응용통계연구 제 **20**권 3호, 2007년, pp. 499–514

공간-시계열 모형을 이용한 결측대체 방법에 대한 연구*

이진희[1] 신기일[2]

요 약

표본조사에서 항목무응답 발생 시 결측대체에 사용되는 일반적인 방법은 결측변수와 관계있는 보조변수를 이용하는 것이다. 최근 이진희 등 (2006)은 2002년 강원지역의 농가경제 자료를 이용하여 표본조사에서 공간통계를 이용한 결측대체 (missing imputation) 방법을 비교하였으며, 자료들 사이에 지역적 상관이 존재할 때 이를 이용한 결측대체가 효율적임을 보였다. 본 논문에서는 이를 확장한 개념으로, 강원지역의 2000–2002까지의 월별 자료가 공간상관과 시계열상관이 존재함을 확인하고 이 관계를 결측대체에 이용하였다. 또한 공간상관과 시계열상관이 모두 존재할 경우 공간시계열 모형을 이용한 결측대체 방법이 공간모형을 이용하였을 때에 비해 더 효율적임을 모의실험을 통해 확인하였다.

주요용어: 결측자료, STAR 모형, 이웃 정보, 공간시계열 대체.

1. 서론

농가경제조사에서 무응답과 표본교체는 불가피하게 발생한다. 일반적으로 표본대상가구가 부적격이거나 조사 불응으로 유고가 발생하면 조사구내에서 동일한 영농형태를 갖는 농업소득이 가장 유사한 가구로 교체 (substitution)하여 조사의 결측치를 방지한다. 그럼에도 불구하고 교체가 원활히 이뤄지지 않거나 장기부재자의 발생 또는 최근 자주 발생하는 기상이변등으로 인하여 1–2개월 정도로 일시적인 결측이 발생하게 된다 (통계청, 2003). 이 때 무응답 가구에 대한 실제적인 값들이 결측 되는데, 이 경우 가구의 기본 정보 이외에 추가적인 정보를 이용한다면 무응답 대체에 큰 도움을 줄 것이다. 예를 들어 이진희 등 (2006)에서와 같이 지역 (regions) 정보를 이용하거나 시계열이 유지된 경우에는 과거시점 자료를 이용하는 것이다. 이렇듯 지역 간 또는 시점들 간에 상관관계가 있을 경우 이들 상관관계 정보를 이용한다면 결측대체에 도움을 줄 수 있을 것이다. 본 논문에서는 관심변수의 결측과 보조정보의 부족으로 결측대체에 어려움이 있을 때 공간정보와 시계열 정보를 이용하여 결측을 대체하는 방법을 살펴보았다. 물론 지역과 시간적으로 얻어진 자료라 할지라도 지역 간 혹은 시점 간에 상관관계가 존재하지 않으면 이들 정보를 활용할 수 없게

* 이 연구는 2007년도 한국외국어대학교 교내연구비에 의해 수행되었음.

1) (122-701) 서울시 은평구 통일로 194, 질병관리본부 국립보건연구원 에이즈종양 바이러스팀, 선임연구원
E-mail: jhlee@cdc.go.kr

2) (교신저자)(449-791) 경기도 용인시 모현면 산 79, 한국외국어대학교 정보통계학과, 교수
E-mail: keyshin@hufs.ac.kr

된다. 그러므로 시간 또는 공간 정보를 가지고 있다면 먼저 이들 각각이 시점과 지역들 간에 상관관계가 존재하는지를 확인해야 한다. 그동안 표본조사에서 시계열의 필요성은 많이 공감하고 있는 상태이며, 통계청 등에서 시계열 자료를 이용할 목적으로 연동표본 등을 도입하였다. 그러나 공간기법의 적용은 공간정보의 부족 등으로 아직 이용되지 못하고 있는 실정이다. 최근 공간통계에 대한 관심이 높아지면서 표본조사에서도 공간통계 분석기법들이 연구되기 시작하였다. 이진희 등 (2006)은 2002년도 강원지역 농가 수입 자료를 이용하여 일반적인 대체방법 (칸평균 대체, 회귀대체) 등과 이웃정보를 이용한 공간 대체방법의 효율성을 비교하였으며 결측이 발생하였을 경우 설명변수가 충분하지 않고 공간상관관계가 존재할 때 이웃정보를 이용한 공간대체가 일반 칸평균 대체나 회귀대체 방법보다 효율적임을 보였다. 농가수입의 경우 한번 표본가구로 선정 되면 같은 가구에 대하여 여러 해 동안 계속 조사를 해 나가게 되므로 시계열적인 관계도 고려할 수 있을 것이다. 이에 본 논문에서는 공간상관과 시계열상관 관계를 동시에 고려한 공간시계열 (space-time autoregressive: STAR) 모형을 이용하여 결측대체를 하였으며 이를 이진희 등 (2006)의 공간모형 (simultaneously autoregressive: SAR)을 이용한 결과와 비교해 보았다. 본 논문에서 사용된 자료는 2000–2002년에 얻어진 강원지역의 월별 농가경제 자료이다.

본 논문의 구성은 다음과 같다. 2장에서는 결측 대체 방법에 대해서 간략하게 살펴보았다. 3장에서는 본 논문에서 비교한 두 가지 결측대체 모형인 공간모형과 공간시계열 모형에 대하여 살펴보았으며 4장에서는 두 모형을 이용한 분석 결과를 제시하였다. 5장에서는 두 모형을 이용한 결측대체의 효율성 비교를 위해 모의실험을 실시하였으며 6장에 전체적인 결론이 있다.

2. 무응답 대체 방법

농가표본 조사에서 무응답 발생을 원천적으로 없앨 수는 없으므로 표본교체 및 무응답 발생을 최소화 하는 것이 무엇보다 중요하다. 무응답이 발생할 경우 편의를 줄이는 가장 일반적인 방법은 보조변수를 이용한 무응답 대체방법이다. 이중 자료기반 대체 방법이 칸평균대체 (cell mean imputation), 핫덱 대체 (hot-deck imputation), 최근방 대체 (nearest neighbor imputation) 등이고 모형기반 대체 방법이 회귀 대체법 (regression imputation method)이다. 이들 대체방법 각각에 대한 장단점 및 더 자세한 내용은 이진희 등 (2006)을 참고하기 바란다. 또한 응답자들의 자료를 가지고 무응답자에 대한 추론을 하기 위해서는 먼저 응답자와 무응답자들 간에 나타나는 응답패턴에 대한 관련성을 알고 있어야 하며, 여기서 무응답 패턴이 무시 가능한 형태인지, 혹은 무시하면 안되는지를 살펴보아야 한다. 농가자료에서 주로 발생하는 무응답은 표본설계 시에는 농가였으나 전업 등으로 일반 가구로 전환되는 경우, 표본설계 시에는 2인 가구였으나 사망이나 전출로 인하여 1인가구로 바뀌어 표본가구로 적절하지 못한 경우, 타 지역으로 이동하는 경우, 응답의 피로감으로 인하여 응답을 거절하는 경우 그리고 천재지변이나 산업재해로 조사를 못하는 경우 등이 있다. 이러한 경우 조사의 연속성 및 조사 결과의 신뢰도 유지를 위하여 표본이탈 농가 대신 다른 농가로 교체된다. 그러나 본 연구에서 사용된 농가 자료의 경우는 결측농가에 대하여

