응용통계연구 제 19권 1호, 2006년, pp. 57-67

표본조사에서 공간 변수(SPATIAL VARIABLE)를 이용한 결측 대체(MISSING IMPUTATION)의 효율성 비교

이진희[1] 김 진[2] 이기재[3]

요 약

표본조사에서 무응답은 여러 가지 이유로 발생하며, 이 때 응답자들의 정보로만 분석을 실시한다면 편향된 결과를 산출할 수 있어 보조변수를 이용한 많은 무응답 대체 방법들이 연구되고 있다. 만일 결측자료 대체를 위한 보조변수들이 충분하지 않고 응답자들과 무응답자들 사이에 지역적 상관관계가 존재한다면 이를 결측자료 대체(missing data imputation)에 이용 할 수 있을 것이다. 본 논문에서는 2002년 강원지역의 농가경제 자료를 예제로 하여 공간상관을 이용한 무응답 대체 방법을 살펴보았으며, 공간상관이 존재할 경우 공간 대체 방법이 효율적임을 확인하였다.

주요용어: 결측자료, 공간 상관관계, SAR 모형, 가장 가까운 이웃,공간대체.

1. 서론

표본조사에서는 여러 가지 이유로 결측자료가 발생한다. 그럼에도 이들 결측자료를 무시하고 분석을 하였을 경우 편향된(biased) 추정결과를 줄 수 있으며, 이때 적절한 값으로 결측자료를 대체한다면 더 정확한 추정치를 얻을 수 있을 것이다. 결측자료가 발생하였을 경우 이를 대체하기 위한 가장 일반적인 방법은 보조정보를 이용하는 것이며 이러한 보조정보의 사용은 분산과 편향(bias)을 줄일 수 있게 해 준다. 그러나 이러한 장점에도 불구하고 사용할 보조정보가 한정되어 있을 경우 결측자료를 대체하는데 어려움이 따른다(Son *et al.*, 2001). 무응답 대체를 위한 보조정보 등의 이용이 만족스럽지 못할 때 만일 얻어진 자료가 공간자료이고, 각 지역들 간에 상관관계가 존재한다면 이를 이용하여 결측자료를 대체 할 수 있을 것이다.

그동안 표본조사에서 공간분석은 공간정보의 부족 등으로 별로 이용되지 못하였으나 공간통계에 대한 관심이 높아지면서 표본조사에서도 공간통계분석기법들이 활용되기 시작하였다. 최근 Shin과 Lee(2003)는 설명변수가 한정되어 있고 이를 이용한 추정의 정도가 만족스럽지 못할 경우 공간 상관관계가 존재한다면 공간 상관관계를 이용한 분석을 실시

1) (411-769) 경기도 고양시 일산구 마두동 809, 국립암센터 국가암관리사업지원 평가연구단, 암등록연구과, 연구원 E-mail : jhlee@ncc.re.kr
2) (302-701) 대전광역시 서구 선사로 139, 통계청 지역통계과, 사무관 E-mail : jink@nso.go.kr
3) (110-791) 서울시 종로구 동숭동 169, 한국방송통신대학교 정보통계학과, 교수 E-mail : kjlee@knou.ac.kr

할 것을 제안하였으며, 2001년 경제활동인구조사 자료를 이용하여 우리나라 16개 시도에 대한 소지역 추정에서 이를 이용하여 더 좋은 추정결과를 얻을 수 있었다.

통계청에서는 전국 농가 중 표본으로 추출된 3,200가구를 대상으로 농가소득과 관련된 주요지표 항목들을 매월 조사·집계하여 1년에 한번씩 농가경제에 대한 통계를 작성하여 발표하고 있다. 통계작성 과정에서는 12개월 모두 조사가 이루어진 가구의 조사결과만을 이용하고 있다. 농가경제조사에서는 표본대상가구가 부적격이거나 조사 불응으로 유고가 발생하면 조사구 내에서 동일한 영농형태를 갖는 가구들 중에서 농업소득이 가장 유사한 가구로 바로 교체(substitution)하여 조사가구에 대한 결측자료를 방지하고 있어 무응답의 발생은 그리 많지 않다. 그러나 교체가 원활히 이루어지지 않거나, 표본대상가구의 단기적인 출타, 자연재해 등으로 인한 1 ~ 2개월의 결측자료가 발생할 가능성은 늘 존재한다 (통계청, 2003). 만일 결측자료들 사이에 공간 상관이 존재 한다면 이를 결측자료 대체에 사용할 수 있을 것이다. 본 논문에서는 2002년 결측자료가 발생한 강원도지역 농가자료를 이용하여 공간상관이 있는지 알아보고 공간상관이 있을 경우 공간상관을 이용한 무응답 대체의 효율성을 살펴보았다. 2장에서는 무응답 대체에 관한 여러 가지 방법들을 살펴보았으며 3장에서는 각 조사구들 간의 공간 상관관계의 유무를 Moran's I (Cressie, 1993)를 이용하여 살펴보고 공간상관이 존재할 경우 사용할 수 있는 두 가지 대체 방법을 살펴보았다. 4장에서는 여러 가지 무응답 대체 방법에 대한 효율성비교를 실시하였으며 전체적인 결론은 5장에 있다.

