 이에 관한 내용은 황희진과 신기일 (2008)을 참조하기 바란다.

4.1. 회귀모형을 이용한 방법

회귀모형을 이용한 진단 방법은 직접추정량이 불편추정량임을 활용한 것으로 직접추정량을 종속변수로 하고 각각의 다른 추정량을 독립변수로 하는 단순선형회귀모형을 만드는 것이다. 이 때 얻어지는 두 통계량, 기울기 $\hat{\beta}_1$과 결정계수 R^2을 이용하여 각 소지역 추정량의 특징을 비교한다. 만약 얻어진 추정량이 정확성을 유지하고 있다면 기울기에 대한 추정값은 $\hat{\beta}_1 \approx 1$을 만족하게 될 것이며 R^2도 "1"에 가까운 값을 얻게 될 것이다.

4.2. 커버리지(Coverage)

직접추정량은 큰 분산을 갖고 있으나 불편추정량이라고 알려져 있다. 커버리지는 직접추정량의 95% 신뢰구간을 구하고 이 구간에 각각의 추정량에서 얻어진 추정치의 몇 퍼센트가 포함되는가를 살펴보는 것이다. 만약 불편이고 분산이 작다면 추정값은 신뢰구간 안에 100% 포함이 될 것이고 반대로 편향이 있으며 분산이 작다면 겹치는 부분이 작거나 거의 없는 경우도 발생할 수 있다. 또한 편향은 없으나 분산이 크면 직접추정량과 겹치는 부분은 작게 될 것이다. 물론 대부분의 모형기반 추정량은 분산이 작기 때문에 이러한 결과는 발생하지 않을 것이다.

4.3. 캘리브레이션(Calibration)

직접추정량은 소지역을 합쳐 지역이 커지게 되면 그 지역에 포함된 표본 수가 증가하게 되고 따라서 일반적으로 정도가 높아지게 된다. 이러한 특징을 이용하여 여러 소지역을 합쳐가면서 각 모형기반 추정량에서 얻어진 추정치들을 비교하게 되는데 이렇게 비교하는 것을 캘리브레이션이라 한다. 본 논문에서 연구된 지역인 전라북도는 크게 시와 군으로 나눌 수 있다. 따라서 시와 군, 두개의 그룹으로 소지역을 확대한 후 직접추정량과 다른 추정량을 비교한다면 의미가 있을 것이다. 즉 만약 지역이 커졌음에도 불구하고 직접추정치와 다른 추정량에서 얻어진 추정치가 큰 차이가 난다면 그 추정량은 주어진 자료를 잘 설명한다고 할 수 없을 것이다.

4.4. MSE

MSE는 추정량 비교에 기본적으로 사용되는 비교 통계량으로 정의는 다음과 같다.

$$\mathrm{MSE}_i = \frac{1}{R}\sum_{r=1}^{R}\left(Y_i - \hat{Y}_i^{(r)}\right)^2,$$

여기서 $i = 1, \ldots, n$은 i번째 소지역을 의미하며 본 논문에서는 13개의 소지역이 있으므로 $n = 13$이 된다. 모의실험에서 사용된 반복수는 $R = 5,000$이 사용되었다. 이제 MSE를 구하기 위해서는 참값이 필요하게 된다. 그러나 참값을 알 수 없기 때문에 본 논문에서는 김정오와 신기일 (2006)에서 사용했던 방법을 사용하였다. 즉 1,128개의 자료에서 얻어진 $\hat{Y}_{DE}$를 참값 Y_i라 가정하였다. 다음으로 1,128개의 자료에서, 500개와 700개를 랜덤 추출한 후 각 소지역 추정량을 이용하여 계산한 결과를 $\hat{Y}_i^{(r)}$이라 하였다. 이렇게 얻어진 값을 R번 반복한 후 평균을 구하게 되면 13개 소지역별로 MSE_i를 얻게 된다. 또한 최종적으로 각 소지역을 평균한 값도 표에 작성하였다.