다른 농가로 교체가 되지 못하여 결측 대체가 필요한 경우이다. 결측의 원인 또한 김규성 등 (2005)의 농어가경제조사에서 주로 발생하는 무응답과 같은 원인에서 발생되었으므로 본 연구에서 사용된 농가 자료에서 발생하는 무응답의 패턴도 무시가능 (ignorable)한 무응답으로 가정 할 수 있다. 본 논문의 결측 대체 방법에 대하여 간략히 설명하면, 먼저 사용된 자료의 결측패턴이 무시될 수 있다는 가정 하에서 대체군에 사용될 보조변수는 영농형태, 공간상관 정보 그리고 시계열상관 정보이다. 여기서 영농형태는 농가경제에 영향을 미치는 변수로 알려져 있어 보조 정보로 사용하는데 별 무리가 없으나, 공간정보와 시계열정보는 보조 정보로 사용할 수 있는지 확인을 해야만 한다. 본 논문의 목적은 공간상관과 시계열 상관이 있음을 확인하고, 이들 상관이 존재할 때 공간상관 만을 보조정보로 사용하는 경우와 시계열상관을 동시에 보조정보로 사용하는 경우의 효율성 비교에 있다. 이를 위하여 공간 상관관계의 존재를 확인한 후 공간상관을 이용한 추정값을 보조정보로 이용하였으며, 그와 동일하게 시계열상관의 확인 후 시계열상관을 이용하여 모형을 설정하고, 이 시계열모형을 이용한 추정값을 보조정보로 사용하였다. 그 첫 번째 단계로 다음 장에서 공간모형과 시계열 모형에 대하여 소개하고자 한다.

3. SAR 모형과 STAR 모형

본 논문에서 결측대체를 위해 사용된 공간 모형 (SAR model)과 공간시계열 모형 (STAR model)을 살펴보기로 하자. 여기서 j번째 지역 (조사구)에 속한 i번째 가구의 소득을 Y_{ij}라 표시하였으며 Y_{ijt}는 j번째 지역, i번째 가구, t월의 소득을 나타내었다. Y_{ij}는 공간 모형을 위한 자료 표현이고 Y_{ijt}는 공간시계열 모형을 위한 자료 표현 방법이다. 공간 모형을 사용하는데 있어 가장 중요한 단계는 이웃을 결정하는 것이다. 본 논문에서는 "경계를 공유할 때 이웃으로 정한다"로 이웃을 결정하였다. 이와 같이 경계를 공유하는 지역을 이웃으로 하여 모형을 분석한 경우가 공간 모형에 기본적으로 사용되고 있으며 이에 관한 내용은 Cressie (1993)에 자세히 설명되어 있다. 또한 본 논문은 전체 자료를 이용한 공간모형과 함께 농가소득 Y_{ij}에 영향을 준다고 알려진 보조변수들 중 영농형태별 분석을 함께 실시하였다. 일반적으로 결측대체에 주로 사용되는 보조변수는 농가소득이 결측일 경우 함께 결측일 경우가 많기 때문에 결측자료에 대한 보조변수를 얻기가 어렵다. 그러나 영농형태나 농지 크기 등은 조사 초기 이미 조사된 자료이기 때문에 이를 이용하는 것은 크게 어렵지 않으며 이러한 이유로 본 논문에서도 영농형태를 보조변수로 사용 하였다. 보조변수에 대한 더 자세한 내용은 이진희 등 (2006)을 참고하기 바란다. 농가조사에서 정의하는 9가지의 영농형태를 본 논문에서는 한 조사구에 10개 가구가 조사되었기 때문에 자료 수 등의 문제로 인하여 3개 층으로 재 그룹화 하였으며 이에 대한 더 자세한 설명은 다음 장에 있다. 이제 공간변수 $S_{ij}^{(k)}$를 $S_{ij}^{(k)} = \sum_{i \in H_i} Y_{ij}^{(k)} / m_{ij}^{(k)}$라 하자. 여기서 (k)는 영농형태를 나타내며, H_j는 i번째 가구가 속한 j번째 지역과 경계를 같이하는 지역들의 집합, 즉 이웃 집합들이고, $m_{ij}^{(k)}$는 영농형태별로 i번째 가구가 속한 j번째 지역의 이웃의 수이다. 따라서 각 영농형태 (k)에서 i번째 가구가 속한 j번째 지역의 이웃으로 알려진 지역에서 얻어진 값을 모두 더한 후 평균 낸 값이 $S_{ij}^{(k)}$가 되며, 이를 이용한 SAR 모형은 다음과 같다.

$$Y_{ij}^{(k)} = \rho_j^{(k)} S_{ij}^{(k)} + \varepsilon_{ij}. \tag{3.1}$$

위 (3.1)식은 각 조사구에 따라 다른 모수를 고려한 모형으로 시간적 요인을 고려하지 않고 있는 공간 모형이다. 즉 위의 모형에서는 가구가 속해있는 지역 (조사구)의 공간상관 관계만을 고려한 모형이다. 그러나 본 논문에서 사용된 자료는 매월 얻어진 시계열 자료이다. 따라서 위의 모형을 이용하기 위해서는 특정 년월을 정하여 분석을 하거나, 월로 얻어진 자료를 평균한 후 얻어진 평균자료를 이용하거나 아니면 시계열을 무시하고 반복적으로 얻어진 것이라 가정하여 분석하는 방법이 있을 것이다. 본 논문에서는 이러한 방법들 중 특정 년월과 평균을 이용한 자료를 이용하여 분석하였다.

다음으로 본 논문에서 사용된 공간시계열 모형은 (3.2)식과 같다.

$$Y_{ijt}^{(k)} = \rho_j^{(k)} S_{ijt}^{(k)} + \phi_j^{(k)}(B) Y_{ijt}^{(k)} + \varepsilon_{ij}. \tag{3.2}$$

여기서 $\phi_j^{(k)}(B) = 1 + \phi_1^{(k)} B + \phi_2^{(k)} B^2 + \cdots + \phi_p^{(k)} B^p$, p는 시계열 모형의 차수를 나타내고 B는 후진 연산자 (Backward shift operator)를 나타낸다. 물론 위 시계열 모형도 j번째 지역마다 다른 모수 추정값을 얻은 경우이다.

