응용통계연구 제 **20**권 2호, *2007*년, *pp. 229–244*

공간 통계 활용에 따른 소지역 추정법의 평가*

신기일[1] 최봉호[2] 이상은[3]

요 약

국내외에서 소지역 추정에 관한 많은 연구가 진행되고 있다. 보조 자료가 충분히 있는 경우 모형기반 추정법을 사용하는 것이 일반적이며 이 중에서 계층적 베이지안(Hierarchi -cal Bayesian: HB) 추정법이 가장 좋은 것으로 알려져 있다. 그러나 보조 자료가 충분하지 않은 경우에는 모형기반 추정법의 사용은 제한적이다. 최근 충분한 보조 자료가 없는 경우 공간 정보를 보조 자료로 사용하는 방법이 제안되었다. 본 논문에서는 공간통계량과 베이즈 접근방법을 활용한 모형기반의 소지역 통계량들을 모형 검진방법(Diagnostic method)들을 이용하여 비교분석하였다. 분석에 사용된 자료는 2005년도 경제활동인구조사이며 소지역(시, 군, 구)통계를 추정하여 비교하였다.

주요용어: 소지역 통계, 공간 통계, 경험적 베이지안 추정법, 계층적 베이지안 추정법.

1. 서론

표본조사 자료를 이용하여 총계를 추정하거나 평균을 추정하는 것은 매우 일반적이다. 표본의 크기가 충분히 큰 경우에 직접 추정량 또는 조사 설계 추정량(Design based estimator)은 원하는 수준의 정도(Precision)를 유지할 수 있다. 그러나 소지역 추정인 경우 상황은 달라진다. Rao(2003)의 소지역 정의는 '조사 설계 당시 일정 오차를 갖는다고 생각되지 않는 비계획적인 지역'이다. 따라서 소지역 추정에는 계획에 없던 지역을 추정해야 하는 어려움이 있으며 가장 일반적으로 발생하는 어려움은 해당 소지역에 할당된 표본 수가 부족하다는 것이다. 이를 극복하기 위한 연구와 여러 방법이 국외 뿐 아니라 국내에서도 활발히 진행되고 있다. 소지역 추정량은 직접 추정량, 합성 추정량, 복합 추정량 등의 자료기반 추정량(Data based estimator)과 설명 변수가 충분히 존재할 때 사용하는 모형기반 추정량(Model based estimator)이 있다. 모형기반 추정량은 자료기반 추정량에 비해 우수한 결과를 주는 것으로 알려져 있으며 회귀분석 방법, 경험적 베이지안(Empirical Baysian: EB) 추정법, 계층적 베이지안(Hierarchical Bayesian: HB) 추정법 등이 있다. 이 중에서 일반적

* 이 논문은 2007년도 한국외국어대학교 학술연구비 지원에 의해 수행 되었음

1) (442-791) 경기도 용인시 모현면 왕산리 산 79, 한국외국어대학교 통계학과, 교수
E-mail: keyshin@hufs.ac.kr
2) (302-701) 대전광역시 서구 선사로 139 정부대전청사3동, 통계청 지역통계과, 과장
E-mail: bonghoo.choi@nso.go.kr
3) (442-702) 경기도 수원시 영통구 이의동 산 94-3, 경기대학교 응용통계학과, 부교수
E-mail: sanglee62@kyonggi.ac.kr

으로 계층적 베이지안 방법이 우수한 것으로 알려져 있으며 김달호와 김재광(2004)이 통계청 연구 용역으로 수행한 소득 자료 분석 결과에서도 이를 확인할 수 있다. 그러나 모형기반 추정량의 기본 가정은 '충분한 설명변수가 있을 경우'이며 이 가정이 만족되지 않을 경우 모형기반 추정량의 사용은 제한적일 수밖에 없다. 이와 같이 충분한 설명변수는 분석 방법 및 분석 결과를 좌우하는 중요한 요인이 된다. 따라서 충분한 보조변수를 확보하는 것이 매우 중요하다. 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) 그리고 Lee와 Shin(2006)을 참조하기 바란다. 이상과 같이 소지역 추정을 위해 자료기반 추정량 뿐 아니라 모형기반 추정량 등 여러 추정량이 제안되었으나 상대적으로 이들을 비교하기 위한 비교통계량의 개발은 이루어지지 않고 있었다. 최근 소지역 추정량의 비교를 위한 여러 방법이 제안되었으며 본 논문에서는 제안된 비교 통계량을 이용하여 현재 많이 사용하고 있는 추정량인 직접 추정량과 모형기반 추정량 중에서 주로 사용되고 있는 회귀 추정량, 경험적 베이즈 추정량, 계층적 베이즈 추정량, 공간 추정량을 비교하였다. 또한 직접 추정량과 모형기반 추정량을 선형 결합하여 얻은 선형추정량도 함께 비교하였다. 비교에 사용된 자료는 2005년도 경제활동인구이다.

2절에서는 각 추정량의 비교에 사용된 진단 방법을 설명하였으며 3절에서는 소지역 추정에 사용된 추정량을 설명하였다. 4절에서는 각 소지역 추정량을 2절에서 설명한 진단 방법을 이용하여 비교하였으며 최종 결론은 5절에 있다.

2. 진단 방법(Diagnostic Method)

소지역 통계는 소규모의 자료에 의존하여 통계를 생산하기 때문에 소지역 통계 생산에 기본적으로 요구되는 몇 가지 조건이 있다. 대부분의 소지역 통계에서는 모형기반 추정량이 사용되므로 모형기반 추정량에 관한 내용을 살펴보자. 먼저 모형기반 추정량의 MSE는 직접 추정량보다 작아야한다. 그러나 직접 추정량과 모형기반 추정량 사이에는 어느 정도의 일치성이 보여야한다. 그리고 모형기반 추정량은 시간의 변화에 직접 추정량 보다 민감하지 않아야한다. 마지막으로는 소지역 추정량의 결과가 소지역에 사는 주민들 혹은 자료를 이용하는 사람들에게 받아들여지는 수치여야 한다는 것이다. 이와 같은 요구 사항들은 소지역 통계를 활용하고자 하는 나라들의 통계청에서 이미 지켜지고 있다. 본 논문에서는 최근에 제안된 모형 검진 통계량 또는 비교 통계량 중에서 직접 추정량을 기준으로 제안된 통계량을 살펴보았다. 즉 논문에서 사용된 비교 통계량은 Brown 등(2001)에서 연구된 절편이 없는 단순회귀식의 기울기와 R^2, 커버리지(Coverage) 그리고 캘리브레이션(Calibration) 방법 등이다.

