The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 637-645

Model-Data Based Small Area Estimation

Key-Il Shin[1], Sang Eun Lee[2]

Abstract

Small area estimation had been studied using data-based methods such as Direct, Indirect, Synthetic methods. However recently, model-based such as based on regression or time series estimation methods are applied to the study. In this paper we investigate a model-data based small area estimation which takes into account the spatial relation among the areas. The Economic Active Population Survey in 2001 are used for analysis and the results from the model based and model-data based estimation are compared with using MSE(Mean squared error), MAE(Mean absolute error) and MB(Mean bias).

Keywords : Moran's I, Simultaneous Autoregressive model, Spatial Correlation, Freeman-Tukey transformation.

1. 서론

소지역 추정은 표본설계 시 고려되지 않은 지역 또는 영역에 관해 추정하고자할 때 이용되는 통계적 추정방법이다. 그러므로 일반적으로 표본의 크기가 작은 지역(small area)이나 나이, 연령 등과 같은 변수의 특정 분류된 소영역(small domain)에 대한 통계를 직접조사 된 자료와 모집단의 특성과 구조를 통한 예측값을 적용, 표본수가 적음을 보안하여 신뢰성을 높일 수 있는 추정기법을 소지역 추정법이라 한다. 많은 통계 선진국에서는 이미 수년 동안 소지역 추정에 관한 연구 결과를 축적하고 있으며 다양한 분야에 이를 활용하고 있다. 미국이나 캐나다 등에서는 센서스 중간연도에 지방자치정부의 요청에 따라 센서스 자료나 일반 행정자료를 이용한 소지역 추정을 활용하여 인구추정, 노동력 추정등을 하고 있다. 최근 국내에서도 소지역 추정에 관한 논문과 분석이 발표되고 있다. 예를 들어 이 계오(2002)는 소지역 추정을 이용한 실업 통계에 관한 분석 보고서를 발표하였다. 또한 통계기획국 조사관리과 (2001)에서는 소지역 통계 추정법에 관한 내용을 발표하였다. 이와 같이 최근 소지역 추정에 관한 많은 연구가 발표되고 있다. 그러나 Rao (2001)의 발표내용을 살펴보아도, 국내외 적으로, 소지역 추정에 공간통계 기법을 이용한 분석은 많지

1) Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin-Si, Kyonggi, 449-791, Korea
E-mail: keyshin@stat.hufs.ac.kr
2) Associate Professor, Department of Applied Statistics, Kyonggi University, Suwan-Si, Kyonggi, 442-720, Korea

않은 것으로 생각된다.

국내에서 게재된 최근 논문 중에서 공간 통계 분석 기법을 이용한 자료분석 논문은 유 성모와 엄 익현(1999)의 강우강도 데이터 분석, Baek 와 Bae(2001)의 K-function을 이용한 한반도의 지진에 관한 연구 등이 있고, 공간 통계학의 이론을 다룬 논문으로는 Lee(1998), Kim 등(1999) 그리고 이 윤동(2001) 등이 있다. 이와 같이 공간 통계학에 관한 이론 및 자료 분석에 많은 학자들이 관심을 보이고 있으며 향후 이에 관한 연구 또한 활발히 진행되리라 생각된다. 그러나 공간 통계를 이용한 소지역 추정은 아직 국내에서는 시도되지 않은 것으로 판단된다.

최근 박종태와 이상은(2001)은 1998년도 경기도 실업자수를 소지역 추정법으로 분석하였다. 소지역 추정에 사용되고 있는 여러 방법들, 직접추정량, 합성추정량, 복합추정량 그리고 베이즈 추정량의 효율성을 비교하여 베이지안 추정법의 우수함을 보였다. 이는 최근 소지역 추정의 여러 방법 중에서 모형-기반 추정법이 많이 사용되고 있는 것과 같은 맥락으로 이해할 수 있다. 즉 특정지역의 실업자 수를 소지역 추정법으로 추정할 경우 이와 관련이 있다고 생각되는 변수를 선택하고, 이들 변수를 이용, 설정된 모형을 기반으로 하여 소지역을 추정하는 방법이 좋은 결과를 주고 있다.

이제 설명 변수가 한정되어 있고 추정의 정도가 만족스럽지 못하다고 가정하자. 그리고 각 지역에서 얻어진 자료가 공간상관관계를 갖고 있다고 가정하자. 이러한 가정 하에서는 지역에 관한 정보인 공간상관관계를 분석에 이용하여야 할 것이다.