5. 자료 분석 및 추정량 비교

5.1. 자료 분석

본 논문에서는 2001년 4월 실업자 수 자료 중 공간상관관계가 높은 것으로 나타났던 전라북도 지역 자료만을 사용하였다. 총계 추정량을 구하기 위해서는 각 자료의 가중치(Weight)가 있어야 한다. 황희진과 신기일 (2008)에서는 가중치가 있어 소지역 추정에 사용하였다. 그러나 본 논문에서 사용한 자료에는 가중치가 존재하지 않는다. 따라서 추정량 비교는 김정오와 신기일 (2006)에서 사용한 방법을 사용하였다. 즉 가중치를 사용하지 않은 표본 자료를 사용하여 추정량을 비교하였다. 먼저 각 이웃정보시스템을 사용한 공간상관관계의 정도를 살펴보기 위하여 Moran's I를 구하였다.

표 5.1. 전라북도 지역 실업자수에 대한 MORAN'S I

	경계기준($\hat{Y}_{SP1}$)	거리기준($\hat{Y}_{SP2}$)	$k=3$개($\hat{Y}_{SP3}$)
Moran's I	0.6171	0.7484	0.6512

표 5.2. 추정량별 기울기와 R^2

추정량	기울기	R^2
경계기준 ($\hat{Y}_{SP1}$)	1.1021	0.9519
거리기준 ($\hat{Y}_{SP2}$)	1.1822	0.9332
$k=3$개 ($\hat{Y}_{SP3}$)	1.2400	0.9345
선형결합 ($\hat{Y}_{DESP1}$)	1.0935	0.9751
선형결합 ($\hat{Y}_{DESP2}$)	1.1547	0.9660
선형결합 ($\hat{Y}_{DESP3}$)	1.1995	0.9674

일반적으로 Moran's I가 0.2 이상이면 분석에서 유의미한 결과를 주는 것으로 알려져 있다. Moran's I를 구하는 방법은 Kaluzny 등 (1998)을 참조하기 바란다. 따라서 본 논문에서 사용된 자료는 높은 공간상관관계가 있으며 결과에 큰 영향을 줄 것으로 생각된다.

5.2. 추정량 비교

4장에서 소개한 비교 통계량을 이용하여 소지역 추정량들을 비교하였다.

5.2.1. 기울기와 R^2 직접추정량 $\hat{Y}_{DE}$를 종속변수로 공간추정량을 독립변수로 하여 회귀적합한 후 구해진 기울기와 R^2를 비교하였다. 표 5.2를 살펴보면 기울기가 "1"에서 가장 많이 벗어난 경우는 이웃을 $k=3$개로 정의한 $\hat{Y}_{SP3}$의 경우이고 가장 "1"에 가까운 경우는 경계를 기준으로 이웃을 정의한 $\hat{Y}_{SP1}$과 $\hat{Y}_{DESP1}$이다. R^2값 역시 $\hat{Y}_{DESP1}$가 가장 높았으며 거리기준으로 구한 추정량 $\hat{Y}_{SP2}$가 가장 낮았다. 하지만 대체로 높은 값을 나타내며 큰 차이는 보이지 않았다.

5.2.2. 커버리지 직접추정량 $\hat{Y}_{DE}$를 기준으로 구한 95% 신뢰구간을 구하고 여기에 다른 추정량들의 추정값이 어느 정도 포함되는지 표 5.3에서 살펴보았다. 대체로 모든 지역에서 90% 이상 신뢰구간에 포함되는 것으로 나타났으나 군산시의 경우 거리기준 이웃정보시스템 추정량 $\hat{Y}_{SP2}$와 $k=3$개인 이웃정보시스템 추정량 $\hat{Y}_{SP3}$은 커버리지가 낮았다. 그리고 직접추정량의 실업자 수가 "0"인 진안군, 장수군, 임실군 그리고 순창군 등 4개 군의 커버리지는 구하지 않았다.

5.2.3. 캘리브레이션 각 지역을 시와 군 등 두 그룹으로 나눈 후 추정값들을 합쳐 직접추정량과 비교하였다. 비교를 간단히 하기위해 직접추정값과의 비율도 함께 작성하였다. 표 5.4의 결과를 살펴보면, 경계기준 이웃정보시스템을 사용한 추정량 $\hat{Y}_{SP1}$과 $\hat{Y}_{DESP1}$이 가장 직접추정량에 근접한 것으로 나타났으며 $k=3$일 경우의 $\hat{Y}_{SP3}$가 직접추정량과 가장 큰 차이가 있어 편향이 존재하는 것으로 보인다. 전체적으로는 모든 추정량이 직접추정량에 비해 과소추정되고 있다.