4. 자료 설명 및 예비 분석

4.1. 자료설명

본 연구에서 사용된 자료는 강원지역에서 얻어진 2000년 1월부터 2002년 9월까지의 월별 농가경제 자료이다. 농가경제 자료는 각 조사구당 10가구로 구성되어 있고 강원지역의 경우 36개 조사구가 조사되어 총 360개 가구가 조사되었다. 그러나 시계열 모형을 이용하기 위해서는 같은 가구가 연속적으로 조사되어 시계열이 유지되어야 하므로 중간에 교체되거나 조사되지 못한 23개 가구를 제외한 337가구만이 분석에 사용되었다. 분석에 사용된 자료 중 2002년 8월과 9월 자료는 폭우로 인하여 조사결과를 얻지 못해 결측이 발생하였으며, 이에 대한 대체가 필요하게 되었다. 최근 잦은 기상이변으로 인한 결측이 많이 발생하기 때문에 본 연구는 매우 중요하다고 하겠다. 일반적으로 결측대체 방법의 효율성비교를 위해서는 대체 자료와 실제 자료가 있어야 한다. 그러나 8월과 9월 자료는 결측자료에 대한 실제 자료가 존재하지 않기 때문에 8월과 9월 자료에 대한 결측대체의 효율성 비교는 불가능하다. 결국 효율성 비교를 위하여 2002년 7월 자료를 8월과 9월을 대신하는 검정자료로 남겨두었으며 2000년 1월 – 2002년 6월 자료를 이용하여 모형을 구축하였다. 본 논문에서 사용된 농가 수입자료의 경우 매우 긴 꼬리를 갖고 있어 이를 보정하기 위해 변환을 실시하였다. 농가수입의 경우 수입은 없고 지출만 있는 (즉, 음의 수입) 월도 존재한다. 이렇듯 실제 값이 양수와 음수가 동시에 존재하는 자료에 대한 로그변환은 음의값을 없애기 위해 적당한 값을 더해준 후 변환을 해야 한다. 이에 본 논문에서는 Box-Cox 변환에서 음의 값이 존재할 경우에도 사용할 수 있는 Yeo-Johnson (2000) 변환을 사용하여 변환을 실시한 후 변환된 자료를 분석에 사용하였다.

본 논문에서 사용된 Yeo-Johnson 변환은 (4.1)식과 같다.

$$Z_{(\lambda)} = \begin{cases} ((Y+1)^{\lambda}-1)/\lambda, & Y \geq 0, \lambda \neq 0 \\ \log(Y+1), & Y \geq 0, \lambda = 0 \\ -((-Y+1)^{2-\lambda}-1)/(2-\lambda), & Y < 0, \lambda \neq 2 \\ -\log(-Y+1), & Y < 0, \lambda = 2. \end{cases} \tag{4.1}$$

즉, 농가수입이 양수이면 $\log(Y+1)$변환을, 음수인 경우는 $-\log(-Y+1)$을 사용하게 되므로 모든 자료를 양수로 바꾸는 수고를 덜 수 있다.

4.2. 공간 상관관계

격자 자료 (Lattice data)에서 공간 상관관계를 살펴보기 위해서는 공간 상관관계를 모형에 적용할 이웃이 정의되어야 한다. 본 논문에서는 3장에서 설명하였듯이 경계를 공유하는 조사구를 이웃으로 정의하였다. 먼저 공간분석을 위하여 조사구별, 월별 자료 즉, 각 가구의 지역별 30개월 자료 (2000년 1월에서 2002년 6월)를 평균하여 조사구별 평균 자료로 만들었다. 이 경우도 시계열 자료분석 결과와 비교하기 위해 30개월 모두 결측이 없는 가구만을 분석에 사용하였다. 또한 각 가구는 그 특징에 따라 영농형태가 다르기 때문에 3개의 영농형태로 층화 (논벼, 2종겸업, 기타)를 한 다음 각 층별로 공간 상관관계를 살펴보았으며 이를 위하여 조사구별, 층별 (영농형태별) 평균을 구하였다. 그러나 이 경우 많지 않은 가구 (10가구)를 다시 3개의 층으로 나누기 때문에 각 층의 각 조사구에 속하는 가구가 없거나 한 가구만 존재 하는 경우가 발생하게 되는데 이 경우는 영농형태에 상관없이 전체 자료의 각 조사구 평균값을 사용하였다. 여기서 영농형태는 논벼, 과수, 채소, 특작, 화훼, 전작, 축산, 기타, 2종겸업 등 9개 층이나 강원지역 농가의 경우 논벼가구와 2종겸업가구를 제외한 나머지 가구들은 각 층에 속하는 자료가 너무 적어 하나의 층으로 병합하여 "기타층"으로 두어, 영농형태를 3개 층으로 나누었다. 여기서 농가경제 결측 시 보조변수로 사용되는 또 하나의 변수인 경지규모는 이진희 등 (2006)에서와 같은 이유로 제외하였다. 공간상관관계의 존재 유무는 Moran's I값을 이용하였으며 공간상관을 위한 귀무가설과 대립가설은 H_0 : "공간상관관계가 없다" 대 H_1 : "공간상관관계가 있다"이다. 분석에 사용된 모듈은 S-plus의 SpatialStat이며 그 결과는 표 4.1과 같다.

표 4.1: 전체와 영농형태에 따른 공간상관관계 검정결과

영농형태	Moran's I	표준오차	p-값
논벼층	0.3835	0.1100	<.0001
2종겸업층	0.3316	0.1136	0.0015
기타층	0.0357	0.1514	0.5637
전체	0.2219	0.0713	0.0181

표 4.1의 결과를 살펴보면 영농형태를 구분하지 않은 전체 자료의 공간상관 관계가 존재함을 알 수 있으며, 영농형태에 따라 살펴보면 논벼와 2종겸업층은 공간상관이 존재하나 기타

층은 공간상관이 존재하지 않는것으로 나타났다. 표 4.2에서는 결측대체가 필요한 2002년 8월과 9월 자료에 대하여 공간상관이 있는지를 살펴보았다. 또한 8월과 9월을 대신하여 효율성 비교에 사용될 2002년 7월 자료도 함께 살펴보았다.