지역에 관한 정보를 분석에 사용하는 방법이 합성추정법이다. 즉 합성추정법은 "소규모 지역은 대규모 지역과 같은 특성을 갖고 있다 "는 가정 하에서 이루어진다. 즉 기본적인 개념은 자료의 공간적인 관계를 분석에 사용하는 것이다. 그러나 합성추정법은 공간 통계분석에서처럼 자료의 특성은 파악한 후 사용하는 것이 아니라 단순히 가정을 한 다음 이를 적용하는 것이다. 따라서 지역에 관한 가정이 맞지 않는 경우에는 좋은 분석 결과를 얻을 수 없게 된다. 따라서 분석의 효율을 높이기 위해서는 공간 통계 분석을 통해 소규모 지역과 대규모 지역에서 얻어진 자료의 공간상관관계를 분석한 후 그 결과를 소지역 추정 분석에 추가하여야 할 것이다.

본 논문에서는 박 종태와 이상은(2001)의 분석에서 사용되었던 것처럼 각 시도별 실업자수를 소지역 추정법을 사용하여 분석하였다. 먼저 2절에서는 단순 회귀분석 기법을 사용하여 독립변수의 효율을 살펴보았고 3절에서는 공간 상관관계의 존재 유무를 Moran's I, (Cressie: 1993)을 이용하여 검정하였으며, Lattice 분석에서 많이 사용되는 SAR(Simultaneous Autoregressive) 모형을 이용하여 모형의 적합성을 살펴보았다. 4절에서는 기존의 단순 회귀모형과 공간 상관관계를 이용한 중회귀 모형을 분석, 비교하였다. 5절에는 결론이 있다.

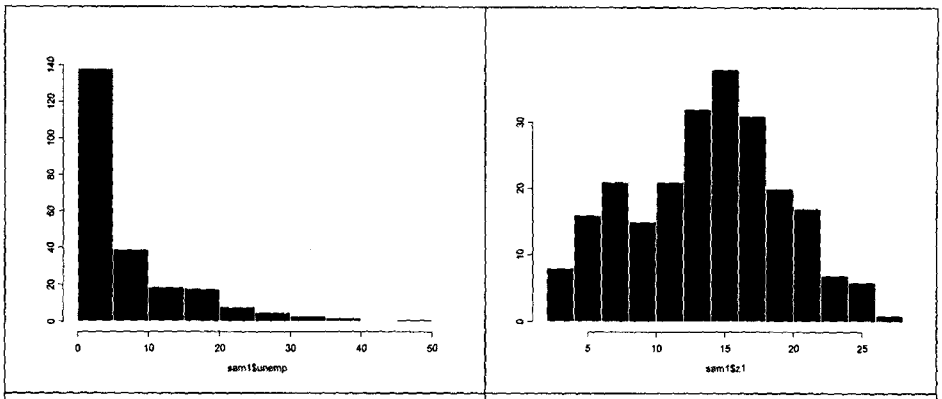
2. 공간 상관관계를 배제한 소지역 추정법

이 연구에서는 2001년 4월 경제활동인구조사에서 전국(제주도 제외)의 29899가구에서 얻어진 각 가구의 전체 실업자수, 취업자수, 경제활동인구수를 시군구별로 합친 자료를 이용하였다.

먼저 각 시군구별(행정구역분류 5자리기준) 자료를 공간 상관관계를 고려하지 않고 분석하였다. 다음 <그림 2.1>의 (a)는 시군구별 실업자수를 히스토그램으로 나타낸 것이다.

그림에서도 알 수 있듯이 우측으로 긴 꼬리가 나타나고 있다. 이는 조사된 시군구의 조사구 수를

감안하지 않은 결과로 판단된다. 본 논문에서는 조사구 수를 고려하기 위하여 Freeman-Tukey (1950) 변환을 실시하였다. 그 결과는 (b)에 나와 있다.