2.1. 회귀모형을 이용한 방법

회귀모형을 이용한 진단방법은 직접 추정량이 불편 추정량임을 활용하여 모형기반 추정량의 불편성을 진단하는 것으로 내용은 다음과 같다. 먼저 직접 추정량을 종속변수로 하고 모형을 이용하여 얻은 소지역 추정량을 독립변수로 하는 절편이 없는 단순선형회귀모

형을 만든다. 즉 각 소지역에서 얻어진 직접 추정치와 모형기반 추정치를 이용하여 단순회귀식을 구한다. 이 때 얻어지는 두 통계량, 기울기 $\hat{\beta}_1$와 결정계수 R^2이 비교 통계량이 된다. 먼저 기울기가 "1"에 가까우면 직접 추정량과 모형기반 추정량은 같은 크기의 편의를 갖는 것으로 판단할 수 있다. 이때 직접 추정량은 불편 추정량이므로 기울기가 "1"에 가까우면 모형기반 추정량도 불편 추정량이라 생각할 수 있다. 따라서 모형기반 추정량이 정확성을 유지하고 있다면 기울기 추정값 $\hat{\beta}_1 \approx 1$을 만족하게 될 것이다. 또한 R^2도 1에 가까운 값을 얻게 될 것이다. 그러나 정확히 $R^2 = 1$ 인 경우는 모형기반 추정량과 직접 추정량이 같다는 의미이므로 이 모형기반 추정량은 문제가 있다. 또한 R^2가 작은 모형기반 추정량의 경우에는 직접 추정량과 많은 차이를 보이는 경우이므로 이 또한 좋은 소지역 추정량이라 할 수 없다. 여기서 참고해야 할 것은 회귀분석을 이용한 소지역 추정량은 이론적으로 $\hat{\beta}_1 = 1$이 된다는 것이다. 즉, $\hat{Y}_{DE}$를 종속변수로 $\hat{Y}_{REG}$을 독립변수로 하는 절편이 없는 회귀분석에서 $\hat{Y}_{REG}$=X$(X'X)^{-1}X'\hat{Y}_{DE}$=$P_x\hat{Y}_{DE}$이므로

$$\hat{\beta}_1 = \frac{S_{\hat{Y}_{DE}\hat{Y}_{REG}}}{S_{\hat{Y}_{REG}\hat{Y}_{REG}}} = \frac{\hat{Y}'_{DE}P_x\hat{Y}_{DE}}{\hat{Y}'_{DE}P'_xP_x\hat{Y}_{DE}} = \frac{\hat{Y}'_{DE}P_x\hat{Y}_{DE}}{\hat{Y}'_{DE}P_xP_x\hat{Y}_{DE}} = 1$$

이다.

2.2. 커버리지(Coverage)

커버리지는 "직접 추정량은 큰 분산을 갖고 있으나 불편 추정량이다"라는 사실을 이용하여 비교하는 방법이다. 먼저 커버리지는 직접 추정량의 95% 신뢰구간을 구하고 이 구간에 각각의 모형기반 추정량에서 얻어진 추정치의 몇 퍼센트가 포함되는가를 살펴보는 것이다. 만약 모형기반 추정량의 편의가 없고 또한 분산이 작다면 이 추정량에서 얻어진 추정치는 불편 추정량의 신뢰구간 안에 95% 이상이 포함될 것이고 반대로 편의가 있으며 분산이 작다면 신뢰구간에서 겹치는 부분이 작거나 없을 경우도 발생하여 낮은 %의 커버리지를 보일 것이다. 또한 편의는 없으나 분산이 크면 직접 추정량의 신뢰구간과 겹치는 부분이 작게 되어 낮은 커버리지를 보일 것이다. 물론 대부분의 모형기반 추정량은 분산이 작기 때문에 이러한 결과가 발생할 가능성은 높지 않다.

2.3. 캘리브레이션(Calibration)

직접 추정량은 소지역을 합쳐 지역이 커지게 되면 그 지역에 포함된 표본의 수가 증가하게 되고 따라서 일반적으로 정도가 높아지게 된다. 이러한 특징을 이용하여 여러 소지역을 합쳐 가면서 각 모형기반 추정량에서 얻어진 추정치들을 비교하게 되는데 이렇게 비교하는 것을 캘리브레이션이라 한다. 만약 지역이 커져 자료의 수가 많아졌음에도 불구하고 직접 추정치와 모형기반 추정량에서 얻어진 추정치가 큰 차이가 난다면 그 모형기반 추정량은 주어진 자료를 잘 설명한다고 할 수 없을 것이다.

3. 소지역 추정량

본 논문에서 모형진단을 통해 비교되어지는 소지역 추정량들은 다음과 같다. 여기서

3.4절과 3.5절의 모형 표시법은 베이지안 통계에서 사용하는 표시방법을 사용하였다.

3.1. 직접 추정량(Direct estimator: 기호 $\hat{Y}_{DE}$)

소지역 추정량에서 가장 근간이 되는 추정량은 직접 추정량이다. 이 추정량은 불편성을 만족하나 분산이 큰 것으로 알려져 있으며, 본 논문에서는 모형기반 추정량 비교의 기준으로 사용되고 있다. 다음은 본 논문에서 사용되어진 직접추정량 식이다.

$$\hat{Y}_{DE} = \hat{Y}_i = \sum_j \omega_{ij} y_{ij}.$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값을 의미하고 ω_{ij}는 추출 가중치를, y_{ij}는 i 지역 j번째 자료값을 나타낸다.

3.2. 회귀분석 추정량(Regression estimator: 기호 $\hat{Y}_{REG}$)

회귀모형을 이용한 추정법은 소지역 추정에서 가장 많이 사용하고 있는 방법 중의 하나이다. 이는 일반적인 회귀모형에서 회귀계수를 추정하고 이를 이용하여 소지역 추정량을 회귀모형의 예측 통계량으로 이용하는 것으로 그 식은 다음과 같다.

$$\hat{Y}_{REG} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}.$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값이며 $x_{1i}, \ldots, x_{ki}$는 이에 해당되는 설명변수이고 $\hat{\beta}_i$는 추정된 회귀계수이다.

