응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 95–107*

농촌생활지표조사에서 무응답 대체 : 사례

조영숙[1] 천영민[2] 황대용[3]

요 약

농촌생활지표조사는 2000년부터 농촌자원개발연구소에서 매년 실시하는 조사로서 통계청 승인통계이다. 본 연구에서는 2005년 농촌생활지표조사에 사용된 원자료를 이용하였다. 원자료에 대한 에디팅 과정을 거친 후 무응답이 포함된 개체를 제거하여 얻어진 1,582 가구를 대상으로 하였으며 총 146문항 중에서 최종 선택되어진 15문항을 중심으로 무응답 대체를 실시하였다. 실험에 사용된 대체법과 각 대체법의 효율성은 자료의 종류에 따라 다르게 적용되었다. 먼저 연속형 자료에 대해서는 평균대체, 회귀대체, 수정된 그레이 기반 k-NN 대체(DU, DW, WU, WW) 방법을 사용하여 무응답을 대체하고 RMSE를 이용하여 실험결과를 비교하였으며, 범주형 자료에 대해서는 최빈값 이용, 확률 대체, 조건부 최빈값 이용, 조건부 확률 대체, 단순 임의 핫덱 대체 방법을 사용하여 무응답을 대체하고 정확도(Accuracy)를 이용하여 실험결과를 비교하였다. 실험결과에 의하면 연속형 자료에 대해서는 회귀대체 또는 그레이 기반 k-NN 대체가 적절하고, 범주형 자료에 대해서는 핫덱 대체가 가장 적절한 것으로 나타났다.

주요용어: 대체법, 무응답, 정확도(Accuracy), RMSE(Root Mean Square Error).

1. 서론

통계청을 비롯한 정부부처 및 관련기관에서는 자료의 효율적 관리 및 공표의 표준화 등의 다양한 목적을 위해 관련 사회지표를 개발하여 사용하고 있다. 선진국에서 1960년대부터 본격적으로 사용된 사회지표는 1970년대에 국내에 소개되기 시작하였고, 최근에는 전문분야의 필요와 목적에 맞춰 변화되고 개발되어 사용되고 있다. 박대식과 이영대 (1997)는 일반 사회지표들은 전국단위나 일반 국민을 대상으로 하기 때문에 농촌생활의 실태를 파악하고 분석하여 농촌생활에 관련된 문제 해결을 위해 사용되기에는 문제가 있다고 보았다. 따라서 농촌생활을 위한 지표개발의 필요성에 따라 새로운 지표 체계를 개발하여 정책수립의 기초자료로 활용할 필요성이 대두되었다. 우리나라의 근대화 과정에서 산업화에 치우친 개발정책으로 도시 인구의 폭발적 증가와 더불어 도시 계획 등의 관련 정책개발에 치우친 경향을 보였다. 지난 수십여 년 간 상대적으로 외면 받아 온 농촌에 대한 시각이 최근

1) (441-853) 경기도 수원시 권선구 서둔동 88-2, 농촌진흥청 농촌자원개발연구소, 연구사.
E-mail: joys@rda.go.kr
2) (150-093) 교신저자. 서울시 영등포구 문래동, 한국고용정보원 고용조사분석센터, 부연구위원.
E-mail: zzari90@work.go.kr
3) (441-853) 경기도 수원시 권선구 서둔동 88-2, 농촌진흥청 농촌자원개발연구소, 연구사.
E-mail: hdy@rda.go.kr

재조명되고 있는 시점에서 농촌 생활의 만족도와 삶의 질 향상의 문제는 중요한 관심사이다. 이에 따라 농촌자원개발연구소에서는 1994년에 처음으로 농촌생활지표를 발간하였다. 하지만 이 지표집은 다른 기관에서 발표한 자료 중에서 농촌생활과 관련한 자료를 발췌하여 정리한 후 지표로 제시한 것이다. 실질적으로 농촌생활과 관련된 독립적인 지표 생산 및 발굴은 2000년 조사에서 처음으로 이루어졌다고 볼 수 있다. 조영숙 등 (2004)이 지적했듯이 농촌생활지표는 농업·농촌의 여건 변화에 따른 농촌생활의 변화를 제시하는 한편 안정적이고 체계적인 지표생산이 이루어져야 하며, 특히 몇 가지 부문에 제한되기보다는 보다 포괄적이며 종합적인 위상정립이 요청되어진다고 할 수 있다. 또한 조영숙 등은 농촌생활지표에 대해 농촌생활에 관련된 전반적인 사항에 관한 한정되고 대표적인 항목의 계량화된 척도라고 정의하고 있다.