(a) 원 자료 히스토그램 (b) Freeman-Tukey 변환 후 그림

<그림 1> 원 자료와 Freeman-Tukey 변환 후의 히스토그램

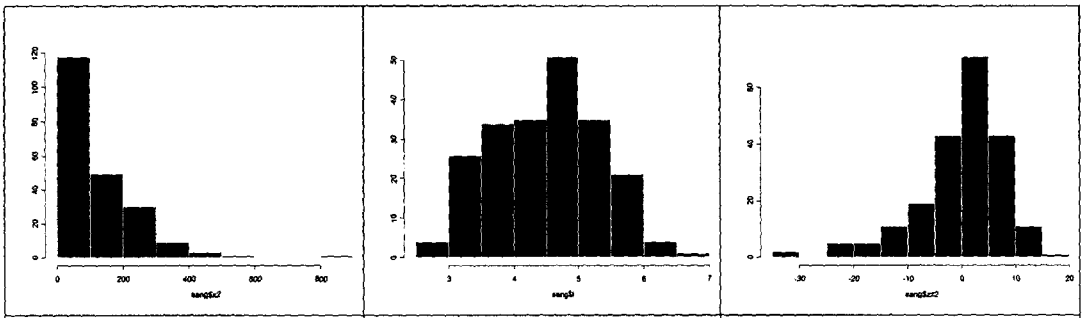
여기서 사용된 Freeman-Tukey 변환은 다음과 같다.

$$Z_i = (1000 \times X_i / n_i)^{1/2} + (1000 \times (X_i + 1)/n_i)^{1/2}$$

위 식에서 X_i는 실업자수이고 n_i는 조사구수 이다. 원 자료와 Freeman-Tukey 변환을 살펴보면 분석에 Freeman-Tukey 변환이 적당한 것으로 판단된다. 따라서 자료 분석은 Freeman-Tukey 변환된 자료를 이용하여 이루어졌다.

다음으로 실업자수에 관련이 있을 것으로 생각되는 독립변수인, 비활동인구를 살펴보자. 먼저 <그림 2>의 (a)가 비활동인구의 히스토그램이며 (b)가 log 변환된 자료의 그림이다.

또한 (c)가 Freeman-Tukey 변환을 이용한 그림이다.



(a) 원 자료의 히스토그램 (b) log 변환 후 히스토그램 (c) Freeman-Tukey 변환 후 히스토그램

<그림 2> 비경제활동인구의 원 자료와 변환 후 히스토그램

비경제활동인구는 실업자수와 관련이 있는 것으로 알려져 있다. 따라서 실업자수를 종속변수로 비경제활동인구를 독립변수로 한 단순회귀분석을 실시하도록 하자. 박종태 와 이상은(2001)이 연구한 것처럼 본 논문에서도 베이즈 추정법을 사용하여 분석할 수 있으나 본 논문의 목적이 각 지역간의 공간상관관계가 소지역 추정에 도움이 된다는 것을 밝히는 것이므로 간단한 회귀분석 방법을 선택하여 분석하였다. 여기에 실지 않았지만 단순회귀분석결과 독립변수로 (b) 또는 (c) 어떤 것을 사용하여도 같은 결과가 나왔다. <표 1>이 단순회귀분석 결과이며 지역별로 나누지 않고 전국을 하나의 지역으로 고려하였다. 여기서는 독립변수로 (c)를 이용한 분석결과를 보였다. 분석에 앞서 각 자료는 평균을 "0"으로 만들었으며 이는 다음절에 나오는 분석 결과와 쉽게 비교하기 위한 것으로 회귀분석결과에는 아무런 영향을 미치지 않는다.

변수	추정값	표준오차	p-값	R^2
비경제활동인구	0.28759	0.04355	<.0001	0.1719

<표 1> 단순회귀 분석결과

위의 결과를 살펴보면 비경제활동인구가 실업자 수와 관계가 있음을 알 수 있다. 그러나 전국 자료를 이용하였음에도 불구하고 설명력은 약하다.

3. 모형-자료기반 소지역 추정법

합성추정법은 소지역과 그 소지역을 포함하는 대지역이 같은 특성을 갖는다는 가정 하에서 대지역의 정보를 이용하여 분석한다. 합성추정법에서 사용한 가정을 공간 통계분석의 관점에서 살펴보면, 이는 특정 지역을 제외한 다른 전체 지역을 이웃으로 정하여 분석하는 것과 같다. 이러한 가정은 일반적으로 쉽게 검정될 수 있으며, 합성추정법의 사용 여부를 결정 할 수 있다. 따라서 이 절에서는 소지역의 근접지역의 정보를 이용하는 공간통계분석의 타당성을 살펴보기로 한다.
각 지역에서 얻어진 자료는 지역을 대표하는 Lattice 자료라고 생각하자. 그리고 각 지역과 같은 경계를 갖고 있는 지역을 이웃으로 정하자. 예를 들면 서울시 종로구의 이웃은 성북구, 은평구, 서대문구, 중구, 동대문구이다. 이상과 같이 전국 각 시군구의 이웃을 정하였다.
주어진 자료를 이용하여 공간 상관관계가 있는 지를 Moran's I를 이용하여 검정하였다. 물론 합성추정법에서 사용하는 가정을 검정할 수 있으나 본 논문에서는 생략하였다. 공간 상관관계 검정은 S-PLUS를 이용하였으며 그 결과는 <표 2>에 나와 있다.

