

## 공간 시계열 모형을 이용한 소지역 추정\*

김재두<sup>1)</sup> 신기일<sup>2)</sup> 이상은<sup>3)</sup>

### 요약

시계열 모형과 공간통계 모형이 결합된 공간 시계열 모형에 관한 많은 연구가 발표되고 있다. 최근 de Luna와 Genton(2003, 2005)은 공간 시계열 모형을 이용한 실업률의 소지역 추정을 발표하는 등 공간 시계열 기법을 이용한 소지역 추정방법의 연구도 빠르게 진행되고 있다. 본 논문에서는 공간 시계열 모형을 이용한 소지역 추정기법을 설명하고 이 기법과 신기일과 이상은(2003)이 발표한 공간모형을 이용한 소지역 추정을 비교하였다.

주요용어: Moran's I, 이웃정보, Simultaneous autoregressive(SAR) model.

### 1. 서론

소지역 추정은 표본설계 시 고려되지 않은 지역 또는 영역에 관해 추정하고자할 때 이용되는 통계적 추정방법이다. 선진국에서는 이미 수년동안 소지역 추정에 관한 연구 결과를 축적하고 있으며 다양한 분야에 이를 사용하고 있다. 국내에서도 소지역 추정에 관한 다수의 논문이 발표되고 있다. 소지역 추정에 관한 자세한 내용은 Rao(2001)을 참조하기 바란다. 최근 신기일과 이상은(2003)은 2001년도 전국 실업자수를 이용하여 소지역 추정의 여러 방법 중 모형-기반 추정법으로 많이 사용되고 있는 회귀모형과 Lattice 자료의 분석에 이용되는 공간통계 모형의 하나인 SAR(Simultaneous Autoregressive) 모형을 이용한 소지역 추정법을 비교하였다. 비교 결과 그들은 모형-기반 소지역 추정을 할 때, 충분한 설명변수가 존재하지 않고 공간상관관계가 존재할 경우, 공간상관관계를 고려한 SAR 모형을 사용하게 되면 일반적으로 사용하는 회귀모형 보다 효율적인 소지역 추정을 할 수 있음을 보였다.

그러나 실업자수를 비롯한 많은 자료는 매월 또는 매년 얻어지는 경우가 대부분이기 때문에 이러한 시계열적 요인을 분석에 포함하는 것이 타당할 것이다. 따라서 이러한 경우에는 시계열적인 관계와 공간적 관계를 이용한 공간시계열(Space-Time Autoregressive: STAR) 모형이 일반적인 분석 모형이 된다. 이에 본 논문에서는 공간상관관계를 고려한 시계열 모형인 STAR 모형을 소지역 추정에 적용하였다.

\* 이 논문은 2005년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임

1) (442-702) 경기도 수원시 영통구 의의동 산 94-3, 경기대학교 응용통계학과, 석사과정

E-mail: luckjd@dreamwiz.com

2) (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 교수

E-mail : keyshin@hufs.ac.kr

3) (442-702) 경기도 수원시 영통구 의의동 산 94-3, 경기대학교 응용통계학과, 부교수

E-mail : sanglee62@kyonggi.ac.kr

이제 본 논문에서 사용할 모형인 SAR 모형과 STAR 모형을 살펴보자. 먼저  $Z \sim N_n(\mu, \Sigma)$  라 하고  $e$ 를  $E(e) = 0, \text{Var}(e) = \sigma^2 I$  인  $n \times 1$  랜덤 벡터라 하면, Lattice 자료 분석을 위한 SARMA(Spatial Autoregressive Moving Average) 모형은 다음과 같이 표현된다.

$$B(Z - \mu) = Ce \quad (B, C \text{는 행렬})$$

분석에서는 행렬  $B$ 와  $C$ 가 자료를 잘 설명할 수 있도록 선택되어야 하며 행렬  $B$ 와  $C$ 의 형태가 모형을 결정하게 된다. 공간 AR 모형은 시계열 분석에서처럼 각 자료간에 존재하는 공간적 상관관계를 모형으로 나타낸 것으로, 모형은 자료들 간의 선형결합을 모형화 할 수 있도록 자료 벡터 앞에 행렬을 곱해 주는 형식을 취하게 된다. 즉, 위의 모형에서  $C = I, B = (I - N)$ 을 선택함으로써 SAR 모형을 만들 수 있게 된다.

