제2장

농업총조사 무응답 대체기법 연구(II) -과수, 작물, 가축 부문을 중심으로-

최 필 근

제1절 서 론

1. 연구배경 및 목적

농업총조사의 주요 목적은 농업정책 수립·평가 및 국가경제 주요지표의 작성, 농업관련 학술연구 및 각종 농업통계 개선을 위한 모집단 자료 확보, 지방화시대에 요구되는 소지역 자료 생산, 국제간 자료 교류 및 분석을 통한 농업부문 국가경쟁력 강화에 기여하는 데 있다. 특히, 본 연구에서 다루고자하는 항목들은 다음과 같은 조사목적을 가지고 있다. 과수원 면적은 과수원 경영구조를 파악하여 주산지 지원, 재해보험사업, 과수 선도농 육성 및 폐원 지원사업 등 과수정책에 활용되며, 노지재배 수확 및 판매작물 면적은 지역별 작물 재배동향을 파악하여 지역특화농업정책과 농산물 유통가공 계획 수립에 활용되고 있다. 그리고 시설재배 수확작물 면적은 고부가가치 농업인 시설원예작물별 재배현황을 파악하여 시설농가 지원정책수립 및 농업인의 농업계획 수립에 활용되며, 가축 마리수는 농가의 가축사육 구조와 지역별 사육현황을 파악하여 가축관련 정책에 활용되고 있다. 이와 같이 농업총조사는 국가농업부분정책을 위한 계획수립에 매우 중요한 역할을 하고 있으므로 조사 자료의 품질을 높이기 위한 끊임없는 노력이 이루어져야 할 것이다.

무응답을 처리하기 위한 연구는 최근의 조사환경을 고려할 때 통계조사의 품질을 향상시킬 수 있는 매우 중요한 노력이라고 할 수 있다. 한국보다 조사환경이 훨씬 열악하다고 할 수 있는 미국, 캐나다, 호주 등에서는 이러한 연구를 오래전부터 장기적으로 계획을 세워 진행하고 있다. 우리나라의 경우, 이들 나라에 비해 무응답률이 낮게 나타나고 있어서인지 지금까지의 무응답 연구는 통계 선진국들에 비해서는 상당히 미흡한 실정이다. 특히, 국가통계의 많은 부분을 생산하고 있는 통계청에서도 무응답 처리에 관한 연구가 활발히 진행되지 못한 것이 사실이다. 하지만 향후 무응답으로 인해 생겨나는 문제들은 심각하게 고려되어야 하며, 이를 극복하기 위한 연구도 각 조사 자료에 맞게 세부적으로 진행되어야 할 것이다.

무응답의 발생원인은 조사내용, 자료수집방법, 응답자의 가구구조 및 태도 등에 많이 좌우되며, 발생형태는 조사자로부터 얻은 정보가 전혀 없는 단위 무응답(unit nonresponse)과 특정 항목값에 응답을 하지 않는 항목 무응답(item nonresponse)으로 나눌 수 있다. 무응답 발생시 정확한 통계분석을 하기 위해서는 발생형태별로 적절한 무응답 처리를 해야 하는데 단위 무응답의 경우는 가중치 조정 방법을 사용하고, 항목 무응답의 경우는 적절한 값을 채워 넣기 위한 여러 가지 대체법을 이용하게 된다.

항목 무응답 대체연구는 크게 두 가지 과정으로 나눌 수 있다. 하나의 과정은 대체하고자 하는 항목에 대한 대체군(보조변수)을 결정하는 것이다. 이는 대체하고자 하는 항목과 연관성이 높은 항목들을 사용함으로써 대체의 정확도를 향상시키기 위함이다. 실제로 가장 적합한 대체기법을 적용하더라도 대체군의 선택이 잘못된다면 대체결과는 좋지 않을 것이라고 판단된다. 따라서 대체항목과 관련이 높은 항목들을 통계적 기법을 이용하여 찾아야 할 것이다. 주로 의사결정나무 분석의 CHAID 및 CART 알고리즘, 카이제곱 독립성검정, 회귀분석의 변수선택 방법들이 많이 이용된다. 농업총조사 항목의 대체군 연구는 『무응답 처리를 위한 방법론 연구(Ⅰ)(통계개발원, 2009)』의 『농업총조사 항목간 연관성 분석 및 대체군(보조변수) 개발(최필근)』을 참조하면 자세하게 알 수 있을 것이다. 다른 하나의 과정은 자료(항목)의 특성에 가

장 적합한 대체기법을 개발 또는 적용하는 것이다. 지금까지 주로 사용되고 있는 대체방법들은 일반적인 사회조사처럼 관심의 대상이 되는 항목이 범주형일 경우에는 핫덱대체와 이와 유사한 최근방 기증자(donor) 대체방법이, 경제관련조사와 같이 연속형 항목들이 주를 이루는 경우는 회귀대체, 최근방대체, 평균대체 방법이 적용되고 있다. 하지만 이러한 방법들은 조사 자료의 특성을 잘 파악하고 적용해야 할 것이다. 그리고 더 정확한 대체가 가능하다면 기존의 방법들을 응용하든지 아니면 새로운 방법을 개발할 수도 있을 것이다. 본 연구에서 사용하고자 하는 응용 핫덱 대체방법도 핫덱 대체방법을 응용한 것으로 농업총조사 항목들의 특성을 고려하여 개발한 것이다. 또한 과수, 작물 및 가축부문의 항목들이 모두 연속형 항목인데도 불구하고 응용 핫덱 대체방법을 적용하는 것도 농업총조사의 항목별 특성 때문이다. 이러한 내용들은 본문에 자세하게 기술할 것이다.

본 연구는 지난 2008년에 진행이 되었던 『농업총조사 무응답 대체기법 연구(Ⅰ)(최필근)』의 후속연구로서 과수, 작물, 가축부문에 대한 무응답 대체결과를 제시하고자 함에 있다. 본 연구를 통해서 체계적이며 정확도가 높은 농업총조사 무응답 대체기반을 마련함으로써 농업총조사의 통계품질향상에 기여하고, 더 나아가 다른 조사통계의 무응답 대체를 위한 선행연구로서도 활용하고자 한다.

2. 연구내용 및 방법

본 연구에서 가장 중점적으로 다루어지는 내용은 2005년 농업총조사의 과수, 작물 및 가축부문의 항목들에 대하여 대체군을 개발하고 사전에 개발된 응용 핫덱 대체방법을 적용하여 대체결과를 검토하는 것이다. 그리고 주로 많이 사용되고 있는 대표적인 방법들과 비교분석을 통하여 본 연구에서 사용된 방법의 우수성을 보여주고자 한다.

각 항목에 대한 대체군은 가구원 및 가구부문과 동일하게 의사결정나무 방법인 CHAID 알고리즘을 이용하여 연관성분석을 실시하고, 이를 토대로 결정할 것이다. 또한 대체군에 속한 항목들의 가중치도 이전의 방법과 동일하게 적용하도록 한다. 무응답을 대체하기 위한 방법도

가구원 및 가구부문에 적용하였던 응용 핫덱 대체방법을 적용할 것이다. 이는 응용 핫덱 대체방법은 목표변수와 대체군(보조변수)의 자료형태가 범주형 또는 연속형에 관계없이 사용할 수 있게 개발되었으며, 다른 대체방법에 비해서 농업총조사에 가장 적합하다고 판단되기 때문이다. 세부적으로 말하면 첫째, 농업총조사의 항목들은 범주형과 연속형이 골고루 섞여있으며, 대체군에도 역시 혼합되어 있다. 따라서 핫덱 대체방법을 연속형 항목에도 적절하게 사용하기 위해서 항목값에 신뢰구간 개념을 도입하였으며, 대체군을 사용함에 있어서도 계층적 핫덱 대체방법의 단점을 보완하기 위해 대체군에 가중치를 부여해서 최종 대체군을 결정하게 하였다. 둘째, 농업총조사의 연속형 자료의 특성은 일정한 부분만 값이 존재하고 나머지는 0의 값을 갖는 형태로 되어있다. 이러한 자료의 형태에 일반적으로 사용되는 회귀 또는 비 대체방법을 적용한다면 대체전후의 평균에서는 큰 차이가 없겠지만 아마도 대체전후의 분포(구성비) 변화는 매우 클 것으로 판단된다.

따라서 이러한 내용을 검토하기 위하여 다음과 같은 비교 연구를 실시하고자 한다. 첫째는 과수, 작물 및 가축부문의 항목들은 모두가 연속형이므로 일반적으로 연속형 항목에 많이 적용되는 회귀 대체방법과의 비교, 둘째는 현재 인구주택총조사에서 사용하고 있는 계층적 핫덱 대체방법과의 비교·분석을 실시할 것이다. 이를 통하여 각 방법들에 대하여 대체전후의 평균차이, 분포(구성비) 변화, 대체 정확도 등을 검토하고자 한다. 그리고 마지막으로 응용 핫덱 대체방법을 이용하여 과수, 작물 및 가축부문의 72개 항목에 대하여 모의실험을 실시하고 대체결과를 제시하고자 한다.

본 연구를 위하여 제2절에서는 연관성 분석을 위한 CHAID 알고리즘과 응용 핫덱 대체방법을 간략하게 설명한다. 제3절에서는 본 연구에서 사용할 농업총조사의 과수, 작물 및 가축부문의 조사항목을 살펴본다. 제4절에서는 본 연구에서 사용하고자 하는 응용 핫덱 대체방법과 회귀 대체방법, 계층적 핫덱 대체방법을 비교·분석할 것이다. 제5절에서는 과수, 작물 및 가축부문 항목들에 대해서 모의실험을 실시하여 본 연구에서 제시한 방법의 효율성을 다양하게 검토할 것이다. 마지막으로 제6절에서는 연구의 최종적인 결론과 더불어 향후 연구되어야 할 내용들을 제시하고자 한다.

제2절 CHAID 알고리즘과 응용 핫덱 대체방법

농업총조사 대체군은 CHAID 알고리즘을 이용하여 연관성분석을 실시하여 대체군을 결정한다. 이 알고리즘은 대용량의 자료로부터 항목들 간의 의미 있는 관계를 탐색하는 데 효과적이라고 알려져 있다. 이 절에서는 CHAID 알고리즘과 무응답 대체에 적용할 응용 핫덱 대체방법에 대하여 간략하게 소개한다. 이 내용은『무응답 처리를 위한 방법론 연구(Ⅰ)(통계개발원, 2009)』에도 소개되어 있으니 참조하기 바란다.

1. CHAID 알고리즘

의사결정나무의 분리 알고리즘 중의 하나인 CHAID는 목표변수가 범주형 자료인 경우에는 χ^2 통계량에 의한 분할, 연속형 자료인 경우에는 F검정을 이용한 분할을 수행하는 분석방법이다. 구체적인 알고리즘을 살펴보면 다음과 같다.

step 1 : 각 설명변수에 대하여, 목표변수와 가장 유사성(값으로 측정)이 큰 범주의 짝을 찾는다. 값을 계산하는 방법은 목표변수의 자료특성에 의해 결정된다. p p

이 때 목표변수가 범주형인 경우는 2×d 분할표를 통한 χ^2 검정을 사용한다. 여기서 d는 목표변수의 범주 수이다.

(예시) 2×d 분할표에서의 χ^2 값 계산

	범주 1	범주 2	…	범주 d	합계
범주 1			…		
범주 2	f_{11}	f_{12}	…	f_{1d}	$f_{1.}$
합계	f_{21}	f_{22}	…	f_{2d}	$f_{2.}$
	$f_{.1}$	$f_{.2}$		$f_{.d}$	$f_{..}$

분할표에서 유사성 검정을 위한 카이제곱 통계량은

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수는

$$e_{ij} = \frac{f_{i.} \times f_{.j}}{f_{..}}$$

와 같이 계산된다. 이 때 카이제곱의 값이 클수록 각 범주에 의하여 목표변수를 분리할 가능성이 커진다. 성별에 의한 선호도를 나타내고 있는 간단한 예를 살펴보면 다음과 같다.

case 1

	찬성	반대	계
남	40 (20)	10 (30)	50
여	0 (20)	50 (30)	50
계	40	60	100

( ) 안의 값은 각 셀에서의 기대도수를 나타냄

case 2

	찬성	반대	계
남	30 (25)	20 (25)	50
여	20 (25)	30 (25)	50
계	50	50	100

· case 1의 카이제곱 통계량 :

$$\chi^2 = \frac{(40-20)^2}{20} + \frac{(10-30)^2}{30} + \frac{(0-20)^2}{20} + \frac{(50-30)^2}{30}$$

$$= 66.67$$

· case 2의 카이제곱 통계량 :

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25}$$

$$= 4$$

따라서 case 1의 경우의 카이제곱 통계량이 크기 때문에 case 2의 경우보다 성별이 분리될 가능성이 커짐을 알 수 있다. 그리고 목표변수가 연속형인 경우에는 2개 이상의 그룹의 평균차이를 검정하는 분산분석표의 F검정을 사용하여 분리한다.

step 2 : 가장 큰 p값을 가지는 설명변수 범주의 짝에 대하여 그 p값과 미리 정해놓은 α값을 비교한다.

- p값이 α값보다 클 경우에는 짝을 이루는 설명변수의 범주를 통합하고, 새로 생성된 범주에 대하여 step 1을 다시 실행한다.
- p값이 α값보다 작을 경우에는 step 3으로 간다.

step 3 : 조정된 각 설명변수의 범주에 대하여 새로운 p값을 계산하고, 가장 작은 p값을 가지는 설명변수를 선택하여 그 p값과 미리 정해놓은 α값을 비교한다.

- p값이 α값보다 작거나 같을 경우에는 설명변수의 범주에 근거한 노드를 분리한다.
- p값이 α값보다 클 경우에는 노드를 분리하지 않으며 이 노드는 최종노드가 된다.

step 4 : 더 이상 분리할 노드가 없거나 정해진 정지규칙이 만족할 때까지 위의 과정을 독립적으로 반복한다.

2. 응용 핫덱 대체방법

응용 핫덱 대체방법은 농업총조사 자료의 특성에 맞게 범주형과 연속형 항목 모두에 적용될 수 있게 개발되었다. 기본 알고리즘을 적당한 예제를 가지고 설명하고자 한다.