통계량	추정량	표준오차	p-값
Moran's I	0.3773	0.0483	0

<표 2>공간 상관관계 검정

위의 결과에서 알 수 있듯이 각 시군구 자료는 전국을 하나의 지역으로 생각하였을 때 양의 공간상관관계를 갖고 있음을 알 수 있다. 따라서 각 자료의 공간상관관계를 분석에 이용하는 것이 타당함을 알 수 있다.

이제 공간통계모형을 소지역 추정법에 적용하기로 하자. 2001년 4월 경제활동인구조사에서 얻은 전국 자료를 하나의 소지역이라 생각하고 전체 실업자수를 공간통계모형으로 추정하기로 하자.

공간 통계모형은 SAR(Simultaneous Autoregressive)모형을 사용하였다. 이웃이 미치는 영향력은 모두 같다고 가정하면 SAR 모형식은 다음과 같다.

$$Z_i = \beta X_i + \rho S_i + \epsilon_i$$

여기서 i는 행정구역 5자리에 해당하는 각 지역(시군구)을 나타낸다. X_i는 지역 i의 비경제활동인구이고, S_i는 지역 i의 이웃 자료값 합이며, ϵ_i는 각 지역의 오차항으로 독립이고 평균이 "0"이다. 또한 2절에서와 같이 자료의 평균을 "0" 으로 만들었다. SAR모형의 모수의 추정은 S-PLUS를 이용하였으며 분석 결과는 <표 3>과 같다.

설명 변수	추정값	표준오차	p-값	공간상관계수	추정값
X_i	0.2156	0.0438	0.000	ρ	0.094

<표 3> SAR 모형을 이용한 모수 추정

SAR 모형을 이용한 모수 추정 결과, 상관계수 추정값은 유의한 것으로 판단되며 각 이웃간에는 양의 공간상관관계가 존재함을 알 수 있다. 이는 이웃간의 이웃정보가 하나의 새로운 독립변수로 사용될 수 있음을 나타내고 있다. 다음절에서는 이웃정보를 하나의 새로운 독립변수로 하는 중회귀분석을 이용하여 두 분석의 효율을 비교하였다. 전국을 하나의 소지역으로 본 경우의 비교뿐 아니라 15개 시도를 하나의 소지역으로 생각한 분석도 비교하였다.

4. 두 방법의 비교

단순회귀방법과 3절에서 얻어진 SAR 모형에 기초한 중회귀모형을 비교하자. SAR 모형을 살펴보면 지역간에 양의 상광관계가 존재하므로 각 지역의 근접지역 자료 값들의 합을 다른 독립변수로 사용할 수 있다. 2절의 비경제활동인구만을 독립변수로하는 단순회귀 모형과 이에 추가로 각 이웃지역 자료값의 합을 독립변수로 하는 중회귀모형을 적합시키기로 하자. <표 4>는 전국 자료를 이용하여 중회귀 모형을 적합시킨 결과이다.

변수	추정값	표준오차	p-값	R^2
X_i	0.21630	0.04216	<.0001	0.2909
S_i	0.14829	0.02505	<.0001	

<표 4> 중회귀모형을 이용한 모수 추정

중회귀 모형을 살펴보면 이웃 자료의 합인 S_i가 실업자 수와 관계가 있음을 알 수 있다. 또한 R^2 값도 0.1719에서 0.2909로 많이 증가하였음을 알 수 있다. 본문에는 실지 않았지만 Root MSE 값도 5.02에서 4.66으로 작아졌다.

다음으로 위의 분석 결과는 변환된 자료에서 얻어진 것이므로, 재 변환을 이용하여 원 실업자수를 예측하고 이를 비교하도록 하자. 먼저 Freeman-Tukey 변환

$$Z_i = (1000 \times X_i/n_i)^{1/2} + (1000 \times (X_i+1)/n_i)^{1/2}$$

에서

$$X_i = \frac{n_i}{1000} \times \left(\frac{Z_i^2 - 1000/n_i}{2Z_i}\right)^2$$

를 얻게 되고 이식을 이용하여 재변환된 예측값을 얻는다. 그리고 예측값과 원 자료값을 비교하면 다음과 같은 결과 <표 5>을 얻는다. 표에서 사용된 공식은 다음과 같다. 여기서 i는 지역을 n은 지역의 개수를 나타낸다.