5.2.4. MSE 각 추정량들의 우수성 비교를 위해서 MSE를 비교하였다. 앞에서 소개한 것처럼 직접추정량 $\hat{Y}_{DE}$를 참값으로 간주하고 1,128개 자료 중 500개를 랜덤추출하여 각 소지역 추정량에 대한 추

표 5.3. 추정량별 지역별 커버리지

지역	경계기준 ($\hat{Y}_{SP1}$)	거리기준 ($\hat{Y}_{SP2}$)	$k=3$개 ($\hat{Y}_{SP3}$)	선형결합 ($\hat{Y}_{DESP1}$)	선형결합 ($\hat{Y}_{DESP2}$)	선형결합 ($\hat{Y}_{DESP3}$)
전주시	96.77	97.91	96.90	98.48	99.66	99.16
군산시	90.02	79.04	81.76	97.98	97.10	97.56
익산시	98.56	98.52	98.48	99.68	99.88	99.88
정읍시	98.18	94.01	94.27	99.42	97.92	98.08
남원시	99.69	99.78	99.97	99.75	99.91	99.97
김제시	100.00	100.00	100.00	100.00	100.00	100.00
완주군	100.00	100.00	100.00	100.00	100.00	100.00
고창군	99.97	100.00	100.00	99.97	100.00	100.00
부안군	98.35	94.87	94.40	99.13	97.98	98.04
90%이하인 개수	0	1	1	0	0	0
최소값	90.02	79.04	81.76	97.98	97.10	97.56

표 5.4. 추정량별 캘리브레이션 결과

지역	직접 ($\hat{Y}_{DE}$)	경계기준 ($\hat{Y}_{SP1}$)	거리기준 ($\hat{Y}_{SP2}$)	$k=3$개 ($\hat{Y}_{SP3}$)	선형결합 ($\hat{Y}_{DESP1}$)	선형결합 ($\hat{Y}_{DESP2}$)	선형결합 ($\hat{Y}_{DESP3}$)
시	33	30.2953	28.1845	26.7752	30.5763	28.9687	27.6823
군	4	3.8085	4.3409	4.6552	3.5368	3.8974	4.2100
전체	37	34.1038	32.5254	31.4303	34.1131	32.8661	31.8922
지역	직접 추정값과의 비율 $\hat{Y}_*/\hat{Y}_{DE}$						
시		0.9180	0.8540	0.8110	0.9270	0.8780	0.8390
군		0.9520	1.0850	0.1640	0.8840	0.9740	1.0520
전체		0.9220	0.8790	0.8490	0.9220	0.8880	0.8620

표 5.5. 추정량별 MSE 비교(500개 추출)

지역	직접 ($\hat{Y}_{DE}$)	경계기준 ($\hat{Y}_{SP1}$)	거리기준 ($\hat{Y}_{SP2}$)	$k=3$개 ($\hat{Y}_{SP3}$)	선형결합 ($\hat{Y}_{DESP1}$)	선형결합 ($\hat{Y}_{DESP2}$)	선형결합 ($\hat{Y}_{DESP3}$)
전주시	3.8759	2.8336	2.6313	2.6922	2.1681	2.0325	2.0549
군산시	17.1652	28.7308	35.2765	32.2245	18.8881	22.3004	20.3458
익산시	20.6451	21.7436	22.8719	21.2428	13.6980	14.4294	14.3796
정읍시	1.2620	0.9310	1.0683	1.1236	0.5959	0.7048	0.7068
남원시	1.2545	0.9266	0.9980	0.8437	0.7623	0.8194	0.7041
김제시	8.7846	7.3355	7.6668	8.3975	5.4518	5.7213	6.8451
완주군	2.5452	1.5888	1.5254	1.3358	1.3942	1.3529	1.1094
진안군	0.0000	0.0140	0.0114	0.0164	0.0035	0.0029	0.0041
장수군	0.0000	0.1602	0.1422	0.1464	0.0400	0.0356	0.0366
임실군	0.0000	0.1857	0.0794	0.0624	0.0464	0.0198	0.0156
순창군	0.0000	0.1289	0.1433	0.1275	0.0322	0.0358	0.0319
고창군	1.2598	0.3046	0.3190	0.3959	0.2984	0.3103	0.3753
부안군	1.2560	0.4447	0.7745	0.8679	0.4118	0.5867	0.6301
평균	4.4652	5.0252	5.6545	5.3443	3.3685	3.7194	3.6338

정값을 계산하고 지역별 MSE_i를 구하였다. 또한 각 지역별 MSE_i를 다시 평균하여 최종적으로 얻은 MSE 평균값도 표 5.5에 작성하였다.

