

인구주택총조사
무응답 처리기법 연구 I

이현정

목 차

제1절 서론	1
제2절 2005년 인구주택총조사 무응답 처리방법에 대한 심층분석	3
1. 인구주택총조사 무응답 처리에 관한 선행연구	3
2. 2005년 인구주택총조사 무응답 처리 현황	3
3. 무응답 처리방법의 문제점	4
4. 무응답 처리방법의 개선방향	8
제3절 인구주택총조사 무응답 처리방법의 적용	11
1. 2008년 제2차 시험조사의 개요	11
2. 무응답 처리방법의 적용을 위한 자료처리	12
3. 항목 간 연관성 분석 실시	14
4. 연관성 분석 결과를 이용한 항목무응답 대체	15
제4절 무응답 처리방법의 효율성 비교	17
1. 항목무응답 대체결과 비교	17
2. 무응답 처리방법에 대한 제안	75
제5절 결론	80

표 목 차

<표 1> 가구원에 관한 사항의 무응답 처리	4
<표 2> 가구에 관한 사항의 무응답 처리	6
<표 3> 주택에 관한 사항의 무응답 처리	7
<표 4> 제2차 시험조사 조사항목	11
<표 5> 성별에 대한 무응답 현황	14
<표 7> 성별에 대한 [방법1]의 대체군 및 대체결과	6
<표 8> 성별에 대한 대체결과	17
<표 9> 만 나이에 대한 대체결과	18
<표 10> 가구주와의 관계에 대한 대체결과	19
<표 11> 교육정도에 대한 대체결과	20
<표 12> 교육상태에 대한 대체결과	21
<표 13> 아동보육에 대한 대체결과	22
<표 14> 출생지에 대한 대체결과	23
<표 15> 출생지 행정구역코드에 대한 대체결과	23
<표 16> 1년 전 거주지에 대한 대체결과	24
<표 17> 1년 전 거주지 행정구역코드에 대한 대체결과	25
<표 18> 5년 전 거주지에 대한 대체결과	25
<표 19> 5년 전 거주지 행정구역코드에 대한 대체결과	26
<표 20> 통근통학 여부에 대한 대체결과	27
<표 21> 통근통학 장소에 대한 대체결과	27
<표 22> 통근통학 장소 행정구역코드에 대한 대체결과	28
<표 23> 이용교통수단에 대한 대체결과	29
<표 24> 통근통학 소요시간에 대한 대체결과	30
<표 25> 취업여부에 대한 대체결과	31
<표 26> 구직여부에 대한 대체결과	31
<표 27> 취업가능성에 대한 대체결과	32
<표 28> 종사상 지위에 대한 대체결과	33
<표 29> 근로장소에 대한 대체결과	33

<표 30> 혼인상태에 대한 대체결과	34
<표 31> 혼인년도에 대한 대체결과	35
<표 32> 혼인월에 대한 대체결과	36
<표 33> 남자 출생이수에 대한 대체결과	37
<표 34> 여자 출생이수에 대한 대체결과	37
<표 35> 남자 사망이수에 대한 대체결과	38
<표 36> 여자 사망이수에 대한 대체결과	38
<표 37> 고령자 생활비 원천에 대한 대체결과	39
<표 38> 가구 구분에 대한 대체결과	40
<표 39> 가구원수에 대한 대체결과	41
<표 40> 거주 기간에 대한 대체결과	42
<표 41> 침실수에 대한 대체결과	42
<표 42> 침실이외의 방수에 대한 대체결과	43
<표 43> 거실수(대청마루)에 대한 대체결과	4
<표 44> 식당수(부엌이 딸린 식당 포함)에 대한 대체결과	5
<표 45> 건물층수에 대한 대체결과	46
<표 46> 거주층 구분에 대한 대체결과	47
<표 47> 거주층수에 대한 대체결과	47
<표 48> 부엌 형태에 대한 대체결과	48
<표 49> 부엌 사용여부에 대한 대체결과	49
<표 50> 수도 형태에 대한 대체결과	49
<표 51> 화장실 형태에 대한 대체결과	50
<표 52> 화장실 사용여부에 대한 대체결과	51
<표 53> 목욕시설 형태에 대한 대체결과	51
<표 54> 목욕시설 사용여부에 대한 대체결과	52
<표 55> 난방시설에 대한 대체결과	53
<표 56> 식수 사용 형태에 대한 대체결과	54
<표 57> 개인용 컴퓨터 보유에 대한 대체결과	54
<표 58> 인터넷 회선 보유에 대한 대체결과	55
<표 59> 케이블TV 보유에 대한 대체결과	56
<표 60> 위성방송 수신기 보유에 대한 대체결과	56