표 4.2: 각 월에 따른 공간 상관관계 검정결과

월	Moran's I	표준오차	p-값
7	0.3933	0.1010	<.0001
8	0.0567	0.1102	0.3080
9	0.3786	0.1017	<.0001

표 4.2의 결과를 살펴보면 결측이 있는 8월과 9월 자료 중 8월 자료는 공간상관이 없어 공간 결측대체 방법을 이용하기 어려우나 9월 자료는 공간상관이 존재하여 공간모형을 이용한 결측대체를 할 수 있음을 알 수 있다. 또한 효율성 비교를 위한 검정집합으로 사용되는 7월 자료도 9월과 비슷한 공간상관을 보이고 있어 공간대체방법을 위한 검정집합으로 타당함을 확인할 수 있다. 표 4.3은 결측월인 8월과 9월 그리고 효율성 비교에 사용될 7월에 대한 영농형태별 공간 상관관계 검정결과이다. 전체 자료에 대한 영농형태별 공간상관 결과는 표 4.1에서 이미 살펴보았다. 이상의 결과를 종합해 보면 공간상관 관계가 있다고 판

표 4.3: 각 월에 따른 영농형태별 공간 상관관계 검정결과

월	영농형태	Moran's I	표준오차	p-값
7	논벼층	0.4594	0.0755	<.0001
	기타층	0.1011	0.1068	0.1913
	2종겸업층	0.3062	0.1035	<.0001
8	논벼층	0.0920	0.1055	0.2160
	기타층	0.0677	0.1054	0.3065
	2종겸업층	0.1591	0.1017	0.0503
9	논벼층	0.4124	0.1156	<.0001
	기타층	0.1068	0.1054	0.0897
	2종겸업층	0.1995	0.0981	0.0125

단되며 따라서 결측대체를 위해 공간모형을 사용하는 것은 타당하다는 결론을 내릴 수 있다. 이에 본 논문에서는 다음의 식 (4.2)을 이용하여 공간대체를 실시하였고 모의실험을 통하여 대체 결과의 효율성을 비교하였다.

$$Z_{ij}^{(k)} = \rho_j^{(k)} T_{ij}^{(k)} + \varepsilon_{ij}. \tag{4.2}$$

여기서 $Z_{ij}^{(k)}$는 농가수입을 나타내고 $Y_{ij}^{(k)}$의 Yeo-Johnson 변환을 한 자료, 그리고 $T_{ij}^{(k)}$는 변환 후 자료를 평균한, 즉 (3.1)식의 $S_{ij}^{(k)}$에 해당되는 공간변수이다.

4.3. 공간시계열 상관관계

다음은 공간시계열 상관관계를 이용한 분석의 타당성을 살펴보기 위하여 36개 조사구 각각에 대한 월별 시계열도를 그려 보았으며 그 결과가 그림 4.1에 있다.

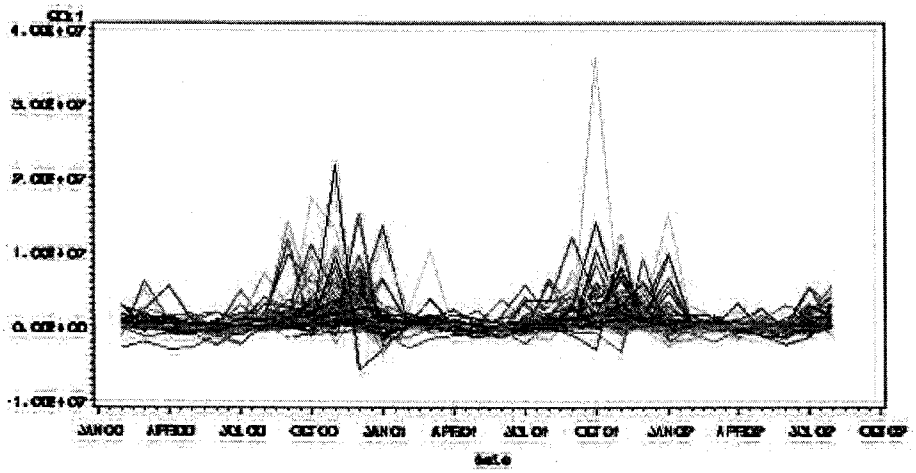

그림 4.1 조사구별 시계열도

그림을 살펴보면 음의 값을 갖는 경우를 발견할 수 있으며 일반적으로 10월경에 큰 값을 갖는 것으로 파악된다. 또한 이상점에 가까운 매우 큰 값도 발견된다. 그러나 본 논문에서는 시계열이 짧고, 또한 간단한 모형을 사용하기 위해, 이상점 분석이나 개입분석, 비정상성(Nonstationary)을 위한 정상차분 (Regular difference) 또는 계절차분 (Seasonal difference)과 승법 계절모형 (Multiplicative Seasonal Model)을 고려하지 않고 일반 ARMA(p,q) 모형을 사용하였다. 적합결과는 강원지역 36개 조사구 각각에 대한 시계열 모형과 강원지역 전체를 하나의 지역으로 생각하고 전체 조사구를 평균한 자료에 대한 시계열 모형을 적합하였다.

표 4.4에 제시한 바와 같이 각 조사구별 모형식별과 함께 전체 조사구 평균에 대한 모형식별도 실시하였다. 모형식별은 최소의 AIC와 SBC를 주는 모형을 기준으로 하였다. 식별결과를 보면 몇몇 조사구를 제외하고는 거의 AR(1,12)모형을 따르고 있으며, 조사구 전체에 대한 모형식별 결과도 AR(1,12) 모형으로 적합 되었다. 아래 표 4.5는 표 4.4 모형을 이용하여 얻은 모수 추정결과이다. 추정 결과는 강원지역 전체를 하나의 지역으로 생각하고 얻은 시계열 모형과 36개 조사구 각각에 대한 모형을 이용하여 얻었다. 그러나 본 논문에서는 각 조사구에 대한 추정결과 모두를 제시하지 않고 강원지역 전체 자료에 대한 추정 결과만을 제시하였다. 또한 영농형태에 따른 조사구 전체에 대한 평균자료를 이용하여 모형적합한 경우 각 층에 따라 최적모형이 다르게 적합 되었으며, 적합결과와 모수추정 결과는 표 4.6에 있다.