2. 무응답 대체 방법(imputation method)

인구통계학적인 조사에서 자주 발생하는 무응답은 이를 무시하고 추정할 경우 편향된 결과를 주기 때문에 무응답 대체 방법에 대한 꾸준한 연구가 진행되어 왔다. 조사단위 자체가 무응답일 경우 일반적으로 사용하는 대체 방법은 보조변수 등을 이용한 가중조정 방법이고, 항목이 무응답일 경우는 랜덤 또는 유일하게 결정되는 방법들을 사용하여 결측자료를 대체하게 된다. 여기서 한 개의 값으로 결측자료를 대체하는 방법을 단일대체법 그리고 여러 개의 값으로 대체하는 방법을 다중대체법 (Rubin, 1987)이라 한다. 이 장에서는 공간 상관관계를 이용한 무응답 대체법을 소개하기에 앞서 실제 조사에서 널리 사용되고 있는 단일대체 형태의 무응답 대체방법들에 대해서 살펴본다.

2.1. 자료기반 결측자료 대체 방법

항목 무응답의 대체 방법 중 자료기반 대체 방법은 칸평균대체, 핫덱 대체(hot-deck imputation), 최근방 대체(nearest neighbor imputation), 연역적 대체(deductive imputation), 콜드덱 대체(cold-deck imputation)등 많은 방법들이 있다. 먼저, 칸평균 대체법은 조사변수에 대해서 유사하도록 조사대상자들을 칸(cell)을 구성하여 구분한 후에 각 칸에 있는 응답자들의 평균값으로 그 칸에서 발생한 결측자료를 대체하는 방법이다. 이 방법은 평균이나 총합과 같은 일변량 모수에 대한 점 추정량의 편의를 감소시키고, 사용이 간편하여 실제적인 사회조사에서 많이 사용되고 있다. 그러나 각 칸에 결측자료가 많을 경우 모집단의

분포가 왜곡될 가능성이 크다는 점이 단점으로 지적되고 있다. 핫덱 대체법은 통계조사에서 발생한 결측자료의 처리에 널리 사용되고 있는 방법으로 보조변수를 이용하여 층을 나눈 후 각 층 내에서 결측값이 있는 경우에는 해당 층 내의 응답값 중에서 랜덤하게 추출된 자료로 대체하는 방법이다. 이 방법은 결측자료 대체 후 표본의 분포가 그대로 유지될 수 있다는 장점이 있으나 분산이 다른 방법에 비해서 상대적으로 클 수 있다는 점이 지적되어 왔다. 최근방 대체는 결측자료가 발생하는 조사단위와 이 조사단위에 대한 보조변수와의 거리가 가장 유사한 응답값으로 결측자료를 대체하는 방법으로 보조변수가 적절하지 않을 경우 사용하는데 어려움이 따른다. 연역적 대체는 동일 자료 내 다른 항목을 이용하여 대체하는 방법으로, 결측자료 대체에 사용하고자 하는 항목에서 결측자료와 분명한 논리관계가 성립하는 값을 결측자료에 대체하는 방법이다. 이 방법은 결측자료와 대체하고자 하는 항목 간에 분명한 논리성이 성립될 때 적용가능한 방법이다. 콜드덱(cold-deck) 대체법은 기존의 조사나 다른 정보들로부터 얻어진 다른 자료집합에 있는 자료로부터 유사한 항목의 응답 값으로 대체하는 방법으로 과거의 자료 정보를 이용하므로 과거와 현재의 시간적 차이로 인한 응답의 일치성에 있어 문제점을 가질 수 있다 (Lohr, 1999).