<표 61> 팩스 보유에 대한 대체결과	57
<표 62> 정보통신기기 보유여부에 대한 대체결과	57
<표 63> 개인용 컴퓨터 보유대수에 대한 대체결과	58
<표 64> 승용차 보유대수에 대한 대체결과	58
<표 65> 승합차 보유대수에 대한 대체결과	59
<표 66> 화물 및 기타자동차 보유대수에 대한 대체결과	59
<표 67> 자동차 보유여부에 대한 대체결과	60
<표 68> 주차시설에 대한 대체결과	60
<표 69> 주거영업구분에 대한 대체결과	61
<표 70> 점유형태에 대한 대체결과	62
<표 71> 사글세월수에 대한 대체결과	63
<표 72> 전세보증금에 대한 대체결과	63
<표 73> 월세에 대한 대체결과	64
<표 74> 가구소득에 대한 대체결과	64
<표 75> 타지 주택 소유에 대한 대체결과	65
<표 76> 주인가구 여부에 대한 대체결과	66
<표 77> 거처의 종류에 대한 대체결과	67
<표 78> 단독주택의 종류에 대한 대체결과	68
<표 79> 주거용 연면적에 대한 대체결과	68
<표 80> 대지 면적에 대한 대체결과	69
<표 81> 방수에 대한 대체결과	70
<표 82> 거실수(대청마루)에 대한 대체결과	71
<표 83> 식당수(부엌이 딸린 식당 포함)에 대한 대체결과	71
<표 84> 건축년도에 대한 대체결과	72
<표 85> 부엌수에 대한 대체결과	73
<표 86> 화장실수에 대한 대체결과	74
<표 87> 독립된 출입구수에 대한 대체결과	74
<표 88> 가구원에 관한 사항의 무응답 처리	75
<표 89> 가구에 관한 사항의 무응답 처리	77
<표 90> 주택에 관한 사항의 무응답 처리	79

인구주택총조사 무응답 처리기법 연구 I

- 제2차 시험조사를 대상으로 -

제1 절 서 론

인구주택총조사는 우리나라의 인구·가구·주택에 대해 규모·구조·분포 및 특성을 파악하기 위한 조사이다. 2005년에 실시한 인구주택총조사는 조사원면접, 응답자 직접기입, 인터넷조사방법을 병행하여 실시하였고 조사항목은 전수 21개 항목과 표본 23개 항목으로 총 44개의 항목이었다. 이 조사의 결과로 47,278,951명의 인구수와 15,988,274 가구수, 그리고 13,222,641 주택수를 얻었다. 이 결과에서도 알 수 있듯이 인구주택총조사는 전수 및 표본조사로 이루어지는 대규모 조사이기 때문에 조사불응 및 부재, 일부 항목의 누락 등의 무응답이 발생할 가능성도 높다고 할 수 있다.

이렇게 발생한 무응답을 분류하면 조사 단위 자체가 응답을 하지 않는 단위무응답(unit non-response)과 일부 항목에 대해서만 응답을 하지 않거나 질문과 무관한 응답을 해서 불필요한 자료가 되는 항목무응답(item non-response)으로 나누어질 수 있다(Kalton & Kasprzyk 1986). 단위 무응답의 경우 재조사나 가중값 조정방법을 통하여 처리하고 항목무응답의 경우에는 무응답대체(imputation)¹⁾를 통하여 처리하는 것이 일반적이다.