3.3. 공간 추정량(Spatial estimator: 기호 $\hat{Y}_{SP}$)

모형기반 소지역 추정량에서 보조변수의 역할은 매우 중요하다. 그러나 적절한 보조변수가 없어 관심 변수를 충분히 설명하지 못하는 경우가 있을 수 있으며 이때 관심 변수의 공간 상관관계를 활용할 수 있는 방법이 있다. 이러한 방법을 공간 통계를 이용한 소지역 추정이라 하며 이때 얻어진 추정량을 본 논문에서는 공간 추정량이라 부르겠다. 또한 본 논문에서 사용한 식은 다음과 같다.

$$\hat{Y}_{SP} = \hat{Y}_i = \hat{\rho} S_i.$$

여기서 $\hat{Y}_i$는 i번째 소지역 추정값이다. 공간 추정량은 이웃을 정하는 방법에 따라 다른 결과를 줄 수 있으며 이웃을 정하는 방법에 관한 연구는 Cressie(1993)을 참조하기 바란다. 위 식에서 사용된 S_i는 다음의 방법을 이용하여 구한다. 먼저 직접 추정치에서 전체 평균을 뺀다. 다음으로 각 지역의 이웃을 결정한다. 이웃으로 정해진 소지역들의 평균 제거 실업자 수를 더한다. 이렇게 더해진 수가 S_i이다. 자세한 내용은 김정오와 신기일(2006)을 참조하기 바란다.

3.4. 경험적 베이즈 추정량: EB(Empirical Bayes estimator)

모형기반 소지역 추정에서 베이지안적 접근 방법을 이용한 것으로 경험적 베이즈 추정량(EB)의 표본 오차 모형은 다음과 같이 정의된다.

$$\hat{\theta}_i = \theta_i + \epsilon_i.$$

여기서 $\epsilon_i|\theta_i \sim^{ind} N(0, \Psi_i)$이고 Ψ_i는 미지이다. 또한 θ의 사전함수로 연결 모형은 다음과 같다.

$$\theta_i = x_i^T\beta + v_i.$$

여기서 x_i, β는 각각 $p \times 1$ 벡터이고 $v_i \sim^{ind} (0, \sigma_v^2)$를 따른다. 이제 위의 두 식을 결합한 결합 모형은 다음과 같다.

$$\hat{\theta}_i = x_i^T\beta + v_i + \epsilon_i.$$

이때 경험적 베이즈 추정량, $\hat{Y}_{EB}$은

$$\hat{Y}_{EB} = \hat{\theta}_i^{EB} = (1-\gamma_i)\hat{\theta}_i + \gamma_i x_i'\hat{\beta}$$

으로 구해진다. 여기서 $\gamma_i = \Psi_i/(\Psi_i + \sigma_\nu)$이며 회귀계수 $\hat{\beta}$는 GLS에 의해 구해진다.

3.5. 계층적 베이즈 추정량: HB(Hierarchical Bayes estimator)

계층적 베이즈 추정량을 위한 모형은 다음과 같다.

$$\hat{Y}_i|\theta_i \sim \ N(\theta_i, \sigma_i^2).$$

여기서 $\hat{Y}_i$는 지역 i의 직접 추정량이며 이때 사전함수는 다음과 같다.

$$\begin{aligned} \theta_i &\sim N(x_i^T\beta, \tau^2), \\ \sigma_i^2 &\sim INGAM(a, b), \\ \beta_{p\times 1} &\sim MVN(\beta^*, \Omega). \end{aligned}$$

또한 a, b, β^*, Ω는 초모수(hyperparameter)로 알려진 값이다. 계층적 베이즈 추정량 $\hat{Y}_{HB}$는 θ_i의 사후분포의 기대값, $\hat{Y}_{HB(i)} = E(\theta_i|x)$,으로 구해지며 이는 MCMC 등과 같은 알고리즘을 이용하여 계산되어 진다.

3.6. 선형결합 추정량

선형결합 소지역 추정량은 일반적으로 불편 추정량인 직접 추정량과 분산을 줄일 수 있

는 모형기반 추정량의 선형결합으로 이루어진다. 다음은 직접 추정량과 모형기반 추정량을 가중 평균하여 구한 추정량이다.

1. $\hat{Y}_{DESP} = \alpha_{SP}\hat{Y}_{DE} + (1-\alpha_{SP})\hat{Y}_{SP}$,
2. $\hat{Y}_{DEREG} = \alpha_{REG}\hat{Y}_{DE} + (1-\alpha_{REG})\hat{Y}_{REG}$,
3. $\hat{Y}_{DEEB} = \alpha_{EB}\hat{Y}_{DE} + (1-\alpha_{EB})\hat{Y}_{EB}$,
4. $\hat{Y}_{DEHB} = \alpha_{HB}\hat{Y}_{DE} + (1-\alpha_{HB})\hat{Y}_{HB}$,
5. $\hat{Y}_{DEREGSP} = \alpha_{REGSP}\hat{Y}_{DE} + (1-\alpha_{REGSP})\hat{Y}_{REGSP}$,
6. $\hat{Y}_{DEHBSP} = \alpha_{HBSP}\hat{Y}_{DE} + (1-\alpha_{HBSP})\hat{Y}_{HBSP}$.