2.2. 모형기반 결측자료 대체 방법

무응답 대체 방법 중 모형기반의 대표적인 방법은 회귀 대체법(regression imputation method)으로 회귀 대체법은 결측값이 있는 주 관심 변수와 조사된 보조변수 사이의 상관관계를 이용하여 모형을 설정한 후 이 모형을 이용하여 결측값을 대체하는 방법 즉, 무응답이 있는 항목 y에 응답이 있는 y의 보조변수 $x_1, x_2, \cdots, x_k$를 회귀모형에 적합 시켜 이를 이용하여 결측값을 대체하는 방법이다 (Little and Rubin, 1987). Greenless *et al.*(1982)은 미국 인구조사 (Current Population Survey; CPS)에서 발생하는 결측값을 대체하기 위해서 회귀대체 방법을 이용하였으며 그 결과 대체된 값과 실제값의 평균절대편차(mean absolute deviation)를 비교할 때 다른 대체 방법에 비해 매우 적합함을 보였다. 농가경제조사 자료에도 회귀 대체법의 적용을 고려할 수 있는데, 이를 위한 회귀모형은 식 (2.1)과 같이 표현될 수 있다.

$$Z_i^{(j)} = \beta X_i^{(j)} + \epsilon_i \tag{2.1}$$

여기서 j 는 각각의 조사구, i는 j번째 조사구에 속하는 가구를 나타내고 β는 회귀계수이다. Z_i^j는 j번째 조사구에 속하는 i번째 가구의 수입을 나타내고 $X_i{}^j$는 j번째 조사구에 속한 i번째 가구의 영농형태, ϵ_i는 각 가구에 대한 오차항으로 독립이고 평균이 0 분산이 "σ^2"이라 가정한다. 여기서 만일 보조변수 $x_i{}^{(j)}$와 $z_i{}^{(j)}$사이에 공간 상관이 있을 경우 일반적인 회귀 추정법을 사용하는 것보다 공간 상관관계를 이용한다면 더 효율적인 대체값을 얻을 수 있을 것이다. 다음 장에서는 공간 상관을 이용한 공간 대체 방법을 살펴본다.

3. 공간상관모형(spatial autocorrelation model)

지역적으로 얻은 자료에 결측자료가 발생하였을 경우 만일 보조정보가 부족하고 지역

적 상관관계가 존재한다면 공간상관모형을 결측자료 대체에 이용할 수 있을 것이다. 다음 절에서 공간상관관계를 나타내는 통계량과 이를 이용한 모형 등을 살펴본다.

3.1. 공간 상관관계(spatial autocorrelation : Moran's I, Garry's G)

공간통계에서 지역적 상관관계를 알아보기 위해서 다음의 식 (3.1)의 일반 교차합 통계량(cross-product statistic)을 이용한다.

$$C = \sum_i \sum_j W_{ij} Y_{ij}, \qquad Y_{ij} = (Z_i - Z_j)^2 \tag{3.1}$$

여기서 W_{ij}는 i번째 지역과 j번째 지역이 이웃인지 아닌지를 나타내는 척도를 나타내고, Y_{ij}는 조사 지역 i와 조사지역 j가 유사 또는 비유사를 나타내는 척도를 나타낸다. 또한 Z_i와 Z_j는 각각 i번째 지역과 j번째 지역의 자료를 나타낸다. 위 (3.1)의 교차합 통계량 C는 $N(E(C), Var(C))$를 따름이 알려져 있다(Cressie, 1993). 여기서 $S_0, S_1, S_2, R_0, T_1, T_2$를 다음과 같이 정의 하였을 때

$$S_0 = \sum\sum_{i \neq j} W_{ij}, \; S_1 = \frac{1}{2} \sum\sum_{i \neq j} (W_{ij} + W_{ji})^2 \, , S_2 = \sum_i (W_{i\cdot} + W_{\cdot i})^2$$

$$T_0 = \sum\sum_{i \neq j} Y_{ij}, \;\; T_1 = \frac{1}{2} \sum\sum_{i \neq j} (Y_{ij} + Y_{ji})^2, \;\; T_2 = \sum_i (Y_{i\cdot} + Y_{\cdot i})^2$$

다음과 같은 평균과 분산을 갖는다.

$$E(C) = \frac{S_0 T_0}{n(n-1)},$$

$$Var(C) = \frac{S_1 T_1}{n(n-1)} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n(n-1)(n-2)} + \frac{({S_0}^2 + S_1 - S_2)({T_0}^2 + T_1 - T_2)}{n(n-1)(n-2)(n-3)} - [E(C)]^2$$

또한 공간상관관계를 나타내는 Moran's I 와 Garry's C는 각 각 식 (3.2)와 식 (3.3)으로, 이 통계량을 이용하여 공간 상관을 확인 할 수 있다.