농촌생활지표와 같은 사회지표 뿐만 아니라 각종 조사에서 얻어진 자료에는 상당히 많은 무응답이 존재하며 이에 대한 처리에 대한 많은 연구들이 있었다. 먼저 김규성 (2000)은 사회·경제조사에서 흔히 발생하는 무응답 문제에 대해 무응답 대체방법과 대체효과에 대해 살펴보고 분산의 과소추정을 알아보기 위한 여러 가지 방법을 살펴보았다. 김규성 등 (2005a)은 통계청에서 실시하는 농어가경제조사에서 무응답이 발생한 농어가를 대상으로 가중핫덱 무응답 대체법을 적용하였고 수정된 잭나이프 분산추정법을 사용하여 신뢰도를 나타내었다. 김규성 등 (2005b)은 패널조사에서 무응답으로 발생하는 추출확률을 보정하기 위해 가중치를 부여하는 방법을 설명하였다. 김재광 등 (2004)은 가계조사에서 발생하는 무응답 처리를 위해 가중치 조정방법을 사용하였고 잭나이프를 통해 분산추정을 실시하였다. 김진 (2004)은 경남지역의 농가경제조사 자료에서 발생한 무응답 자료를 대체하기 위해 핫덱대체법, 최근접 대체법, 군집내 랜덤대체법을 사용하였고 수정된 잭나이프분산을 구하여 각 대체법의 효율성을 비교하였다. 김영원과 이주원 (2003)은 2000년 인구주택총조사 자료에서 혼인 문항에 한해 무응답 대체를 실시하고자 하였는데, 이 때 사용된 대체법은 CART에서 얻어지는 최대예측확률을 이용하는 통계적 예측방법과 CART를 대체층 구성에 활용한 순차적 핫덱 대체법이며 대체법의 성능을 평가하기 위해 재조사 방법을 통해 오분류를 비교하였다. 김영원과 조선경 (1996)은 표본조사에서 항목 무응답에 사용가능한 여러 가지 대체법들을 소개하고 모의실험을 통해 대체법들의 성능을 비교하고자 하였다. 도세록과 이관제 (2006)는 국민건강·영양조사에서 무응답이 심각한 22개 문항에 대해 다중대체법을 이용하여 대체를 실시하였다. 박태성과 이승연 (1998)은 범주형 자료에서 발생한 무응답을 대체하기 위해 최대우도추정법을 소개하고 베이지안 추정법을 제안하였으며, 두 가지 방법의 성능을 비교하기 위해 1948년 미국 여론조사기관에서 실시한 미국 대통령 후보의 지지도에 관한 예비조사 자료 (Baker와 Laird, 1988)를 사용하였다. 신형원과 손소영 (2002)은 범주형 자료에 대한 대체방법으로 로지스틱 회귀분석, 연관규칙, 최빈범주법을 사용한 후 각 결측값에 대한 대체값 후보로 선정한 후에 이 대체값들을 신경망의 입력으로 사용하여 결합된 예측을 하는 신경망 융합기법을 제안하였다. 이진희 등 (2006)은 2002년 강원지역 농가경제조사 자료에서 변수들 사이에 공간 상관이 존재하는 경우 공간 SAR 모형을 이용한 대체법인 자료기반 공간대체방법을 이용하여 대체를 실시하였으며, 성능을 비교하기 위해 대체값의 MSE와 MAE를 이용하였다. Park (2002)은 패널조사에서 웨이브 무응

답이 발생한 경우에 단위 무응답의 가중치 조정 방법과 항목 무응답의 대체법의 장단점을 모두 고려한 자료 보정방법을 제안하였다. Kim 등 (2003)은 2000년 인구주택총조사 자료를 이용하여 CART 모형기반 대체법, CART 기반 핫덱 대체법, 통계청에서 개발한 논리적 핫덱 대체법 등 세 가지 방법을 적용하여 오분류율을 비교한 결과, CART 기반 핫덱 대체법이 가장 우수함을 밝히고 있다. 신민웅과 백정용 (2005)은 아웃바운드 캠페인에서 발생하는 항목 무응답을 대체하기 위해 조건부 평균 대체를 활용한 ML 추정과 베이지안 추정, 단순랜덤대체, 근사적 베이즈 붓스트랩 대체 방법을 추정하였다. 네 개의 대체방법을 층화 다중대체 방법으로 다시 추정한 후 평균과 표준편차를 이용하여 결과를 비교하였다. 한편 김주환 (2004)은 단위 무응답을 줄이기 위해 전화조사에서 응답자의 성별 및 나이에 따라 면접원의 성별을 고려해야 한다고 설명하고 있다.

농촌진흥청 농촌자원개발연구소에서 실시하는 농촌생활지표 조사 (황대용 등, 2005)는 농촌생활의 실태를 파악하고 분석하여 농촌생활에 관련된 문제 해결을 위해 2000년부터 본격적인 설문조사를 실시하고 있다. 지난 2005년 조사의 경우에는 농촌으로 분류되는 전국의 읍·면을 대상으로 하여 가구수 크기 비례 추출법(Probability Proportional to Size: PPS)에 따라 1,870가구를 표본으로 추출하여 설문조사를 실시하였다. 설문조사 실시 과정에서 발생하는 무응답을 줄이기 위해 최초 방문 면담 조사후 수차례 재방문 면담조사 및 전화조사를 실시했는데, 이 과정에서 적지 않은 노동과 시간이 투입되고 있으며 문항에 따라 무응답률에 차이는 있으나 최종 무응답률이 약 39.5%에 이르는 문항도 있는 것으로 나타났다. 무응답의 존재는 각 지표간의 총합 불일치의 문제를 나타내며, 이로 인한 추가 분석시에 개체들의 일부가 분석에서 제외되는 문제가 발생되기도 한다. 따라서 본 연구에서는 2005년에 수집한 자료를 중심으로 결측값 대체법을 적용했을 경우에 RMSE 또는 정확도가 각 대체법에 따라 차이가 있는지를 알아보고 자료의 종류에 따라 최적의 대체법을 발견하여 추후 조사에 적용가능한지를 알아보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 무응답의 종류 및 발생형태에 대해 알아보고, 본 연구를 위해 사용된 대체법을 중심으로 무응답 대체법에 대해 설명한다. 3장에서는 실험을 위한 조건 설정 및 실험 방법에 대해 알아보고 실험결과를 살펴본다. 4장에서는 실험결과에 대해 설명하고 본 논문의 성과 및 결론에 대하여 설명한다.