1) 무응답 오차(조사통계연구회, 2000, 자유아카데미) 중 제5장 항목무응답 참조

2 인구주택총조사 무응답 처리기법 연구 I

인구주택총조사와 같은 대규모 조사의 경우, 자료처리 시간을 단축하기 위해서는 전체 총조사 과정에 있어서 상당히 오랜 기간이 소요되었던 제조사 및 자료편집과정을 효율적으로 수행하는 것이 필요하다(이재원, 2000). 인구주택총조사의 항목에 대한 무응답 처리방법은 이미 선행연구(송순관, 2005)를 거쳐 2005년 인구주택총조사에서는 자료처리과정의 일부분으로 직접 적용되었다.

본 연구는 항목 간의 상관관계를 구하여 대체군(imputation class)을 개발하고, 여러 가지 방법들을 직접 적용해 보고자 한다. 그 결과 여러 가지 원인으로 발생하는 무응답들을 적절한 무응답 처리기법을 적용하여 완전한 자료로 구성하는 데 연구의 목적이 있다고 할 수 있다.

본 연구는 이론연구보다는 실제 적용분석 연구이므로 이론소개 및 내용에 대해서는 제외하기로 한다. 그리고 무응답 처리방법을 적용하기 위한 사전의 자동내검 단계는 꼭 필요한 부분이기는 하지만 전체적인 내용만 제시하고 항목별 적용방법에 대한 예시 및 상세한 내용 등은 언급하지 않도록 한다.

이번 연구에서는 2008년에 실시한 제2차 시험조사 자료를 활용하여 모의실험으로 무응답 처리를 실시한 후 무응답 처리기법의 효율성을 분석해 보고자 한다. 2절에서는 2005년 인구주택총조사 무응답 처리방법의 문제점과 개선방향에 대하여 알아보도록 한다. 3절에서는 2008년 4월에 실시한 제2차 시험조사의 개요에 대하여 간략히 소개하고 실제적인 무응답 대체기법을 적용하기 위하여 자료의 변환 및 일부 자동내검에 대한 과정을 거쳐 자료처리를 한다. 처리된 자료에 대하여 항목 간 연관성 분석을 하고 무응답을 대체하기 위한 대체군을 개발하여 기존의 대체군과 비교해 보고자 한다. 각각의 대체군을 이용하여 가구원, 가구 및 주택에 대한 항목무응답 대체를 실시한다. 4절에서는 앞절에서 모의실험한 무응답 처리결과에 대해 자료의 적합률 및 표본평균의 차이를 비교하여 그 효율성을 비교해 보고 무응답 처리방법을 제안해 보고자 한다. 5절에서는 제2차 시험조사에 대하여 무응답 처리기법을 적용한 후의 결론과 향후 연구방향을 제시하고자 한다.

제 2 절 2005년 인구주택총조사 무응답 처리방법에 대한 심층분석

1. 인구주택총조사 무응답 처리에 관한 선행연구

무응답 처리를 위한 기법개발은 통계학의 이론뿐만 아니라 다년간의 축적된 경험과 실무적인 지식을 필요로 한다.

미국, 일본, 캐나다 등의 인구센서스 실시기관에서는 무응답 및 오류 항목 처리라는 현실적인 문제해결을 위하여 자체적으로 대체기법을 개발·적용하고, 적용결과에 대한 경험 축적을 통하여 지속적인 개선작업을 실시하여 오고 있다(이재원, 2000).

우리도 인구주택총조사의 항목에 관한 무응답 처리기법에 대하여 논하고 있다. 따라서 본 연구에서는 이에 대한 연구의 시작으로 무응답 처리에 관한 선행연구를 소개하고자 한다.

이재원(2000)은 “무응답 및 오류자료의 Imputation 적용 결과”에서 2000년 인구주택총조사를 위해 1999년 11월에 실시한 시험조사 자료를 이용하여 출생지, 5년 전 거주지, 혼인상태의 3개 항목에 연역적 방법과 Hot-deck방법 등을 적용하여 무응답 처리를 했다.

김영원과 이주원(2003)은 “CART를 활용한 결측값 대체방법”에서 2000년 인구주택총조사를 위해 1999년 11월에 실시한 시험조사 자료를 이용하여 혼인상태 1개 항목에 Sequential Hot-deck방법 등을 적용하여 무응답 처리를 했다.

송순관(2005)은 “Imputation 방법 개발을 위한 자료 검토”에서 2005년 인구주택총조사를 위해 5차 시험조사 자료를 이용하여 가구원, 가구, 주택과 관련한 주요한 항목에 Probability, Hot-deck, Hierarchical Hot-deck 방법 등을 적용하여 무응답 처리를 했다.