영농형태에 따른 시계열 모수 추정결과 기타층과 2종겸업층은 영농형태를 고려하지 않은 자료와 같은 모형인 AR(1,12) 모형이 적합되었으나 논벼층의 경우 AR(2,12)로 전체 모형과 다른 모형이 적합되었다. 각 영농형태에서 조사구 각각에 대한 모수 추정결과도 산출

표 4.4: 전체와 각 조사구에 따른 시계열 모형 적합

조사구번호	시계열 모형	조사구번호	시계열 모형
1	AR(1,12)	19	AR(1,12)
2	AR(12)	20	AR(2,3)
3	AR(1,12)	21	AR(1,12)
4	AR(1,12)	22	AR(1,12)
5	AR(6,12)	23	AR(1,12)
6	AR(1,12)	24	AR(1,12)
7	AR(6,12)	25	AR(1,12)
8	AR(1)	26	AR(1,12)
9	AR(1,12)	27	AR(1,12)
10	AR(1,12)	28	AR(1,12)
11	AR(1,12)	29	AR(1,12)
12	AR(1,12)	30	AR(2,4,12)
13	AR(1,12)	31	AR(1,12)
14	AR(1,12)	32	AR(1,12)
15	AR(1,12)	33	AR(1,12)
16	AR(1,12)	34	AR(1,2,4)
17	AR(1,12)	35	AR(1,12)
18	AR(1)	36	AR(1,12)
전체	AR(1,12)		

표 4.5: 전체평균자료를 이용한 시계열 모수 추정결과

모수	추정량	표준오차	p-값
ϕ_1	0.4580	0.1017	<.0001
ϕ_{12}	0.9516	0.0135	<.0001

하였으나 본 논문에는 제시하지 않았다. 물론 각 영농형태에 따른 조사구별 모형은 각 층에 따라 모든 조사구가 같은 모형을 갖는 것은 아니다. 먼저 논벼층의 경우 주로 AR(2,12) 모형이 적합되나 AR(1.12) 모형도 여럿 적합 되었으며 그 밖의 모형도 적합되었다. 2종 겸업층의 경우 많은 조사구가 AR(1,12) 모형이 적합되었고 AR(1,12)가 아닌 모형도 적합되었다. 이러한 이유로 본 논문에서는 각 조사구에 다음과 같은 식 (4.3)과 식 (4.4)을 적합모형으로 하였다. 먼저 표 4.4에서와 같이 영농형태에 관계없이 조사구별 시계열 모형의 거의 대부분에서 AR(1,12) 모형을 주었고, 또한 강원지역을 하나의 지역으로 생각하여 얻은 전체 조사구 평균자료에 대한 시계열 분석결과도 AR(1,12) 모형으로 식별되었다. 그러므로 시계열 상관모형의 타당성을 살펴보기 위하여 본 논문에서는 계산의 편의상 모든 조사구에 대하여 AR(1, 12) 모형을 갖는 STAR 모형을 적용하였으며 모형식은 다음 식 (4.3) 모

표 4.6: 각 영농형태에 따른 시계열 모형과 모수 추정결과

영농형태	시계열 적합모형	모수	추정량	표준오차	p-값
논벼층	AR(2,12)	ϕ_2	0.6120	0.0755	<.0001
		ϕ_{12}	0.9794	0.0075	<.0001
기타층	AR(1,12)	ϕ_1	0.6368	0.1033	<.0001
		ϕ_{12}	0.9104	0.0447	<.0001
2종겸업층	AR(1,12)	ϕ_1	0.5458	0.1074	<.0001
		ϕ_{12}	0.9189	0.0376	<.0001

형과 같다. 또한 영농형태를 고려한 경우 기타층과 2종겸업층은 AR(1,12) 모형으로 적합하였으며, 논벼층의 경우 (4.3)식과 같이 AR(2,12)로 모형을 적합하였다.

$$Z_{ijt}^{(k)} = \rho_j^{(k)} T_{ijt}^{(k)} + \phi_1^{(k)} Z_{ijt-1}^{(k)} + \phi_2^{(k)} Z_{ijt-12}^{(k)} + \varepsilon_{ijt}, \tag{4.3}$$

$$Z_{ijt}^{(k)} = \rho_j^{(k)} T_{ijt}^{(k)} + \phi_1^{(k)} Z_{ijt-2}^{(k)} + \phi_2^{(k)} Z_{ijt-12}^{(k)} + \varepsilon_{ijt}. \tag{4.4}$$

여기서 $Z_{ijt}^{(k)}$ 와 $T_{ijt}^{(k)}$에 관한 내용은 (4.2)식의 내용을 살펴보기 바란다. 특히 기타층의 경우 표 4.2의 결과에 제시된 바와 같이 공간상관이 존재하지 않으므로 비교에 의미가 없어 제외하였다.

5. 자료분석 및 모의실험

표본조사에서 결측자료가 발생하여 결측대체를 하려 할 때 충분한 설명변수가 존재하지는 않지만 자료들 간에 공간상관 관계와 시계열상관 관계가 존재한다면 이를 결측대체에 사용하는 것은 타당하다. 본 논문에서 사용된 강원지역 자료는 특정 조사구에 속하는 표본가구의 농가수입이 결측인 경우로 4.2절과 4.3절에서 확인한 바와 같이 각 조사구들 사이에 공간상관 관계와 시계열상관 관계가 있는 것으로 판단되었다. 이 절에서는 공간모형과 공간시계열 모형을 사용하여 결측자료를 대체한 후 이들의 우수성을 비교하였다.

모의 실험에 앞서 결측대체값을 얻기 위한 방법을 간단히 설명하면 다음과 같다. 첫째 농가경제에 영향을 미치는 변수를 고려한다 (본 논문에서는 영농형태와 농지크기 변수중 자료의 수와 공간상관등을 고려하여 영농형태를 층화변수로 선택함). 둘째 전체자료, 층별 자료 각각에 대하여 공간 상관이 있는지를 확인한다. 여기서 공간 상관이 존재하면 공간 모형을 이용할 수 있고 존재하지 않으면 이용할 수 없다.

5.1. 모의실험

모의실험에 앞서 4.3절에서 설명한 Yeo-Johnson 변환을 실시하였다. 전술하였듯이 결측자료는 8월과 9월의 농가소득 자료이나 7월 자료를 효율성 비교를 위해 선택하였으며 7월