여기서 $\hat{Y}_{DE*}$는 $\hat{Y}_{DE}$와 $\hat{Y}_{*}$의 선형결합으로 $\hat{Y}_{DESP}$는 $\hat{Y}_{DE}$와 $\hat{Y}_{SP}$를 결합한 것이다. 또한 $\hat{Y}_{REGSP}$는 회귀분석 모형의 독립변수에 공간 통계항이 들어간 경우이며 $\hat{Y}_{HBSP}$ 또한 모형에 공간 통계항을 추가한 추정량이다. 특별히 주목해야할 내용은 $\hat{Y}_{EB}$와 $\hat{Y}_{HB}$등 베이지안 추정량은 $\hat{Y}_{DE}$의 선형결합으로 얻어진다는 것이다. 이에 관한 내용은 Press(1989)를 참조하기 바란다. 즉 이미 $\hat{Y}_{DE}$와 선형결합이 되어 있는 상태이다. 그러나 본 논문에서는 추가로 $\hat{Y}_{DE}$와 선형결합을 다시 함으로써 $\hat{Y}_{DE}$의 영향을 더 많이 받게 만들었다. 물론 그 영향력은 추정된 α_* 값에 좌우되나 $\hat{Y}_{EB}$ 또는 $\hat{Y}_{HB}$의 분산이 작아(표 4.1) α_* 값은 "0"에 가깝게 되어 일반적으로 $\hat{Y}_{DEEB}$와 $\hat{Y}_{EB}$, $\hat{Y}_{DEHB}$와 $\hat{Y}_{HB}$는 큰 차이를 보이지 않는다. 위의 6가지 추정량이 모형진단 방법에 의해 비교되었다.

4. 자료 분석 및 추정량 비교

분석에 사용된 자료는 2005년도 실업자 자료이며 보조변수로 사용된 자료는 비경제활동인구수이다. 이 자료를 선택한 이유는 선행연구인 신기일과 이상은(2003) 그리고 김재두 등(2005)에서 실업자 수를 소지역 추정하였을 때 보조변수로 비경제활동인구수를 사용하였고 또한 현실적으로 추가적인 보조변수를 얻는 것이 불가능하였기 때문이다. 먼저 3.6절의 선형결합 추정량을 구하기 위해서는 각 추정량의 가중값인 α_*들을 계산하여야 한다. 이를 위하여 붓스트랩 방법이 사용되었다. 먼저 9000여개의 2005년도 실업자 자료를 복원 추출하여 붓스트랩 샘플을 생성하였다. 생성한 자료에 3.1절–3.5절에서 정의한 소지역 추정량을 적용하여 추정값을 구하였다. 1,000번의 반복이 사용되었고 구해진 추정값을 이용하여 분산을 구한다. 가중치 α_*는 구해진 분산을 이용해 구하게 되며 이에 관한 내용은 4.1절에서, 각 추정량 비교는 4.2절–4.4절에서 설명하였다.

4.1. 가중치 α_*의 추정

일반적으로 α_*는 선형결합 추정량 $\hat{Y}_{DE^*}$의 MSE를 최소로 하는 값으로 정해지며 $\hat{Y}_{DE^*}$의 MSE를 최소로 하기 위하여 MSE를 구하면 다음과 같다.

$$\mathrm{MSE}(\hat{Y}_{DE^*}) = \alpha_*^2\mathrm{MSE}(\hat{Y}_{DE}) + (1-\alpha_*)^2\mathrm{MSE}(\hat{Y}_*) + 2\alpha_*(1-\alpha_*)E(\hat{Y}_{DE}-Y_i)(\hat{Y}_*-Y_i).$$

여기서 $\hat{Y}_{DE^*}$는 선형결합 추정량을, $\hat{Y}_{DE}$는 직접 추정량 그리고 $\hat{Y}_*$는 모형기반 추정량을 의미한다. 위의 식을 최소로 하는 α_*를 구하면 다음과 같다.

$$\alpha_* = \frac{\mathrm{MSE}(\hat{Y}_*) - E(\hat{Y}_{DE} - Y_i)(\hat{Y}_* - Y_i)}{\mathrm{MSE}(\hat{Y}_{DE}) + \mathrm{MSE}(\hat{Y}_*) - 2E(\hat{Y}_{DE} - Y_i)(\hat{Y}_* - Y_i)}$$

이제 $E(\hat{Y}_{DE} - Y_i)(\hat{Y}_* - Y_i)$가 무시될 정도로 작다고 가정하면, 최적의 가중치

$$\alpha_* = \frac{\mathrm{MSE}(\hat{Y}_*)}{\mathrm{MSE}(\hat{Y}_*) + \mathrm{MSE}(\hat{Y}_{DE})}$$

로 구해진다. 이에 관한 내용은 Falosi 등(1994)을 참조하기 바란다. 그러나 MSE를 추정하는 것이 어렵기 때문에 본 논문에서는 MSE의 추정값 대신 붓스트랩에서 구한 분산의 추정값을 사용하였다. 즉 본 논문에서는 다음식을 사용하였다.

$$\hat{\alpha}_* = \frac{\hat{\mathrm{Var}}(\hat{Y}_*)}{\hat{\mathrm{Var}}(\hat{Y}_*) + \hat{\mathrm{Var}}(\hat{Y}_{DE})}.$$

각 소지역별로 얻어진 $\hat{\alpha}_*$ 중 일부를 정리하면 다음과 같다.

표 4.1: 소지역별 가중치 α_*의 추정치

지역	α_{DESP}	α_{DEREG}	$\alpha_{DESPREG}$	α_{DEEB}	α_{DEHB}	α_{DEHBSP}
1	0.653	0.027	0.647	0.037	0.032	0.160
2	0.532	0.072	0.530	0.044	0.085	0.204
3	0.288	0.072	0.281	0.045	0.086	0.098
⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	0.458	0.260	0.424	0.178	0.069	0.415

표 4.1을 살펴보자. 먼저 지역 1에서 $\hat{\alpha}_{DESP}$는 0.653으로 그 값이 다른 추정량의 값에 비해 큰 것을 알 수 있다. 또한 $\hat{\alpha}_{DESPREG}$, $\hat{\alpha}_{DEHBSP}$ 등도 다른 값에 비해 상대적으로 큰 값을 보이고 있다. 그러나 그 값이 0.5 이상인 경우는 많지 않고 공간 추정량을 사용한 경우 몇 개 정도가 얻어졌다. 이에 반하여 $\hat{\alpha}_{DEEB}$와 $\hat{\alpha}_{DEHB}$인 경우 매우 작은 숫자를 보이고 있다. 이러한 현상은 40개 지역을 다 표로 만들지 않았지만 전 지역에 걸쳐 발견되고 있다. 이는 공간 관계가 분석에 사용되는 경우, 공간 관계가 분석에 사용되지 않은 것에 비해 상대적으로 큰 분산을 갖고 있음을 말해주고 있다.