위와 같은 선행연구에서도 알 수 있듯이 인구주택총조사 항목에 관한 무응답 처리기법에 대해서는 연구가 계속 진행 중이며, 실제적으로도 2005년 인구주택총조사 자료 처리 시 대부분의 항목에 대하여 무응답 처리기법을 적용했다.

2. 2005년 인구주택총조사 무응답 처리현황

가. 가구원에 관한 사항

나이는 만 나이를 계산하여 무응답을 처리하는 것으로 하고, 실제나 이 및 생년월일에 대해서는 무응답 처리를 하지 않도록 한다. 대체군은 무응답 대체를 할 때 정보제공자(donor)의 그룹을 형성하기 위한 변수들의 집합이다. 2005년 가구원 관련 항목의 무응답 처리는 <표 1>과 같다.

<표 1> 가구원에 관한 사항의 무응답 처리

조사항목	무응답 처리여부	대체군	대체방법	
성명	X	-	-	
성별	○	시도, 시군구, 읍면동, 조사구번호	Probability	
나이	○	교육정도, 혼인상태, 시도, 시군구, 읍면동, 조사구번호, 성별	Hierarchical Hot-deck	
가구주와의 관계	○	gacount, 시도, 시군구, 읍면동, 조사구번호	"	
교육정 도	교육정도	○	시도, 만 나이	Probability
	교육상태	○	시도, 교육정도	"
종교	X	-	-	
남북이산가족	X	-	-	
아동보육	○	시도, 시군구, 읍면동, 만 나이	Hierarchical Hot-deck	
5년 전 거주지	X	-	-	
활동계약	X	-	-	
통근·통학 여부	○	읍면동, GOSchool, in_tong12, 성별, 만 나이, 교육정도	Hierarchical Hot-deck	
통근· 통학 장소	통근·통학 장소	○	통근·통학 여부, 읍면동, GOSchool, 성별, 만 나이, 교육정도	"
	통근·통학 장소 행정구역코드	○	읍면동, 통근·통학 여부, GOSchool, 이용교 통수단, 통근·통학 소요시간	"
이용교통수단	○	읍면동, 통근·통학 여부, 통근·통학 장소, 이용교통수단, GOSchool, 성별, 만 나이, 교 육정도	"	

통근·통학 소요시간		○	읍면동, 통근·통학 여부, 통근·통학 장소, 이용교통수단, GOschool, 성별, 만 나이, 교육정도	Hierarchical Hot-deck
경제활동상태	취업여부	○	시군구, 읍면동, 성별, 만 나이, 교육정도	"
	구직여부	○	성별, 만 나이, 교육정도, 시군구, 읍면동	"
	취업가능성	○	성별, 만 나이, 교육정도, 시군구, 읍면동	"
종사상 지위		○	성별, 만 나이, 교육정도, 시군구, 읍면동	"
산업		X	-	-
직업		X	-	-
근로장소		○	성별, 만 나이, 교육정도, 시군구, 읍면동	Hierarchical Hot-deck ²⁾
혼인상태		○	시도, 시군구, 성별, 만 나이	Probability
혼인년월	혼인년도	○	성별, 만 나이	"
	혼인월	○	성별, 만 나이	"
총출생아수	남자 출생아수	○	대체군 없음	무응답은 0으로 처리
	여자 출생아수	○	대체군 없음	"
	함께 살고있는 남자자녀	○	대체군 없음	"
	함께 살고있는 여자자녀	○	대체군 없음	"
	남자 사망아수	○	대체군 없음	"
	여자 사망아수	○	대체군 없음	"
추가 계획 자녀 수		○	대체군 없음	"
고령자 생활비 원천		○	시도, 시군구, 성별, 만 나이	Hierarchical Hot-deck

주) gacount는 가구원수를 의미하며 가구원전체가 무응답인지의 여부를 비교하기 위한 변수, GOschool은 재학여부를 의미하는 변수, in_tong12는 통근·통학자를 의미하는 변수임