가. 설명을 위한 자료

본 연구는 농업총조사의 과수, 작물 및 가축부문에 대한 연구로 알고리즘을 사과면적 항목으로 설명하고자 한다. 앞에서 설명한 CHAID 알고리즘에 의하여 분석한 결과 사과면적 항목의 대체군과 목표변수와의 연관도(중요도)는 영농형태(상세)(100), 밭면적(78), SS분무기수(66), 영농형태(29)로 결정되었다.

〈표 2-1〉 2005년 농업총조사 자료일부

개체번호	사과면적	영농형태 상세 ( )	밭면적	분무기 SS	영농형태
1	2350	1	2350	1	2
2	0	-1	150	0	1
3	1200	1	1200	1	2
4	1200	3	3350	0	2
5	0	-1	280	0	1
6	0	1	540	0	4
7	3000	1	4000	1	2
8	0	-1	300	0	6
9	0	-1	1430	0	1
10	600	1	1200	1	2
11	300	-1	400	0	6
12	0	-1	600	0	1
13	100	3	503	0	2
14	2800	1	3800	1	2
15	**520(A)**	1	520	1	2
16	0	6	300	0	2
17	0	2	5500	0	2
18	1500	2	5000	1	2
19	500	6	8000	1	2
20	0	1	500	0	4
21	10000	1	10000	1	2
22	200	1	530	1	4
23	0	-1	0	0	1
24	0	2	1500	0	4
25	0	-1	0	0	1
26	0	-1	0	0	1
27	100	4	2750	0	2
28	0	-1	500	0	1
29	510	1	510	1	2
30	0	-1	600	0	1

<표 2-1>은 실제 농업총조사 자료 중에서 일부를 발췌하여 작성한 것으로 설명을 위한 자료로 사용하고자 한다. 총 자료의 수는 1,272,908 가구이지만 임의로 30가구의 자료만 가지고 설명을 할 것이다. 영농형태상세 항목의 값들 중에는 (-1)의 값이 존재한다. 이것은 연구자가 임의로 만든 값으로 허용되는 무응답을 표시하기 위함이다. 즉, 영농형태가 논벼(1), 특용작물(3), 화훼(5), 일반밭작물(6), 양잠기타(8)이면 영농형태(상세)는 기입하지 않도록 되어 있다.

나. 예제를 이용한 알고리즘 설명

<표 2-1>은 임의의 30가구의 자료이며 완전하게 조사가 된 내용들이다. 무응답 대체과정을 설명하기 위하여 15번째 가구에서 사과면적을 응답하지 않았다고 가정하자. 따라서 (A)항목은 무응답임을 표시하기 위해 (-2)의 값을 부여하게 된다.

step 1 : 첫 번째 가구부터 사과면적 항목의 값이 (-2)인지 아닌지를 체크한다. 만일 (-2)가 아니면 다음 가구로 이동하고 맞으면 사과면적과 가장 연관성이 높은 영농형태(상세) 항목과의 일치여부를 판단하여 점수화한다.

- 영농형태(상세)가 1인 가구에는 100점을 부여한다.
- 영농형태(상세)가 1이 아닌 가구에는 0점을 부여한다.

step 1의 과정이 끝나고 나면 <표 2-2>와 같이 가장 점수가 높은 총 10가구의 예비 도너가 선택된다. 현재의 점수는 100점이 되며 영농형태(상세)가 1인 것만이 선택된 것을 볼 수 있다.

step 2 : 첫 번째 절차가 끝난 후 두 번째로 사과면적과 연관성이 높은 밭면적과의 일치여부를 판단하여 이전에 획득한 점수와 합산한다.

앞의 절차와는 달리 밭면적은 연속형 항목이다. 따라서 일치여부를 같은 값으로 판단하지 않고 신뢰구간의 개념을 사용한다.

<표 2-2> step 1 절차 후의 예비도너(donor)

개체번호	사과면적	영농형태 상세( )	밭면적	분무기 SS	영농형태	점수
1	2350	1	2350	1	2	100
3	1200	1	1200	1	2	100
6	0	1	540	0	4	100
7	3000	1	4000	1	2	100
10	600	1	1200	1	2	100
14	2800	1	3800	1	2	100
15	**520(A)**	1	520	1	2	
20	0	1	500	0	4	100
21	10000	1	10000	1	2	100
22	200	1	530	1	4	100
29	510	1	510	1	2	100

대체하려는 가구의 밭면적은 520평으로 이 값으로부터 5%의 오차 범위 안에 있다면 같은 값으로 간주하고자 하는 것이다. 즉, 밭면적이 (520)*(0.95)=494 (평)보다 많고 (520)*(1.05)=546 (평)보다 적으면 520평과 같은 값으로 인정을 할 것이다. 이와 같이 연속형 자료도 범주형 자료처럼 사용가능하며, 임의로 구간을 나누어 사용하는 것보다 훨씬 오차가 작을 것으로 판단된다. 이 과정을 거쳐 밭면적의 일치여부를 판단하여 78점을 부여하고 <표 2-3>과 같이 합산한 점수가 가장 높은 개체를 예비도너로 선택하게 된다. 결과적으로 4가구가 선택된 것을 볼 수 있다.

step 3 : 두 번째 절차가 끝난 후 세 번째로 사과면적과 연관성이 높은 SS분무기수 항목과의 일치여부를 판단하여 이전에 획득한 점수와 합산한다.

앞의 경우처럼 SS분무기수도 연속형 항목이므로 이 값으로부터 5%의 오차 범위 안에 있다면 같은 값으로 간주하기로 한다. SS분무기수가

$(1)*(0.95)=0.95$(대)보다 많고 $(1)*(1.05)=1.05$(대)보다 적으면 1대와 같은 값으로 인정을 할 것이다. 이 경우는 1대인 경우와 일치하므로 범주형 항목과 동일하게 적용된다. 이 과정을 거쳐 SS분무기수의 일치여부를 판단하여 66점을 부여하고 <표 2-4>와 같이 합산한 점수가 가장 높은 개체를 예비도너로 선택하게 된다. 결과적으로 2가구가 선택된 것을 볼 수 있다.

〈표 2-3〉 step 2 절차 후의 예비도너(donor)

개체번호	사과면적	영농형태 상세 ( )	밭면적	분무기 SS	영농형태	점수
6	0	1	540	0	4	178
15	**520(A)**	1	520	1	2	
20	0	1	500	0	4	178
22	200	1	530	1	4	178
29	510	1	510	1	2	178

〈표 2-4〉 step 3 절차 후의 예비도너(donor)

개체번호	사과면적	영농형태 상세 ( )	밭면적	분무기 SS	영농형태	점수
15	**520(A)**	1	520	1	2	
22	200	1	530	1	4	244
29	510	1	510	1	2	244

step 4 : 세 번째 절차가 끝난 후 네 번째로 사과면적과 연관성이 높은 영농형태 항목과의 일치여부를 판단하여 이전에 획득한 점수와 합산한다.

- 영농형태가 2(과수)인 가구에는 29점을 부여한다.
- 영농형태가 2(과수)가 아닌 가구에는 0점을 부여한다.

step 4의 과정이 끝나고 나면 <표 2-5>와 같이 합산한 점수가 가장 높은 개체가 최종적으로 선택된다.

〈표 2-5〉 step 4 절차 후의 최종도너(donor)

개체번호	사과면적	영농형태 상세 ( )	밭면적	분무기 SS	영농형태	점수
15	**520(A)**	1	520	1	2	
29	510	1	510	1	2	273

step 5 : 최종적으로 선택된 도너의 사과면적의 값을 무응답 가구의 사과면적에 대체한다.

- 최종적으로 선택된 도너가 1개이면 그 값을 대체한다.
- 최종적으로 선택된 도너가 2개 이상이면 임의로 뽑아서 선택된 도너의 값을 대체한다.

본 예제에서는 <표 2-5>와 같이 29번 가구가 최종도너로 선택이 되어 510평을 무응답 가구에 대체하게 된다. 이 가구의 실제값은 520평으로 정확도가 높은 대체가 되었음을 알 수 있다.

step 6 : 사과면적 항목의 값이 ( )가 나오지 않을 때까지 **step 1**로 돌아가 계속적으로 반복한다.

제3절 과수, 작물 및 가축부문 항목검토

이 절에서는 본 연구에서 사용되는 2005년 농업총조사의 과수, 작물 및 가축부문 항목들을 검토하고자 한다. 세부적으로 살펴보면 과수 13개 항목, 노지재배 수확작물 15개 항목, 노지재배 판매작물 12개 항목, 시설재배 수확작물 20개 항목, 가축 16개 항목으로 구성되어 있다.

[과수원에 관한 사항]

과수원에 관한 사항은 <표 2-6>과 같이 13개 항목으로 구성되어 있다. 사과, 배, 복숭아, 포도, 단감, 감귤 항목은 조사초기부터 지속적으로 조사해오고 있으며, 떫은감, 자두, 키위, 매실 항목은 1990년 이후 조사되고 있다. 그리고 살구와 유자는 2005년에 새로 추가된 항목으로써 연속성 여부를 알 수 없으나 본 연구에서는 포함시킬 것이다. 하지만, 기타 항목은 무응답대체의 의미가 없으므로 연관성 분석을 실시하지 않고 제외시키기로 한다.

〈표 2-6〉 과수원면적 항목의 연도별 조사여부

조사항목	조사년도						
사과	1960	1970	1980	1990	1995	2000	2005
(1)배	○	○	○	○	○	○	○
(2)복숭아	○	○	○	○	○	○	○
(3)포도	○	○	○	○	○	○	○
(4)단감	○	○	○	○	○	○	○
(5)떫은감					○	○	○
(6)감귤		○	○	○	○	○	○
(7)자두				○	○	○	○
(8)참다래 키위				○	○	○	○
(9) 매실 ( )					○	○	○
(10)살구							○
(11)유자							○
(12)기타	○	○	○	○	○	○	○
(13)							

[노지재배 수확작물에 관한 사항]

노지재배 수확작물에 관한 사항은 <표 2-7>과 같이 논벼, 보리, 옥수수, 콩, 팥, 감자, 고구마, 김장무, 김장배추, 고추, 양파, 대파, 마늘, 참깨, 인삼 15개 항목으로 구성되어 있다. 모든 항목들은 조사초기 또는

1980년 이후부터는 지속적으로 조사되고 있다. 따라서 본 항목들에 대한 연관성 분석 및 무응답대체를 실시하기로 한다.

〈표 2-7〉 노지재배 수확작물면적 항목의 연도별 조사여부

조사항목	조사년도						
	1960	1970	1980	1990	1995	2000	2005
논벼	○	○	○	○	○	○	○
(1)보리	○	○	○	○	○	○	○
(2)옥수수		○	○	○	○	○	○
(3)콩	○	○	○	○	○	○	○
(4)팥	○	○	○	○	○	○	○
(5)감자	○	○	○	○	○	○	○
(6)고구마	○	○	○	○	○	○	○
(7)김장무			○	○	○	○	○
(8)김장배추			○	○	○	○	○
(9) 고추	○	○	○	○	○	○	○
(10)양파		○	○	○	○	○	○
(11)대파		○	○	○	○	○	○
(12)마늘	○	○	○	○	○	○	○
(13)참깨		○	○	○	○	○	○
(14)인삼		○		○	○	○	○
(15)							

[노지재배 판매작물에 관한 사항]

노지재배 판매작물에 관한 사항은 <표 2-8>과 같이 12개 항목으로 구성되어 있다. 시금치, 상추, 쑥갓, 오이, 수박, 당근, 들깨, 땅콩 항목은 1980년 이후부터 지속적으로 조사해오고 있다. 양배추, 호박, 화훼 항목은 2005년에 다시 추가된 항목으로써 연구에 포함을 시키고 이전의 경우와 동일하게 기타 항목은 제외시킨다.

〈표 2-8〉 노지재배 판매작물면적 항목의 연도별 조사여부

조사항목	조사년도						
	1960	1970	1980	1990	1995	2000	2005
양배추		○	○				○
(1)시금치			○	○	○	○	○
(2)상추			○	○	○	○	○
(3)쑥갓			○	○	○	○	○
(4)오이			○	○	○	○	○
(5)수박	○		○	○	○		○
(6)호박							○
(7)당근				○	○	○	○
(8)들깨		○	○	○	○	○	○
(9) 땅콩		○	○	○	○	○	○
(10)화훼				○			○
(11)기타							○
(12)							

[시설재배 수확작물에 관한 사항]

시설재배 수확작물에 관한 사항은 <표 2-9>과 같이 20개 항목으로 구성되어 있다. 상추, 토마토, 오이, 수박, 참외, 고추, 버섯, 화훼 항목은 1970년부터, 배추, 시금치, 딸기 항목은 1980년부터, 무, 서양채소, 메론은 1990년 이후 조사되고 있다. 그리고 호박, 대파, 감자 항목은 2005년에 새로 추가된 항목으로써 연속성 여부를 알 수 없으나 본 연구에서는 포함시킬 것이다. 하지만, 기타 항목은 무응답대체의 의미가 없으므로 연관성 분석을 실시하지 않고 제외시키기로 한다.

[가축에 관한 사항]

가축에 관한 사항은 <표 2-10>과 같이 16개 항목으로 구성되어 있다. 한우, 육우, 돼지, 산란계, 육계, 젖산양, 염소, 토끼, 오리, 꿀벌 항목은 1960~70년 조사초기부터 조사되고 있으며, 멧돼지, 곰, 고라니 항목은

2005년에 새로 추가된 항목이다. 앞의 경우와 마찬가지로 주어진 모든 항목들에 대하여 연관성 분석 및 무응답대체를 실시하기로 한다. 연속성 여부는 2010년 농업총조사 항목들이 결정되어야 알 수 있으므로 이후에 다시 조정하면 될 것으로 판단된다.