4.2. 회귀분석 기법을 이용한 방법

기울기가 "1"에서 많이 떨어져 있거나 또는 R^2가 "1" 보다 많이 작다면 좋은 모형기반 추정량이라 할 수 없다. 전체적으로 분석된 모든 모형기반 추정량의 기울기가 "1"에 가깝

다고 할 수 있다. 그러나 R^2는 추정량에 따라 차이가 큰 것을 확인할 수 있다. 비교된 추정량 중에서 기울기와 R^2을 비교해 보면 $\hat{Y}_{DESPREG}$와 $\hat{Y}_{DESP}$가 가장 우수한 것으로 나타났다. 이에 관한 내용을 표 4.2에 정리하였다.

표 4.2: 추정량별 기울기와 R^2

추정량	기울기	R^2
$\hat{Y}_{SP}$	0.907	0.951
$\hat{Y}_{REG}$	1	0.787
$\hat{Y}_{SPREG}$	0.904	0.953
$\hat{Y}_{EB}$	1.089	0.794
$\hat{Y}_{HB}$	1.012	0.787
$\hat{Y}_{HBSP}$	1.017	0.881
$\hat{Y}_{DESP}$	1.007	0.990
$\hat{Y}_{DEREG}$	1.076	0.850
$\hat{Y}_{DESPREG}$	1.003	0.991
$\hat{Y}_{DEEB}$	1.154	0.866
$\hat{Y}_{DEHB}$	1.084	0.847
$\hat{Y}_{DEHBSP}$	1.084	0.936

4.3. 커버리지(Coverage)

커버리지는 직접 추정량 $\hat{Y}_{DE}$를 기준으로 구한다. 먼저 붓스트랩 방법을 이용, i번째 지역의 95% 신뢰구간을 구하였다. 구해진 신뢰구간을 기준으로 각 모형기반 추정량에서 얻어진 추정치가 직접 추정량에 의해 구해진 신뢰구간에 얼마나 포함되는지를 이용하여 커버리지를 구하였다. 커버리지에서 90% 이하로 겹치는 지역을 살펴보면 $\hat{Y}_{REG}$와 $\hat{Y}_{HB}$가 각각 10지역, $\hat{Y}_{HBSP}$는 11개 지역으로 가장 많아 이들 추정량의 커버리지가 나쁜 것으로 나타났다. 반면 90% 이하로 겹치는 지역의 수가 가장 적은 추정량은 $\hat{Y}_{DESPREG}$와 $\hat{Y}_{DESP}$로 각각 1개와 2개인 것으로 나타났다. 또한 가장 작은 커버리지를 살펴보면 이 두 추정량이 각각 89%와 88%로 다른 추정량에 비해 매우 높게 나타난 것을 확인 할 수 있다. 반면 $\hat{Y}_{REG}, \hat{Y}_{EB}, \hat{Y}_{HB}$ 그리고 $\hat{Y}_{DEEB}$의 가장 작은 커버리지는 "0"으로 나타난 것을 확인 할 수 있다. 결론적으로 커버리지를 기준으로 했을 경우 $\hat{Y}_{DESP}$와 $\hat{Y}_{DESPREG}$가 가장 우수한 결과를 주는 것을 확인할 수 있다. 이 결과는 표 4.2의 R^2를 이용한 결과와 일치하고 있음을 확인할 수 있다.

표 4.3: 소지역별 커버리지

지역	$\hat{Y}_{SP}$	$\hat{Y}_{REG}$	$\hat{Y}_{SPREG}$	$\hat{Y}_{EB}$	$\hat{Y}_{HB}$	$\hat{Y}_{HBSP}$
1	80	47	81	63	49	99
2	90	100	91	100	100	100
3	99	100	99	100	100	100
			⋮			
39	85	99	85	100	100	85
40	97	99	98	14	100	88
90%밑인 개수	6	10	6	5	10	11
최소값	53	0	55	0	0	20
지역	$\hat{Y}_{DESP}$	$\hat{Y}_{DEREG}$	$\hat{Y}_{DESPREG}$	$\hat{Y}_{DEEB}$	$\hat{Y}_{DEHB}$	$\hat{Y}_{DEHBSP}$
1	91	59	91	78	60	99
2	93	100	93	100	100	99
3	98	100	98	100	100	100
			⋮			
39	92	100	92	100	100	93
40	97	99	98	47	100	95
90%밑인 개수	2	7	1	7	8	7
최소값	88	12	89	0	24	78

4.4. 캘리브레이션(Calibration)

캘리브레이션은 지역이 커지면서 자료가 많아 질 때 직접 추정량은 불편이면서 분산이 작아진다는 특징을 이용하여 모형기반 추정량을 비교하는 방법이다. 먼저 지역을 크게 만들기 위해 40개의 소지역을 3-4개씩 묶어 11개의 조금더 큰 소지역으로 만들었다. 이렇게 만든 것을 그룹 1로 하였다. 다음으로 11개의 소지역을 다시 1개, 2개, 3개 그리고 5개씩 묶어 4개의 더 큰 소지역으로 만들었다. 이렇게 만든 소지역을 그룹 2라 하였다. 그룹을 나누는 방법을 표 4.4에 작성하였다. 또한 구해진 새로운 큰 소지역의 모집단 크기는 그룹 1은 표 4.5 또는 표 4.6에 그리고 그룹 2는 표 4.9 또는 표 4.10에 작성하였다.

표 4.4: 캘리브레이션을 위한 그룹

소지역 ID		그룹 1	그룹 2
From	End		
1	4	1-1	2-1
5	7	1-2	
8	10	1-3	
11	13	1-4	
14	16	1-5	
17	20	1-6	2-2
21	24	1-7	
25	28	1-8	
29	32	1-9	2-3
33	36	1-10	
37	40	1-11	2-4

각각의 소지역에서 얻어진 추정치를 표 4.4를 기준으로 합치면 그룹 1에서는 11개의 소지역 추정치가 그룹 2에서는 4개의 소지역 추정치가 얻어진다. 그룹 1의 결과는 표 4.5와 표 4.6에 그룹 2에 관한 결과는 표 4.9과 표 4.10에 나타냈다. 모형기반 추정량의 수가 많아 표 4.5와 표 4.9에는 공간 분석이 포함된 결과를 그리고 표 4.6과 표 4.10에는 공간 분석이 포함되지 않은 결과를 나타냈다.