자료의 경우 공간상관이 9월과 비슷하여 공간대체 방법의 효율성 비교를 위해 적절할 것으로 판단된다. 또한 효율성 비교를 위한 모의실험에서는 8월과 9월의 결측자료 수가 전체 자료의 10%를 전 후 하고 있어 이와 비슷하게 10%의 결측자료를 생성하였다. 결측 생성방법은 36개 조사구에서 각각 1개씩 총 36개의 가구표본을 SRS (simple random sample) 방법으로 뽑아, 뽑힌 자료를 결측 값으로 사용하였다. 영농형태를 고려하지 않은 자료분석은 2000년 1월부터 2002년 6월까지의 자료를 이용하여 모형식별을 한 후 이 모형에서 얻어진 추정값을 가지고 2002년 7월 자료를 이용하여 대체값을 얻었으며, 이때 이용된 식은 (4.3)식이다. 또한 영농형태를 고려한 자료분석은 영농형태를 고려하지 않은 방법과 동일하나 각 층에 따라 모형이 다르게 적합 되었으므로 2종겸업층은 (4.3)식을 이용하였으며, 논벼층의 경우 (4.4)식을 이용하였다. 추정값을 얻는데 있어 이진희 등 (2006)은 GLM을 이용하였으나, 본 논문에서는 GLM 추정값과 회귀분석을 이용한 추정값이 별 차이를 주지 않아 사용이 간편한 회귀분석을 이용하여 추정하였다. 각 방법으로 얻어진 추정치는 재변환, 즉 $Z_i = \log(Y_i + 1)$이면 $Y_i = \exp(Z_i) - 1$로 $Z_i = -\log(-Y_i + 1)$이면 $Y_i = 1 - \exp(-Z_i)$을 이용하여 얻은 추정값을 결측값 각각에 대체하였다. 효율성 비교를 위하여 공간모형과 공간시계열모형을 이용하여 얻은 대체값과 실제값의 MSE (mean square error)와 MAE (mean absolute error)를 계산하였으며, 반복은 5,000번을 실시한 후 5,000번에 대한 평균을 구하였다. 시계열 모형적합을 위해서는 SAS/ETS를 이용하였으며, 각 층에서의 반복적인 표본추출과 추정값에 대한 계산은 각각 SAS/STAT과 SAS/MACRO을 사용하였다. 또한 2가지 결측대체 방법의 효율성 비교를 위하여 사용된 공식인 MSE와 MAE는 (5.1)식과 (5.2)식에 나타내었다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{36}(Y_i^{(j)} - \hat{Y}_i^{(j)})^2, \tag{5.1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{36}|Y_i^{(j)} - \hat{Y}_i^{(j)}|. \tag{5.2}$$

여기서 i는 결측 되었다고 가정된 가구를 나타내며 $n = 5,000$은 반복수를 나타낸다. 또한 $Y_i^{(j)}$는 j번째 조사구 내에서 i번째 가구의 실제 수입을 나타내고 $\hat{Y}_i^{(j)}$는 j번째 조사구내에서 i번째 가구가 결측 되었다는 가정하에서 SAR 모형과 STAR 모형을 이용하여 얻은 추정값이다.

5.2. 모의실험 결과

결측대체 방법의 효율성 비교를 위하여 두 가지 방법으로 얻어진 추정값을 결측값 대신 대체한 후 얻은 MSE와 MAE 결과를 표 5.1과 표 5.2에 나타내었다. 편의상 공간대체에 대한 MSE를 SMSE (Spatial MSE), 공간시계열대체에 대한 MSE를 STMSE (Space Time MSE)라 하였으며 같은 방법을 MAE에도 적용하였다. 또한 비교를 쉽게 하기 위하여 $RMSE = STMSE/SMSE$, $RMAE = STMAE/SMAE$도 함께 계산하였다. 즉 $RMSE$ 또는 $RMAE$가 1보다 적은 경우는 공간시계열 모형을 이용한 경우, 그리고 1보다 큰 경우

에는 공간모형을 이용한 경우가 효율성이 더 좋음을 나타낸다. 표 5.1과 표 5.2는 영농형태에 상관없이 조사구 각각에 대하여 얻어진 결과로 3개 조사구에서만 공간모형이 더 좋게 나오고 있으며 이 경우에도 $RMSE$ 또는 $RMAE$ 모두 미미한 차이를 보이고 있다. 즉 전체 자료분석과 조사구 각각에 대한 비교 결과 모두 공간시계열 모형을 이용한 분석 결과가 더 좋은 효율을 주고 있다. 표 5.3부터 표 5.6은 영농형태에 따른 각 조사구별 비교 결과로 표 5.3과 표 5.4는 논벼층에 대한 결과를, 표 5.5와 표 5.6은 2종겸업층에 대한 결과를 나타낸다. 기타층에 대한 결과는 표 4.1에서 확인한 바와 같이 기타층은 공간상관이 존재하지 않아 공간상관이 존재하는 영농형태인 논벼층과 2종겸업층에 대한 결과만 비교하였다.

표 5.3에서 표 5.6의 영농형태를 고려한 결측대체 결과 또한 영농형태를 고려하지 않은 전체 결과와 비슷한 결과를 준다. 단지 영농형태에 따른 대체의 경우 시계열상관의 효과가 영농형태를 고려하지 않은 자료에 비하여 더 떨어지는 경향이 있다. 이는 공간상관은 없고 시계열 상관은 있는 기타층을 분석에서 제외하였기 때문인 것으로 판단되며 최적의 시계열 모형을 사용한다면 공간시계열 모형을 이용한 대체 결과가 현재 결과보다 더 많이 좋아질 것으로 예상된다.

표 5.1: 결측대체 후 MSE (전체)

조사구	SMSE	STMSE	RMSE	조사구	SMSE	STMSE	RMSE
1	1.355E+12	1.353E+12	0.998	19	1.765E+12	1.417E+12	0.803
2	1.119E+12	1.119E+12	1.000	20	1.558E+12	1.588E+12	1.019
3	2.495E+11	2.373E+11	0.951	21	2.545E+12	2.497E+12	0.981
4	1.217E+10	1.197E+10	0.983	22	2.714E+11	2.661E+11	0.980
5	9.710E+12	9.708E+12	1.000	23	2.420E+13	2.407E+13	0.995
6	8.955E+11	8.917E+11	0.996	24	2.309E+12	1.992E+12	0.863
7	2.569E+12	2.569E+12	1.000	25	1.050E+13	1.004E+13	0.956
8	9.802E+11	9.801E+11	1.000	26	1.378E+13	1.362E+13	0.988
9	2.712E+11	2.680E+11	0.988	27	1.642E+12	1.418E+12	0.863
10	1.148E+13	1.140E+13	0.993	28	1.487E+12	1.433E+12	0.964
11	7.530E+12	6.388E+12	0.848	29	1.399E+12	1.375E+12	0.983
12	1.225E+12	9.783E+11	0.799	30	8.635E+11	8.715E+11	1.009
13	6.462E+11	6.233E+11	0.965	31	2.309E+12	2.217E+12	0.960
14	3.628E+12	3.515E+12	0.969	32	2.247E+12	2.149E+12	0.956
15	7.540E+10	7.522E+10	0.998	33	2.518E+10	2.451E+10	0.973
16	1.139E+13	1.113E+13	0.977	34	2.376E+12	2.406E+12	1.013
17	1.673E+12	1.672E+12	0.999	35	1.687E+12	1.652E+12	0.979
18	2.413E+13	2.412E+13	1.000	36	3.159E+10	2.906E+10	0.920
전체	4.175E+12	4.028E+12	0.965				

표 5.2: 결측대체 후 MAE (전체)