$$I = \frac{n}{2T} \frac{\sum_i \sum_j \delta_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_i (Z_i - \bar{Z})^2} = \frac{\sum_i \sum_j \delta_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})/2T}{\sum_i (Z_i - \bar{Z})^2 / n} \tag{3.2}$$

$$C = \frac{n-1}{4T} \frac{\sum_i \sum_j \delta_{ij} (Z_i - Z_j)^2}{\sum_i (Z_i - \bar{Z})^2} \tag{3.3}$$

위 공간상관을 나타내는 통계량 중 일반적으로 많이 사용하는 통계량은 Moran's I이며, 본 논문에서는 Moran's I를 사용하여 공간 상관관계를 확인하였다. 다음 절에서는 가장 가까운 이웃을 정의하고 정의된 이웃행렬을 이용하여 공간 상관관계의 정도를 살펴본다.

3.2. 가장 가까운 이웃(nearest neighborhood)

격자 자료(lattice data)에서 공간 상관관계를 살펴보기 위해서는 먼저 이웃행렬이 정의되어야 한다. 이웃행렬은 정의에 따라 여러 가지로 나타낼 수 있으나 일반적으로 가장 많이 사용하는 이웃행렬은 경계를 공유하는 이웃을 가장 가까운 이웃으로 정의하는 경우이다. 본 논문에서는 여러 가지 방법들 중 2가지 방법으로 이웃을 정의하였으며, 조사구 단위로 자료가 얻어졌으므로 가장 가까운 이웃의 주체는 각 조사구이다. 정의된 두개의 이웃행렬 중 첫 번째는 결측값을 포함한 가구가 있는 조사구와 같은 경계면을 갖는 조사구를 가장 가까운 이웃으로 정의한 경우로 이를 "이웃행렬1"로 표기하였다. "이웃행렬2"는 "이웃행렬1"에 같은 경계를 가지지는 않지만 같은 군(또는 동) 내에서 같은 영농형태일 때 가장 가까운 이웃으로 추가한 경우이다. 정의된 이웃행렬 중 "이웃행렬1"이 공간상관을 위한 이웃행렬로 주로 사용되는 방법이다. 일반적으로 농가경제 조사에서는 결측자료가 별로 발생하지 않으나 2002년 강원지역 360개 조사구에서 얻어진 농가자료의 경우 8월과 9월에 많은 결측자료가 발생하였으며 이에 대한 결측값 대체가 필요하게 되었다. 8월의 경우 360가구 중 41개 가구(11.4%), 9월 자료는 26가구(7.2%)의 자료를 얻지 못하였다. 물론 얻어지지 못한 이 자료를 대체하기 위한 많은 방법들이 있지만 만일 조사구들 간에 지역적 상관관계가 존재한다면 이를 이용하여 결측값 대체를 할 수 있을 것이므로 먼저 지역적 상관관계가 존재하는지를 살펴보았다.

본 논문에서 사용된 농가자료는 각 조사구에 10가구가 조사되며 강원도는 36개 조사구가 있으므로 총 360가구가 조사되었다. 먼저 해당 월에 대하여 전체 조사구간의 수입에 대한 공간상관을 살펴보았으며 이 경우 조사구들 간에 상관관계를 보이지 않았다. 이 때문에 농가 경제 조사에서 영향을 많이 미친다고 알려져 있는 영농형태(3개 층)와 경지규모로 층화를 한 후 각 조사구들의 공간상관관계를 살펴보았다. 그 결과 경지 규모로 층화를 하였을 경우 각 조사구들의 공간상관은 영농형태로 층화를 하였을 경우에 비하여 크지 않았으며, 영농형태와 경지규모 두 특성 모두를 이용하여 층화를 할 경우 각 조사구에 해당하는 자료의 수가 없거나 너무 적어 본 논문에서는 영농형태만을 층화변수로 활용하였다. 그러나 이 경우 모든 층이 공간 상관관계를 보이고 있는 것은 아니어서 공간 상관관계가 있는 층은 공간 상관관계를 이용하여 대체를 실시하고 공간 상관관계가 없는 층은 칸 평균 대체를 이용하였다. 이와 같은 절차에 의하여 결측자료가 발생한 8월과 9월 자료에 대하여 360개 조사구에서 얻어진 농가경제조사 자료의 이웃행렬에 대한 공간 상관관계를 Moran's I값을 통하여 살펴보았으며 그 결과는 표 3.1에 나타내었다.