표 4.5: 그룹 1에서 공간 분석이 있는 경우

Group	$\hat{Y}_{DE}$	$\hat{Y}_{SP}$	$\hat{Y}_{SPREG}$	$\hat{Y}_{HBSP}$	$\hat{Y}_{DESP}$	$\hat{Y}_{DESPREG}$	$\hat{Y}_{DEHBSP}$	Population
1-1	20962	23723	23513	22103	21736	21655	20156	812839
1-2	15763	15562	15688	19385	15627	15663	17728	797840
1-3	26046	21942	22180	20029	23210	23422	20542	903360
1-4	8933	8345	8548	10423	8365	8500	9898	684030
1-5	27825	24490	23719	17271	25713	25218	17848	646361
1-6	36137	37873	39357	31759	35545	36408	31907	1476117
1-7	16321	15374	15271	19400	15982	15901	18804	1035461
1-8	21630	27807	27274	25710	22602	22504	22042	974970
1-9	10765	11317	11340	16795	11099	11060	14833	706105
1-10	9753	8421	8419	20633	9071	9061	15441	730659
1-11	1318	1877	1669	5238	1365	1271	1987	260207

표 4.6: 그룹 1에서 공간 분석이 없는 경우

Group	$\hat{Y}_{DE}$	$\hat{Y}_{REG}$	$\hat{Y}_{EB}$	$\hat{Y}_{HB}$	$\hat{Y}_{DEREG}$	$\hat{Y}_{DEEB}$	$\hat{Y}_{DEHB}$	Population
1-1	20962	17285	15947	17358	13727	14146	13774	812839
1-2	15763	17423	14782	17305	15683	14621	15974	797840
1-3	26046	19462	17156	19241	19883	17647	19675	903360
1-4	8933	13796	11886	13775	12936	11703	12957	684030
1-5	27825	12812	17484	12867	13520	18055	13573	646361
1-6	36137	35441	26998	34788	34870	25576	34595	1476117
1-7	16321	21836	23345	21734	20684	22369	20826	1035461
1-8	21630	18795	21406	18761	18049	21006	17694	974970
1-9	10765	16053	17562	16159	15137	16577	15118	706105
1-10	9753	16065	17508	16160	14809	16607	14824	730659
1-11	1318	6487	11380	6910	1549	3319	1713	260207

표 4.5와 표 4.6에서 추정량의 비교를 위해 기준이 되는 것은 직접 추정치인 $\hat{Y}_{DE}$이다. 표를 쉽게 보기 위해 다음과 같은 비, $R_* = \hat{Y}_*/\hat{Y}_{DE}$을 정의하였다. 예를 들면 $R_{SP} = \hat{Y}_{SP}/\hat{Y}_{DE}$가 된다. R_*가 "1"에 가까우면 직접 추정량 $\hat{Y}_{DE}$와 모형기반 추정량 $\hat{Y}_*$가 비슷한 값을 갖는다는 것을 뜻하며 "1"에 비해 매우 큰 값을 취하거나 또는 매우 작은 값을 취한다면 이는 직접 추정치와 차이가 크다는 것을 말해 주고 있는 것이다. 이에 관한 결과는 표 4.7과 표 4.8에 정리하였다.

표 4.7: 그룹 1에서 공간 분석이 있는 경우의 비율 $R_* = \frac{\hat{Y}_*}{\hat{Y}_{DE}}$

Group	R_{SP}	R_{SPEG}	R_{HBSP}	R_{DESP}	$R_{DESPREG}$	R_{DEHBSP}
1-1	1.132	1.122	1.054	1.037	1.033	0.962
1-2	0.987	0.995	1.230	0.991	0.994	1.125
1-3	0.842	0.852	0.769	0.891	0.899	0.789
1-4	0.934	0.957	1.167	0.936	0.952	1.108
1-5	0.880	0.852	0.621	0.924	0.906	0.641
1-6	1.048	1.089	0.879	0.984	1.007	0.883
1-7	0.942	0.936	1.189	0.979	0.974	1.152
1-8	1.286	1.261	1.189	1.045	1.040	1.019
1-9	1.051	1.053	1.560	1.031	1.027	1.378
1-10	0.863	0.863	2.115	0.930	0.929	1.583
1-11	1.423	1.266	3.973	1.035	0.964	1.507

표 4.8: 그룹 1에서 공간 분석이 없는 경우의 비율 $R_* = \frac{\hat{Y}_*}{\hat{Y}_{DE}}$

Group	R_{REG}	R_{EB}	R_{HB}	R_{DEREG}	R_{DEEB}	R_{DEHB}
1-1	0.825	0.761	0.828	0.655	0.675	0.657
1-2	1.105	0.938	1.098	0.995	0.928	1.013
1-3	0.747	0.659	0.739	0.763	0.678	0.755
1-4	1.544	1.331	1.542	1.448	1.310	1.450
1-5	0.460	0.628	0.462	0.486	0.649	0.488
1-6	0.981	0.747	0.963	0.965	0.708	0.957
1-7	1.338	1.430	1.332	1.267	1.371	1.276
1-8	0.869	0.990	0.867	0.834	0.971	0.818
1-9	1.491	1.631	1.501	1.406	1.540	1.404
1-10	1.647	1.795	1.657	1.518	1.703	1.520
1-11	4.920	8.631	5.241	1.175	2.517	1.299

다음은 같은 방법을 이용하여 그룹 2에 관하여 정리하였다.