〈표 2-9〉 시설재배 수확작물면적 항목의 연도별 조사여부

조사항목	조사년도						
	1960	1970	1980	1990	1995	2000	2005
무				○	○	○	○
(1)배추			○	○	○	○	○
(2)시금치			○	○	○	○	○
(3)상추		○	○	○	○	○	○
(4)토마토		○	○	○	○	○	○
(5)오이		○	○	○	○	○	○
(6)딸기			○	○	○	○	○
(7)수박		○		○	○	○	○
(8)참외		○	○	○	○	○	○
(9) 호박							○
(10)고추		○	○	○	○	○	○
(11)대파							○
(12)서양채소				○	○	○	○
(13)포도					○	○	○
(14)감귤					○	○	○
(15)메론				○	○	○	○
(16)버섯		○			○	○	○
(17)감자							○
(18)화훼		○	○	○	○	○	○
(19)기타					○	○	○
(20)							

〈표 2-10〉 가축 항목의 연도별 조사여부

조사항목	조사년도						
	1960	1970	1980	1990	1995	2000	2005
(1) 한우	○	○	○	○	○	○	○
(2) 육우	○	○	○	○	○	○	○
(3) 젖소암컷		○	○	○	○	○	○
(4) 돼지	○	○	○	○	○	○	○
(5) 멧돼지							○
(6) 산란계	○	○	○	○	○	○	○
(7) 육계	○	○	○	○	○	○	○
(8) 젖산양	○	○	○	○	○	○	○
(9) 엽소	○			○	○	○	○
(10) 사슴				○	○	○	○
(11) 토끼	○	○	○	○	○	○	○
(12) 오리	○		○	○	○	○	○
(13) 꿀벌	○	○	○	○	○	○	○
(14) 곰							○
(15) 고라니							○
(16) 기타					○	○	○

제4절 대체방법 및 대체군 비교

이 절에서는 농업총조사의 무응답을 대체하기 위하여 개발된 응용 핫덱 대체방법과 현재 인구주택총조사에서 사용되고 있는 계층적 핫덱 대체방법, 그리고 연속형 항목에 주로 사용하는 회귀 대체방법을 비교하고자 한다. 간단한 모의실험을 통하여 본 연구에서 대체할 항목을 선택하여 대체전후의 평균차이 및 분포변화, 대체의 정확도를 비교함으로써 응용 핫덱 대체방법이 농업총조사에 적절한 대체방법임을 보이고자 한다. 또한 대체군 형성과

정에 있어서 자료전체를 사용한 경우와 지역별로 나누어 적용한 경우의 결과를 비교하여 전체자료를 이용해서 개발된 대체군을 각 항목에 사용해도 문제점이 없는지 확인하고자 한다.

1. 회귀 대체방법과의 비교

회귀 대체방법은 목표변수와 보조변수의 관계가 절편이 있는 직선관계이고 목표변수의 분산이 동일할 때 유용하게 사용할 수 있는 하나의 방법이다. 즉,

$$y_k^* = \begin{cases} y_k, & k \in R, \quad \text{응답값} \\ \bar{y}_R + b(x_k - \bar{x}_R), & k \in R^c, \quad \text{대체값} \end{cases}$$

여기서 $b = \sum_{k \in R}(x_k - \bar{x}_R)(y_k - \bar{y}_R) / \sum_{k \in R}(x_k - \bar{x}_R)^2$ 이며, 여기서 R은 응답 집합, R^c은 무응답 집합이며 $\bar{y}_R$은 응답값의 평균이다. 이 방법은 연속형 항목의 대체에 많이 이용되는데 대체 값과 실제 값의 평균절대편차를 비교할 때 다른 대체 방법에 비해 매우 적절하다고 알려져 있다. 하지만, 조사의 특성을 고려하지 않고 일반적인 방법을 적용하는 것은 잘못된 결과를 초래할 수 있다. 본 연구에서 다루는 과수, 작물 및 가축의 항목들 모두가 연속형 항목이므로 회귀 대체방법을 이용할 수 있으나, 농업총조사 특성상 분포(구성비)의 문제가 발생하게 된다. 이를 확인하기 위해서 간단한 모의실험을 실시하고자 한다.

가. 사과면적 항목에의 적용

응용 핫덱 대체방법과 회귀 대체방법을 비교하기 위하여 사과면적에 대체를 실시하여 결과를 검토할 것이다. 먼저 모의실험에 앞서 CHAID 알고리즘에 사용하여 구한 사과면적 항목의 대체군은 <표 2-11>과 같다. 영농형태(상세) 항목이 사과면적과 가장 큰 연관성을 가지며, 밭면적, SS분무기, 영농형태 순으로 연관성이 높게 나타나고 있다. 따라서 대체군으로 이 4가지 항목을 사용하여 대체를 실시할 것이다. <표 2-11>에 대한 자세한 설명은 『무응답 처리를 위한 방법론 연구(Ⅰ)(통

계개발원, 2009)』를 참조하기 바란다.

〈표 2-11〉 사과 항목에 관한 연관성 분석의 세부내용

깊이 연관성 ( )	분리변수	사과 분리지점 좌 ( )	분리지점 우 ( )	연관비중
1	영농형태 상세 ( )	2~7 미만	1 이상	1.0000
2	SS 분무기 밭면적 (1) (2)	0.5 미만 6750	0.5 이상 6750	0.6598 0.7797
3	영농형태 밭면적 (3) (4)	2 미만 13200	1, 3-8 이상 13200	0.2878

다음으로 실제 적용에 대해서 살펴보고자 한다. 응용 핫덱 대체방법의 적용은 제2절에서 자세하게 설명을 하였으므로, 회귀 대체방법의 적용에 대해서 설명할 것이다. 사과면적을 추정하기 위해서 대체군(보조변수)을 사용하는데 이 대체군에는 연속형 항목인 밭면적과 SS분무기, 범주형 항목인 영농형태와 영농형태(상세)가 포함되어 있다. 따라서 회귀모형으로 추정시 영농형태와 영농형태(상세)에 의한 셀을 만들고 각 셀 안에서 추정이 이루어지도록 한다. 이는 가변수를 써서 추정하는 것과 같은 개념으로 볼 수 있을 것이다. 2005년 농업총조사의 조사표를 근거로 <표 2-12>와 같은 셀을 구성할 수 있다.

〈표 2-12〉 회귀 대체방법 적용을 위한 셀 구성

영농형태 상세	영농형태 1	2	3	4	5	6	7	8
1		●		●			●	
2		●					●	
3	●	●	●		●	●	●	●
4		●					●	
5		●					●	
6							●	
7								

영농형태는 8가지의 세부항목으로 구성되어 있으며, 영농형태(상세)는 영농형태에 따라서 0~7개의 세부항목을 가진다. <표 2-12>에서처럼 영농형태가 1(논벼), 3(특용작물), 5(화훼), 6(일반밭작물), 8(양잠기타)은 영농형태(상세) 항목을 가지지 않으며, 영농형태가 2(채소)는 6개, 4(채소)는 2개, 7(축산)은 7개의 영농형태(상세) 세부항목을 가진다. 따라서 총 20개의 셀을 구성할 수 있으며 이 셀에 대해서 회귀모형을 이용한 추정을 하였다. 그리고 총 50,000가구에 대해서 임의로 무응답으로 간주하고 무응답 대체를 실시하였다.

나. 적용결과

응용 핫덱 대체방법과 회귀 대체방법에 의하여 대체된 결과가 <표 2-13>에 정리되어 있다.

〈표 2-13〉 두 방법의 대체결과 비교

평균 전체가구 평 평균 면적보유가구 평

(69.77) (2279.99)

대체방법 (응용 핫덱 대체방법 (			회귀 대체방법( )	
	전체가구	면적보유가구	전체가구	면적보유가구
대체후 평균				
절대차이	69.27	2263.57	69.04	83.01
오차비율	0.50	16.42	0.73	2196.98
대체정확도)	**(0.72%)**	**(0.72%)**	**(1.05%)**	**(96.36%)**
	98.52(%)		**20.98(%)**	

임의로 선택된 50,000가구의 실제 사과면적 평균은 69.77평이며, 이 중에서 실제로 사과면적을 보유하고 있는 가구의 사과면적 평균은 2279.99평이다. 전체가구에 대해서 대체전후의 사과면적 평균의 차이를 보면 응용 핫덱 대체방법은 0.50평으로 0.72%의 오차를 보이며, 회귀 대체방법은 0.73평으로 1.05%의 오차를 가진다. 이 경우 두 방법의 차이는 크게 나타나지 않는 것처럼 보인다. 하지만, 면적보유가구에 대해서 살펴보면 매우 큰 차이가 있음을 발견하게 된다. 응용 핫덱 대체방법은 오차가 0.72%로 전체가구의 경우와 차이가 없는 반면에 회귀 대체방법은 96.36%의 오차로써 터무니없는 결과를 보이고 있다. 이러한 이유는 회

귀 대체방법의 문제가 아닌 농업총조사 항목의 특성 때문이다. 다시 말하면, 사과면적의 경우 면적을 소유하지 않은 가구는 0의 값을 가지는데 회귀 대체방법으로 대체시 0과 가까운 값으로 추정되는 경우가 상당히 많아진다. 그러므로 대체되는 값들은 큰 차이가 나지 않지만 대체의 정확도는 98.52%와 20.98%로 매우 큰 차이가 남을 알 수 있다. 이러한 문제는 심각한 분포(구성비)의 변화를 가져온다. 여기서 말하는 대체의 정확도는 면적보유유무에 대한 것으로 회귀모형으로 추정하는 방법자체가 좋지 않다는 것은 아니므로 오해가 없길 바란다.

〈표 2-14〉 대체전후의 분포변화

범주	없음 (0)	1-999	1000-1999	2000-2999	3000-3999	4000-4999	5000-
실제 총가구수	48470	376	443	271	177	101	162
응용 핫덱 대체방법							
대체후 총가구수							
절대차이	48492	367	434	266	174	103	164
분포변화비율 ( )	22 **(0.044%)**	9 **(0.018%)**	9 **(0.018%)**	5 **(0.010%)**	3 **(0.006%)**	2 **(0.004%)**	2 **(0.004%)**
회귀 대체방법							
대체후 총가구수							
절대차이	8958	40184	168	219	146	110	99
분포변화비율 ( )	39512 **(79.02%)**	39808 **(79.62%)**	27 **(0.054%)**	52 **(0.104%)**	31 **(0.062%)**	9 **(0.018%)**	63 **(0.126%)**

<표 2-14>에는 두 방법에 대해서 대체전후의 분포변화의 결과가 주어져있다. 총 50,000가구의 대체전의 분포를 살펴보면 사과면적을 보유하지 않은 가구가 48,470가구, 1-999평이 376가구, 1000-1999평이 443가구, 2000-2999평이 271가구, 3000-3999평이 177가구, 4000-4999평이 101가구, 마지막으로 5000평 이상은 162가구이다. 이 50,000가구에 대해서 응용 핫덱 대체방법으로 대체한 후의 분포가 다음 칸에 제시되어 있다. 전체적으로 살펴보면 분포의 변화가 0.004%~0.044% 정도로 거의 일어나지 않음을 알 수 있다. 이러한 결과와는 달리 회귀 대체방법으로 대체한 후의 분포를 보면 사과면적 없음과 1-999평을 가진 가구에서 약 79%

정도의 분포변화가 일어난다. 이것은 면적이 없어야 하는 가구의 80%정도가 면적이 있는 것으로 추정됨으로 일어나는 현상이다. 앞에서 언급하였듯이 사과면적은 연속형 항목이지만 연속형 항목에 주로 적용하는 회귀 대체방법을 사용할 수 없는 가장 큰 이유가 대체전후의 분포변화 때문이다. 평균 대체방법이나 비 대체방법도 같은 결과를 보이게 된다. 이러한 결과를 통해서 농업총조사의 과수, 작물 및 가축 항목의 경우에는 조사 특성상 연속형 항목에 주로 적용되는 대체방법을 사용하면 심각한 오류가 발생할 수 있다는 것을 알 수 있다. 따라서 대체전후의 평균 및 분포의 변화가 매우 작은 응용 핫덱 대체방법을 농업총조사의 무응답 대체에 적용하는 것이 적절하다고 판단된다.

2. 계층적 핫덱 대체방법과의 비교

계층적 핫덱 대체방법은 응용 핫덱 대체방법과 마찬가지로 핫덱 대체방법을 응용하여 만들어진 방법으로써 현재 인구주택총조사에 적용되고 있다. 이 방법은 대체군에 속한 모든 항목들에 대하여 중요도에 따라서 일렬로 대체군을 형성하는 것이다. 그리고 형성된 최단 셀로부터 시작하여 무응답 항목은 응답자의 그룹으로부터 목표변수의 값을 할당받는다. 이 때 최단 셀까지의 정보를 모두 만족하지 않는 무응답 항목이 존재하면 마지막에 배치된 대체군 항목을 제거하고 이전의 작업을 다시 실행하는 것이다.

응용 핫덱 대체방법도 이와 같은 원리로 모든 대체군을 만족하지 않는 무응답 항목이 존재하면 사용가능하지 않는 대체군의 항목을 제거시킨 후 새로운 대체군 항목만을 가지고 대체를 하게 된다. 하지만 계층적 핫덱 대체방법과의 가장 큰 차이점은 제거되었던 대체군의 항목들이 필요에 따라 다시 사용될 수 있다는 점이다. 따라서 무응답 대체시 대체군의 정보손실을 최소로 했다고 할 수 있다. 연구자의 생각으로는 일반적으로 많은 차이는 없을 것으로 보이지만 응용 핫덱 대체방법이 조금이나마 더 정확할 수 있을 것으로 판단된다. 이를 확인하기 위해서 간단한 모의실험을 실시하고자 한다.

가. 고추면적 항목에의 적용

응용 핫덱 대체방법과 계층적 핫덱 대체방법을 비교하기 위하여 고추면적에 대체를 실시하여 결과를 검토할 것이다. 고추면적 항목의 대체군은 <표 2-15>에 주어져있다. 밭면적 항목이 고추면적과 가장 큰 연관성을 가지며, 건조기수, 비닐하우스 면적, 영농형태(상세), 영농형태, 시도, 시군구 순으로 연관성이 높게 나타나고 있다. 따라서 대체군으로 이 7가지 항목을 사용하여 대체를 실시할 것이다.

〈표 2-15〉 고추 항목에 관한 연관성 분석의 세부내용

깊이 (연관성)	분리변수	분리지점 (좌: 미만)	분리지점 (우: 이상)	연관비중
1	밭면적	2763.5 미만	2763.5 이상	1.0000
2	건조기수 (1) 영농형태 상세 (2)	447.5 미만 0.5	447.5 이상 0.5	0.8954
3	시도 (1) 비닐하우스 (3) 영농형태 (4)	2~7 미만 13	1 이상 13	0.5070 0.3821 0.5100
4	시군구 (1) (5)	4	2, 7	0.4739 0.2145

다음으로 실제 적용에 대해서 살펴보고자 한다. 두 방법은 대체군의 사용에서 차이가 발생하므로 대체군 사용에 따른 대체결과를 비교하면 될 것이다. 계층적 핫덱 대체방법의 경우 중요한 항목 순으로 대체군을 일렬로 정렬해야 한다. 이 중요도는 <표 2-15>의 연관비중의 값을 사용하면 되며, 무응답이 전혀 발생하지 않는 항목들을 가장 앞에 놓고 정렬하면 된다. 고추면적 대체군을 근거로 <표 2-16>과 같이 정렬된다.