$$(I - N)(Z - \mu) = e$$

이 모형에서 행렬  $N$ 은 이웃 정보를 행렬로 표시한 이웃행렬을 나타내며, 여기서  $\mu = 0$ 을 가정하면, 평균이 '0'인 모형은 다음과 같이 된다.

$$(I - N)Z = e \text{ 또는 } Z = NZ + e$$

이제 행렬  $N$ 의 원소를  $N = \{N_{ij}\} = \{\rho_{ij}\}$  라 하면, 이웃행렬의 정의에 의해  $N_{ii} = \rho_{ii} = 0$  이 되고 따라서 위 모형은

$$Z_i = \sum_{j, j \neq i} \rho_{ij} Z_j + e_i$$

가 된다. 이 모형을 살펴보면  $i$ 번째 조사지역에서 얻어진 자료  $Z_i$ 는 이 지역을 제외한 이웃의 모든 조사지역에서 얻어진 자료의 선형결합으로 표시된다. 많은 분석에서 이웃이 미치는 영향력이 같다는  $\rho_{ij} = \rho$  가정을 사용하게 되며 이를 이용하면 다음의 모형이 된다.

$$Z_i = \rho \sum_{j, j \neq i} Z_j + e_i \quad (1.1)$$

이제 위의 모형은 시계열 요인이 포함된 자료를 분석할 수 있는 모형으로 확장될 수 있다. 이러한 모형을 STAR(Space Time Autoregressive) 모형이라 부른다. STAR 모형은 시계열 요인과 공간 요인이 동시에 있는 모형을 분석하기 위한 모형으로, 이웃이 미치는 영향력을 모두 같다고 가정하면 STAR 모형식은 다음과 같다.

$$Z_{i,t} = \rho \sum_{j, j \neq i} Z_{j,t} + \phi_i(B)Z_{i,t} + e_{i,t} \quad (1.2)$$

여기서  $i, t$ 는 시점  $t$ 의  $i$ 번째 조사지역을 나타내며  $\phi_i(B) = (\phi_{i,1}B + \phi_{i,2}B^2 + \dots + \phi_{i,p}B^p)$ 이다. 이때  $p$ 는 AR(p) 모형의 차수  $p$ 를 나타내며  $p = 1$ 인 경우, 즉 AR(1) 모형인 경우  $\phi_i(B)Z_{i,t} = \phi_{i,1}BZ_{i,t} = \phi_{i,1}Z_{i,t-1}$ 이 된다. STAR 모형에 관한 내용은 Cressie(1993)을 참조하기 바란다. 본 논문에서는 STAR 모형과 신기일과 이상은(2003)의 SAR 모형을 비교하였다.

## 2. 자료분석

### 2.1. 자료소개

이 연구에서는 1999년 5월부터 2001년 4월까지의 경제활동인구조사에서 전국(제주도 제외)의 시군구별 조사구수와 가구에서 얻어진 실업자수, 비경제활동인구수의 시군구별 자료를 이용하였다. 신기일과 이상은(2003)이 사용한 29,899가구 자료 중에서 시계열 유지를 위해 24개월의 조사기간 동안 매월 조사가 이루어진 16,095가구를 분석대상으로 하였다. SAR 모형과 STAR 모형의 비교에는 그중 2001년 4월 자료를 이용하였다. 또한 특정 시군구의 조사구수가 많으면 실업자수도 많게 되므로 이를 보정하기 위해 신기일과 이상은(2003)에서 사용하였던 Freeman-Tukey(1950) 변환을 사용하였다. 여기서 사용된 Freeman-Tukey 변환은 다음과 같다.

$$Z_i = (1000 \times Y_i/n_i)^{1/2} + (1000 \times (Y_i + 1)/n_i)^{1/2} \quad (2.1)$$

위 식에서  $Y_i$ 는  $i$ 번째 시군구의 실업자수이고  $n_i$ 는  $i$ 번째 시군구의 조사구수이다.