2) Hierarchical Hot-deck방법에 있어서 대체군의 배열은 관심변수(imputing variable) X와 상관관계가 높은 카테고리 변수들을 사용해야 하고 상관성이 가장 높은 변수는 상위 class(예: level1)로 시작하고 상관성이 가장 약한 변수는 하위 class(예: level5)로 지정하여 최단 class변수 제거가 필요한 경우 제일 먼저 상관성이 가장 약한 변수를 제거 대상이 되게 함으로써 가장 강한 상관관계의 변수는 최후까지 대체군 cell형성에 기여하도록 한다. 이 class변수의 level순서의 결정은 통계적인 방법을 사용하기는 용이하지 않으므로 관심변수와 class변수에 상당한 지식이 있는 subject matter 전문가와 상의하여 순서를 결정하기를 권하고자 한다(송순관, 2005).

6 인구주택총조사 무응답 처리기법 연구 I

나. 가구에 관한 사항

2005년 가구 관련 항목의 무응답 처리는 <표 2>와 같다.

<표 2> 가구에 관한 사항의 무응답 처리

조사항목		무응답 처리여부	대체군	대체방법
가구 구분	가구 구분	○	시도, 시군구, 읍면동	Probability
	가구원수	○	시도, 시군구, 읍면동, 가구 구분	Hot-deck
거주 기간		X	-	-
사용 방 수	침실수	○	ga_wonsu_group	Hot-deck
	침실 이외의 방수	○	ga_wonsu_group	"
	거실수	○	ga_wonsu_group	"
	식당수	○	ga_wonsu_group	"
주거시 설 형태	부엌 형태	○	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck
	부엌 사용여부	○	거처의 종류, 시도, 시군구, 읍면동, 부엌 형태	"
	수도 형태	○	거처의 종류, 시도, 시군구, 읍면동	"
	화장실 형태	○	거처의 종류, 시도, 시군구, 읍면동	"
	화장실 사용여부	○	거처의 종류, 시도, 시군구, 읍면동, 화장실 형태	"
	목욕시설 형태	○	거처의 종류, 시도, 시군구, 읍면동	"
	목욕시설 사용여부	○	거처의 종류, 시도, 시군구, 읍면동, 목욕시설 형태	"
거주층	거주층 구분	X	-	-
	거주층수	X	-	-
자동차 보유 대수	승용차 보유대수	○	대체군 없음	무응답은 0으로 처리
	승합차 보유대수	○	대체군 없음	"
	화물 및 기타 자동차 보유대수	○	대체군 없음	"
	자동차 보유여부	○	대체군 없음. 승용차, 승합차, 화물 및 기타자동차 보유대수가 모두 0대이면 자동차 보유여부를 '1 (없음)'으로 변경하고 보유대수가 1대라고 있으면 무응답은 '0(있음)'으로 변경	무응답은 0과 1로 처리
주차 시설		X	-	-
난방 시설		○	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck

점유 형태	주거영업구분	○	거처의 종류	Probability
	점유형태	○	주거영업구분	"
	사글세월수	X	-	-
임차료	전세보증금	X	-	-
	월세(사글세)	X	-	-
주인 가구 및 주택 소유 여부	주인가구 여부	○	점유형태	Probability
	타지 주택 소유 여부	○	주인가구 여부	"

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수임

다. 주택에 관한 사항

2005년 주택 관련 항목의 무응답 처리는 <표 3>과 같다.

<표 3> 주택에 관한 사항의 무응답 처리

조사항목	무응답 처리여부	대체군	대체방법
거처 의 종류	거처의 종류	○ 시도, 시군구, 읍면동, Type	Probability
	단독주택의 종류	○ Size_m, 주거용 연면적, 방수, 대지면적, 시도, 시군구, 읍면동, 조사구번호	Hierarchical Hot-deck
	건물층수	X -	-
연건평 (연면적)	○	거처의 종류, 방수, 부엌수, 화장실수, 독립된 출입구수	Hierarchical Hot-deck
대지 면적	○	Size_m, 주거용 연면적, 방수, 시도, 시군구, 읍면동, 조사구번호	"
총 방수	방수	○ 거처의 종류, Size_m, 주거용 연면적, 부엌수, 화장실수, 독립된 출입구수	"
	거실수	○ 거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 화장실수, 독립된 출입구수	"
	식당수	○ 거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 화장실수, 독립된 출입구수	"
건축년도	○	거처의 종류, 시도, 시군구, 읍면동, 조사구번호	"
편의 시설 수	부엌수	○ 거처의 종류, Size_m, 주거용 연면적, 방수, 화장실수, 독립된 출입구수	"
	화장실수	○ 거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 독립된 출입구수	"
	독립된 출입구수	○ 거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 화장실수	"