〈표 2-16〉 계층적 핫덱 대체를 위한 정렬된 층

층 번호	1	2	3	4	5	6	7
대체군 항목	시도	시군구	밭면적	건조기	비닐하우스	영농형태 상세 ( )	영농형태

층1과 층2는 시도와 시군구로서 이 항목들은 무응답이 없으므로 항

상 사용하게 된다. 그리고 나머지 5개의 항목을 이용하여 대체를 실시하는데 두 방법에서의 대체군 활용의 차이는 여기에서 발생하게 된다. 예를 들면, 고추면적을 대체하려고 하는데 대체군에 포함된 항목들 중에서 비닐하우스 면적도 무응답일 가능성이 있다. 따라서 이 경우에는 비닐하우스 면적을 대체군에서 제외하고 대체를 실시하게 된다. 그러면 두 방법에서 제외되는 대체군은 <표 2-17>과 같이 될 것이다. 계층적 핫덱 대체방법에서는 비닐하우스 면적, 영농형태(상세), 영농형태가 제외되는 반면 응용 핫덱 대체방법은 비닐하우스 면적만 제외된다.

〈표 2-17〉 비닐하우스 면적이 무응답인 경우의 대체군

층 번호	1	2	3	4	5	6	7
계층적 핫덱	시도	시군구	밭면적	건조기	×	×	×
응용 핫덱	시도	시군구	밭면적	건조기	×	영농형태 상세 ( )	영농형태

또 하나의 예로써 고추면적과 가장 연관성이 높은 밭면적을 알 수 없을 때에는 계층적 핫덱 대체방법의 경우 대체시 시도와 시군구 정보만을 사용하므로 많은 정보의 손실이 일어난다. 반면에 응용 핫덱 대체방법의 경우에는 밭면적을 제외한 나머지 대체군을 모두 사용할 수 있다. <표 2-18>을 참조하기 바란다.

〈표 2-18〉 밭면적이 무응답인 경우의 대체군

층 번호	1	2	3	4	5	6	7
계층적 핫덱	시도	시군구	×	×	×	×	×
응용 핫덱	시도	시군구	×	건조기	비닐 하우스	영농형태 상세 ( )	영농형태

제시된 예를 통해서 알 수 있듯이 대체를 위한 정보손실이 적은 응용 핫덱 대체방법은 계층적 핫덱 대체방법과 비교할 때 대체군 활용의 측면에서 보완되었다고 판단된다.

나. 적용결과

대체군 중에서 비닐하우스 면적이 무응답일 경우의 대체결과가 <표 2-19>에 정리되어 있다. 임의로 선택된 50,000가구의 실제 고추면적 평균은 165.94평이며, 이 중에서 실제로 고추면적을 보유하고 있는 가구의 고추면적 평균은 294.21평이다. 전체가구에 대해서 대체전후의 고추면적 평균의 차이를 보면 응용 핫덱 대체방법은 0.92로 0.55% 오차를 보이며, 계층적 핫덱 대체방법은 1.55로 0.93%의 오차를 가진다. 그리고 면적보유가구에 대해서 살펴보면 응용 핫덱 대체방법은 0.70%, 계층적 핫덱 대체방법은 0.88%의 오차를 보이며, 대체의 정확도는 78.26%와 75.88%를 나타내고 있다. 결과에서 알 수 있듯이 응용 핫덱 대체방법이 오차비율은 조금 낮고 대체의 정확도는 조금 높게 나타난다. 이러한 이유는 이전에 설명하였듯이 응용 핫덱 대체방법의 대체군에 영농형태와 영농형태(상세) 항목의 정보가 더 포함되었기 때문이다.

〈표 2-19〉 대체군에서 비닐하우스 면적이 무응답인 경우의 대체결과

평균 전체가구 165.94평 평균 면적보유가구 294.21평

대체방법 (응용 핫덱 대체방법)			(계층적 핫덱 대체방법)	
	전체가구	면적보유가구	전체가구	면적보유가구
대체후 평균				
절대차이	166.86	292.15	167.49	296.80
오차비율	0.92	2.06	1.55	2.59
대체정확도)	**(0.55%)**	**(0.70%)**	**(0.93%)**	**(0.88%)**
	78.26(%)		**75.88(%)**	

<표 2-20>에는 두 방법에 대해서 대체군 중에서 비닐하우스 면적이 무응답일 경우의 대체전후의 분포변화 결과가 주어져있다. 총 50,000가구의 대체전의 분포를 살펴보면 고추면적을 보유하지 않은 가구가 21,794가구, 1-99평이 6,020가구, 100-199평이 7,549가구, 200-299평이 5,020가구, 300-399평이 3,274가구, 400-499평이 1,368가구, 마지막으로 500평 이상은 4,975가구이다. 이 50,000가구에 대해서 대체를 하고난 후의 분포변화를 보면 대부분이 0.5%가 넘지 않는 정확한 대체가 되고 있음을 알 수 있다. 그리고 두 방법에서의 차이도 거의 보이지 않는데 이

러한 이유는 두 방법 모두가 분포가 유지되는 장점을 가진 핫덱 대체방법에 근간을 두고 있기 때문이다.

〈표 2-20〉 대체전후의 분포변화(비닐하우스 면적이 무응답인 경우)

범주	없음 (0)	1-99	100-199	200-299	300-399	400-499	500-
실제 총가구수	21794	6020	7549	5020	3274	1368	4975
응용 핫덱 대체방법							
대체후 총가구수	21635	5977	7625	5006	3321	1391	5045
절대차이 분포변화비율 ( )	159 **(0.318%)**	43 **(0.086%)**	76 **(0.152%)**	14 **(0.028%)**	47 **(0.094%)**	23 **(0.046%)**	70 **(0.140%)**
계층적 핫덱 대체방법							
대체후 총가구수	21564	6038	7576	5072	3289	1401	5060
절대차이 분포변화비율 ( )	230 **(0.460%)**	18 **(0.036%)**	27 **(0.054%)**	52 **(0.104%)**	15 **(0.030%)**	33 **(0.066%)**	85 **(0.170%)**

두 번째로 대체군 중에서 고추면적과 가장 연관성이 높은 항목인 밭면적이 무응답일 경우의 대체결과가 <표 2-21>에 정리되어 있다.

〈표 2-21〉 대체군에서 밭면적이 무응답인 경우의 대체결과

평균 전체가구 평 평균 면적보유가구 평

대체방법 (응용 핫덱 대체방법 **165.94**( ), 계층적 핫덱 대체방법 **294.21**( )

	전체가구	면적보유가구	전체가구	면적보유가구
대체후 평균				
절대차이	168.99	298.64	171.34	301.50
오차비율	3.05	4.43	5.40	7.29
대체정확도)	**(1.84%)**	**(1.51%)**	**(3.25%)**	**(2.48%)**
	62.77(%)		**55.06(%)**	

전체가구에 대해서 대체전후의 고추면적 평균의 차이를 보면 응용 핫덱 대체방법은 3.05로 1.84% 오차를 보이며, 계층적 핫덱 대체방법은 5.40로 3.25%의 오차를 가진다. 그리고 면적보유가구에 대해서 살펴보면 응용 핫덱 대체방법은 1.51%, 계층적 핫덱 대체방법은 2.48%의 오차

를 보이며, 대체의 정확도는 62.77%와 55.06%를 나타내고 있다. 이전의 결과와 패턴이 같은 것을 볼 수 있으며, 계층적 핫덱 대체방법은 대체군 중에서 시도와 시군구 정보만 이용하게 되므로 정보의 손실이 상당히 큰 것을 볼 수 있다. <표 2-22>에는 두 방법에 대해서 대체군 중에서 밭면적이 무응답일 경우의 대체전후의 분포변화 결과가 주어져있다. 앞의 경우와 마찬가지로 핫덱 대체방법의 특성상 분포변화의 차이는 두 방법에서 나타나지 않음을 알 수 있다.

〈표 2-22〉 대체전후의 분포변화(밭면적이 무응답인 경우)

범주	없음 (0)	1-99	100-199	200-299	300-399	400-499	500-
실제 총가구수	21794	6020	7549	5020	3274	1368	4975
응용 핫덱 대체방법							
대체후 총가구수	21679	6115	7451	4996	3308	1411	5040
절대차이	115	95	98	24	34	43	65
분포변화비율 ( )	**(0.230%)**	**(0.190%)**	**(0.196%)**	**(0.048%)**	**(0.068%)**	**(0.086%)**	**(0.130%)**
계층적 핫덱 대체방법							
대체후 총가구수	21522	6153	7492	5002	3347	1400	5084
절대차이	272	133	57	18	73	32	109
분포변화비율 ( )	**(0.544%)**	**(0.266%)**	**(0.114%)**	**(0.036%)**	**(0.146%)**	**(0.064%)**	**(0.218%)**

응용 핫덱 대체방법은 대체군의 활용 측면에서 계층적 핫덱 대체방법의 단점을 보완한 것이다. 하지만 밭면적을 무응답으로 가정한 모의실험은 모든 무응답 가구의 밭면적을 알지 못한다고 가정한 극단적인 실험으로 실제로는 두 방법의 차이가 본 결과보다는 많은 부분 줄어들 것이라고 생각되며, 인구주택총조사처럼 큰 조사에서는 차이가 거의 없을 가능성도 있을 것이다. 계층적 핫덱 대체방법을 사용하는 연구자들의 오해가 없었으면 하는 바람이다. 하지만 대체의 정확도를 높이기 위해서는 사용할 수 있는 정보를 최대한 활용해야 하므로 그러한 측면에서 개선된 방법이라 할 수 있다. 앞의 검토결과를 고려할 때 농업총조사 모든 항목들에 대하여 응용 핫덱 대체방법을 적용하는 것은 매우 적절하다고 판단된다.

3. 지역이 고려된 대체군들의 비교

본 연구에서는 CHAID 알고리즘을 이용하여 각 항목에 대한 대체군을 개발하였다. 이 과정에서 사용된 자료는 농업총조사 전체자료이다. 이 때 각 항목에서의 대체군은 지역내 특성이 고려된다면 지역별로 다소 차이가 생길 수도 있을 것이다. 이 경우 만약 16개 시도로만 나눈다고 하더라도 각 항목에서의 대체군은 16개가 되며, 전체항목으로 본다면 대략 2000개가 넘어 상당히 복잡하게 될지도 모른다. 하지만 지역을 고려하여 개발된 대체군을 사용했을 때의 대체결과가 현재의 결과보다 상당부분 좋다면 이러한 문제를 고려해야 할 것이다. 따라서 자료전체를 사용할 경우와 시도별로 분리하여 사용할 경우의 대체군을 비교하고, 각 대체군을 사용하였을 경우의 대체결과를 모의실험을 통하여 살펴볼 것이다. 비교할 항목은 연속형 항목 중에서 사과면적을, 범주형 항목 중에서 영농형태 항목을 이용하고 지역은 경기도, 전라북도, 경상북도, 충청북도 4곳에 대하여 기존의 대체결과와 비교하고자 한다.

가. 사과면적 항목에서의 비교

<표 2-23>에는 지역별 사과면적에 대한 대체군의 결과가 정리되어 있다. 각 지역별로 분리한 후 분석을 통해 선택된 대체군과 자료전체를 사용한 결과는 거의 차이가 나지 않음을 알 수 있다. 경기도와 경상북도는 시군구 항목이 추가되었고, 전라북도는 시군구와 SS분무기 항목이 교체되었으며, 충청북도는 동일한 대체군이 선택된 것을 볼 수 있다.

〈표 2-23〉 사과 항목에 대한 지역별 대체군

전체	영농형태 상세 밭면적 분무기 영농형태
경기도	**영농형태 (영농형태 상세, SS밭면적** ,시군구 **분무기**
전라북도	**영농형태 상세 밭면적 영농형태** ,시군구 , **SS**
경상북도	**영농형태 상세**), **분무기**, **밭면적 영농형태** 시군구
충청북도	**영농형태 상세**),**SS밭면적 분무기 영농형태** ,
	( ), , **SS** ,

사과면적에 대하여 전체 대체군을 사용한 경우와 지역별 대체군을 분리하여 사용한 경우의 대체결과가 <표 2-24>에 정리되어 있다. 표의 내용 중 실제값은 모의실험에 사용된 50,000가구의 실제 평균을 의미하며, 괄호안의 값들은 대체를 하고난 후의 평균을 말한다. 경기도의 경우 전체가구의 실제 평균은 8.04평, 면적보유가구는 1658.35평이다. 여기에서 전체 대체군을 사용했을 경우의 대체값은 7.84평, 1703.30평, 그리고 대체의 정확도는 99.33%로 나타났으며 경기도에서의 대체군을 사용했을 경우의 대체값은 7.92평, 1691.90평, 대체의 정확도는 99.43%로 조금의 개선이 있는 것으로 보인다. 하지만 이 차이는 매우 미미하게 보이며 전반적으로 다른 지역에서는 전체 대체군과 지역별 대체군의 사용에서 발생하는 차이는 거의 없는 것으로 판단된다.

〈표 2-24〉 지역별 대체군 적용결과(사과면적)

지역		전체 대체군 사용			지역별 대체군 사용		
		전체 가구	면적보유 가구	대체 정확도	전체 가구	면적보유 가구	대체 정확도
전체	실제값						
	대체후값	**70.58**	**2375.13**	**98.36%**			**98.40%**
	실제값	(70.56)	(2381.65)		(70.55)	(2380.72)	
경기도	대체후값	**8.04**	**1658.35**	**99.33%**			**99.43%**
	실제값	(7.84)	(1703.30)		(7.92)	(1691.90)	
전라북도	대체후값	**28.99**	**2850.16**	**99.07%**			**99.07%**
	실제값	(28.84)	(2817.29)		(28.84)	(2817.64)	
경상북도	대체후값	**260.97**	**2341.87**	**94.89%**			**94.59%**
	실제값	(261.68)	(2349.81)		(262.33)	(2350.54)	
충청북도	대체후값	**142.82**	**2574.83**	**96.30%**			**96.71%**
		(143.05)	(2580.99)		(142.11)	(2574.95)	

나. 영농형태 항목에서의 비교

<표 2-25>에는 지역별 영농형태에 대한 대체군의 결과가 정리되어 있다. 사과면적과 유사하게 중요 항목(영농형태(상세), 논벼면적, 한우수, 사과면적, 배면적)들은 대체군에 대부분 포함되어 있는 것을 알 수 있으며, 지역별 특성에 따라 몇몇 항목의 차이를 보이고 있다.