2. 무응답 대체법

조사와 실험과 같은 자료 수집 과정에서 여러 가지 다양한 원인들로 인해 무응답이 발생하게 된다. 조사의 가장 대표적인 설문조사의 경우에는 응답자의 응답 거부, 조사자의 능력 부족, 설문 문항의 잘못된 구성, 실험 대상이나 자료의 누락 등으로 인해 무응답이 발생하게 되고, 실험의 경우에는 측정과정에서 발생하는 측정도구의 오차, 실험자의 능력 차이, 피실험자의 실험에서의 탈락, 피실험자의 실험에 대한 비협조 등으로 인해 무응답이 발생하게 된다.

2.1. 무응답의 종류와 무응답 발생 형태

무응답의 종류는 크게 단위 무응답(unit nonresponse)과 항목 무응답(item nonresponse)으로 나뉜다. 먼저 단위 무응답은 표본으로 추출된 조사 대상에서 응답을 얻지 못하는 경우로써 다른 조사 대상을 찾는 방법, 해당되는 대상을 제거하는 방법, 가중치를 조정하는 방법 등을 이용하게 된다 (조사통계 연구회, 2000). 한편 항목 무응답은 특정한 문항에 대해 조사 대상자가 응답을 하지 않거나 질문과는 무관한 응답을 하는 경우로써 대체법을 이용하여 자료를 대체하게 된다.

한편 무응답이 발생하는 형태는 통계적 특성에 따라 완전임의결측(MCAR), 임의결측(MAR) 그리고 비임의결측(NMAR)으로 분류된다. 완전 임의 결측(MCAR; Missing Completely At Random)이란 어느 문항에 대한 결측이 자신의 문항 뿐만 아니라 다른 문항에 의존하지 않고 완전하게 임의로 정해지는 경우를 말한다. 즉 특정문항이나 항목에서 무응답이 발생하는 것이 아니라 문항과 항목에 관계없이 임의로 발생하는 경우가 해당된다. 임의결측(MAR; Missing At Random)이란 특정 문항의 무응답 발생 형태가 다른 문항에 따라 결정되는 경우이다. NMAR(Not Missing At Random)는 세 가지 무응답 발생 형태 중에서 가장 복잡한 경우로써 특정 문항의 무응답 발생 형태가 자신의 분포에 의해 영향을 받을 뿐만 아니라 다른 문항과 연관되어 있어 영향을 받는 경우를 말한다 (Little과 Rubin, 2002).

2.2. 대체법

조사 자료에 존재하는 무응답은 연구자의 연구에 제약을 주게 된다. 따라서 연구자는 무응답의 존재를 인정하고 분석을 실시하는 방법이 있지만, 자료 분석 과정에서 문항 선택마다 무응답의 포함여부에 대한 결정을 해야 하고 이에 따라 총합 불일치의 문제도 생각해야 한다. 또한 무응답의 존재는 기존에 개발된 학습 알고리즘의 적용을 어렵게 하는데, 이럴 경우 학습 알고리즘 자체를 무응답을 처리할 수 있도록 변경해야 한다. 학습 알고리즘의 변경은 학습 알고리즘에 대한 완벽한 이해가 선행되어야 하기 때문에 상당히 어려운 문제가 될 것이다. 또한 항목 무응답이 발생한 대상을 제거하는 전통적인 방법은 다른 항목에서 얻어진 자료까지 사용할 수 없게 되고 이에 따라 자료에 대한 신뢰도 하락을 초래하게 되는 문제가 발생한다. 따라서 학습 알고리즘의 변경이나 무응답이 포함된 대상을 제거하지 않고 무응답을 어떤 다른 값으로 창조하는 방법을 이용하면 되는데, 우리는 이 방법을 '대체(imputation)'라고 한다.