### 2.2. SAR(Simultaneous Autoregressive)모형을 이용한 분석

SAR 모형을 이용한 분석에서는 이웃 정보가 필요하며 본 논문에서 사용된 이웃은 각 지역(시군구)과 같은 경계를 갖고 있는 지역을 이웃으로 정하였다. 예를 들면 서울시 종로구의 이웃은 성북구, 은평구, 서대문구, 중구, 동대문구이다. 이와 같은 방법으로 전국 각 시군구의 이웃을 정하였다. 다음의 표 2.1은 서울 지역의 이웃을 나타낸 것으로 이중 일부지역만 표에 나타내었다.

표 2.1: 경계를 기준으로 한 이웃 정보

행정구역번호	시·군구	지역번호	이웃정보
11010	종로구	1	2 6 8 12 13
11020	중구	2	1 3 4 13 14
11030	용산구	3	2 4 14 19 20 22 23
11040	성동구	4	2 3 5 6 23
11050	광진구	5	4 6 7 23 24 25
11060	동대문구	6	1 4 5 7 8
11070	중랑구	7	5 6 8 11
11080	성북구	8	1 6 7 9 11
11090	강북구	9	8 10 11
11100	도봉구	10	9 11
11110	노원구	11	7 8 9 10
11120	은평구	12	1 13 14
11130	서대문구	13	1 2 12 14
11140	마포구	14	2 3 12 13 16 19

주어진 자료와 이웃 정보를 이용하여 공간 상관관계가 있는 지를 Moran's I를 이용하여 검정하였다. Moran's I는 ' $H_0$ :공간상관관계가 없다 대  $H_1$ : 공간상관관계가 있다'를 검정하는데 일반적으로 사용되는 통계량이다. S-PLUS를 이용하여 얻은 결과는 다음과 같다.

표 2.2: 공간 상관관계 검정 결과

통계량	추정량	표준오차	p-value
Moran's I	0.2096	0.04939	0.00001414

위의 결과에서 알 수 있듯이 각 시군구 자료는 양의 공간상관관계를 갖고 있다. 따라서 자료의 공간상관관계를 분석에 이용하는 것이 타당할 것이다.

먼저 전국을 시 및 도의 구별없이 하나의 소지역이라 생각하고 공간 상관관계를 고려한 다음의 모형을 고려하자.

$$Z_i = \beta X_i + \rho S_i + \epsilon_i \quad (2.2)$$

전국을 하나의 지역이라 생각하였으므로 공간상관관계를 나타내는 모수는 시도와 상관없이 하나만 사용된다. 여기서  $i$ 는  $i$ 번째 시군구를 나타낸다.  $X_i$ 는 지역  $i$ 의 비경제활동인구이고,  $i$ 지역과 이웃하고 있는 지역의 자료를  $Z_{ik}$ 라 하면  $S_i = \sum_k Z_{ik}$ , 즉  $S_i$ 는 지역  $i$ 와 이웃하고 있는 시군구 자료값의 합이다. 그리고  $\epsilon_i$ 는 각 지역의 오차항으로 독립이고 평균이 '0'이다. 모형 (2.2)는 모형 (1.1)의 이웃에 해당된 변수를 간단히 표현한 것이며 실업자수에 관련이 있는 것으로 판단되는 비경제활동인구수를 독립변수 항에 포함시킨 모형이다. 이 모형은 신기일과 이상은(2003)에서도 사용된 모형이다. 이제 (2.2)의 모수추정 결과를 표 2.3에 정리하였다.