〈표 2-25〉 영농형태 항목에 대한 지역별 대체군

	대체군
전체	영농형태 상세 논벼면적 한우 수 사과면적 배면적 ( ), ( ), , 판매처구분 시설화훼면적 ,
경기도	**영농형태 상세 논벼면적 배면적 한우 수** 포도면적 ( ), **시설화훼면적** 복숭아면적 육우 수 , ( ), ( )
전라북도	**영농형태 상세** , **논벼면적 한우 수** 밭면적 **배면적** 인삼면적 ( ), **사과면적** 화훼면적 , ( ), , ,
경상북도	**영농형태 상세 논벼면적 사과면적 배면적 판매처구분** ( )
충청북도	**영농형태 상세 논벼면적** 과수활동여부 밭면적 **한우 수** ( ) **사과면적 판매처구분 배면적** , , ( ) , ,

영농형태에 대하여 전체 대체군을 사용한 경우와 지역별 대체군을 분리하여 사용한 경우의 대체결과가 <표 2-26>에 정리되어 있다. 영농형태는 범주형 항목이므로 대체의 정확도로 살펴볼 수 있다. 경상북도와 충청북도는 지역별 대체군을 사용했을 경우 정확도가 조금 높아진 반면 경기도와 전라북도는 낮아진 것을 볼 수 있다. 이는 모의실험을 할 때마다 약간의 차이가 생길 수도 있지만 종합적으로 살펴보면 지역을 고려한 대체군의 효과는 거의 나타나지 않음을 알 수 있다. 따라서 농업총조사 항목에 대한 대체군은 기존의 방식대로 전체 대체군을 사용해도 큰 문제가 없을 것으로 판단된다. 또한 대체전후의 분포변화 차이도 발생하지 않음을 확인하였다.

〈표 2-26〉 지역별 대체군 적용결과(영농형태)

지역	전체 대체군 사용	지역별 대체군 사용
	대체정확도	대체정확도
전체		
경기도	**88.05%**	**88.13%**
전라북도	**90.34%**	**89.91%**
경상북도	**88.46%**	**86.24%**
충청북도	**86.01%**	**86.65%**
	80.53%	**82.82%**

제5절 응용 핫덱 대체방법의 적용

앞에서 제시한 응용 핫덱 대체방법을 적용하여 과수, 작물 및 가축 항목들에 대하여 모의실험을 실시한다. 실험에서 사용할 2005년 농업총조사 자료는 총 1,272,908가구 중에서 약 50%에 해당하는 635,779가구이다. 먼저 CHAID 알고리즘을 이용하여 대체군을 결정하고 이를 근거로 모의대체를 실시하여 대체기법의 정확성을 검토한다. 본 실험을 위해서 각 항목에 대하여 임의로 50,000가구의 무응답을 발생시키고, 항목별로 50,000가구의 무응답을 모두 대체를 하고난 후에 표본평균의 차이정도와 범주화된 분포 변화 비율을 살펴본다.

본 연구에서 적용되는 항목은 72개 항목이지만 같은 형태의 분석이 반복되므로 모든 항목이 아닌 몇몇 항목에 대한 분석내용만 보고서에 작성할 것이다. 하지만 모든 항목에 대한 분석결과는 표를 통해서 제시하고자 한다.

1. 과수원에 관한 사항

가. 배면적

배면적에 대한 연관성 분석의 결과 밭면적이 가장 연관성이 높은 것으로 판단되며, 다음으로 영농형태(상세), SS분무기수, 영농형태의 순으로 배면적과 연관성이 높은 것으로 나타나 이 항목들을 대체군으로 사용한다. <표 2-27>에 제시된 분석결과의 해석은 『무응답 처리를 위한 방법론 연구(Ⅰ)(통계개발원, 2009)』에 자세하게 설명되어 있으니 참조하기 바란다.

배면적에 대한 종합적인 대체결과가 <표 2-28>에 정리되어 있다. 임의로 선택된 50,000가구의 실제 배면적 평균은 50.11평이며, 이 중에서 실제로 배면적을 보유하고 있는 가구의 배면적 평균은 1753.43평이다. 전체가구에 대해서 대체전후의 배면적 평균의 차이를 보면 0.07로 0.14%의 오차를 보이며, 면적보유가구에 대해서는 평균차이가 4.25로 0.15%의 오차가 나타난다. 따라서 대체로 인하여 일어날 수 있는 평균

의 변화는 거의 없는 것으로 판단된다. 그리고 실제 배면적을 보유한 가구가 제대로 파악이 되는지의 여부를 판단하기 위하여 배의 면적보유 유무에 대한 대체정확도를 살펴본 결과 97.37%로 상당히 높은 정확도를 가짐을 알 수 있다.

〈표 2-27〉 배 항목에 관한 연관성 분석의 세부내용

길이 연관성 ( )	분리변수	배 분리지점 좌 미만( )	분리지점 우 이상 ( )	연관비중
1	영농형태 상세 ( )(1)	6750	6750	1.0000
2	영농형태 상세 ( )(2)	1, 3~7	2	0.9628
	분무기 SS	미만	이상	
3	영농형태 (1)	0.5	0.5	0.6760
	영농형태 (2)	4, 7	2	0.4701
	(3)	2~4, 6, 8	1, 5, 7	

〈표 2-28〉 배면적에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 평 평균 면적보유가구 평 정확도

( ) : 50.11( ), ( ) : 1753.43( ), : 98.37%

무응답 가구수	대체후 평균 전체가구 ( )	절대차이 오차비율 ( )	대체후 평균 면적보유가구 ( )	절대차이 오차비율 ( )
50,000	50.04	0.07**(0.14%)**	1756.02	4.25**(0.15%)**

배 면적 실제 총가구수	없음 (0)	1-499	500-999	1000-1499	1500-1999	2000-2999	3000-
대체후 총가구수	48571	337	313	234	113	161	271
절대차이	48599	325	303	230	116	158	269
분포변화비율 ( )	28 **(0.056%)**	12 **(0.024%)**	10 **(0.020%)**	4 **(0.008%)**	3 **(0.006%)**	3 **(0.006%)**	2 **(0.004%)**

한 가지 더 고려할 사항은 대체전후의 분포변화이다. 50,000가구에 대해서 대체전후의 분포변화비율을 살펴본 결과 배면적을 보유하지 않은 가구는 0.056%, 1-499평의 가구는 0.024%, 500-999평의 가구는 0.020%, 1000-1499평의 가구는 0.008%, 1500-1999평의 가구는 0.006%, 2000-2999평의 가구는 0.006%, 3000평 이상의 가구는 0.004% 정도의 분

포가 변화되는 것으로 나타나 분포상의 문제는 거의 일어나지 않음을 알 수 있다.

나. 포도면적

포도면적에 대한 연관성 분석의 결과는 <표 2-29>에 제시되어 있다. 영농형태(상세)가 포도면적과 가장 연관성이 높으며, 다음으로 밭면적, 영농형태, 시설포도면적, 시도, 콩면적의 순으로 나타났다.

〈표 2-29〉 포도 항목에 관한 연관성 분석의 세부내용

포도

깊이 연관성	분리변수	분리지점 좌	분리지점 우	연관비중
( )	영농형태 상세	( )	( )	
1	콩면적 ( )	1~4, 5~7	이상	1.0000
2	밭면적 (1)	95미만	95이상	0.1349
	영농형태 (2)	2265	2265	0.7704
3	시도 (3)	2	7	0.4333
	시설포도(4)	미만	이상	0.1721
4	(5)	915	915	0.2741

〈표 2-30〉 포도면적에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 평 평균 면적보유가구 평 정확도

무응답) : 대체후 평균 (절대차이), 대체후 평균 (절대차이), 98.25%

가구수 가구	전체가구 ( )	오차비율 ( )	면적보유가구 ( )	오차비율 ( )
50,000	37.95	0.12(0.32%)	1291.34	5.07(0.39%)

포도 면적 없음

실제 총가구수	(0)	1-499	500-999	1000-1499	1500-1999	2000-2999	3000-
대체후 총가구수	48488	250	454	308	162	200	138
절대차이	48458	232	480	318	182	184	146
분포변화비율 ( )	34 (0.068%)	18 (0.036%)	26 (0.052%)	10 (0.020%)	20 (0.040%)	16 (0.032%)	8 (0.016%)

<표 2-30>에 제시된 결과를 살펴보면 실제 포도면적 평균은 37.83평이며, 이 중에서 실제로 포도면적을 보유하고 있는 가구의 포도면적 평

균은 1286.27평이다. 전체가구에 대해서 대체전후의 포도면적 평균의 차이를 보면 0.12로 0.32%의 오차, 면적보유가구에 대해서는 평균차이가 5.07로 0.39%의 오차가 나타났다. 또한 포도의 면적보유 유무에 대한 대체정확도는 98.25%로 정확한 대체가 되고 있음을 알 수 있다. 그리고 대체전후의 분포변화비율을 살펴보면 포도면적을 보유하지 않은 가구는 0.068%, 1-499평의 가구는 0.036%, 500-999평의 가구는 0.052%, 1000-1499평의 가구는 0.020%, 1500-1999평의 가구는 0.040%, 2000-2999평의 가구는 0.032%, 3000평 이상의 가구는 0.016% 정도로 다른 항목들처럼 대체한 후에 분포상의 문제는 없는 것으로 보인다.

다. 과수원 항목의 대체군 및 대체결과

<표 2-31>에는 배, 포도면적을 제외한 과수원면적 항목들의 대체군 및 대체결과가 주어져있다. 대체결과는 전체가구 및 면적보유가구의 평균변화와 대체정확도로 구성되어있다. 결과의 가장 위의 값은 실제평균값이며 괄호안의 값은 대체후의 평균값이고, 가장 아랫부분은 오차비율을 나타내고 있다. 과수원 항목들은 대체적으로 정확한 대체가 되고 있음을 알 수 있다. 그리고 과수원 항목과 연관성이 높은 항목들은 밭면적, 영농형태, 영농형태(상세) 등으로 이 항목들은 대체시 매우 중요한 정보를 제공한다는 것을 보여준다. 또한 대체전후의 분포변화는 거의 나타나지 않아 대체한 후에 분포상의 문제는 없는 것으로 보인다.

2. 노지재배 수확작물에 관한 사항

가. 콩면적

콩면적에 대한 연관성 분석의 결과 밭면적이 가장 연관성이 높은 것으로 판단되며, 다음으로 영농형태, 시도, 시군구, 고추면적, 감자면적, 마늘면적, 영농형태(상세) 순으로 콩면적과 연관성이 높은 것으로 나타나 이 항목들을 대체군으로 사용할 것이다. 세부 분석내용은 <표 2-32>를 참조하기 바란다.

〈표 2-31〉 과수원 항목의 대체군 및 대체결과

항목	전체	면적 보유	대체 정확도	대체군
사과	70.58 (70.31) **0.38%**	2375.13 (2370.88) **0.18%**	**98.41%**	영농형태 상세 밭면적 분무기 영농형태 ( ), , SS
복숭아	34.44 (34.28) **0.46%**	1311.50 (1307.19) **0.33%**	**97.74%**	영농형태 상세 밭면적 과수활동여부 영농형태 시군구, 시도 , , ,
단감	34.60 (34.34) **0.75%**	968.01 (960.80) **0.74%**	**95.74%**	밭면적 시도 영농형태 상세 임가구분 시군구 , , ( ),
떫은감	18.34 (18.09) **1.36%**	788.36 (783.51) **0.62%**	**97.85%**	자기밭면적 영농형태 상세 임가구분 밭면적 시군구, 시도 영농형태 판매처 , , , ,
감귤	43.94 (43.96) **0.05%**	2588.69 (2591.51) **0.11%**	**99.80%**	시도 밭면적 영농형태 상세 판매금액 시설감귤면적 과수활동여부 콩면적, , ,
자두	11.70 (11.59) **0.94%**	820.88 (814.98) **0.72%**	**98.21%**	밭면적 시군구 시도 영농형태 상세 읍면동 영농형태, , , ( ) ,
키위	1.66 (1.63) **1.81%**	936.84 (930.21) **0.71%**	**99.74%**	시군구 시도 읍면동 영농형태 상세 자기밭면적 조사구번호, , ( ) , , ,
매실	8.82 (8.66) **1.81%**	554.54 (550.44) **0.74%**	**98.42%**	자기밭면적 영농형태 상세 시군구 밭면적 시도 , 읍면동 판매처구분 단감면적 , ( ), 자두면적 시도 , 밭면적 , 분무기
살구	0.46 (0.459) **0.22%**	307.82 (313.34) **1.79%**	**99.83%**	영농형태 상세 복숭아면적 자기밭면적 배실면적 , , , SS , ( ), ,
유자	2.37 (2.38) **0.42%**	571.20 (574.60) **0.60%**	**99.44%**	시군구 밭면적 시도 자기밭면적 영농형태 상세 , , , , ( )

<표 2-33>에 제시된 결과를 살펴보면 실제 콩면적 평균은 190.13평이며, 이 중에서 실제로 콩면적을 보유하고 있는 가구의 콩면적 평균은 397.94평이다. 전체가구에 대해서 대체전후의 콩면적 평균의 차이를 보면 0.53로 0.28%의 오차, 면적보유가구에 대해서는 평균차이가 0.04로 0.01%의 오차가 나타났다. 또한 콩의 면적보유 유무에 대한 대체정확도는 75.81%로 다른 항목들에 비해서는 다소 정확성이 떨어지는 것을 알 수 있다.