조사구	SMAE	STMAE	RMAE	조사구	SMAE	STMAE	RMAE
1	614301.7	611003.3	0.995	19	1106185.1	913608.0	0.826
2	747062.6	747054.0	1.000	20	904589.9	923523.7	1.021
3	346985.6	331667.9	0.956	21	1041081.7	1020081.5	0.980
4	94053.0	92889.8	0.988	22	315234.9	307520.9	0.976
5	1807798.8	1807106.6	1.000	23	1869890.0	1852981.7	0.991
6	606232.7	606106.6	1.000	24	940867.4	820101.5	0.872
7	1200704.6	1200652.5	1.000	25	1479736.5	1386595.6	0.937
8	709970.1	709935.7	1.000	26	1806158.8	1779414.7	0.985
9	359445.2	356110.0	0.991	27	1130323.4	984888.8	0.871
10	2426102.5	2404837.4	0.991	28	729336.3	712390.0	0.977
11	1946011.4	1700150.0	0.874	29	742820.0	732415.8	0.986
12	996385.8	824579.7	0.828	30	426336.2	436337.3	1.023
13	544004.6	524307.2	0.964	31	1112601.4	1074475.2	0.966
14	1048808.0	998570.8	0.952	32	1164891.7	1121256.1	0.963
15	144899.3	142767.2	0.985	33	128756.0	126671.8	0.984
16	2191281.1	2145847.1	0.979	34	883734.8	898668.3	1.017
17	1954703.0	1954333.7	1.000	35	819722.8	805488.6	0.983
18	2545219.9	2545125.3	1.000	36	151562.3	144039.6	0.950
전체	1096481.0	1063586.6	0.970				

표 5.3: 결측대체 후 MSE (논벼층)

조사구	SMSE	STMSE	RMSE	조사구	SMSE	STMSE	RMSE
1	8.280E+11	8.193E+11	0.989	19	1.010E+11	9.999E+10	0.990
2	5.123E+11	5.107E+11	0.997	20	5.196E+11	5.196E+11	1.000
3	4.403E+11	4.282E+11	0.973	21	1.650E+12	1.598E+12	0.968
4	5.595E+10	5.594E+10	1.000	22	9.047E+10	8.456E+10	0.935
5	3.547E+11	3.597E+11	1.014	23	1.998E+11	1.959E+11	0.980
6	6.128E+11	6.096E+11	0.995	24	1.869E+12	1.812E+12	0.969
7	1.062E+11	9.996E+10	0.941	25	5.096E+10	5.124E+10	1.006
8	6.256E+09	6.265E+09	1.001	26	1.573E+12	1.613E+12	1.025
9	9.512E+11	9.505E+11	0.999	27	7.532E+11	7.513E+11	0.998
10	1.470E+12	1.398E+12	0.951	28	1.880E+11	1.803E+11	0.959
11	4.199E+12	3.765E+12	0.897	29	5.409E+10	5.507E+10	1.018
12	8.576E+11	8.575E+11	1.000	30	2.825E+11	2.700E+11	0.956
13	5.387E+11	4.763E+11	0.884	31	3.115E+11	3.051E+11	0.980
14	1.144E+12	1.092E+12	0.955	32	1.041E+11	1.035E+11	0.994
15	8.859E+10	7.980E+10	0.901	33	1.266E+11	1.364E+11	1.077
16	3.072E+11	2.719E+11	0.885	34	7.270E+10	6.763E+10	0.930
17	8.266E+11	8.216E+11	0.994	35	3.984E+10	3.663E+10	0.919
18	1.706E+09	1.706E+09	1.000	36	5.398E+09	5.272E+09	0.977
전체	5.732E+11	5.600E+11	0.977				

표 5.4: 결측대체 후 MAE (논벼층)

조사구	SMAE	STMAE	RMAE	조사구	SMAE	STMAE	RMAE
1	752903.2	746470.5	0.991	19	314524.3	312960.3	0.995
2	420614.8	419591.9	0.998	20	529511.7	529507.2	1.000
3	364329.7	352903.0	0.969	21	1236333.2	1201205.9	0.972
4	76146.9	76139.0	1.000	22	221939.4	208252.4	0.938
5	224988.5	227968.3	1.013	23	319952.8	304417.9	0.951
6	857619.8	856434.2	0.999	24	3020081.7	2953165.4	0.978
7	993626.6	974638.6	0.981	25	224750.3	225348.6	1.003
8	74464.7	74756.3	1.004	26	2299129.8	2408996.5	1.048
9	1135320.7	1134564.9	0.999	27	948138.0	946065.6	0.998
10	1721858.9	1647279.8	0.957	28	398882.1	400727.9	1.005
11	2226591.1	2105237.8	0.945	29	163448.5	173345.4	1.061
12	869169.1	869160.5	1.000	30	332410.9	318904.6	0.959
13	223164.6	198178.1	0.888	31	651962.9	639100.6	0.980
14	1007916.5	972380.6	0.965	32	373542.4	370465.2	0.992
15	227183.7	207098.1	0.912	33	519357.7	547272.8	1.054
16	514147.7	457318.3	0.889	34	258533.9	249436.7	0.965
17	1179691.6	1159685.4	0.983	35	199386.4	184928.8	0.927
18	90158.9	89962.2	0.998	36	47963.7	46943.2	0.979
전체	697253.5	684702.9	0.982				

표 5.5: 결측대체 후 MSE (2종겸업층)

조사구	SMSE	STMSE	RMSE	조사구	SMSE	STMSE	RMSE
1	1.220E+11	1.171E+11	0.959	19	5.379E+10	5.226E+10	0.972
2	2.519E+11	2.363E+11	0.938	20	1.606E+12	1.550E+12	0.965
3	5.545E+10	5.498E+10	0.991	21	9.983E+10	9.679E+10	0.970
4	1.223E+10	1.080E+10	0.883	22	4.033E+10	3.885E+10	0.963
5	1.517E+12	1.480E+12	0.975	23	9.502E+11	8.983E+11	0.945
6	8.975E+11	8.917E+11	0.994	24	3.731E+10	3.698E+10	0.991
7	5.999E+10	5.527E+10	0.921	25	8.365E+11	8.171E+11	0.977
8	9.295E+11	9.263E+11	0.997	26	5.573E+10	5.294E+10	0.950
9	7.662E+12	7.692E+12	1.004	27	1.710E+12	1.708E+12	0.999
10	8.154E+11	8.164E+11	1.001	28	2.215E+11	2.153E+11	0.972
11	8.198E+10	7.911E+10	0.965	29	1.746E+11	1.746E+11	1.000
12	3.050E+10	3.049E+10	1.000	30	5.938E+11	5.976E+11	1.006
13	5.195E+11	4.984E+11	0.959	31	1.169E+12	9.965E+11	0.852
14	1.133E+12	1.069E+12	0.943	32	3.146E+10	3.139E+10	0.998
15	1.817E+11	1.728E+11	0.951	33	5.187E+11	5.213E+11	1.005
16	6.098E+11	6.078E+11	0.997	34	3.780E+11	3.702E+11	0.979
17	3.343E+11	3.342E+11	1.000	35	2.649E+10	2.576E+10	0.972
18	2.005E+11	1.912E+11	0.954	36	7.217E+09	6.716E+09	0.931
전체	6.755E+11	6.559E+11	0.971				