표 2.3: SAR 모형에 의한 모수 추정 결과

변수	추정값	표준오차	p-value	$R^2$
$X_i$	0.22265	0.04763	<.0001	0.1856
$S_i$	0.07325	0.02776	0.0089	

계수 추정값은 (2.2)식을 중회귀 모형이라 간주한 후 최소제곱추정법을 이용하여 얻어졌다. 일반적으로는 독립변수가 있는 SAR 모형으로 모수를 추정해야 정확한 모수 추정값을 얻을 수 있다. S-Plus에서는 독립변수가 있는 SAR 모형의 모수 추정값을 계산해 준다. 그러나 중회귀모형을 이용하여 얻은 추정값과 S-Plus를 이용하여 얻은 추정값의 차이가 별로 없고 또한 본 논문의 목적이 STAR 모형과 SAR 모형의 비교에 있으므로 본 논문에서는 간단히 중회귀 모형을 이용하여 모수를 추정하였다. 또한 신기일과 이상은(2003)은 (2.2)모형을 이용한 분석 기법이 소지역 추정에 사용될 수 있음을 보였으며 표 2.3의 모수 추정결과를 보아도 모수 추정값이 유의한 것으로 나타나 같은 결론을 내릴 수 있다.

### 2.3. STAR(Space Time Autoregressive) 모형을 이용한 분석

먼저 전국과 시도별 실업자수의 시계열 분석을 실시하였다. 분석은 ARIMA 모형을 이용하였으며, 2년간 얻어진 자료가 분석에 사용되었기 때문에 자료 수가 충분하지 않다고 판단되어 계절요인은 고려하지 않았다. 전국과 시도별 모형 식별 결과는 다음과 같다.

표 2.4: 시계열 모형 식별 결과

지역	시계열 모형	지역	시계열 모형
서울 (11)	AR(1)	강원 (32)	AR(1)
부산 (21)	AR(1)	충북 (33)	AR(1)
대구 (22)	AR(1)	충남 (34)	AR(2)
인천 (23)	AR(1)	전북 (35)	AR(2)
광주 (24)	AR(1)	전남 (36)	AR(1)
대전(25)	AR(1)	경북 (37)	AR(1)
울산 (26)	AR(1)	경남 (38)	AR(2)
경기(31)	AR(1)	전국	AR(1)

이제 2.2절에서 언급한 SAR 모형과 시계열 요인을 추가한 모형인 STAR 모형을 살펴보자. 먼저 SAR 모형과 비교하기 위하여 모형 (2.2)에서와 같이 전국을 시 및 도의 구별 없이 하나의 지역이라 생각하고 분석하였다. 따라서 공간상관관계를 나타내는 모수는 시도와 상관없으므로 하나만 사용되었으며, 전국을 하나의 지역으로 생각하였을 때 시계열 분석결과 AR(1) 모형이 구해졌으므로 AR(1) 모형을 적용하였다.

$$Z_{i,t} = \beta X_{i,t} + \rho S_{i,t} + \phi Z_{i,t-1} + \epsilon_{i,t} \quad (2.3)$$

여기서  $i, t$ 는  $t$ 시점의  $i$ 번째 시군구를 나타내며,  $Z_{i,t-1}$ 은 시점  $t-1$ 에서의  $i$ 지역의 실업자수이고,  $i$ 지역과 이웃하고 있는 지역의 자료를  $Z_{i,k,t}$ 라 하면  $S_{i,t} = \sum_k Z_{i,k,t}$ , 즉  $S_{i,t}$ 는 지역  $i$ 와 이웃하고 있는 시군구 자료값의 합이다. 그리고  $\epsilon_{i,t}$ 는 각 지역의 오차항으로 독립이고 평균이 '0'이다. 모형 (2.3)의 모수추정 결과는 아래 표와 같다.