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

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

<표 3>과 <표4>에서 얻어진 모수를 이용하여 두 방법을 소지역 즉, 15개 시도와 전국에 적용한 결과는 다음과 같다.

지역	MSE		MAE		MB	
	모형기반	모형자료기반	모형기반	모형자료기반	모형기반	모형자료기반
서울(25)	49.5283	40.0528	5.5343	4.6477	3.9858	1.4158
부산(16)	56.4809	42.3481	5.0437	4.8041	4.2003	1.7090
대구(8)	63.7937	45.5906	6.7934	6.0030	2.5709	0.005
인천(8)	78.2403	29.5094	7.39972	3.96025	7.39972	2.99824
광주(5) *	18.0387	23.0963	3.2096	4.3051	2.0398	0.0108
대전(5) *	12.2444	13.6080	2.9304	3.2053	0.34761	-0.2310
울산(5) *	23.2905	30.9908	3.9283	4.5624	-0.1989	0.4173
경기(26)	5.0506	4.7594	1.7873	1.6645	0.5876	0.2055
강원(17)	13.680	4.9223	2.4803	1.5302	-1.9872	-0.5659
충북(10) *	10.2812	11.9684	2.4230	2.2677	0.36473	0.6412
충남(15) *	10.7843	11.2633	2.6577	2.4639	-0.3752	0.5322
전북(13)	2.3865	1.0313	1.0824	0.7255	-0.8323	-0.1614
전남(17) *	17.5835	27.3197	1.8090	2.0263	0.7830	1.3071
경북(21) *	15.5822	17.0924	2.4206	2.2512	0.4044	0.9512
경남(19)	4.3966	3.1293	1.6583	1.3010	0.0219	-0.0611
전국(210)	23.3641	19.1388	3.16076	2.75880	1.15576	0.63891

<표 5> 단순 소지역 추정법과 공간 통계를 이용한 소지역 추정 비교

전술한 데로 <표 5>는 전국자료를 이용하여 얻은 상관관계를 각 지역에 적용하여 얻은 결과이다. 결과를 살펴보면 광주, 대전, 울산, 충북, 충남, 전남 그리고 경북 등 7개 지역은 단순회귀분석을 이용한 분석이 우수한 것으로 나왔으며 나머지 8개 지역은 공간상관관계를 이용한 분석이 우수한 것으로 나타났다. 이는 전국자료를 이용하여 얻은 상관관계를 각 지역에 적용하는 것이 무리임을 말해주고 있다. 물론 전국을 하나의 소지역으로 생각할 경우에는 공간상관관계를 이용하는 것이 더 효율적임을 알 수 있다.

다음 <표 6>은 각 지역을 나눈 후 각 지역에서 얻어진 모수를 이용한 분석 결과이다.

지역	MSE		MAE		MB	
	모형기반	모형자료기반	모형기반	모형자료기반	모형기반	모형자료기반
서울(25)	18.9690	18.6471	3.3683	3.3469	0.2347	0.0738
부산(16)	38.1134	38.1814	4.15588	3.86998	1.36716	1.20063
대구(8)	28.5606	27.4634	4.19117	4.20725	1.10062	1.10112
인천(8)	29.6872	28.2240	3.66293	3.54674	-1.16902	-0.48404
광주(5)	26.5782	26.1622	4.42725	4.41717	-1.29328	-1.26663
대전(5)	7.7117	5.3431	2.4040	2.0055	-0.1198	-0.02051
울산(5)	28.7620	1.18757	4.18667	0.82687	1.60622	0.052749
경기(26) *	4.14835	4.45011	1.57968	1.68685	0.020388	0.032225
강원(17)	5.43618	5.20968	1.43056	1.38863	0.76746	0.72968
충북(10) *	13.5979	14.0623	2.38405	2.44125	0.41209	0.37945
충남(15)	13.5958	12.5730	2.51197	2.43167	0.50149	0.46209
전북(13)	2.57347	0.97658	1.03788	0.61489	0.29645	0.18052
전남(17)	26.1645	18.9398	2.46699	1.94627	0.31608	0.78002
경북(21)	18.1106	17.7794	2.37588	2.34721	1.10159	0.75843
경남(19)	4.20871	3.01282	1.56075	1.22021	0.32993	0.084107
전국(210)	23.3641	19.1388	3.16076	2.75880	1.15576	0.63891