표 **5.6.** 추정량별 MSE 비교(500개 추출 = $\mathrm{MSE}_*/\mathrm{MSE}_{DE}$)

지역	경계기준 ($\hat{Y}_{SP1}$)	거리기준 ($\hat{Y}_{SP2}$)	$k=3$개 ($\hat{Y}_{SP3}$)	선형결합 ($\hat{Y}_{DESP1}$)	선형결합 ($\hat{Y}_{DESP2}$)	선형결합 ($\hat{Y}_{DESP3}$)
전주시	0.731	0.679	0.695	0.559	0.524	0.530
군산시	1.674	2.055	1.877	1.100	1.299	1.185
익산시	1.053	1.108	1.029	0.663	0.699	0.697
정읍시	0.738	0.847	0.890	0.472	0.559	0.560
남원시	0.739	0.796	0.673	0.608	0.653	0.561
김제시	0.835	0.873	0.956	0.621	0.651	0.779
완주군	0.624	0.599	0.525	0.548	0.532	0.436
고창군	0.242	0.253	0.314	0.237	0.246	0.298
부안군	0.354	0.617	0.691	0.328	0.467	0.502
평균	1.125	1.266	1.197	0.754	0.833	0.814

표 5.5을 보면 추정값이 상대적으로 큰 지역인 군산시와 익산시의 MSE가 다소 크게 나타났다. 이는 추정값 자체가 큰 값이기 때문에 이에 해당되는 분산이 큰 값으로 나온 것으로 생각할 수 있으나, 공간추정량이 갖고 있는 문제점을 보여주고 있는 결과라고 할 수 있다. 즉 군산시와 익산시는 전라북도에서 실업자수가 많은 두 지역이다. 그러나 그 이웃지역의 실업자 수는 많지 않다. 결국 실업자 수가 많지 않은 이웃정보를 이용하여 실업자 수가 많은 지역을 예측하게 되면 과소추정 될 수밖에 없으며 따라서 참값과 차이를 보이게 된다. 물론 이러한 현상은 공간추정량뿐만 아니라 일반적인 추정량에서도 나타나는 공통된 현상이다.

이제 본 논문의 목적인 제안된 추정량들을 비교하자. 선형결합을 하지 않은 경우에는 $\hat{Y}_{SP1}$이 가장 좋은 결과를 주는 반면 $\hat{Y}_{SP2}$가 가장 나쁜 결과를 주고 있다. 이 결과는 Moran's I 결과와는 다른 결과이다. 선형결합을 한 결과를 비교하면, 우수성의 순서에는 변화가 없지만 전체적인 MSE는 매우 향상된 것을 확인 할 수 있다. 비교를 간단히 하기 위해 표 5.6에는 각 추정량의 MSE 값을 직접추정량의 MSE 값으로 나눈 결과를 작성하였다. 군산시의 경우 모든 추정량의 MSE가 직접추정량의 MSE에 비해 크게 나타났으나, 다른 소지역의 경우 공간추정량이 우수한 것을 확인 할 수 있다. 특히 선형결합추정량의 경우는 군산시를 제외한 모든 소지역에서 직접추정량보다 우수한 것을 확인 할 수 있다.

또한 표 5.7과 5.8에는 700개의 자료를 추출하여 구한 MSE의 결과와 직접추정량과의 비율을 구한 결과를 각각 작성하였다.