표 2.5: STAR 모형의 모수추정 결과

변수	추정값	표준오차	p-value	$R^2$
$X_i$	0.08773	0.03772	0.0193	0.5451
$S_i$	0.08214	0.02141	0.0078	
$Z_{i,t-1}$	0.65161	0.05106	0.0001	

추정값은 2.2 절에서와 마찬가지로 모형 (2.3)을 중회귀모형이라 간주하고 최소제곱추정법을 이용하여 얻어진 추정결과이다. 이제 표 2.3과 표 2.5의 결과를 비교하자. 먼저  $R^2$ 를 살펴보면 시계열 요인을 추가한 모형인 모형 (2.3)의  $R^2$ 가 매우 높은 것을 알 수 있다. 이는 추

가된 시계열 요인이 많은 정보를 보유하고 있음을 말해주는 것으로 전달 또는 전전달과 같은 과거 시점의 실업자수가 하나의 새로운 독립변수로 사용될 수 있음을 보여주는 것이다.

다음절에서는 과거시점의 실업자수를 독립변수로 추가한 모형인 모형 (2.3)과 신기일과 이상은(2003)에서 사용된 모형인 모형 (2.2)를 모의실험을 통하여 비교하였다. 전국을 하나의 소지역으로 생각하고 계수를 추정할 경우의 비교와 15개 시도를 하나의 소지역으로 생각하여 각각의 계수를 추정할 분석도 비교하였다.

### 3. 두 방법의 비교 및 모의실험 결과

#### 3.1. 두 방법의 비교

공간통계 기법을 이용한 모형인 (2.2)와 공간 시계열 기법을 이용한 모형인 (2.3)에서 모수 추정은 최소제곱추정법이 사용되었다. 모형의 비교에서도 최소제곱추정법이 사용되었으며, 이 때 얻어진 추정값을 이용하여 얻어진 예측값과 실제값이 비교에 사용되었다. 자료 분석이 Freeman-Tukey 변환을 한 후 이루어졌으므로 비교를 위하여 재변환을 실시하였다. 즉 Freeman-Tukey 변환

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

에서

$$Y_i = \frac{n_i}{1000} \times \left( \frac{Z_i^2 - 1000/n_i}{2Z_i} \right)^2 \quad (3.1)$$

를 얻게 되고 이 식을 이용하여 재변환된 각 시군구의 예측값이 얻어지며 이 값이 비교에 사용되었다. SAR 모형인 (2.2)와 STAR 모형인 (2.3)의 비교를 위하여 다음의 세 통계량을 사용하였다. 이 통계량들은 신기일과 이상은(2003)에서도 사용되었다.

$$\begin{aligned} MSE &= \frac{1}{N} \sum (Y_i - \hat{Y}_i)^2 \\ MAE &= \frac{1}{N} \sum |Y_i - \hat{Y}_i| \\ MB &= \frac{1}{N} \sum (Y_i - \hat{Y}_i) \end{aligned}$$

여기서  $Y_i$ 는 16,095 모집단 가구에서 얻어진 모집단 값이며  $\hat{Y}_i$ 는 시군구의 실업자 수 추정값이고  $N$ 은 각 시 및 도 또는 전국에 포함된 시군구 수이다. 예를 들어 서울의 경우  $N = 25$ 이고  $i$ 는 서울에 있는 각각의 구이다.

이 논문에서는 모수추정에 사용된 자료수에 따라 두 가지 경우로 나누어 비교하였다. 첫 번째 경우는 2절에서 사용한 분석 방법을 그대로 적용하는 것이다. 즉 모형 (2.2)와 (2.3)의 모수 추정을 전국 209개의 자료를 모수 추정에 사용하는 것이다. 두 번째 경우는 각 시 및 도 별 자료만을 이용하여 모수를 추정하는 것이다. 예를 들면 서울의 경우 25개의 시군구 자료를 이용하여 서울 소지역의 모수를 추정하였다. 각 경우에서 얻어진 추정값을 이용하여 소지역 추정이 이루어졌다. 두 경우에 대한 결과는 각각 표 3.1과 표 3.2에 나와있다. 참고로 시도와 달리, 전국의 경우는 209개의 시군구를 동시에 사용하여 추정을 하므로 표 3.1과 표 3.2의 전국 결과는 같게 된다. 또한 현재의 시군구 수는 234개이나 신기일과 이상은(2003)에서 사용한 시군구를 본 논문에서도 사용하였다.