〈표 2-32〉 콩 항목에 관한 연관성 분석의 세부내용

값이 연관성 ( )	분리변수	분리지점 좌 ( )	분리지점 우 ( )	연관비중
1	콩 밭면적	미만 4950	이상 4950	1.0000
2	영농형태 (1)	1~5, 7, 8	6	0.5498
	시도 (2)	미만	이상	0.5416
3	고추면적 (1)	190 미만	190 이상	0.2703
	감자면적 (2)	110	110	0.2630
	시군구 (3)	미만	이상	0.4246
4	마늘면적 (1)	1425	1425	0.1962
	영농형태 상세 ( )(2)	1	2~7	0.1457

그리고 대체전후의 분포변화비율을 살펴보면 콩면적을 보유하지 않은 가구는 0.856%, 1-99평의 가구는 0.392%, 100-199평의 가구는 0.176%, 200-299평의 가구는 0.204%, 300-399평의 가구는 0.028%, 400-499평의 가구는 0.084%, 500평 이상의 가구는 0.028% 정도로 다른 항목들처럼 대체 후에 분포상의 문제는 발생하지 않는 것으로 판단된다.

〈표 2-33〉 콩면적에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 평 평균 면적보유가구 평 정확도

무응답) : **190.13**( ), ( ) **397.94**( ), **75.81%**

가구수 가구	전체가구 ( )	대체후 평균	오차비율 ( )	절대차이	면적보유가구 ( )	대체후 평균	오차비율 ( )	절대차이
50,000	189.60		0.53**(0.28%)**		397.98		0.04**(0.01%)**	

콩 면적 실제 총가구수	없음 (0)	1-99	100-199	200-299	300-399	400-499	500-
대체후 총가구수	26018	5398	5848	3970	2404	1188	5174
절대차이	26446	5202	5760	3868	2390	1146	5188
분포변화비율 ( )	428 **(0.856%)**	196 **(0.392%)**	88 **(0.176%)**	102 **(0.204%)**	14 **(0.028%)**	42 **(0.084%)**	14 **(0.028%)**

나. 감자면적

감자면적에 대한 연관성 분석의 결과 밭면적이 가장 연관성이 높은 것으로 판단되며, 다음으로 영농형태, 시도, 시군구, 콩면적, 읍면동, 남

의밭면적 순으로 감자면적과 연관성이 높은 것으로 나타났다. 세부 분석내용은 <표 2-34>를 참조하기 바란다.

〈표 2-34〉 감자 항목에 관한 연관성 분석의 세부내용

깊이 연관성 ( )	분리변수	감자 분리지점 좌 ( )	분리지점 우 ( )	연관비중
1	밭면적	8800 미만	8800 이상	1.0000
2	밭면적 (1) 영농형태 (2)	3999 미만 1~5, 7, 8	3999 이상 6	0.7085
3	시군구 (1) 읍면동 (2) 콩면적 (3) 남의밭 (4)	미만 75 미만	이상 75이상	0.6418 0.5418 0.2642 0.4499
4	(3)	19900 미만	19900 이상	0.2206

<표 2-35>에 제시된 결과를 살펴보면 감자의 면적보유 유무에 대한 대체정확도는 80.48%이며 실제 감자면적 평균은 62.10평, 이 중에서 실제로 감자면적을 보유하고 있는 가구의 감자면적 평균은 297.24평이다. 두 경우에 대하여 대체전후의 감자면적 평균의 오차비율은 0.35%와 0.17%로 다른 항목들과 큰 차이를 보이지 않는다.

〈표 2-35〉 감자면적에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 평 평균 면적보유가구 평 정확도

무응답) :대체후 평균 (절대차이), 대체후 평균), 절대차이 **80.48%**

가구수 가구	전체가구 ( )	오차비율 ( )	면적보유가구 ( )	오차비율 ( )
50,000	61.88	0.22**(0.35%)**	296.73	0.51**(0.17%)**

감자 면적

실제 총가구수	(0)	1-24	25-49	50-99	100-199	200-299	300-
대체후 총가구수	39354	2502	1050	2626	2144	750	1574
절대차이	39546	2422	990	2580	2074	748	1640
분포변화비율 ( )	192 **(0.384%)**	80 **(0.160%)**	60 **(0.120%)**	46 **(0.092%)**	70 **(0.140%)**	2 **(0.004%)**	66 **(0.132%)**

대체전후의 분포변화비율을 살펴보면 감자면적을 보유하지 않은 가구는 0.384%, 1-24평의 가구는 0.160%, 25-49평의 가구는 0.120%, 50-99평의 가구는 0.092%, 100-199평의 가구는 0.140%, 200-299평의 가구는 0.004%, 300평 이상의 가구는 0.132%로 대체 후의 분포변화는 최대 0.4%가 넘지 않는 것으로 보인다. 그러나 50,000가구가 아닌 모든 가구를 대상으로 생각하면 분포변화는 이 수치보다 훨씬 더 줄어드는 것을 알 수 있다.

다. 노지재배 수확작물 항목의 대체군 및 대체결과

<표 2-36>에는 콩, 감자면적을 제외한 노지재배 수확작물면적 항목들의 대체군 및 대체결과가 주어져있다. 노지재배 수확작물 항목들은 대체적으로 과수원 항목들에 비해서는 대체의 정확도가 다소 낮은 것을 볼 수 있다. 이는 과수원 항목들과 비교할 때 면적보유가구의 비율이 높기 때문에 발생하는 현상으로 판단된다. 그리고 밭면적과 지역적인 정보가 노지재배 수확작물면적을 대체하는 데 상대적으로 중요한 역할을 하고 있음을 알 수 있다. 또한 대체전후의 분포변화도 크지 않음을 확인하였다.

3. 노지재배 판매작물에 관한 사항

가. 시금치면적

시금치면적에 대한 연관성 분석의 결과는 <표 2-37>에 제시되어 있다. 시군구항목이 시금치면적과 가장 연관성이 높으며, 다음으로 이모작논면적, 읍면동, 밭면적, 농업특성조사구, 논면적, 시도 항목들과 연관성이 높은 것으로 나타났다.

<표 2-38>에 제시된 대체결과를 살펴보면 실제 시금치면적 평균은 7.66평이며, 이 중에서 실제로 면적을 보유하고 있는 가구의 시금치면적 평균은 392.36평이다. 전체가구에 대해서 대체전후의 시금치면적 평균의 차이를 보면 0.04로 0.52%의 오차, 면적보유가구에 대해서는 평균차이가 2.35로 0.60%의 오차가 나타났다. 또한 시금치의 면적보유 유무에 대한 대체정확도는 97.64%로 상당히 높은 정확도를 보여준다.

〈표 2-36〉 노지재배 수확작물 항목의 대체군 및 대체결과

항목	전체	면적 보유	대체 정확도	대체군
논벼	2240.79 (2238.34) **0.11%**	3050.32 (3047.14) **0.10%**	**99.47%**	논면적
보리	95.72 (95.58) **0.15%**	2700.56 (2701.28) **0.03%**	**96.97%**	이모작논면적 시군구 시도 논벼면적 영농형태 논면적 , , ,
옥수수	24.09 (23.68) **1.70%**	269.361 (266.85) **0.93%**	**90.49%**	시도 밭면적 영농형태 남의밭면적 감자면적 콩면적 시군구 , , ,
팥	10.41 (10.43) **0.19%**	104.80 (105.64) **0.80%**	**85.58%**	밭면적 콩면적 시도 농업종사경력 꿀벌 통 ,옥수수면적 , ( ),
고구마	45.82 (45.24) **1.27%**	195.93 (193.70) **1.14%**	**75.13%**	밭면적 남의밭면적 시도 시군구 영농형태, , , 밭면적 시도 남의밭면적 배추면적
김장무	42.45 (42.04) **0.97%**	98.38 (97.92) **0.47%**	**84.54%**	농업특성조사구, 자기밭면적, 시군구 영농형태 상세 , , ( )
김장배추	59.85 (59.54) **0.52%**	125.53 (125.39) **0.11%**	**88.03%**	시군구 밭면적 영농형태 무면적 읍면동 시도 , , , ,
고추	168.66 (168.76) **0.06%**	300.68 (300.82) **0.05%**	**80.14%**	밭면적 건조기수 비닐하우스면적 영농형태 상세 영농형태 시도 시군구 , ,
양파	33.59 (33.37) **0.65%**	413.25 (411.87) **0.33%**	**90.23%**	시군구 밭면적 읍면동 이모작논면적 영농형태 상세 콩면적 남의밭면적 마늘면적, , ,
대파	20.73 (20.55) **0.87%**	158.43 (156.90) **0.97%**	**82.93%**	시군구 영농형태 상세 밭면적 읍면동 남의밭면적 시도 , ( ), ,
마늘	72.40 (72.07) **0.46%**	241.91 (241.10) **0.33%**	**80.09%**	시군구 이모작논면적 콩면적 시도 영농형태 상세 밭면적 양파면적 읍면동 영농형태, , ,
참깨	53.93 (53.71) **0.41%**	172.77 (172.68) **0.05%**	**79.29%**	( ), , , 콩면적 밭면적 시도 고추면적 시군구 조사구번호 양배추면적 마늘면적 , , , ,
인삼	34.82 (34.47) **1.01%**	3243.50 (3200.99) **1.31%**	**98.58%**	밭면적 영농형태 남의밭면적 임가구분 비닐하우스면적 시도 , , , ,

〈표 2-37〉 시금치 항목에 관한 연관성 분석의 세부내용

깊이 (연관성)	분리변수	시금치 분리지점 좌 ( )	분리지점 우 ( )	연관비중
1	시군구			1.0000
2	조사구특성번호 (1) 이모작논 (2) 시군구	1, A 미만 2176.5	2 이상 2176.5	0.4337 0.7041
3	읍면동 (1) 시군구 (3) 시도 (4)			0.6249
4	논면적 (5) 밭면적 (7) (8)	미만 1610 미만 1589.5	이상 1610 이상 1589.5	0.2115 0.2629 0.4722

〈표 2-38〉 시금치면적에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 평 평균 면적보유가구 평 정확도

( ) : 7.66( ), ( ) : 392.36( ), : 97.64%

무응답 가구수	대체후 평균 전체가구 ( )	절대차이 오차비율 ( )	대체후 평균 면적보유가구 ( )	절대차이 오차비율 ( )
50,000	7.62	0.04**(0.52%)**	390.01	2.35**(0.60%)**

시금치 면적 실제 총가구수	(0)	1-49	50-99	100-199	200-299	300-499	500-
대체후 총가구수	48998	232	184	150	104	116	216
절대차이	49034	222	168	164	96	104	212
분포변화비율 ( )	36 **(0.072%)**	10 **(0.020%)**	16 **(0.032%)**	14 **(0.028%)**	8 **(0.016%)**	12 **(0.024%)**	4 **(0.008%)**

그리고 대체전후의 분포변화비율을 살펴보면 시금치면적을 보유하지 않은 가구는 0.072%, 1-49평의 가구는 0.020%, 50-99평의 가구는 0.032%, 100-199평의 가구는 0.028%, 200-299평의 가구는 0.016%, 300-499평의 가구는 0.024%, 500평 이상의 가구는 0.008% 정도로 다른 항목들과 비슷한 수준의 분포변화를 보인다.

나. 땅콩면적

땅콩면적에 대한 연관성 분석의 결과 시군구 항목이 가장 연관성이 높은 것으로 나타났으며, 다음으로 읍면동, 밭면적, 시도, 참깨면적, 마늘면적, 영농형태 순으로 땅콩면적과 연관성이 높은 것으로 나타났다. 세부 분석내용은 <표 2-39>를 참조하기 바란다.

〈표 2-39〉 땅콩 항목에 관한 연관성 분석의 세부내용

깊이 연관성	분리변수	분리지점 좌	분리지점 우	연관비중
( )	시군구	( )	( )	
1	밭면적	미만	이상	1.0000
2	읍면동 (1)	4900	4900	0.9519
	영농형태 (2)			0.9615
3	참깨면적 (1)	미만	이상	0.2596
	시도 (2)	190	190	0.3883
	마늘면적 (3)	미만	이상	0.6759
4	(5)	750	750	0.3339

〈표 2-40〉 땅콩면적에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 평 평균 면적보유가구 평 정확도

무응답 ( ) : 3.82( ) (절대차이) : 344.13( ), : 98.22%

가구수 가구	전체가구 ( )	대체후 평균	오차비율 ( )	면적보유가구 ( )	대체후 평균	오차비율 ( )
50,000	3.79		0.03**(0.79%)**	343.90		0.23**(0.07%)**

땅콩 면적 없음

실제 총가구수	(0)	1-99	100-199	200-299	300-399	400-499	500-
대체후 총가구수	49470	108	90	134	54	28	116
절대차이	49498	100	86	124	54	32	106
오차비율 ( )	28 **(0.056%)**	8 **(0.016%)**	4 **(0.008%)**	10 **(0.020%)**	0 **(0.000%)**	4 **(0.008%)**	10 **(0.020%)**

종합적인 대체결과는 <표 2-40>에 정리되어 있다. 실제 땅콩면적 평균은 3.82평이며, 이 중에서 실제로 땅콩면적을 보유하고 있는 가구의 평균은 344.13평이다. 두 경우에 대하여 대체전후의 땅콩면적 평균의 오

차비율은 0.79%와 0.07%로 차이가 많지 않음을 알 수 있다. 또한 땅콩의 면적보유 유무에 대한 대체정확도는 98.22%로 높게 나타나고 있다. 대체전후의 분포변화비율을 살펴보면 땅콩면적을 보유하지 않은 가구는 0.056%, 1-99평의 가구는 0.016%, 100-199평의 가구는 0.008%, 200-299평의 가구는 0.020%, 300-399평의 가구는 0.000%, 400-499평의 가구는 0.008%, 500평 이상의 가구는 0.020%로 대체 후에도 분포도의 변화가 거의 없음을 볼 수 있다.