표 5.7를 살펴보면 700개 자료를 추출하였기 때문에 표 5.5에 비해 더 작은 MSE 값을 보이고 있다. 그러나 표 5.5에서 얻어진 결과와 같이 $\hat{Y}_{SP1}$의 MSE가 가장 작았으며 $\hat{Y}_{SP2}$, $\hat{Y}_{SP3}$는 거의 같은 결과를 주고 있다. 또한 선형결합추정량에서도 같은 결과를 주고 있다. 표 5.8을 살펴보면 "1"보다 큰 값들이 많이 있으며 이는 직접추정량의 경우 자료가 커지게 되면 분산이 줄어들기 때문에 다른 추정량에 비해 빠른 속도로 정도가 좋아지기 때문인 것으로 판단된다. 이 결과는 김정오와 신기일(2006)에서도 확인 할 수 있다. 그럼에도 불구하고 $\hat{Y}_{DESP1}$의 MSE는 직접 추정량에 비해 우수한 것을 확인할 수 있다.

6. 결론

본 논문에서는 여러 가지 이웃정보시스템에 따른 소지역 추정량의 효용성을 여러 가지 비교통계량을 기준으로 살펴보았다. Moran's I 결과로는 거리기준 공간추정량인 $\hat{Y}_{SP2}$가 가장 좋은 결과를 줄 것으로 예상되었다. 본 논문에서 사용된 비교 통계량의 결과를 보면 경계공유기준 공간추정량인 $\hat{Y}_{SP1}$이 가

표 **5.7.** 추정량별 MSE 비교(700개 추출)

지역	직접 ($\hat{Y}_{DE}$)	경계기준 ($\hat{Y}_{SP1}$)	거리기준 ($\hat{Y}_{SP2}$)	$k=3$개 ($\hat{Y}_{SP3}$)	선형결합 ($\hat{Y}_{DESP1}$)	선형결합 ($\hat{Y}_{DESP2}$)	선형결합 ($\hat{Y}_{DESP3}$)
전주시	1.8770	2.4039	2.0117	2.1838	1.4242	1.2850	1.3738
군산시	8.7122	21.0349	29.1218	26.7344	13.1128	17.6947	16.1513
익산시	10.3408	10.5791	11.8636	12.2159	6.5242	7.2928	7.9352
정읍시	0.6169	0.7024	1.0458	1.1533	0.3785	0.5352	0.5806
남원시	0.6115	0.2192	0.2714	0.2164	0.1942	0.2317	0.1888
김제시	4.3665	4.4035	4.9211	6.2738	3.1169	3.5025	5.0044
완주군	1.2166	1.2503	1.1177	0.6811	1.0692	0.9668	0.5631
진안군	0.0000	0.0042	0.0035	0.0052	0.0011	0.0009	0.0013
장수군	0.0000	0.0839	0.0679	0.0734	0.0210	0.0170	0.0183
임실군	0.0000	0.1144	0.0608	0.0315	0.0286	0.0152	0.0079
순창군	0.0000	0.0757	0.0949	0.0739	0.0189	0.0237	0.0185
고창군	0.6110	0.2351	0.2575	0.4263	0.2083	0.2261	0.3825
부안군	0.6087	0.4219	0.9965	1.1080	0.2983	0.5795	0.6571
평균	2.2278	3.1945	3.9872	3.9367	2.0305	2.4901	2.5294

표 **5.8.** 추정량별 MSE 비교(700개 추출 = $\mathrm{MSOE}_*/\mathrm{MSE}_{DE}$)

지역	경계기준 ($\hat{Y}_{SP1}$)	거리기준 ($\hat{Y}_{SP2}$)	$k=3$개 ($\hat{Y}_{SP3}$)	선형결합 ($\hat{Y}_{DESP1}$)	선형결합 ($\hat{Y}_{DESP2}$)	선형결합 ($\hat{Y}_{DESP3}$)
전주시	1.281	1.072	1.163	0.759	0.685	0.732
군산시	2.414	3.343	3.069	1.505	2.031	1.854
익산시	1.023	1.147	1.181	0.631	0.705	0.767
정읍시	1.138	1.695	1.869	0.613	0.867	0.941
남원시	0.358	0.444	0.354	0.318	0.379	0.309
김제시	1.008	1.127	1.437	0.714	0.802	1.146
완주군	1.028	0.919	0.560	0.879	0.795	0.463
고창군	0.385	0.421	0.698	0.341	0.370	0.626
부안군	0.693	1.637	1.820	0.490	0.952	1.080
평균	1.434	1.790	1.767	0.911	1.118	1.135