무응답이 포함된 문항의 자료 종류가 무엇인가에 따라 대체 방법도 달라지게 되는데, 범주형과 연속형에 따라 대체 방법을 구분해 보면 다음과 같다. 먼저 무응답이 포함된 문항이 범주형일 경우에는 최빈값 이용, 중앙값 이용, 핫덱 대체, 콜드덱 대체, 확률 대체, 조건부 최빈값 이용, 조건부 확률 대체, 이득비율(gain ratio) 이용, 로지스틱 회귀 대체, k-NN 대체, 다중 대체 등을 이용할 수 있다. 먼저 최빈값 이용은 무응답이 포함된 문항이 명목형 자료일때, 그 문항의 최빈값을 구하여 무응답에 대한 대체값으로 사용한다. 중앙값 이용은 무응답이 포함된 문항이 순서형 자료일 때, 그 문항의 중앙값을 구하여 무응답에 대한 대체값으로 사용한다. 핫덱 대체는 변형된 형태가 많은데, 무응답이 포함된 문항에서 랜덤하게 선택하여 무응답에 대한 대체값으로 사용한다. 본 연구에서는 비복원 랜덤 핫덱 대체법

을 사용하였는데, 무응답에 대한 대체값을 무작위로 선택하여 이용하는데 한번 사용된 대상은 다시 사용하지 않는 방법이다. 콜드덱 대체는 이번 조사와 유사한 과거 조사의 자료를 사용하여 대체하는 방법이다. 확률 대체는 무응답이 발생한 문항의 값들의 비율에 따라 무응답에 대체값으로 사용하는 것이다. 조건부 최빈값 이용은 결정변수의 값에 따라 최빈값을 구해 대체하는 방법이며 조건부 확률 대체는 결정변수의 값에 따라 확률값을 구해 대체하는 방법이다. 이득비율 이용은 Quinlan (1993)이 제안한 이득 비율을 이용하여 대체하는 것이다. 로지스틱 회귀대체는 무응답이 포함된 문항을 로지스틱 회귀모형의 반응변수로 하여 예측값을 추정하여 대체하는 것이다. k-NN 대체는 유클리드 거리를 유사성 척도로 사용하여 가장 유사한 대상들을 찾아 그 정보를 활용하여 대체하는 것이다. 다중 대체는 무응답을 하나의 값이 아닌 여러 개의 대체후보로 보아 통계분석을 실시하여 최적의 대체값을 찾는 것이다.

무응답이 포함된 문항이 연속형일 경우에는 평균 대체, 코헨 방법, 회귀 대체, k-NN 대체, 다중 대체, EM 알고리즘 이용 등을 사용할 수 있다. 평균 대체는 무응답이 포함된 문항의 평균을 구하여 무응답에 대한 대체값으로 사용하면 된다. 코헨 방법은 평균 대체의 과소분산 추정을 해결하기 위해 두 개의 값으로 대체하는 방법이다. 회귀 대체는 무응답이 포함된 문항을 회귀모형의 반응변수로 하여 예측값을 추정하여 대체하는 것이다. k-NN 대체와 다중 대체는 순서형 자료에서 적용하는 것과 동일한 방법이며, EM 알고리즘 방법은 최대우도함수를 구하여 최대값을 반복 추정하여 무응답을 대체하는 방법이다. 한편 Huang과 Lee (2004)가 제안한 그레이 기반 k-NN 대체법은 유사성 척도를 유클리드 거리가 아닌 그레이 관계 등급을 사용하여 무응답을 대체하는 방법으로 연속형 자료나 순서형 자료에 적용 가능하다. 그레이 관계 등급은 Deng (1982)이 제안한 그레이 시스템 이론에서 개체들 사이의 유사성을 재는 척도로 소개된 개념으로 이 값이 1에 가까우면 가까울수록 개체가 서로 유사함을 나타내는 척도이다. Huang과 Lee (2004)의 대체법은 그레이 관계 등급을 근접이웃의 수를 결정하는데 사용한 반면, Chun 등 (2006)이 제안한 수정된 그레이 기반 k-NN 대체법은 그레이 관계 등급을 가중평균의 가중값으로 사용하여 대체 능력을 향상시켰을 뿐만 아니라 Huang과 Lee (2004)가 사용한 Deng (1989)의 그레이 관계 등급 뿐만 아니라 다른 종류의 그레이 관계 등급인 Wen (2004)의 그레이 관계 등급을 사용하였다. Chun 등 (2006)이 제안한 방법은 다음과 같다. 원자료에 대한 그레이 생성여부를 판단한 후에 그레이 생성이 이루어지지 않았다면 Hsia와 Wu (1998)의 방법을 이용하여 0과 1 사이의 값으로 자료 생성을 실시한다. 자료 생성 후에 한 문항을 기준으로 무응답 자료가 발견되면 무응답을 포함한 개체를 기준으로 무응답이 포함되지 않은 개체들과의 유사성을 계산하기 위해 그레이 관계 등급(GRG: grey relational grade)을 계산한다. 이 때, Deng의 방법과 Wen의 방법을 사용할 수 있다. 계산된 그레이 관계 등급을 기준으로 정렬한 후에 근접이웃의 수에 따라 사용할 개체를 결정한다. 결정된 개체들이 갖고있는 값들을 이용하여 산술평균 또는 가중평균을 구하게 된다. 가중평균의 가중값은 그레이 관계 등급을 이용하게 된다. 마지막으로 계산된 가중평균을 무응답에 대한 대체값으로 사용하면 되는데 전체적인 흐름은 그림 2.1과 같다.