표 3.1: 전국 자료를 이용한 소지역 추정 비교

지역	MSE		MAE		MB	
	SAR	STAR	SAR	STAR	SAR	STAR
서울 (25)	5.4328	2.2597	2.0227	1.2103	0.6435	0.0881
부산 (16)	13.7523	5.3647	2.7097	1.9481	1.7354	1.0387
대구 (8)	12.3081	3.6968	2.5951	1.5863	0.7848	0.0950
인천(9)	10.6466	6.1273	2.3158	2.0313	1.0739	0.5897
광주(5)	12.5814	6.9768	3.0985	2.4760	0.0267	-0.4278
대전(5)	11.5568	7.0720	2.7296	2.2973	0.2456	0.5688
울산 (5)	3.0458	5.2859	1.4801	1.8197	-1.2101	-1.4929
경기 (26)	1.6608	1.2525	0.9105	0.7486	0.1112	-0.0829
강원 (17)	3.201	2.5556	1.0778	0.8634	-0.3399	0.1125
충북 (10)	5.7659	0.8735	1.6054	0.8127	0.3605	-0.1355
충남 (15)	1.1816	1.3749	0.8956	0.9109	-0.1558	0.1918
전북 (13)	0.9241	1.3639	0.6945	0.7462	-0.0951	0.0969
전남 (17)	4.9792	1.5955	1.1365	0.846	0.3638	-0.0968
경북 (19)	2.4286	1.5837	0.9473	0.688	0.2302	0.2609
경남(19)	1.4101	0.8085	0.9575	0.6307	0.1357	0.4506
전국 (209)	4.9217	2.4851	1.4741	1.1026	0.3129	0.1557

표 3.2: 각 시도별 자료를 이용한 소지역 추정 비교

지역	MSE		MAE		MB	
	SAR	STAR	SAR	STAR	SAR	STAR
서울 (25)	4.9851	2.0239	1.9323	1.0961	0.6624	0.1588
부산 (16)	12.7582	3.8494	2.6423	1.4585	1.7785	0.8076
대구 (8)	14.9426	3.2311	2.8417	1.3324	1.4071	0.69
인천 (9)	10.6143	6.6753	2.5114	2.0211	0.4654	-0.0113
광주(5)	15.7952	1.0327	3.2365	0.9015	-0.7719	0.1534
대전 (5)	12.2354	3.1909	2.8009	1.6177	0.1562	0.0233
울산 (5)	2.0812	2.1781	1.3288	1.3517	-0.7154	-0.7546
경기 (26)	1.2776	0.8045	0.7586	0.582	0.1883	0.0601
강원 (17)	2.0955	2.2333	0.8649	0.8018	0.1332	0.198
충북 (10)	5.4873	0.5051	1.4186	0.6071	0.5542	0.0025
충남 (15)	2.386	3.7826	1.0776	1.2644	-0.4702	-0.2058
전북 (13)	1.0548	0.1402	0.7692	0.2741	-0.211	-0.0869
전남 (17)	6.3991	2.0223	1.3744	0.9178	-0.4244	-0.1668
경북 (19)	2.6113	1.4801	0.9309	0.6603	0.3444	0.2544
경남 (19)	1.4763	0.7435	0.988	0.5755	0.2516	0.3109
전국 (209)	4.9217	2.4851	1.4741	1.1026	0.3129	0.1557

전술한 대로 표 3.1은 시 및 도의 구분 없이 전국자료를 이용하여 얻은 모수 추정값을 각 지역에 적용하여 얻은 결과이다. 결과를 살펴보면 서울, 부산, 대구 등 대부분의 지역에서 공간상관관계와 시계열 관계를 이용한 STAR 모형이 MSE, MAE를 기준으로 하였을 때 우수한 것으로 나타났다. MB의 경우도 서울, 부산 대구 등 많은 지역에서 우수한 것으로 나타났으나 울산, 충남 등 몇 지역에서는 나쁜 결과를 주는 것으로 나타났다.