장 우수한 결과를 주는 것으로 나타났다. 비록 $\hat{Y}_{SP1}$이 가장 우수한 결과를 주는 것으로 나타났지만 Moran's I 결과도 큰 차이를 보이지 않고 또한 본 논문에서 사용한 여러 비교 통계량의 결과를 살펴보아도 큰 차이는 없는 것으로 판단된다. 결국 일정 수준 이상의 타당한 이웃정보시스템을 이용할 경우에는 이웃정보시스템이 분석 결과에 큰 영향을 미치지 않는 것으로 판단할 수 있다. 그러나 공간추정량만을 사용하는 것보다는 선형결합을 함으로써 매우 좋은 결과를 얻을 수 있었다. 따라서 경계를 공유하는 이웃정보시스템을 사용하여 만들어진 소지역 추정량은 큰 문제없이 사용할 수 있으며 선형결합추정량을 사용하면 더 좋은 결과를 얻을 수 있다고 판단된다.

참고문헌

김달호, 김재광 (2004). 가계조사 지역별 추정기법, <통계청 용역보고서>.

김정오, 신기일 (2006). Comparison of small area estimation by sample sizes, <한국통계학회논문집>, **13**, 669–683.

이강석, 신기일 (2008). 격자자료분석을 위한 이웃정보시스템의 비교, <응용통계연구>, **21**, 387–397.

이상은 (2006). 공간통계량을 활용한 베이지안 자기포아송 모형을 이용한 소지역 통계, <응용통계연구>, **19**, 421–430.

황희진, 신기일 (2008). 축소예측을 이용한 소지역 추정, <응용통계연구>, **21**, 109–123.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods-application to unemployment estimates from the UK LFS, In *Proceedings of Statistics Canada Symposium 2001*.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, John Wiley & Sons, New York.

Falosi, P. D., Falosi, S. and Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian labour force survey, *Survey Methodology*, **20**, 171–176.

Kaluzny, S. P., Vega, S. C., Cardoso, T. P. and Shelly, A. A. (1998). *S+ Spatial Stats: User's Manual for Windows and UNIX*, Springer, New York.

McEwin, M. and Elazar, D. (2006). *Regional Statistics: Small Area Estimation in Official Statistics*, UNESCAP, APEX2.

Rao, J. N. K. (2003). *Small Area Estimation*, John Wiley & Sons, New York.

Comparison of Spatial Small Area Estimators Based on Neighborhood Information Systems

Jeong-Suk Kim[1] · Hee-Jin Hwang[2] · Key-Il Shin[3]

[1]Dept. of Statistics, Hankuk University of Foreign Studies;
[2]Dept. of Statistics, Hankuk University of Foreign Studies;
[3]Dept. of Statistics, Hankuk University of Foreign Studies

(Received June 2008; accepted July 2008)

Abstract

Recently many small area estimation methods using the lattice data analysis have been studied and known that they have good performances. In the case of using the lattice data which is mainly used for small area estimation, the choice of better neighborhood information system is very important for the efficiency of the data analysis. Recently Lee and Shin (2008) compared and analyzed some neighborhood information systems based on GIS methods. In this paper, we evaluate the effect of various neighborhood information systems which were suggested by Lee and Shin (2008). For comparison of the estimators, MSE, Coverage, Calibration, Regression methods are used. The number of unemployment in Economic Active Population Survey(2001) is used for the comparison.

Keywords: Spatial statistics, geographic information system, spatial autoregressive model, direct estimator.

This research was supported by the research fund of Hankuk University of Foreign Studies, 2008.

[1]Graduate student, Dept. of Statistics, Hankuk University of Foreign Studies, San 89, Wangsan, Mohyun, Yongin, Kyonggi Do 449-791, Korea. E-mail: hjs31428@naver.com

[2]Graduate student, Dept. of Statistics, Hankuk University of Foreign Studies, San 89, Wangsan, Mohyun, Yongin, Kyonggi Do 449-791, Korea. E-mail: lshhj01@naver.com

[3]Corresponding Author: Professor, Dept. of Statistics, Hankuk University of Foreign Studies, San 89, Wangsan, Mohyun,Yongin, Kyonggi Do 449-791, Korea. E-mail: keyshin@hufs.ac.kr