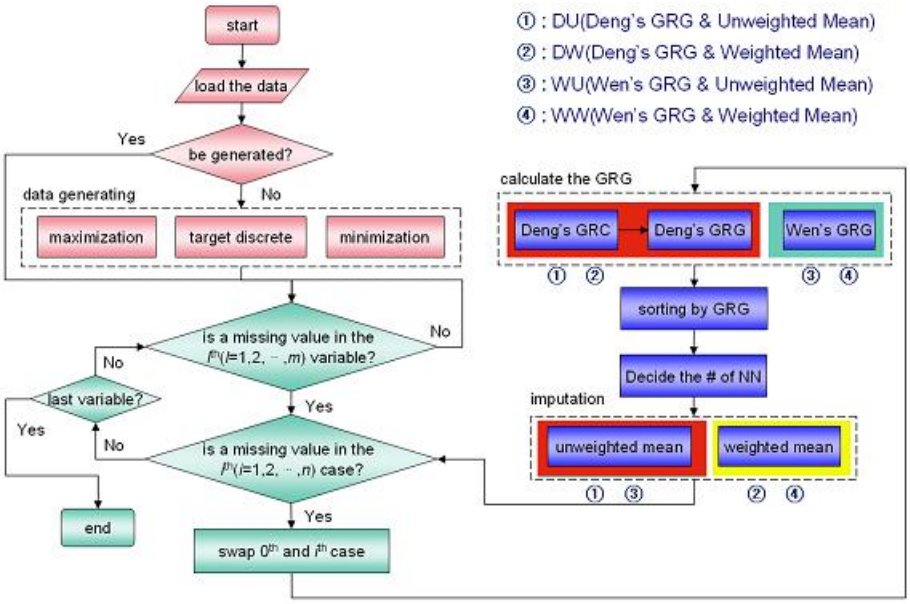

그림 2.1: 수정된 그레이 기반 k-NN 대체법의 절차

3. 모의 실험

3.1. 실험 조건 및 방법

1,870 가구를 대상으로 총 146 문항에 대한 자료에서 일차적으로 실험을 위하여 선택된 문항은 표 3.1과 같이 총 22 문항이다. 1,870 가구는 영농여부에 따라 농가가 1,379 가구, 비농가가 491가구로 구성되어 있다. 따라서 표 3.1과 같이 농가소득은 농가만 응답을 해야 하며 비농가 총가구 소득은 비농가만 응답을 해야 하는데, 소득의 경우에는 다른 조사의 경우에도 무응답률이 가장 많은 문항 중에 하나로써 본 조사에서도 표 3.2와 같이 실제로 가장 많은 무응답을 갖는 것으로 나타났다. 22 문항에 대해 에디팅 과정을 통해 자료 검증 작업을 다시 실시한 후에 연구자들의 토의를 거쳐 최종 15 문항을 실험에 사용하기로 결정하였다. 최종적으로 15 문항에 무응답이 포함된 대상은 자료에서 제외하여 1,582 가구를 실험 대상으로 결정하였다. 총 1,582 가구 중에서 농가는 1,126가구이며, 비농가는 456가구이다. 자료 종류가 연속형인 4개 문항에 대해서는 평균 대체, 회귀 대체, 수정된 그레이 기반 k-NN 대체를 적용하였으며, 다시 범주형으로 변환하여 최빈값 이용, 확률 대체, 조건부 최빈값 이용, 조건부 확률 대체, 비복원 랜덤 핫덱 대체 등을 적용하였다. 이 때, 회귀대체의 경우에는 연속형 문항인 네 문항을 이용하였는데, 무응답이 포함된 문항을 반응변수로하고 무응답이 포함되지 않은 세 문항을 설명변수로 하여 완전모형을 적용하였다. 전체적인 실험 절차를 간단히 정리하면 그림 3.1과 같다.

표 3.1: 1차 선택된 문항

구분	문항명칭	문항	농가	비농가
일반사항	province	도	○	○
	city	시군	○	○
	farm	영농여부	○	○
	rel	가구주와의 관계	○	○
	sex	성별	○	○
	married	결혼여부	○	○
	region	읍면	○	○
개인 및 가구사항	family	가족수	○	○
	period	거주기간	○	○
	age	나이	○	○
	edu	학력	○	○
	job	직업	○	○
	income11	농가소득	○	×
	income12	농업소득	○	×
	income13	농외소득	○	×
	income14	이전소득	○	×
	income2	비농가 총가구소득	×	○
	cost	연간 생활비	○	○
농가의 경우	ratio	농업소득과 농외소득 비율	○	×
	species	영농형태	○	×
	scale1	영농규모	○	×
	scale2	영농규모하위	○	×

표 3.2: 소득관련 무응답률

문항	가구수	무응답 가수수	무응답률
농가소득	1,379	58	4.21
농업 · 농외 · 이전소득	1,379	545	39.52
비농가 총가구소득	491	82	16.70

그림 3.1: 실험 절차

표 3.3: 각 문항 무응답률

문항	family	period	cost	ratio	species	scale1	scale2
무응답수	4	11	94	45	20	12	158
무응답률	0.29	0.80	6.82	3.26	1.45	0.87	11.46