표 3.1을 살펴보면 7월, 9월, 12월은 공간상관이 존재하나 8월 자료는 공간 상관관계가 없음을 확인할 수 있다. 비록 8월자료가 공간상관이 존재하지 않지만 9월 자료는 공간상관을 나타내는 Moran's I값이 비교적 높은 값을 주고 있어 공간 상관관계를 이용한 자료분석이 가능하게 된다. 그런데 공간변수를 이용한 결측자료 대체 방법의 효율성 비교를 위해서는 실제 관측값과 지역적 상관관계를 이용하여 추정한 값이 모두 필요하게 되므로 결측이 발생한 9월 자료를 이용할 경우 결측된 자료를 모두 제외한 나머지 자료를 가지고 결측자료를 생성하여 실제자료와 비교할 수밖에 없다. 이러한 이유로 본 논문에서는 결측자료가 없는 월 중 9월과 상관관계가 비슷한 월을 대안으로 찾았으며 표 3.1에 제시된 7월과 12월

표 3.1: 각 월과 이웃행렬에 따른 공간상관

월	이웃행렬 종류	Moran's I	표준오차	p-값
7	이웃행렬1	0.4384	0.1050	8.371e-6
	이웃행렬2	0.3922	0.1094	1.164e-4
8	이웃행렬1	0.0627	0.1112	0.7649
	이웃행렬2	0.0727	2.8153	1.8504
9	이웃행렬1	0.4450	0.0855	2.863e-8
	이웃행렬2	0.3258	0.0892	6.802e-5
12	이웃행렬1	0.4968	0.1203	1.445e-5
	이웃행렬2	0.4760	0.1207	2.803e-5

의 자료가 9월과 비슷한 공간상관을 가지고 있는 월이다. 이 결과는 9월 자료를 대신하여, 공간상관이 비슷한 7월과 12월의 자료를 이용하여 공간변수를 이용한 결측자료 대체방법의 효율성을 살펴보는데 별 무리가 없다고 생각된다. 또한 각 월별 자료에서 두개의 이웃행렬에 따른 Moran's I값은 거의 비슷한 값을 가지나 대체적으로 "이웃행렬1"이 더 높은 값을 주고 있음도 확인할 수 있다.

3.3. 공간 SAR(Simultaneous Autoregressive) 모형

표본조사에서 결측자료가 발생할 경우 이에 대한 대체 방법은 전술한 바와 같이 많은 방법들이 있으나 얻어진 자료에 공간상관이 존재한다면 이를 이용한 결측자료 대체 방법도 타당하리라 본다. 본 논문에서 사용한 강원지역 자료는 특정 조사구에 속하는 표본가구의 농가수입이 결측인 경우로 3.2절에서 살펴본 바와 같이 각 조사구들 사이에 공간 상관관계가 있어 이를 이용한 공간모형을 사용하여 결측자료를 대체하고자 한다. 격자자료(lattice data)에서 일반적으로 사용되는 공간 통계모형은 SAR(Simultaneous Autoregressive)모형으로 아래의 식 (3.4)와 같으며 가장 가까운 이웃으로 정해진 이웃들이 미치는 영향력은 모두 같다고 가정한다.

$$Z_i^{(j)} = \beta X_i^{(j)} + \rho S_i^{(j)} + \epsilon_i \tag{3.4}$$

여기서 j는 각각의 조사구, i는 j번째 조사구에 속하는 가구를 나타내며, $Z_i^{(j)}$는 표준화 자료로 j번째 조사구에 속하는 i번째 가구의 수입, $X_i^{(j)}$는 j 번째 조사구에 속한 i번째 가구의 영농형태, $S_i^{(j)}$는 i번째 가구가 속한 j번째 조사구에 이웃하는 조사구들의 평균수입, ρ는 공간 상관정도를 나타내는 모수, 그리고 ϵ_i는 각 가구에 대한 오차항으로 독립이고 평균이 "0" 분산이 σ^2이라 가정한다. 본 논문에서는 이 모형을 이용하여 얻은 결과를 모형기반 공간대체라고 명명한다. 일반적인 공간자료에서 공간 정보는 결측된 값에서 가장 가까운 이웃들로 정의 하지만 농가경제조사 자료의 경우 농가수입이 많은 가구 바로 옆 가구도 수입

이 많은 가구가 위치하는 것은 아니며, 또한 현 시점에서 각각의 가구에 대한 공간상관 행렬을 얻는 것이 어렵다. 이러한 이유로 본 논문에서는 결측된 각 가구에 대한 가장 가까운 이웃을 결측된 가구가 속한 조사구의 이웃들로 정의하였다. 공간상관이 존재할 경우 결측자료가 존재하는 공간자료에서 사용될 수 있는 또 하나의 방법은 모형을 이용하지 않으면서 공간상관을 이용하여 결측자료를 대체하는 방법이다. 칸평균대체 방법이 각 칸에서 결측자료 값에 해당 칸의 평균값을 대체하는 반면, 공간상관이 존재할 경우 결측된 가구가 속한 조사구들에 최근접 이웃으로 판단된 조사구들의 평균값을 대체 하는 것이다. 본 논문에서는 식 (3.4)를 이용한 모형기반 대체 방법에 이어 이를 자료기반 공간대체방법이라 명명한다. 지금까지 살펴본 결측대체 방법에 대한 효율성 비교를 위하여 다음 장에서 모의실험을 실시하여 보았다.