또한 지역별 공간상관관계와 시계열 모형을 이용한 분석결과인 표 3.2에서도 공간상관관계와 시계열 모형을 이용한 중회귀분석이 공간상관관계만을 고려한 중회귀분석에 비해 전체적으로 우수함을 알 수 있다. 그러나 울산, 충남 등 일부 지역에서는 표 3.1에서와 마찬가지로 시계열 모형을 이용한 경우 오히려 나쁜 결과를 주는 것을 확인할 수 있다. 이는 시계열 요인이 있어 이 요인을 모형에 추가하여 분석한 경우에도 추정해야 할 시점의 자료 특성이 과거 자료의 특성과 다를 경우 시계열 모형에 의한 예측에는 큰 오차가 발생할 수 있다. 따라서 이러한 경우에는 시계열 요인을 첨가한 STAR 모형의 결과가 시계열 요인을 첨가하지 않은 SAR 모형의 결과보다 나쁘게 나올 수 있다. 결국 자료의 특성과 지역별 특성을 고려한 후 모형을 선택하여야 더 좋은 결과를 얻을 수 있음을 말해 주고 있는 것이다.

이상의 결과를 종합해 보면 시계열 요인을 분석할 수 있는 모형인 STAR 모형이 더 좋은 결과를 주고 있음을 알 수 있다.

### 3.2. 모의실험을 이용한 두 방법의 비교

3.1절의 표 3.1과 표 3.2는 16,095개의 모집단 전체 자료를 이용하여 모수를 추정하고 예측한 실업자 수와 실제 실업자 수를 비교하여 얻어진 표이다. 그러나 실제 소지역 추정에서는 모집단 전체의 자료를 이용하는 것이 아니라 추출된 표본을 통해 추정이 이루어지므로 표본에 의해 추정된 모수를 SAR 모형과 STAR 모형에 적용한 경우에도 앞에서와 같은 결과를 얻을 수 있는지 확인해 보았다.

이를 위하여 먼저, 16,095개의 모집단 자료에서 8000개의 표본을 추출한 후 이 표본을 이용하여 추정된 모수로 SAR 모형과 STAR 모형으로부터 MSE, MAE 그리고 MB를 구하였다. 그리고 이를 1,000번 반복한 후 이 결과를 평균하였다. 이렇게 얻어진 결과를 표 3.3에 수록하였다.

표본 추출된 자료로 지역별 공간상관관계와 시계열 모형을 적용한 분석결과에서도 공간상관관계와 시계열 모형을 이용한 중회귀분석이 공간상관관계만을 고려한 중회귀분석에 비해 전체적으로 우수한 것을 알 수 있다. 또한 전국의 공간상관관계를 적용한 비교도 하여 보았으나 표 3.3과 같은 결과가 나와 본 논문에는 수록하지 않았다. 이처럼 표본 추출된 자료를 이용하여 소지역 추정을 한 경우에도 모집단 자료를 이용한 결과와 마찬가지로 공간상관관계와 시계열적 관계를 함께 적용한 STAR 모형이 SAR 모형에 비해 좋은 추정 결과를 보여주었다.

표 3.3: 표본추출 된 자료에서 각 지역별 상관관계를 이용한 소지역 추정 비교

지역	MSE		MAE		MB	
	SAR	STAR	SAR	STAR	SAR	STAR
서울 (25)	3.3217	1.5966	1.4959	1.0194	0.4581	0.0213
부산 (16)	7.1749	3.7902	1.7719	1.344	0.9606	0.171
대구 (8)	11.5718	4.3701	2.4494	1.3888	1.0641	0.1418
인천 (9)	9.4659	1.5211	2.3268	0.8873	0.5606	-0.0035
광주 (5)	9.8363	7.755	2.3575	2.1681	-0.5543	-0.4262
대전 (5)	5.2804	7.6825	1.8376	2.0958	0.9496	-0.6448
울산 (5)	4.6221	3.9528	1.4824	1.3583	-0.6519	-0.6189
경기 (26)	1.3051	0.6703	0.8462	0.5449	0.2191	-0.0354
강원 (17)	1.4327	0.2884	0.7302	0.3142	-0.2395	-0.1433
충북 (10)	5.1821	1.4656	1.6917	0.9244	-0.0428	0.0984
충남 (15)	1.8786	0.9587	1.0292	0.7241	-0.2993	-0.0318
전북 (13)	0.9428	0.4469	0.5927	0.4539	-0.1777	-0.0427
전남 (17)	4.4709	2.0743	1.0073	0.6934	0.0389	-0.0498
경북 (19)	1.9466	0.7285	0.8294	0.5048	0.1904	0.0299
경남 (19)	1.458	0.8051	0.804	0.5729	0.1338	0.1671
전국 (209)	3.4443	1.29	1.1776	0.7265	0.2491	0.0671