표 4.9: 그룹 2에서 공간 분석이 있는 경우

Group	$\hat{Y}_{DE}$	$\hat{Y}_{SP}$	$\hat{Y}_{SPREG}$	$\hat{Y}_{HBSP}$	$\hat{Y}_{DESP}$	$\hat{Y}_{DESPREG}$	$\hat{Y}_{DEHBSP}$	Population
2-1	99529	94062	93649	89211	94650	94457	86171	3844430
2-2	74089	81054	81902	76869	74129	74813	72753	3486548
2-3	20518	19737	19759	37427	20170	20121	30274	1436764
2-4	1318	1877	1669	5238	1365	1271	1987	260207

표 4.10: 그룹 2에서 공간 분석이 없는 경우

Group	$\hat{Y}_{DE}$	$\hat{Y}_{REG}$	$\hat{Y}_{EB}$	$\hat{Y}_{HB}$	$\hat{Y}_{DEREG}$	$\hat{Y}_{DEEB}$	$\hat{Y}_{DEHB}$	Population
2-1	99529	80777	77256	80545	75748	76171	75954	3844430
2-2	74089	76072	71749	75283	73603	68952	73115	3486548
2-3	20518	32118	35070	32319	29945	33184	29941	1436764
2-4	1318	6487	11380	6910	1549	3319	1713	260207

표 4.11: 그룹 2에서 공간 분석이 있는 경우의 비율 $R_* = \frac{\hat{Y}_*}{\hat{Y}_{DE}}$

Group	R_{SP}	R_{SPEG}	R_{HBSP}	R_{DESP}	$R_{DESPREG}$	R_{DEHBSP}
2-1	0.945	0.941	0.896	0.951	0.949	0.866
2-2	1.094	1.105	1.038	1.001	1.010	0.982
2-3	0.962	0.963	1.824	0.983	0.981	1.475
2-4	1.423	1.266	3.973	1.035	0.964	1.507

표 4.12: 그룹 2에서 공간 분석이 없는 경우의 비율 $R_* = \frac{\hat{Y}_*}{\hat{Y}_{DE}}$

Group	R_{REG}	R_{EB}	R_{HB}	R_{DEREG}	R_{DEEB}	R_{DEHB}
2-1	0.812	0.776	0.809	0.761	0.765	0.763
2-2	1.027	0.968	1.016	0.993	0.931	0.987
2-3	1.565	1.709	1.575	1.459	1.617	1.459
2-4	4.920	8.631	5.241	1.175	2.517	1.299

그룹 1에 해당하는 표 4.7과 표 4.8 그리고 그룹 2에 해당하는 표 4.11과 표 4.12를 그림으로 나타내면 다음과 같다. 여기서 x축은 모집단 수(Population)를 나타내며 y축은 직접 추정치와의 비율인 R_*를 나타낸다.

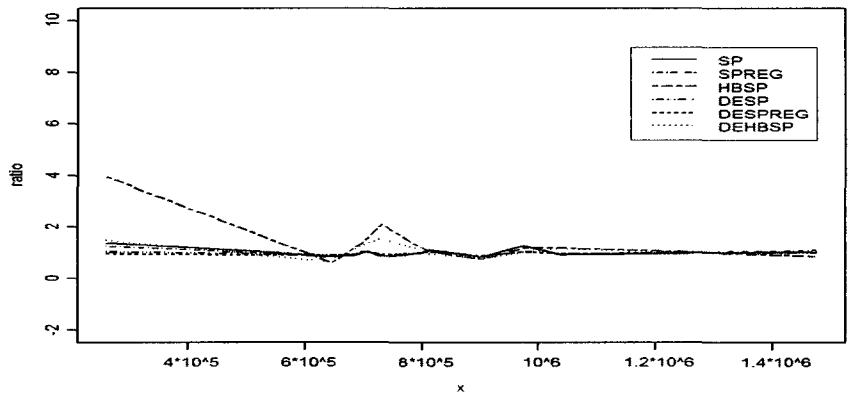

그림 4.1 그룹 1에서 공간 분석이 있는 경우의 비율

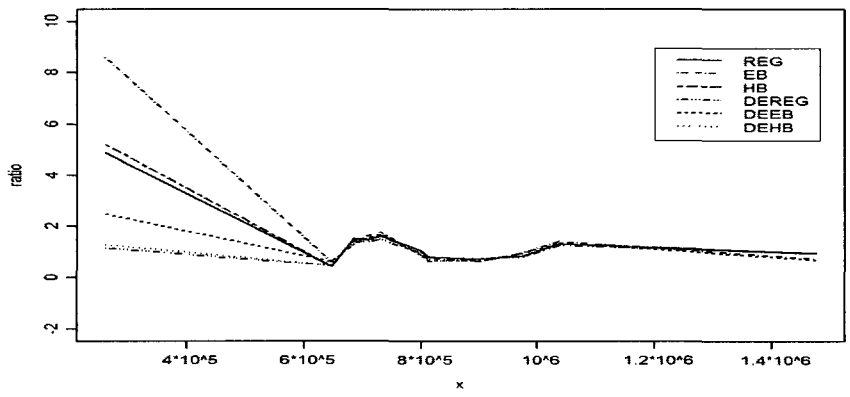

그림 4.2 그룹 1에서 공간 분석이 없는 경우의 비율

그림 4.1과 그림 4.2를 살펴보자. 먼저 그림 4.1은 공간 분석이 있는 추정량을 모아 그림으로 그린 것이다. 역시 모집단 수가 작은 경우, 이때 표본 수도 작기 때문에, 정확한 소지역 추정이 어렵다. 이를 반영하듯 직접 추정치는 다른 모형기반 추정치와 큰 차이를 보이고 있다. 그러나 모집단의 크기가 커질수록 그 차이는 줄어드는 것을 확인할 수 있다. 이러한 결과는 그림 4.2에서도 나타난다. 이제 그림 4.1과 그림 4.2를 비교하면 모집단의 수가 작은 경우에 많은 차이를 보이고 있다. 즉 그림 4.1의 공간 분석이 포함된 분석은 비율이 4를 넘지 않고 있으나 그림 4.2의 공간 분석이 포함 안 된 분석에서는 4를 넘는 수가 여러 개 보이고 있다. 특히 $\hat{Y}_{DESP}$와 $\hat{Y}_{DESPREG}$는 거의 "1"에 가까운 비율을 보이고 있다. 다음으로 그림 4.3과 그림 4.4에서는 그룹 2의 내용인 표 4.11과 표 4.12를 이용하여 그린 것이다.