연속형 자료의 대체에서 각 대체법의 성능을 비교하기 위하여 다음과 같은 RMSE를 비교하였다. 실험방법은 leave-one-out-cross-validation을 이용하여 하나의 대상에 무응답을 발생시킨 후 나머지 1,581개의 정보를 이용하여 대체를 실시하여 원 자료와 대체 자료의 RMSE 평균과 표준편차를 구하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2}.$$

한편 범주형 자료의 대체를 하기에 앞서 연속형 자료의 범주화 방법은 다음과 같다. 가족수(family)는 1)1명, 2)2명, 3)3명, 4)4명, 5)5명 이상으로, 교육기간(period)은 1)24개월 이하, 2)24∼48개월, 3)48∼72개월, 4)73개월 이상으로, 나이(age)는 1)30대 이하, 2)40대, 3)50대, 4)60대, 5)70대 이상으로, 연간 생활비(cost)는 1)200만 이하, 2)201∼500만, 3)501∼1,000만, 4)1,001∼2,000만, 5)2,001만 이상으로 범주화하였다. 연속형 자료를 범주화한 문항 네 개를 포함해 전체 15개 문항 중에서 원자료에 무응답이 포함되어 있는 문항들의 무응답률은 표 3.3과 같고 다른 문항들에는 무응답이 포함되어 있지 않았다. 따라서 1,582가구를 대상으로 표 3.3과 같은 문항들에 대해 각각의 문항들의 무응답 비율을 발생시킨 후, 대체를 실시하였다. 실험은 1,000회 반복하여 다음과 같은 정확도의 평균과 표준편차를 비교하였다.

$$Acccuracy = \sum_{i=1}^{n} A_i,$$

단, A_i는 정확히 대체했을 때는 1, 다른 값을 대체했을 때는 0의 값을 갖는다.

3.2. 실험 결과

가족수, 거주기간, 나이, 연간 생활비 등 네 개의 문항은 연속형 자료이므로 평균 대체, 회귀 대체, 그레이 기반 k-NN 대체를 이용하여 어떤 대체 방법이 좋은지를 알아보기 위한 실험을 실시하였다. 표 3.4 ∼ 표 3.7은 각각 가족수, 거주기간, 나이, 연간 생활비 문항에 대한 RMSE를 나타낸 것이다.

표 3.4의 가족수 문항에 대해 평균 대체의 RMSE가 가장 큰 것으로 나타났다. 회귀 대체의 RMSE는 작은 것으로 나타났으며, 그레이 기반 k-NN 대체는 근접 이웃의 수가 증가함에 따라 회귀 대체보다 작은 RMSE를 갖는 것으로 나타났다. 표 3.5 ∼ 표 3.7의 결과 역시 표 3.4와 비슷한데, 평균대체의 RMSE가 가장 크고 회귀대체는 평균대체보다 작은 RMSE를 갖는 것으로 나타났으며 그레이 기반 k-NN 대체는 근접 이웃의 수가 증가함에 따라 회귀대체보다 작은 RMSE를 갖는 것으로 나타났다. 나이 문항의 경우에는 근접 이웃의 수가

표 3.4: 가족수(family) 문항의 RMSE

k	3	7	11	15	19	23
DU	1.295	1.229	1.195	1.183	1.171	1.169
DW	1.295	1.229	1.195	1.183	1.171	1.169
WU	1.295	1.229	1.195	1.183	1.171	1.169
WW	1.295	1.229	1.195	1.183	1.171	1.169
회귀 대체	1.176					
평균 대체	1.491					

표 3.5: 거주기간(period) 문항의 RMSE

k	3	7	11	15	19	23
DU	19.144	17.493	17.262	17.045	16.912	16.810
DW	19.134	17.493	17.260	17.044	16.911	16.810
WU	19.072	17.501	17.209	16.981	16.888	16.846
WW	19.073	17.501	17.208	16.980	16.887	16.845
회귀 대체	16.610					
평균 대체	21.019					

표 3.6: 나이(age) 문항의 RMSE

k	3	7	11	15	19	23
DU	8.176	7.614	7.498	7.431	7.395	7.365
DW	8.174	7.611	7.495	7.429	7.392	7.362
WU	8.174	7.610	7.519	7.457	7.418	7.394
WW	8.172	7.608	7.517	7.456	7.416	7.392
회귀 대체	8.143					
평균 대체	12.468					

표 3.7: 연간생활비(cost) 문항의 RMSE

k	3	7	11	15	19	23
DU	844.116	793.680	780.441	772.765	766.390	759.633
DW	844.161	793.420	780.185	772.580	766.270	759.612
WU	846.941	793.475	783.107	783.108	768.863	760.375
WW	846.932	793.355	782.943	783.030	768.784	760.377
회귀 대체	761.016					
평균 대체	904.908					

7 정도로 작을 때에 회귀대체보다 작은 RMSE를 갖는 반면에 거주기간 문항은 근접 이웃의 수가 23 이후에서 회귀대체보다 작은 RMSE를 갖는 것으로 나타났다.