4. 모의 실험

4.1. 자료설명

모의실험에 사용된 자료는 2002년 강원지역의 농가소득 자료이다. 결측자료는 8월과 9월의 농가소득 자료이며 이중 공간상관이 존재하는 9월 자료에 대하여 공간변수를 이용한 결측값을 대체하고자 한다. 그러나 전술한 바와 같이 결측대체 방법의 효율성 비교를 위해서는 결측대체 자료에 대한 실제값이 필요하게 되는데 9월 자료를 이용할 경우 결측된 자료를 모두 제외한 자료를 이용해야 하므로 문제가 있게 된다. 이에 공간대체 방법의 효율성 비교를 위해 결측이 없으면서 공간상관이 9월과 비슷한 7월과 12월 자료를 사용하였다.

모의실험에 앞서 먼저 월 자료 각각을 표준화한 후 정규성 만족을 위해 로그변환을 실시하였으며 로그 변환을 할 때 값들이 음의 값을 가지는 것을 방지하기 위하여 각 자료에 각 월에 대한 음의 최소값의 절대값을 더해 주었다. 효율성 비교를 위한 모의실험에서는 8월과 9월의 결측자료 수가 전체 자료의 10%를 전 후 하고 있어 이와 비슷하게 10%의 결측자료 수를 생성하였다. 생성한 방법은 36개 조사구에서 각각 1개씩 총 36개의 표본을 SRS(simple random sample) 방법으로 뽑아 뽑힌 자료를 결측값으로 이용하였다.

결측대체는 자료 기반의 경우 뽑히지 않은 나머지 자료를 이용하여 같은 영농형태이면서 같은 조사구의 평균값과 이웃 조사구들의 평균값을 결측값에 대체하였으며, 모형기반 대체 방법의 경우 일반 회귀분석과 공간상관을 공분산 행렬에 적용한 일반화 선형모형(GLM)을 사용하여 모형을 추정한 후 이를 이용하여 결측지점을 예측하였다. 각 방법으로 얻어진 자료는 다시 더해주었던 음의 최소값의 절대값을 빼준 후 재 변환(지수변환)하여 결측값 각각에 대체 하였다. 또한 어느 특정한 달에 행사(예, 결혼, 장례 등)로 인한 가구 수입이 유난히 크거나 작을 때 이를 이상점으로 판단하고 이를 수정한 후 분석결과도 함께 비교하여 보았다. 본 논문에서 사용된 이상점 탐지와 수정 방법은 이와 신(2004)의 방법을 이용하였다. 효율성 비교를 위하여 실제값과 4가지 방법으로 구한 대체값의 MSE(mean square error)와 MAE(mean absolute error)를 계산하였다. 반복은 5,000번을 실시한 후 5,000번에 대한 평균을 구하였다. 가장 가까운 이웃행렬을 이용한 공간상관관계인 Moran's I 값은 S-plus의 SpatialStat 모듈을 이용하여 산출하였으며 각 층에서의 반복적인 표본 추출과 추

표 4.1: 각모의실험 결과(자료기반 결측자료 대체방법)

통계량	자료의 종류	7월 자료			12월 자료		
		회귀대체	공간대체 (이웃행렬1)	공간대체 (이웃행렬2)	회귀대체	공간대체 (이웃행렬1)	공간대체 (이웃행렬2)
MAE	원자료	967801.51	916131.71	910102.54	921784.29	899530.57	911772.22
	수정자료	710200.26	547161.33	537673.72	505965.94	517686.78	533578.68
MSE	원자료	3.82E+12	3.22E+12	3.22E+12	2.68E+12	2.49E+12	2.63E+12
	수정자료	1.24E+12	6.35E+11	6.15E+11	5.50E+11	5.50E+11	6.07E+11

정값에 대한 계산은 각각 SAS/STAT과 SAS/IML 등을 사용하였다. 또한 4가지 결측대체 방법의 효율성 비교를 위하여 사용된 공식은 평균제곱오차(Mean Square Error)와 평균절대오차(Mean Absolute Error)로 식 (4.1)과 (4.2)이다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Z_i^{(j)} - \hat{Z}_i^{(j)})^2 \tag{4.1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(Z_i^{(j)} - \hat{Z}_i^{(j)})| \tag{4.2}$$

여기서 i는 결측되었다고 가정된 가구를 나타내며 n은 전체 결측된 가구 수를 나타낸다. 또한 $Z_i^{(j)}$는 j번째 조사구 내에서 i번째 가구의 실제 수입을 나타내고 $\hat{Z}_i^{(j)}$는 j번째 조사구내에서 i번째 가구가 결측되었다고 가정하고 위에서 설명한 네 가지 방법으로 얻은 추정 값(estimate value)이다.