다. 노지재배 판매작물 항목의 대체군 및 대체결과

〈표 2-41〉 노지재배 판매작물 항목의 대체군 및 대체결과

항목	전체	면적보유	대체 정확도	대체군
양배추	9.62 (9.63) **0.10%**	1825.87 (1829.19) **0.18%**	**99.37%**	시도 밭면적 감자면적 영농형태 읍면동 영농형태 상세 , 시군구 조사구번호 , ( ), , 밭면적 남의밭면적 시금치면적 시도
상추	2.29 (2.28) **0.44%**	176.23 (178.67) **1.38%**	**98.45%**	시군구 시설상추면적 , 배추면적 , , ,
쑥갓	0.2504 (0.2502) **0.08%**	53.33 (53.41) **0.15%**	**99.54%**	상추면적 시금치면적 시도 오이면적 , , , 읍면동 시도 호박면적 시군구 밭면적
오이	2.64 (2.60) **1.52%**	364.62 (368.74) **1.13%**	**99.02%**	채소활동여부 영농형태 상세 남의밭면적 , , , , ( ), 시도 시군구 읍면동 무면적 밭면적
수박	8.14 (8.07) **0.86%**	1745.53 (1736.57) **0.51%**	**99.41%**	영농형태 조사구번호 , , , , 남의밭면적 시도 밭면적 팥면적 시군구
호박	6.82 (6.76) **0.88%**	487.99 (489.02) **0.21%**	**97.66%**	채소활동여부 한우수 영농형태 상세 , , , , ( ) 읍면동 밭면적 행정리 시도 남의밭면적
당근	6.50 (6.51) **0.15%**	1436.28 (1440.80) **0.31%**	**99.40%**	조사구번호 , , , , 시도 밭면적 콩면적 옥수수면적 시군구
들깨	16.44 (16.30) **0.85%**	259.85 (259.71) **0.05%**	**90.39%**	참깨면적 고구마면적 , , , , 밭면적 영농형태 임가구분
화훼	17.33 (17.18) **0.87%**	2606.59 (2608.64) **0.08%**	**99.29%**	, ,

<표 2-41>에는 시금치, 땅콩면적을 제외한 노지재배 판매작물면적 항목들의 대체군 및 대체결과가 주어져있다. 항목들 대부분이 상당히 높은 대체의 정확도를 가짐을 볼 수 있다. 하지만 들깨 항목은 다른 항목에 비하여 다소 낮은 정확도를 가진다. 대체군을 보면 밭면적과 지역적인 특성이 주로 포함되어 있는 것을 알 수 있다. 또한 대체전후의 분포도 변화가 거의 나타나지 않았음을 알려둔다.

4. 시설재배 수확작물에 관한 사항

가. 상추면적(시설)

상추면적에 대한 연관성 분석의 결과는 <표 2-42>에 제시되어 있다. 비닐하우스면적이 상추면적과 가장 연관성이 높으며, 다음으로 시군구, 영농형태(상세), 읍면동, 시도, 시설시금치면적, 시설토마토면적 순으로 연관성이 높은 것으로 나타났다.

<표 2-42> 상추(시설) 항목에 관한 연관성 분석의 세부내용

상추_시설

깊이	연관성	분리변수	분리지점 좌	분리지점 우	연관비중
	( )	비닐하우스	미만 ( )	이상 ( )	
1		비닐하우스 (1)	775 미만	775 이상	1.0000
2		시군구 (1) 영농형태 상세 (2)	140	140	0.8940
3		읍면동( )(1) 시설시금치면적 (2) 비닐하우스 (3) 시도 (4) 시설토마토면적 (5)	1, 3, 4~7 미만 375 미만 3950	2 이상 375 미만	0.3648 0.3481 0.3206
4		(7)	25	25	0.3412 0.2795

<표 2-43>에 제시된 결과를 살펴보면 실제 상추면적 평균은 8.61평이며, 이 중에서 실제로 면적을 보유하고 있는 가구의 평균은 771.43평이다. 전체가구에 대해서 대체전후의 상추면적 평균의 차이를 보면 0.06으로 0.70%의 오차, 면적보유가구에 대해서는 평균차이가 1.32로 0.17%의 오차가 나타났다. 또한 상추의 면적보유 유무에 대한 대체정확도는

98.64%로 높게 나타나고 있다. 그리고 대체전후의 분포변화비율을 살펴보면 상추면적을 보유하지 않은 가구는 0.146%, 1-9평의 가구는 0.012%, 10-29평의 가구는 0.040%, 30-49평의 가구는 0.010%, 50-99평의 가구는 0.014%, 100-999평의 가구는 0.038%, 1000평 이상의 가구는 0.032% 정도로 분포변화는 거의 일어나지 않음을 알 수 있다.

〈표 2-43〉 상추면적(시설)에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 ( ) : 8.61( ), 평 평균 면적보유가구 ( ) : 771.43( ), 평 정확도 : 98.64%							
무응답 가구수 가구	대체후 평균 전체가구 ( )	절대차이 오차비율 ( )	대체후 평균 면적보유가구 ( )	절대차이 오차비율 ( )			
50,000	8.55	0.06**(0.70%)**	772.75	1.32**(0.17%)**			
상추 시설 면적 ( )	없음 (0)	1-9	10-29	30-49	50-99	100-999	1000-
실제 총가구수	49448	74	100	24	36	184	134
대체후 총가구수	49521	68	80	19	29	165	118
절대차이 오차비율 ( )	73 **(0.146%)**	6 **(0.012%)**	20 **(0.040%)**	5 **(0.010%)**	7 **(0.014%)**	19 **(0.038%)**	16 **(0.032%)**

나. 수박면적(시설)

수박면적에 대한 연관성 분석의 결과 비닐하우스면적 항목과 가장 연관성이 높은 것으로 판단되며, 다음으로 시군구, 시도, 판매처구분, 영농형태(상세) 순으로 수박면적과 연관성이 높은 것으로 나타났다. 이는 수박면적을 추정할 때에는 비닐하우스면적과 지역적인 정보가 상당히 많은 영향을 미치는 것을 알 수 있다. 세부 분석내용은 <표 2-44>를 참조하기 바란다.

종합적인 대체결과는 <표 2-45>에 정리되어 있다. 실제 수박면적 평균은 29.42평이며, 이 중에서 실제로 수박면적을 보유하고 있는 가구의 평균은 1981.30평이다. 두 경우에 대하여 대체전후의 수박면적 평균의 오차비율은 0.34%와 0.14%, 대체의 정확도는 98.64% 정도로 적절한 대체가 되고 있음을 알 수 있다.

〈표 2-44〉 수박(시설) 항목에 관한 연관성 분석의 세부내용

수박 시설 ( )

깊이 연관성 ( )	분리변수	분리지점 회 미만( )	분리지점 우 이상( )	연관비중
1	비닐하우스 영농형태 상세 ( )(1)	1550	1550	1.0000
2	시군구 (2) 시도 (3)	1, 3~7	2	0.2393 0.7277
3	비닐하우스 판매처구분 (4)	미만 3115	이상 3115	0.4824
4	(5)	1~4, 6~11	5	0.4144

〈표 2-45〉 수박면적(시설)에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 ( ) : 29.42( ), 평 평균 면적보유가구 ( ) : 1981.30( ), 평 정확도 : 98.62%

무응답 가구수 가구	대체후 평균 전체가구 ( )	절대차이 오차비율 ( )	대체후 평균 면적보유가구 ( )	절대차이 오차비율 ( )
50,000	29.32	0.10**(0.34%)**	1984.11	2.81**(0.14%)**

수박 시설 면적 없음 ( )

실제 총가구수	(0)	1-499	500-999	1000-1499	1500-1999	2000-2999	3000-
대체후 총가구수	49266	56	116	170	92	130	170
절대차이	49231	66	132	164	106	139	162
오차비율 ( )	35 **(0.070%)**	10 **(0.020%)**	16 **(0.032%)**	6 **(0.012%)**	14 **(0.028%)**	9 **(0.018%)**	8 **(0.016%)**

대체전후의 분포변화비율을 살펴보면 수박면적을 보유하지 않은 가구는 0.070%, 1-499평의 가구는 0.020%, 500-999평의 가구는 0.032%, 1000-1499평의 가구는 0.012%, 1500-1999평의 가구는 0.028%, 2000-2999평의 가구는 0.018%, 3000평 이상의 가구는 0.016%로 분포변화는 거의 일어나지 않는다.

다. 시설재배 수확작물 항목의 대체군 및 대체결과

<표 2-46>과 <표 2-47>에는 상추, 수박면적을 제외한 시설재배 수확작물면적 항목들의 대체군 및 대체결과가 주어져있다. 대부분의 항목들이 높은 정확도를 가짐을 알 수 있다.

〈표 2-46〉 시설재배 수확작물 항목의 대체군 및 대체결과

항목	전체 면적	보유 면적	대체 정확도	대체군
무	3.18 (3.14) **1.26%**	677.05 (683.65) **0.97%**	**99.51%**	비닐하우스면적 시도 시군구 시설배추면적 시설수박면적 배면적 , ,
배추	5.35 (5.30) **0.93%**	634.75 (640.49) **0.90%**	**98.83%**	시설시금치면적 비닐하우스면적 시도 읍면동 시군구 시설무면적 , , ,
시금치	6.39 (6.31) **1.25%**	933.38 (926.30) **0.76%**	**99.10%**	시군구 비닐하우스면적 읍면동 시도 시설배추면적 남의밭면적 , , 시도 농업용난방기수 비닐하우스면적
토마토	10.86 (10.76) **0.92%**	1020.44 (1018.82) **0.16%**	**98.69%**	영농형태 상세 밭면적 시군구 읍면동 영농형태 ( ), , ,
오이	10.39 (10.30) **0.87%**	477.94 (476.02) **0.40%**	**97.78%**	시군구 시도 농업용난방가수 영농형태 상세 읍면동 비닐하우스면적 ( ), ,
딸기	13.25 (13.31) **0.45%**	1243.82 (1239.54) **0.34%**	**99.17%**	비닐하우스면적 시도 시군구 읍면동 판매처 농업용난방기수 영농형태 상세 , , , ( )
참외	13.81 (13.80) **0.07%**	1724.29 (1727.03) **0.16%**	**99.65%**	시군구 비닐하우스면적 시도 읍면동 시설수박면적 , ,
호박	5.31 (5.32) **0.19%**	901.75 (906.11) **0.48%**	**99.26%**	시군구 비닐하우스면적 읍면동 시설수박면적 시도 자기밭면적 채소관련활동 ,
고추	10.39 (10.32) **0.67%**	477.94 (474.46) **0.73%**	**97.82%**	시도 비닐하우스면적 시군구 농업용난방기수 읍면동 건조기수 , , ,
대파	13.25 (13.24) **0.08%**	1243.82 (1238.40) **0.44%**	**99.03%**	배파면적 비닐하우스면적 읍면동 시도 , , ,
서양채소	4.37 (4.33) **0.92%**	1574.18 (1570.60) **0.23%**	**99.65%**	시군구 읍면동 비닐하우스면적 시도 , , ,
포도	4.17 (4.14) **0.72%**	1290.24 (1287.50) **0.21%**	**99.77%**	비닐하우스면적 자동화비닐하우스면적 영농형태 상세 ( ),
감귤	5.02 (5.04) **0.40%**	1523.68 (1524.67) **0.06%**	**99.94%**	비닐하우스면적 자동화비닐하우스면적 시도 영농형태 상세 , , ( )

〈표 2-47〉 시설재배 수확작물 항목의 대체군 및 대체결과(계속)

항목	전체 면적보유		대체 정확도	대체군
메론	3.11 (3.09) **0.64%**	1460.03 (1452.78) **0.50%**	**99.81%**	시도 시군구 비닐하우스면적 읍면동 농업특성조사구, 판매처구분 , ,
버섯	6.18 (6.13) **0.81%**	943.14 (944.69) **0.16%**	**99.37%**	기타시설면적 비닐하우스면적 영농형태 판매금액구분 전겸업수입구분 , ,
감자	3.79 (3.78) **0.26%**	1369.08 (1368.14) **0.07%**	**99.67%**	이모작논면적 영농형태 비닐하우스면적 읍면동 감자면적 시군구 , , ,
화훼	8.35 (8.30) **0.60%**	1318.91 (1325.83) **0.52%**	**99.77%**	영농형태 밭면적 비닐하우스면적 농업용난방기수 ,

5. 가축에 관한 사항

가. 한우

한우수에 대한 연관성 분석의 결과는 <표 2-48>에 제시되어 있다. 판매금액구분이 한우수와 가장 연관성이 높으며, 다음으로 영농형태, 영농형태(상세), 승용차보유여부, 화물차보유여부 등의 순으로 나타났다. 여기서 한우수를 추정할 때에는 판매금액과 영농형태가 중요한 역할을 하는 것을 알 수 있다.

〈표 2-48〉 한우 항목에 관한 연관성 분석의 세부내용

한우

값이 연관성 ( )	분리변수	분리지점 좌 ( )	분리지점 우 ( )	연관비중
1	판매금액구분 영농형태	1~9	10, 11, 12	1.0000
2	승용차보유 (1) 영농형태 상세 (2)	1~6, 8 1	7 M	0.7353 0.2382
3	화물차보유 ( )(2) (3)	1 1	2~7 M	0.7017 0.1987

<표 2-49>에 제시된 결과를 살펴보면 실제 한우수의 평균은 1.275마리이며, 이 중에서 실제로 한우를 보유하고 있는 가구의 평균은 8.73마

리이다. 전체가구에 대해서 대체전후의 한우수의 평균의 차이를 보면 0.002로 0.16%의 오차, 한우보유가구에 대해서는 평균차이가 0.01로 0.11%의 오차가 나타나 매우 정확한 대체가 되고 있음을 보여준다. 또한 한우의 보유유무에 대한 대체정확도는 82.80%로 나타나고 있다.

〈표 2-49〉 한우에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 ( ) : 1.275( ), 마리		**평균 보유가구 ( ) : 8.73( ), 마리**		**정확도 : 82.80%**
무응답 가구수	대체후 평균 전체가구 ( )	절대차이 오차비율 ( )	대체후 평균 보유가구 ( )	절대차이 오차비율 ( )
50,000 가구	1.277	0.002**(0.16%)**	8.74	0.01**(0.11%)**

한우 마리수	(0)	1	2-4	5-9	10-19	20-49	50-
실제 총가구수	42574	1592	2856	1218	854	642	264
대체후 총가구수	42588	1552	2908	1226	824	628	274
절대차이 오차비율 ( )	14 **(0.028%)**	40 **(0.080%)**	52 **(0.104%)**	8 **(0.016%)**	30 **(0.060%)**	14 **(0.028%)**	10 **(0.020%)**

대체전후의 분포변화비율을 살펴보면 한우를 보유하지 않은 가구는 0.028%, 1마리인 가구는 0.080%, 2-4마리의 가구는 0.104%, 5-9마리의 가구는 0.016%, 10-19마리의 가구는 0.060%, 30-49마리 가구는 0.028%, 50마리 이상의 가구는 0.020%로 대체 후에도 분포는 거의 변화가 없는 것으로 판단된다.