표 3.8은 범주형 문항 중에서 원자료에 무응답이 존재하는 7 문항에 대해 서로 다른 다섯 가지 대체법을 적용했을 때 구해진 정확도이다. 이 때 표 3.3의 각 문항의 무응답률을 이용하여 무응답을 발생시킨 후 각각의 대체법을 이용하여 대체를 실시하였는데, 1,000회를

표 3.8: 범주형 문항의 정확도(Accuracy)

문항	최빈값대체	확률대체	조건부최빈값	조건부확률	핫덱대체
family	99.851 ± 0.06	99.812 ± 0.05	99.850 ± 0.06	99.815 ± 0.06	99.813 ± 0.06
period	99.551 ± 0.11	99.480 ± 0.10	99.571 ± 0.11	99.490 ± 0.10	99.500 ± 0.11
cost	95.068 ± 0.28	94.718 ± 0.27	95.220 ± 0.29	94.745 ± 0.28	94.384 ± 0.24
ratio	98.777 ± 0.26	98.265 ± 0.27	–	–	98.777 ± 0.26
species	99.378 ± 0.17	99.075 ± 0.17	–	–	99.102 ± 0.17
scale1	99.693 ± 0.13	99.563 ± 0.13	–	–	99.693 ± 0.13
scale2	98.532 ± 0.33	97.312 ± 0.41	–	–	98.532 ± 0.33

반복실시하여 얻어진 정확도의 평균과 표준편차를 제시하였다. 한편 조건부 최빈값과 조건부 확률을 이용한 대체는 농가 여부에 따라 농가인 개체와 비농가인 개체로 구분해서 얻어진 조건부 최빈값과 조건부 확률을 이용하였다. 따라서 ratio, species, scale1, scale2 등 농가에 대해서만 응답이 얻어진 문항에 대해서는 조건부 최빈값과 조건부 확률을 이용한 대체를 적용하지 않았다. 모든 문항에서 최빈값 대체, 핫덱 대체, 확률 대체의 순서로 정확도가 높고, 최빈값 대체보다는 조건부 최빈값 대체의 정확도가 높으며 확률 대체 보다는 조건부 확률 대체의 정확도가 높은 것으로 나타났다. 대부분의 문항에서 최빈값 대체의 정확도가 높게 나타났으나 최빈값 대체는 원자료의 분포를 왜곡시키는 문제점이 있기 때문에 핫덱 대체를 사용하는 것을 적극적으로 고려해야 할 것으로 생각된다.

4. 결론

농촌생활지표조사는 총 146문항으로 이루어져 있기 때문에 본 연구에서는 일부 문항에 대해서만 대체를 실시하였다. 최종 선택된 15 문항을 자료의 형태에 따라 연속형과 범주형으로 구분하여 적용하였다. 15 문항에 대해 무응답이 존재하는 개체를 제외하고 얻어진 1,582 가구에 대해 최종적으로 실험을 실시하였다. 연속형 자료로 구성된 가족수(family), 거주기간(period), 나이(age), 연간생활비(cost) 문항에 대해서는 평균대체, 회귀대체, 그레이기반 k-NN 대체(DU, DW, WU, WW)를 실시하였고, 회귀대체와 그레이기반 k-NN 대체가 작은 RMSE를 갖는 것으로 나타났다. 이 때 그레이기반 k-NN 대체는 문항에 따라 효과적인 근접이웃의 수가 다르게 나타나는데, 이를 밝히기 위한 연구가 진행될 필요가 있을 것으로 생각된다. 한편 원래 범주형 자료로 구성된 문항과 연속형 문항을 범주형으로 변환한 15 문항 중에서 1,870 가구로 구성된 원자료에 무응답이 포함된 가족수(family), 거주기간(period), 나이(age), 연간생활비(cost), 농업소득과 농외소득 비율(ratio), 영농형태(species), 영농규모(scale1), 영농규모하위(scale2) 문항에 대해서는 최빈값 이용, 확률 대체, 조건부 최빈값 이용, 조건부 확률 대체, 핫덱 대체를 실시하였다. 이 때 최빈값 이용, 핫덱 대체, 확률 대체의 순으로 정확도가 높은 것으로 나타났으며, 최빈값 이용 보다는 조건부 최빈값 이용이 정확도가 높고 확률 대체 보다는 조건부 확률 대체가 정확도가 높은 것으로 나타났다. 최빈값 이용이 정확도가 높은 것으로 나타났지만 최빈값 이용이 원자료의 분포를 왜곡시키는 성질을 감안한다면 핫덱 대체를 적용하는 것이 적극 검토되어야 할 것

이다. 한편 소득 문항에 대해서는 무응답률이 상당히 높은 것으로 나타났는데, 추후 조사에서는 이 문항의 응답률 향상을 위해 많은 노력이 필요할 것으로 생각한다.

참고문헌

김규성 (2000). 무응답 대체 방법과 대체 효과, <조사연구>, **1**, 1–14.

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

김규성, 황영은, 박진우 (2005b). 패널조사에서 가중치 부여 방법 및 효과에 관한 연구, <제6회 한국노동패널 학술대회>.

김영원, 이주원 (2003). CART를 활용한 결측값 대체방법: 인구주택총조사 혼인상태 항목을 중심으로, 조사연구, <조사연구>, **4**, 1–21.