표 5.6: 결측대체 후 MAE (2종겸업층)

조사구	SMAE	STMAE	RMAE	조사구	SMAE	STMAE	RMAE
1	324245.6	311904.1	0.962	19	173804.5	169705.1	0.976
2	437684.4	411940.8	0.941	20	1893216.8	1834281.3	0.969
3	239866.3	239497.0	0.998	21	182688.1	176342.7	0.965
4	91359.5	87150.0	0.954	22	145853.9	141079.1	0.967
5	1002624.1	979178.5	0.977	23	697156.1	668173.5	0.958
6	813232.7	801006.6	0.985	24	180125.0	178864.8	0.993
7	214775.2	198915.0	0.926	25	686109.0	668798.0	0.975
8	2073828.0	2053376.6	0.990	26	232415.5	221362.4	0.952
9	2759787.4	2768537.3	1.003	27	2807098.8	2807106.6	1.000
10	901761.0	901765.5	1.000	28	973781.2	953207.8	0.979
11	182066.9	176753.9	0.971	29	715311.0	715307.7	1.000
12	155013.7	155012.7	1.000	30	951408.5	959253.8	1.008
13	472102.1	462118.1	0.979	31	1439498.1	1341231.0	0.932
14	1265153.1	1204955.2	0.952	32	431376.9	431351.9	1.000
15	127330.0	122293.2	0.960	33	822513.9	828985.2	1.008
16	169465.8	167467.8	0.988	34	696791.0	681321.3	0.978
17	245275.1	245257.3	1.000	35	360030.5	349867.9	0.972
18	194789.3	181333.7	0.931	36	164105.6	153403.9	0.935
전체	701283.3	685153.8	0.977				

6. 결론

조사된 자료에 무응답이 존재할 경우 이를 해결하기 위해 사용되는 대체법에 대하여 많은 연구가 진행되어 왔으며 (Little과 Rubin, 1987; Rao와 Shao, 1992), 보조정보를 이용하여 대체를 할 경우 분산과 편향을 줄일 수 있는 것으로 알려져 있다. 그러나 항목무응답이 발생하였을 경우 결측대체를 위해 필요한 보조정보가 한정되어 있어 결측대체에 어려움이 따른다 (Son 등, 2001). 이때 이진희 등 (2006)에서 제시한 공간상관 관계나 본 논문에서 살펴본 공간시계열 상관의 이용은 자료 분석에 큰 도움이 될 것이다. 그동안 공간정보의 이용은 자료가 얻어진 위치에 대한 정보의 부족 등으로 인하여 상대적으로 널리 활용되지 못하였으나 최근 공간정보를 이용한 자료분석이 꾸준히 진행되고 있다 (이진희와 신기일, 2004). 본 논문에서는 2002년 8월과 9월의 강원지역 자료가 수해로 인해 많은 결측값을 가지게 되었으며 이에 대한 결측대체 방법으로 고려된 공간모형과 공간시계열모형을 이용한 결측대체 방법의 효율성을 살펴보았다. 물론 효율성 비교를 위해서 결측이 발생한 해당 월을 직접 사용하지 못하고 결측이 발생한 해당 월과 공간상관이 비슷한 다른 월을 선택하여 공간모형과 공간시계열모형을 이용한 결측대체 방법의 효율성을 비교하였지만 그 결과의 타당성에는 무리가 없다고 판단된다. 본 논문에서는 농가수입 자료에 대하여 어느 특정 월에 대한 결측이 생겼을 경우 공간상관과 시계열적 상관을 동시에 이용한 결측대체방법에 대한 연구로 공간상관과 시계열 상관이 동시에 존재할 경우 공간시계열 모형을 이용한 결측대체 방법이 효율적임을 확인하였다. 물론 이 경우 공간시계열 상관을 결측대체에 이용

하기 위해서는 공간상관과 시계열 상관이 모두 존재하여야 함을 전재조건으로 한다. 결론적으로 결측대체를 위한 보조정보가 충분하지 않고 공간상관과 시계열상관이 함께 존재할 때 공간시계열 모형을 이용하면 효과적인 결측대체를 할 수 있으리라 판단된다.

참고문헌

김규성, 이기재, 김진 (2005). 농어가경제조사에서 가중핫덱 무응답 대체법의 연구, <응용통계연구>, **18**, 311–328.

이진희, 김진, 이기재 (2006). 표본조사에서 공간변수(spatial variable)를 이용한 결측대체(missing imputation)의 효율성비교, <응용통계연구>, **19**, 57–67.

이진희, 신기일 (2004). 공간통계분석에서 이상점 수정방법의 효율성 비교, <응용통계연구>, **17**, 327–336.

통계청 (2003). 농가경제조사, 농산물 생산비조사 지침서 (2003).

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

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

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, **79**, 811-822.

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.

Yeo, I. and Johnson, R. A. (2000). A new family of power transformation to improve normality or symmetry, *Biometrika*, **87**, 954–959.

[2007년 5월 접수, 2007년 8월 채택]

Imputation Method using the Space-Time Model in Sample Survey*

Jin-Hee Lee[1] Key-Il Shin[2]

ABSTRACT

It is a common practice to use the auxiliary variables to impute missing values from item nonresponse in surveys. Sometimes there are few auxiliary variables for missing value imputation, but if spatial and time autocorrelations exist, we should use these correlations for better results. Recently, Lee *et al.* (2006) showed that spatial autocorrelation could be efficiently used for missing value imputation when spatial autocorrelation existed, using the data from the farm household economy data in Gangwon-do, 2002. In this paper, we present an evaluation of spatial and space-time nonresponse imputation methods when there exist spatial and time autocorrelations using the monthly data during 2000–2002 from the same data previously used by Lee *et al.* (2006). We show that space-time imputation method is more efficient than the other through the numerical simulations.

Keywords: Missing data, STAR model, neighborhood information, space-time imputation.

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

1) Senior researcher, Division of AIDS, Immunology and Pathology Centers, National Institute of Health, Korea
E-mail: jhlee@cdc.go.kr

2) (Corresponding author) Professor, Department of Statistics, Hankuk University of Foreign Studies, Korea
E-mail: keyshin@hufs.ac.kr