<표 6> 각 지역별 상관관계를 이용한 소지역 추정 비교

지역별 공간상관관계를 이용한 분석결과를 살펴보면 경기와 충북 등 2개 지역만이 공간상관관계를 이용한 분석 결과가 나쁜 것으로 나타났다. 이는 <표 5>의 7개 지역에 비하여 많이 향상된 결과이다. 결국 지역을 이용한 공간상관관계 분석이 소지역 추정에 효과적임을 알 수 있다. 그러나 각 지역별 MSE, MAE 그리고 MB를 비교하면 차이가 거의 없는 지역을 볼 수 있다. 이러한 경우는 각 지역간의 공간상관관계가 크지 않거나 없기 때문에 나타나는 결과로 해석될 수 있다. 특히 경기와 충북 지역은 오히려 통계적으로 공간상관관계가 나타나지 않았음에도 불구하고 이를 무리하게 적용한 결과로 오히려 나쁜 분석 결과를 주는 지역이다. 반면 특별히 공간상관관계가 높

은 지역인 울산과 전북의 경우에는 공간상관관계를 이용하였을 때 결과에 많은 향상을 가져왔다. 결론적으로 공간상관관계를 이용한 분석은 먼저 공간상관관계가 있는지를 검정한 후 이루어져야 할 것이다.

5. 결론

현재 사용되고 있는 대부분의 소지역 추정법은 아직 공간통계분석 기법이 사용되고 있지 않다. 본 논문에서 밝힌 바데로 각 지역에서 얻어진 자료들 간에는 공간 상관관계가 존재할 수 있다. 따라서 소지역에서 얻어진 자료를 이용하여 모형-기반 소표본 추정을 할 때 충분한 설명변수가 존재하지 않을 경우에는 설명변수에 의한 설명력이 떨어지므로 이를 극복하기 위한 방법으로 공간상관관계를 이용하는 것은 좋은 방법이 될 것이다. 위의 자료 분석에서도 언급하였듯이 공간상관관계가 존재 하지 않을 경우 이를 이용할 경우 오히려 나쁜 영향을 미칠 수 있기 때문에 충분한 검토 후 적용하는 것이 바른 분석 방법일 것이다. 본 논문에서 연구되지 않았지만 실제 자료 분석에서 베이지안 방법 등과 같은 높은 수준의 추정방법을 사용한다면 더 좋은 결과를 얻을 수 있을 것으로 예상된다.

참고 문헌

1. Baek, J., and Bae, J. S. (2001). K-function Test for the Spatial Randomness among the Earthquakes in the Korean Peninsular, The Korean Communication in Statistics, Vol. 8, No. 2, pp 499-505.
2 Freeman, M. F, and Tukey, J. W. (1950). Transformations related to the Angular and the square root. Annals of Mathematical Statistics, Vol. 21, pp. 607-611.
3. Kim, Y-W., Choi, E-H. and Hwang S. Y. (1999). Partially Observed Data in Spatial Autologistic Models with Applications to Area Prediction in the Plane, Journal of the Korean Statistical Society, Vol. 28, No. 4, pp 457-468.
4. Lee J. (1998). A Doubly Winsorized Poisson Auto-Model, The Korean Communications in Statistics, Vol. 5, No. 2, pp. 559-570.
5. Rao, J. N. K. (2001). Introduction to Small Area Estimation, 2001 ISI Manuscript.
6. 박종태, 이상은 (2001). 소지역 추정법에 관한 비교연구, Journal of the Koeran Data & Information Science Society, Vol. 12, No. 2, pp 47-55.
7. 유성모, 엄익현 (1999). 강우강도 데이터를 이용한 세미베리오그램이 추정과 공간이상치에 관한 연구. 응용통계연구, 12권 1호, pp. 25-141.
8. 이윤동 (2001). 확률난수를 이용한 공간자료의 생성과 베이지안 분석, 응용통계연구, 제 14권, 2호. pp. 379-391.
9. 이계오 (2002). 소지역 추정법에 관한 시.군.구 실업 통계 개발, 최종보고서
10. 통계기획국 조사관리과 (2001). 소지역 통계 추정법

[2003년 7월 접수, 2003년 11월 채택]