4.2. 분석결과

다음은 공간상관을 이용한 결측대체 방법의 효율성 비교를 위하여 4가지 방법으로 모의실험한 결과이다. 표 4.1과 표 4.2는 각 조사구에 랜덤하게 1개씩의 결측을 생성한 후 전술한 4가지 방법을 이용하여 결측자료를 대체하고,결측전의 실제 값과 결측대체를 한 추정값의 MAE(Mean Absolute Error)와 MSE(Mean Square Error)이다. 표 4.1은 자료기반 결측자료 대체 방법 중 공간상관을 이용하지 않은 경우와 공간상관을 이용한 경우의 비교결과이고 표 4.2는 모형기반 결측자료 대체 방법 중 공간상관의 이용유무의 비교결과이다.

자료기반 결측자료의 대체 방법결과인 표 4.2의 경우 공간상관관계가 있을 경우 공간상관관계를 이용한 경우가 공간 상관관계를 이용하지 않은 일반 칸평균대체 방법을 사용한 경우보다 이웃행렬의 선택에 관계없이 더 좋은 효율을 준다. 이웃행렬의 선택에 있어서는

표 4.2: 각모의실험 결과(모형기반 결측자료 대체방법)

통계량	자료의 종류	7월 자료			12월 자료		
		회귀대체	공간대체 (이웃행렬1)	공간대체 (이웃행렬2)	회귀대체	공간대체 (이웃행렬1)	공간대체 (이웃행렬2)
MAE	원자료	1027238.53	908388.3	902149.44	986698.9	897636.81	905856.02
	수정자료	595244.26	533637.54	525257.55	558843.09	511156.5	513543.96
MSE	원자료	3.63E+11	3.26E+11	3.25E+11	2.82E+11	2.47E+11	2.55E+11
	수정자료	6.48E+11	5.97E+11	5.92E+11	6.13E+11	5.52E+11	5.59E+11

별 차이를 주고 있지 않으며, 7월 자료에서는 "이웃행렬2"를 이용하였을 경우 더 좋은 결과를 주나 12월 자료에서는 "이웃행렬1"을 이용하였을 경우에 더 좋은 결과를 주는 등, 월별 자료에 따라 서로 다른 결과를 준다.

모형기반 결측자료 대체 방법결과인 표 4.2의 경우도 자료기반 결측자료 대체 방법과 같이 공간상관이 있을 경우 공간상관을 이용한 경우가 공간상관을 이용하지 않은 일반 회귀대체 방법을 사용한 경우보다 이웃행렬의 선택에 관계없이 더 좋은 효율을 준다. 이웃행렬의 선택에 있어서도 자료기반 결측자료 대체에서와 같은 결과를 준다. 또한 전체적으로 살펴볼 때 자료기반 보다는 모형기반 결측자료 대체 방법이 더 나은 효율을 주며 공간상관이 존재할 경우, 공간상관을 이용하지 않은 경우보다 공간상관을 이용하였을 때 더 좋은 효율을 주고 있다.

5. 결론

조사된 자료에 결측(missing)이 존재할 경우 이를 해결하기 위한 많은 연구가 진행되어 왔다. 그러나 공간 상관관계를 이용한 연구결과는 자료가 얻어진 위치에 대한 정보가 필요하기 때문에 상대적으로 널리 활용되지 못하였다. 특히 조사구 단위의 공간 상관관계를 이용하기 위해서는 조사구들의 위치파악이 선행되어야 하는데 지금까지 연구에서는 각 조사구에 대한 위치파악이 되어있지 않아 본 논문에서도 이웃행렬을 만들기 위한 조사구의 위치 파악이 가장 어려웠다. 또한 조사구에 대한 가장 가까운 이웃행렬을 정의할 때 어떤 특별한 기준을 부여하기 힘들어 가장 가까운 이웃행렬을 구하는 데는 다소 주관적일 수도 있다. 이는 가장 가까운 이웃행렬을 어떻게 정의하느냐에 따라 공간 상관관계가 다르게 나타날 수도 있어 어떤 이웃행렬을 사용할 것인지를 정해야하는 어려움을 갖는다. 그러나 본 논문에서의 결과에서 살펴보았듯이 일반적인 기준에서 크게 벗어나지 않는다면 그 값은 큰 차이를 보이지 않는다. 따라서 본 논문에서 예제로 든 농가경제조사 자료뿐만 아니라 공간 상관관계가 존재하는 다른 자료에서도 이를 분석에 활용한다면 더 좋은 결과를 얻을 수 있으리라 생각된다. 물론 각 조사구에 대한 위치파악이 되어있다 하더라도 공간 상관관계가