김영원, 조선경 (1996). 표본조사에서 항목 무응답 대체 방법, <한국통계학회논문집>, **3**, 145–159.

김재광, 한근식, 윤연옥 (2004). 가계조사 무응답 처리기법 연구, <통계연구>, **9**, 79–102.

김주환 (2004). 인구학적 특성에 따른 단위 무응답률 분석 : 사례, *Journal of the Korean Data Analysis Society*, **6**, 1725–1734.

김진 (2004). 농가경제조사에 대한 대체법 비교, 통계연구, <통계연구>, **9**, 133–145.

도세록, 이관제 (2006). 국민건강 검진조사의 무응답 대체에 관한 연구, *Journal of the Korean Data Analysis Society*, **8**, 139–151.

박대식, 이영대 (1997). 농촌복지지표의 개발에 관한 연구, <한국농촌경제연구원>.

박태성, 이승연 (1998). 무응답을 포함하는 범주형 자료의 분석, <응용통계연구>, **11**, 83–95.

신민웅, 백정용 (2005). 아웃바인드 캠페인의 변경 희망률 추정을 위한 무응답 대체법 비교, *Journal of the Korean Data Analysis Society*, **7**, 1653–1667.

신형원, 손소영 (2002). 범주형 자료의 결측치 추정방법 성능 비교, <응용통계연구>, **15**, 33–43.

이진희, 김진, 이기재 (2006). 표본조사에서 공간변수를 이용한 결측 대체의 효율성 비교, <응용통계연구>, **19**, 57–67.

조사통계연구회 (2000). <무응답 오차>, 자유아카데미.

조영숙, 박은식, 고정숙, 황대용, 강경하 (2004). 농촌생활지표 개발 및 작성에 관한 연구, <농촌자원개발연구>, 농촌진흥청 농업과학기술원, 255–286.

황대용, 박은식, 신덕주, 조영숙, 고정숙, 강경하, 최윤지, 윤순덕, 김효철, 이채식 (2005). <농촌생활지표 조사보고서>, 농촌진흥청 농촌자원개발연구소.

Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse, *Journal of the American Statistical Association,* **78,** 708–717.

Chun, Y. M., Lee, J. W. and Chung, S. S. (2006). A modified grey-based k-NN approach for treatment of missing value, *Journal of the Korean Data & Information Science Society*, **17**, 421–436.

Deng, J. (1982). Control problems of grey systems, *Systems and Control Letters*, **5**, 288–294.

Deng, J. (1989). The basic course of grey system theory, *HUST Publisher*.

Hsia, K. H. and Wu, J. H. (1998). A study on the data preprocessing in grey relational analysis, *Journal of Chinese Grey System*, **1**, 47–54.

Huang, C. C. and Lee, H. M. (2004). A grey-based nearest neighbor approach for missing attribute value prediction, *Applied Intelligence*, **20**, 239–252.

Kim, Y. W., Ryu, J. B., Park, J. W. and Lee, J. W. (2003). Imputation methods for the population and housing census 2000 in Korea, *The Korean Communications in Statistics*, **10**, 575–583.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons, 2nd ed., New York.

Park, Jinwoo (2002). A Combined Method Compensating for Wave Nonresponse, *Journal of the Korean Statistical Society*, **31**, 469–482.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

Wen, K. L. (2004). *Grey systems : Modeling and Prediction*, Yang's Scientific Press, Tucson.

[2007년 7월 접수, 2007년 10월 채택]

An Imputation for Nonresponses in the Survey on the Rural Living Indicators

Young Sook Cho[1)] Young-Min Chun[2)] Dae Yong Hwang[3)]

Abstract

Survey on the rural living indicators was the statistic approved from National Statistical Office and the survey executed by rural resources development institute. This study was used the raw data of survey on the rural living indicators in 2005. After editing procedure for raw data, we were studied 1,582 households which is acquired through elimination of case included nonresponses, and imputed a nonresponses of 15 item selected from 146 item. The imputation methods and efficiency of imputation for simulation was adapted differently from type of data. For continuous data, we imputed the nonresponses with mean imputation, regression imputation, adjusted grey-based k-NN imputation(DU, DW, WU, WW) and compared the results with RMSE. For categorical data, we imputed the nonresponses with mode method, probability imputation, conditional mode method, conditional probability method, hot-deck imputation, and compared the results with Accuracy. By the results, regression imputation and adjusted grey-based k-NN imputation appropriated for continuous data and hot-deck imputation appropriated for categorical data.

Keywords: Accuracy, imputation, nonresponse, RMSE(Root mean square error).

1) Researcher, Rural Resource Development Institute, Suwon 441-853, Korea.
E-mail: joys@rda.go.kr
2) Corresponding author. Associate Research Fellow, Korea Employment Information Service, 77-11, Mullae-dong 3-ga, Yeongdeungpo-gu, Seoul 150-093, Korea.
E-mail: zzari90@work.go.kr
3) Researcher, Rural Resource Development Institute, Suwon 441-853, Korea.
E-mail: hdy@rda.go.kr