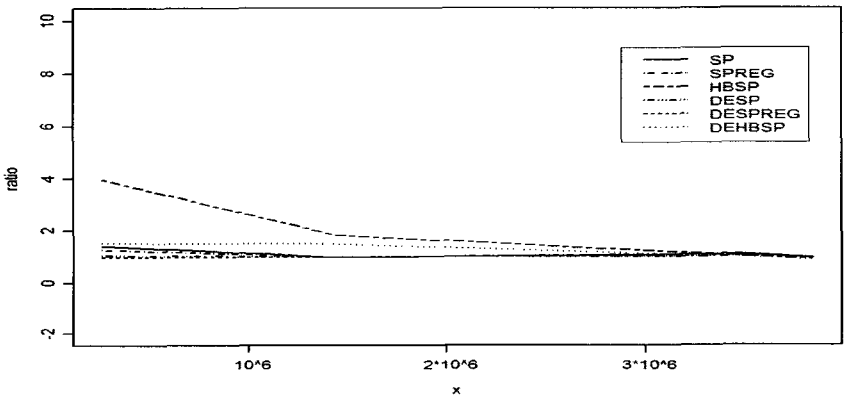

그림 4.3 그룹 2에서 공간 분석이 있는 경우의 비율

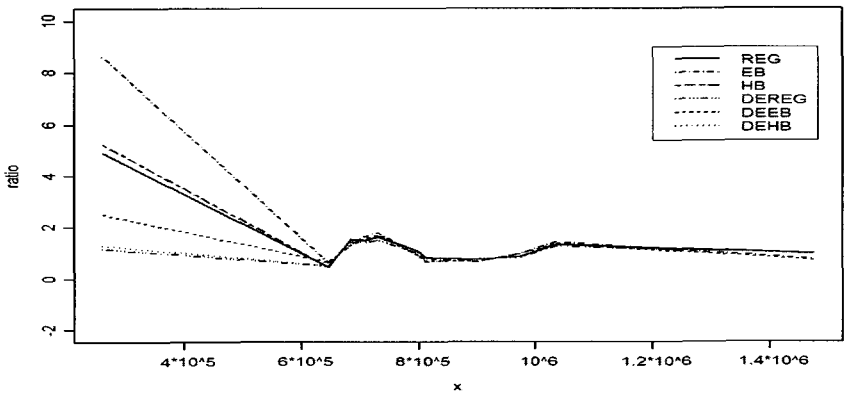

그림 4.4 그룹 2에서 공간 분석이 없는 경우의 비율

그림 4.3과 그림 4.4에서도 그림 4.1과 그림 4.2와 같은 결론을 내릴 수 있다. 특히 모집단의 수가 특정한 값 이상이 되면 많은 추정량들이 거의 같은 값으로 수렴하는 것을 볼 수 있다.

5. 결론

소지역 추정에 관한 많은 연구가 진행되었다. 보조변수가 충분한 경우 모형기반 추정량인 계층적 베이즈 추정량 HB가 우수한 결과를 주는 것으로 알려졌다. 그러나 본 논문에서 사용한 세 비교 방법을 기준으로할 때 $\hat{Y}_{DESP}$와 $\hat{Y}_{DESPREG}$이 가장 우수한 결과를 주는 것처럼 충분한 설명 변수가 존재하지 않은 경우에는 모형기반 추정량을 사용할 때 주의가 필요하다. 즉 모형기반 추정량인 경우 분산은 작지만 매우 큰 편의가 발생할 수 있기 때문이다. 따라서 충분한 설명 변수가 존재하지 않을 때에는, 설명 변수를 찾는 노력뿐만 아니라 공간 분석 등 주어진 설명변수 이외의 추가적인 정보를 찾는데에도 노력을 하여야 할 것이다. 공간 분석에서는 여러 이웃을 정하는 방법을 자료에 적용하여 가장 잘 맞는 이웃을 찾는 노력을 해야 할 것이며 추정에서도 베이지안 기법을 도입하는 등 첨단 추정법을 개발하여야 할 것이다. 또한 소지역 추정량을 비교할 때 전통적인 MSE 기준 이외에도 추가적인 기준을 이용한 비교를 고려해야 할 것이다.

참고문헌

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

김재두, 신기일, 이상은 (2005). 공간 시계열 모형을 이용한 소지역 추정, <응용통계연구>, **18**, 627–637.

김정오, 신기일 (2006). Comparison of Small Area Estimations by Sample Sizes, *The Korean Communications in Statistics*, **13**, 669–683.

신기일, 이상은 (2003). Model-Data Based Small Area Estimation, *The Korean Communications in Statistics*, **10**, 637–645.

이상은 (2006). 공간 통계량을 활용한 베이지안 자기 포아송 모형을 이용한 소지역 통계, <응용통계연구>, **19**, 421–430.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of Small Area Estimation Methods-Application to Unemployment estimations form 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.

Lee, S. E. and Shin, K.-I. (2006). A Review of Small Area Estimation as Official Statistics, *UNESCAP, APEX2.*

Press, S. J. (1989). *Bayeian Statistics: Principles, Models and Application*, John Wiley & Sons, New York.

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

[2006년 12월 접수, 2007년 3월 채택]

Evaluations of Small Area Estimations with/without Spatial Terms*

Key-Il Shin[1] Bong Ho Choi[2] Sang Eun Lee[3]

ABSTRACT

Among the small area estimation methods, it has been known that hierarchical Bayesian(HB) approach is the most reasonable and effective method. However any model based approaches need good explanatory variables and finding them is the key role in the model based approach. As the lacking of explanatory variables, adopting the spatial terms in the model was introduced. Here in this paper, we evaluate the model based methods with/without spatial terms using the diagnostic methods which were introduced by Brown *et al.* (2001). And Economic Active Population Survey(2005) is used for data analysis.

Keywords: Small area estimation methods, spatial terms, Empirical Bayesian(EB) estimation, Hierarchical Bayesian(HB) estimation.

* This research was supported by the research fund of Hankuk University of Foreign Studies, 2007.

1) Professor, Dept. of Statistics, Hankuk University of Foreign Studies, San89, Wangsan, Mohyun, Yongin, Kyonggi Do 449-791, Korea
E-mail: keyshin@hufs.ac.kr

2) Director, KNSO Regional Statistics & Sampling Division, Government Complex Daejon, 139 Seonsaro seo-gu, Daejon 302-701, Korea
E-mail: bonghoo.choi@nso.go.kr

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