나. 산란계

산란계수에 대한 연관성 분석의 결과는 <표 2-50>에 제시되어 있다. 영농형태(상세)가 산란계수와 가장 연관성이 높으며, 다음으로 판매금액구분, 영농형태, 승용차보유여부 등의 순으로 나타났다.

<표 2-51>에 제시된 결과를 살펴보면 실제 산란계수의 평균은 37.13마리이며, 이 중에서 실제로 산란계를 보유하고 있는 가구의 평균은 1030.88마리이다. 전체가구에 대해서 대체전후의 산란계수의 평균의 차이를 보면 0.22로 0.59%의 오차, 산란계보유가구에 대해서는 평균차이

가 6.17로 0.60%의 오차, 그리고 산란계의 보유유무에 대한 대체의 정확도는 93.47%로 나타나고 있다.

〈표 2-50〉 산란계 항목에 관한 연관성 분석의 세부내용

깊이 연관성 ( )	분리변수	산란계 분리지점 좌 ( )	분리지점 우 ( )	연관비중
1	영농형태 상세 ( ) 판매금액구분	1~4, 6, 7	5	1.0000
2	영농형태 (1)	1~11	12	0.7232
	승용차보유 (2)	1~6, 8	7	0.1771
3	(3)	1	M	0.1456

〈표 2-51〉 산란계에 대한 대체전후의 평균차이 및 분포변화

평균 총가구 마리 평균 보유가구 마리 정확도

( ) : 37.13( ), ( ) : 1030.88( ), : 93.47%

무응답 가구수	대체후 평균 전체가구 ( )	절대차이 오차비율 ( )	대체후 평균 보유가구 ( )	절대차이 오차비율 ( )
50,000	37.35	0.22**(0.59%)**	1037.05	6.17**(0.60%)**

산란계 마리수	(0)	1-4	5-9	10-14	15-19	20-24	25-
실제 총가구수	48212	610	483	295	74	95	231
대체후 총가구수	48149	593	505	326	83	110	234
절대차이	63	17	22	31	9	15	3
오차비율 ( )	**(0.126%)**	**(0.034%)**	**(0.044%)**	**(0.062%)**	**(0.018%)**	**(0.030%)**	**(0.006%)**

대체전후의 분포변화비율을 살펴보면 산란계를 보유하지 않은 가구는 0.126%, 1-4마리인 가구는 0.034%, 5-9마리의 가구는 0.044%, 10-14마리의 가구는 0.062%, 15-19마리의 가구는 0.018%, 20-24마리 가구는 0.030%, 25마리 이상의 가구는 0.006%로 대체로 인한 분포왜곡 현상은 발생되지 않는다.

다. 가축 항목의 대체군 및 대체결과

<표 2-52>에는 한우, 산란계수를 제외한 가축 항목들의 대체군 및 대체결과가 주어져있다. 항목들 대부분이 상당히 높은 대체의 정확도를 가짐을 볼 수 있다.

〈표 2-52〉 가축 항목의 대체군 및 대체결과

항목	전체	가축 보유	대체 정확도	대체군
육우	0.142 (0.141) **0.70%**	26.34 (26.30) **0.15%**	**99.31%**	영농형태 상세 영농형태 판매금액구분 ( ), ,
젖소암컷	0.359 (0.358) **0.28%**	50.09 (49.88) **0.42%**	**99.74%**	영농형태 상세 판매금액구분 영농형태 ( ), ,
돼지	6.42 (6.38) **0.62%**	680.54 (676.41) **0.61%**	**99.12%**	영농형태 판매금액구분 영농형태 상세 전겸업수입규분 농업용난방기수 ( ) ,
멧돼지	0.028 (0.027) **1.45%**	42.84 (42.35) **1.14%**	**99.87%**	영농형태 상세 고추면적 옥수수면적 농업용난방기수), 화물차보유여부 ,
육계	106.00 (105.12) **0.83%**	3613.55 (3587.22) **0.73%**	**94.90%**	영농형태 판매금액구분 판매처구분 영농형태 상세 농업용난방기수 ( ),
젖산양	0.050 (0.049) **1.41%**	127.31 (130.04) **2.14%**	**99.10%**	대체군 없음
염소	0.441 (0.442) **0.23%**	11.58 (11.59) **0.09%**	**93.02%**	영농형태 상세 판매금액구분 영농형태 사슴수 경운기수), 시도 ,
사슴	0.084 (0.083) **0.12%**	13.21 (13.22) **0.08%**	**98.82%**	목초지면적 영농형태 상세 판매금액구분 판매처구분 꿀벌 통 영농형태 ( ), 옥수수면적 , ( ),
토끼	0.205 (0.207) **0.98%**	19.96 (20.15) **0.95%**	**98.19%**	영농형태 상세 염소수 밭면적 ~~뽕~~은감면적 (영농형태 들깨면적 육계수), , , , ,
오리	4.97 (4.93) **0.80%**	490.45 (486.48) **0.81%**	**97.96%**	농업용난방기수 시도 판매처구분 밭면적 , , ,
꿀벌	0.679 (0.680) **0.15%**	46.21 (46.62) **0.89%**	**97.49%**	영농형태 상세 판매처구분 판매금액구분 염소수 영농형태 사슴수 농업종사경력 ( ), , , ,
곰	0.001 (0.001) **3.00%**	11.41 (11.52) **2.45%**	**99.85%**	대체군 없음
고라니	0.001 (0.001) **3.15%**	8.51 (8.20) **3.64%**	**99.86%**	대체군 없음

이러한 이유는 가축의 항목 대부분에서 사육하는 가구비율이 상당히 낮기 때문인 것으로 판단된다. 대체군을 보면 영농형태(상세), 판매금액, 판매처 정보가 매우 중요하다는 것을 볼 수 있다. 젖산양, 곰, 고라니 항목은 특별히 연관성이 높은 항목을 없으며 대체시 지역적인 정보만을 고려해도 큰 문제가 없음을 확인하였다. 또한 대체전후의 분포도 변화가 거의 나타나지 않았음을 알려둔다.

제6절 결론 및 향후 연구계획

본 연구에서는 농업총조사 과수, 작물 및 가축부문에 대하여 연관성 분석을 실시하여 대체군을 제시하였고, 이전에 개발된 응용 핫덱 대체방법을 적용하여 효율성 및 대체의 정확성을 검토하였다. 본 연구에서 적용한 과수, 작물 및 가축부문은 모두 연속형 항목이지만, 주로 연속형 항목에 적용되는 방법들로 대체를 할 때에는 자료의 특성상 문제가 발생한다. 각 항목들은 모든 가구에 대하여 일정한 값을 가지지 않기 때문에 대체시 이러한 부분이 고려되어야 한다. 연속형 항목에 사용되는 대표적인 대체방법인 회귀 대체방법과의 비교실험에서 두 방법의 차이를 확실하게 보였다. 전체가구의 평균에 대해서는 두 방법은 차이가 거의 없어 보이지만 구성비(분포)변화 측면에서는 회귀 대체방법에서 상당한 문제가 발생되었다. 이러한 결과는 회귀 대체방법이 좋지 않은 대체방법이라는 의미는 아니며 농업총조사 자료에는 사용상에 문제가 있다는 것을 보여준다. 자세한 비교결과는 제4절을 참조하기 바란다.

이전에 연구되었던 가구원과 가구부문에 이어서 과수, 작물 및 가축부문에도 응용 핫덱 대체방법을 적용함으로써 범주형과 연속형 항목에 관계없이 일률적으로 응용 핫덱 대체방법을 사용하면 될 것으로 판단된다. 또한 이 방법은 농업총조사 이외의 조사에도 충분히 적용될 수 있을 것으로 생각된다. 하지만 다른 조사의 경우 자료의 특성을 잘 파악하고 적용상의 문제는 없는지 검토한 후 사용해야 할 것이다.

본 연구의 모의실험에서는 2005년 농업총조사 1,272,908가구 중에서 약 50%에 해당하는 635,779가구를 이용하였다. 실험을 위해 목표변수에

대해서 임의로 50,000가구의 무응답을 발생시켰으며, 다른 변수들도 동시에 일정비율만큼 무응답을 발생시켰다. 그리고 항목별로 50,000개의 무응답을 모두 대체를 하고난 후에 평균의 차이정도, 임의로 범주화된 분포변화비율, 대체의 정확도를 살펴보았다. 대부분의 항목에서 전체가구 및 보유가구의 대체전후의 평균차이는 매우 작은 것으로 나타났다. 대체의 정확도 측면에서는 부문에 따라 차이를 보이는데 과수원 항목들은 대략 95~99%, 노지재배 수확작물 항목들은 75~99%, 노지재배 판매작물 항목들은 90~99%, 시설재배 수확작물 항목들은 97~99%, 가축 항목들은 83~99%의 정확도를 보인다. 몇 항목들은 다른 항목에 비해서 정확도가 조금 낮으나 큰 문제는 되지 않는 것으로 판단되며, 대체후의 분포변화 역시 거의 나타나지 않음을 확인하였다.

하지만 여기서 한 가지 고려할 사항은 본 연구에서의 실험은 완전한 자료에 대해서 임의로 무응답을 발생시킨 것으로 실제 무응답을 대체하였을 경우 이 결과보다는 다소 정확도가 떨어질 것으로 판단된다. 왜냐하면 모든 항목들을 일정비율로 결측을 시켰으나, 실제 자료는 단위 무응답에 가까운 개체들도 많이 존재하며 대체 환경이 훨씬 더 열악할 수 있기 때문이다. 따라서 연구결과와 실제 자료를 대체한 후의 결과는 다소 차이가 있을 수도 있지만 본 연구에서 제시된 대체군과 대체방법은 현재의 농업총조사 자료를 적절하게 대체하기 위한 하나의 방법임에 틀림없을 것으로 확신한다. 모든 항목들에 대한 자세한 모의실험의 결과는 제5절을 참조하기 바란다.

본 연구를 끝으로 2005년 농업총조사의 모든 항목에 대하여 제안된 대체기법을 적용하여 모의실험을 실시하였다. 따라서 다음 단계인 2010년 농업총조사에 적용하기 위한 사전검토가 필요할 것이다. 이 연구가 필요한 이유는 기존의 연구가 2005년 농업총조사 자료에 의하여 이루어졌기 때문에 2010년의 항목과는 다소 차이가 발생할 수 있기 때문이다. 그러므로 기존 항목에서 삭제되는 항목 및 새로 추가되는 항목들에 대해서 대체군을 개발하고 조정해야 할 것이다. 또한 완성된 자료가 아닌 실제 조사 자료에서 발생할 수 있는 무응답 구조 및 현황을 파악하고 단위 무응답에 대한 처리도 생각을 해야 할 것으로 판단된다. 따라서 현재 실시되고 있는 농업총조사 시험조사 자료를 활용하여 추가적인 연구

가 수행되어야 할 것이다.

통계조사의 환경이 계속적으로 악화되고 있고, 최근에는 농촌의 환경도 점차 도시화 되어 무응답률이 높아지고 있는 현실이다. 그러므로 이러한 무응답 처리에 대한 관심을 가지고 지속적으로 연구가 진행되어야 할 것이다. 이러한 측면에서 본 연구는 향후 농업총조사의 품질을 향상시키는 데 많은 부분 도움이 될 것이며, 또한 다른 조사에서의 무응답 처리연구에도 일조할 수 있게 되기를 기대한다.

참고문헌

김규성 (2000), "무응답 대체 방법과 대체 효과", 「조사연구」 제 1 권 2 호, pp.1-14.

김규성 (2000), "표본 대체 방법과 대체자료의 합리적 이용", 한국은행 지원논문

김규성・이기재・김진 (2005), "농어가경제조사에서 가중핫덱 무응답 대체방법의 활용", 「응용통계연구」 제 18 권 2 호, pp.311-328.

김영원・이주원 (2003), "CART를 활용한 결측값 대체방법 : 인구주택총조사 혼인상태 항목을 중심으로", 「조사연구」 제 4 권 2 호, pp.1-21.

김영원・조선경 (1996), "표본조사에서 항목 무응답 대체 방법", 「한국통계학회논문집」 제 3 권 3 호, pp.145-159.

김재광・한근식・윤연옥, 가계조사 무응답 처리기법 연구 통계청 「통계연구」 제 9 권 1 호, pp.79-103.

김진 (2004), "농가경제조사에 대한 대체법 비교", 통계청 「통계연구」 제 9 권 2 호, pp.133-145.

송순관 (2005), "인구주택총조사 무응답 처리방법 연구 및 읍면동 통계확정 가능성 검토", 통계청 인구조사과

이진희・김진・이기재 (2006), "표본조사에서 공간 변수를 이용한 결측대체의 효율성 비교", 「응용통계연구」 제 19 권 1 호, pp.157-167.

이현정 (2008), "인구주택총조사 무응답 처리기법 연구 I", 연구보고서 통계개발원

최통진 (2006), "농림어업총조사를 위한 무응답 보정에 관한 연구", 석사학위논문

최필근 (2008), "농업총조사 항목간 연관성 분석 및 대체군 보조변수 개발", 연구보고서 통계개발원

최필근 (2008), "농업총조사 무응답 대체기법 연구 I", 연구보고서 통계개발원

통계교육원 「무응답처리 실무론」

통계청 (2005), 「농림어업총조사 조사지침서」

(2005), 2005

_____(2007), 「농림어업총조사 조사항목 변천 자료집」 .

Afifi, A. A. and R. M. Elashoff(1966), "Missing Observations in Multivariate Statistics I : Review of the Literature", *J. Am. Statist. Assoc.*, Vol. 61, pp.595-604.

Agresti, A.(1990), *Categorical Data Analysis*, A Wiley-Interscience Publication.

Berry, M. J. A. and G. S. Linoff(1997), Data Mining Techniques, John Wiley & Sons, New York.

Kalton, G. and D. Kasprzyk(1986), "The Treatment of Missing Survey Data", *Survey Methodology*, Vol. 12, pp.1-16.

Kass, G.(1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, Vol. 29, No. 2, pp.119-127.

Lessler and Kalsbeek(1992), *Nonsampling Error in Surveys*, John Wiley & Sons, New York.

Quinlan, J. R.(1986), "Induction of Decision Tree", Machine Learning, 1, pp.81-106.

Rubin, D. B. and J. A. Little(1986), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Sande, I. G.(1979), "A Personal View of Hot Deck Imputation Procedures", *Survey Methodology*, Vol. 5, pp.238-258.