#### 4. 결론

본 논문에서는 각 지역에서 얻어진 자료들 간에 공간상관관계가 존재하며, 매월 또는 매년 얻어진 시계열 자료를 이용하여 모형-기반 소지역 추정을 할 때 공간상관관계 및 시계열적인 관계를 고려한 분석 방법을 살펴보았다. 예상대로 공간상관관계나 시계열 요인이 존재할 경우 이 요인을 설명할 수 있는 모형이 더 좋은 결과를 주고 있다. 이제 김달호와 김남희(2002)에서 처럼 시계열 요인이 있는 경우의 베이지안 추정법등과 같은 모수 추정 기법을 사용하여 분석한다면 더욱 좋은 결과를 얻을 수 있으리라 생각된다. 또한 소지역 추정에서 중요한 부분의 하나인 MSE 추정에 관한 활발한 연구가 필요하다 하겠다.

#### 참고문헌

- 김달호, 김남희 (2002), 반복조사에서 소지역자료의 베이지안 분석, <응용통계연구>, 15, 119-128  
 신 기일, 이 상은 (2003), Model-data based small area estimation, <한국통계학회논문집> 10, 637-645.  
 Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley and Sons, Inc.

- De Luna, X., and Genton, M.G. (2005). Spatio-temporal autoregressive models for US unemployment rate, *Advances in Econometrics*, To Appear.
- De Luna, X., and Genton, M.G. (2003). Predictive Spatio-temporal models for spaiial sparse environmental data, Institute of Statistics Mimeo Series, 2546.
- Freeman, M. F., and Tukey, J. W. (1950). Transformation related to the angular and the square root, *Annals of Mathematical Statistics*, **21**, 607-611.
- Rao, J. N. K. (2001). Introduction to Small Area Estimation, 2001 ISI Manuscript.

[ 2005년 2월 접수, 2005년 7월 채택 ]

## Space Time Autoregressive Model for Small Area Estimation

Jae Doo Kim<sup>1)</sup> Key-Il Shin<sup>2)</sup> Sang Eun Lee<sup>3)</sup>

### ABSTRACT

Small area estimation has been studied using various methods such as direct, indirect, synthetic and based on regression or time series model. In this paper we investigate a model-based small area estimation which takes into account the space time autoregressive model. The Economic Active Population Surveys in 2001 are used for analysis and the results from space-time autoregressive(STAR) and simultaneous autoregressive(SAR) model are compared with using MSE, MAE and MB.

*Keywords:* Moran's I, Neighborhood information, Simultaneous autoregressive(SAR) model

---

1) Graduate Student, Dept. of Applied Statistics, Kyonggy University, 94-3 san, Eeudong, Youngtonggu Suwon, Kyonggi, 449-791, Korea

E-mail: luckjd@dreamwiz.com

2) Professor, Dept. of Statistics, Hankuk University of Foreign Studies, 89 San, Wangsan, Mohyun, Yongin Kyonggi, 449-791, Korea

E-mail: keyshin@hufs.ac.kr

3) Professor, Dept. of Applied Statistics, Kyonggy University, 94-3 San, Eeudong, Youngtonggu, Suwon, Kyonggi, 449-791

E-mail: sanglee62@kyonggi.ac.kr