존재하지 않는다면 공간 상관관계를 이용한 결측자료 대체방법은 사용하기 어려울 것이므로 결측자료에 대한 공간대체를 사용할 경우 먼저 공간 상관관계에 대한 검토가 선행되어야만 한다.

본 논문에서는 2002년 강원지역 자료가 수해로 인해 특정 월에 많은 결측값을 가지게 되었으며 이에 대하여 공간 변수를 이용한 결측대체 방법의 효율성을 살펴보았다. 물론 효율성 비교를 위해 결측이 발생한 해당 월을 직접 사용하지 못하고 결측이 발생한 해당 월과 공간상관이 비슷한 다른 월을 비교하여 분석하였지만 이 결과를 결측이 있는 월에 사용하여도 무방하리라 본다. 또한 농가경제조사 자료와 같이 가구가 표본으로 선정이 되면 몇 년씩 조사가 계속되므로 농가 수입에 대하여 어느 특정 월에 대한 결측이 생겼을 경우 공간상관과 시계열적 상관을 동시에 이용한 대체 방법 또한 타당하리라 본다. 공간 시계열 상관을 이용한 무응답 대체와 대체법을 적용한 이후의 추정량 분산 추정방법에 대한 연구는 추후 연구과제로 남겨둔다.

참고문헌

통계청 (2003). <농가경제조사>, 농산물 생산비조사 지침서.

이진희, 신기일 (2004). 공간통계분석에서 이상점 수정방법의 효율성비교, <응용통계연구>, **17**, 327-336.

Cressie, N. (1993). *Statistics for spatial data*, John Wiley and Sons, Inc.

Greenless, J. S., Reece, W. S. and Zeischang, K. O. (1982). Imputation of missing values when the probability of response depends upon the variable being imputed, *Journal of the American Statistical Association*, **77**, 251-261.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, John Wiley Y Sons.

Lohr, S. L. (1999). *Sampling : Design and Analysis*, Duxbury Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, John wiley and Sons.

Shin, K. I. and Lee, S. E. (2003). Model-Data Based Small Area Estimation, *The Korean Communications in Statistics*, **10**, 637-645.

Son, C. K., Hong, K. H., and Lee G. S. (2001). The Calibrated Variance Estimator under the Unit Nonresponse. *Korean Computational and Applied Mathematics*, **8**, 975-987.

[2005년 6월 접수, 2005년 12월 채택]

Missing Imputation Methods Using the Spatial Variable in Sample Survey

Jin-Hee Lee[1)] Jin Kim[2)] Kee-Jae Lee[3)]

ABSTRACT

In sampling survey, nonresponse tend to occur inevitably. If we use information from respondents only, the estimates will be baised. To overcome this, various nonresponse imputation methods have been studied. If there are few auxiliary variables for replacing missing imputation or spatial autocorrelation exists between respondents and nonrespondents, spatial autocorrelation can be used for missing imputation. In this paper, we apply several nonresponse imputation methods including spatial imputation for the analysis of farm household economy data of the Gangwon-Do in 2002 as an example. We show that spatial imputation is more efficient than other methods through the numerical simulations.

Keywords: Missing data; Spatial autocorrelation; SAR model; Nearest neighborhood; Spatial imputation.

1) Researcher, Cancer Registration Branch, Research Institute for National Cancer Control and Evaluation. National Cancer Center, 809, Madu-dong, Ilsan-gu, Goyang, Kyunggi, 411-769, Korea.
E-mail: jhlee@ncc.re.kr

2) Deputy Director, Regional Statistics and Sampling Division, Korea National Statistical Office, 139, Seonsaro, Seo-gu, Daejeon, 302-701, Korea.
E-mail : jink@nso.go.kr

3) Professor, Department of Information Statistics, Korea National Open University, 169, Dongsung-dong, Jongno-gu, Seoul, Korea
E-mail : kjlee@knou.ac.kr