주) Type은 주택관련 항목이 모두 무응답인지 아닌지를 판단하는 변수, Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

3. 무응답 처리방법의 문제점

가. 대체군(보조변수)의 구성

2005년 인구주택총조사 자료에 무응답 처리방법을 적용함에 있어서 대체군 형성에 관한 명확한 설명이 없다. 오류내용을 제거하는 과정에서 무응답으로 인한 오류는 경험과 통계적 기법을 접목하여 추정하는 기법으로 해결해야 하지만, 적용 대체군에 있어서는 항목 간의 상관관계를 증명해 줄 이론적인 배경이 부족하다고 할 수 있다(통계청, 2006).

경험적으로 이 항목과 저 항목이 상관관계가 있을 것이라는 막연한 생각으로 작성하는 것보다 이론적 배경과 관련된 검토보고서가 필요하다. 항목 간의 상관관계가 높을 것이라는 가정보다는 증거자료가 제시될 필요가 있다.

나. 항목무응답 대체방법의 선택

항목무응답 대체방법에 있어서 Probability, Hot-deck, Hierarchical Hot-deck방법 등 여러 가지 적당한 방법을 제시하였으나 그에 대한 비교 연구는 없다. 예를 들면 성별에 Probability방법을 사용하였지만, 다른 방법을 사용했을 경우와 어떤 방법이 더 좋은지에 대한 비교 자료가 없는 것이다. Probability방법과 Hot-Deck방법은 정보제공자(donor)가 충분하지 못할 경우는 대체값을 발견하지 못하고 그대로 결측값을 남겨놓는 단점이 있으므로 좀 더 효율적인 방법을 선택할 수 있도록 모의실험을 통한 효율성 비교가 필요하다.

다. 무응답 대체항목의 선정

5년 전 거주지, 거주 기간, 임차료 등 항목무응답 대체를 하지 않는 항목에 대하여 그 이유를 제시하고 있지 않다. 항목에 따라 무응답을 처리하지 않을 경우는 명확한 이유나 원칙이 필요하다.

4. 무응답 처리방법의 개선방향

가. 대체군(보조변수)의 개발

대체군 개발을 위한 알고리즘은 여러 가지가 있다. Breiman 등(1984)이 제안한 CART(Classification and Regression Tree)알고리즘은 김영원·이주원(2003)이 혼인상태 항목의 대체군 개발에 이용하였고, Kass(1980)가 제안한 CHAID(Chi-squared Automatic Interaction Detection)알고리즘은 김재광·한근식·윤연옥(2004)이 가계조사 무응답 처리기법 연구에서 사용하고 있다. 그리고 카이제곱 독립성 검정(test of independence)은 이 내성(2007)이 사회통계조사 항목무응답 처리를 위한 대체군 개발에서 이용하고 있다.

본 연구에서는 대체군을 개발하는 데 있어서 통계학에서 범주형 자료분석의 한 방법인 카이제곱 독립성 검정을 실시하고자 한다. 두 변수 간의 독립성을 검정하고 연관성 측도를 이용하여 두 변수 간의 연관성을 찾아내어 대체군을 개발하도록 한다. 독립성 검정은 자료에 포함된 두 가지 특성 사이에 어떠한 연관관계가 있는지를 검정하는 방법으로 독립성 검정에서의 귀무가설은 “두 가지 특성들이 서로 독립(independent)이다”라는 것이다. 즉, 두 특성 사이에 연관관계가 없다는 것이다.

나. 항목무응답 대체방법의 효율성 비교

본 연구에서는 2005년 인구주택총조사에서 사용한 기존의 항목대체 방법과 주로 Hierarchical Hot-Deck방법을 이용하여 그 효율성을 비교해보고자 한다. 송순관(2005)의 견해로는 이 방법은 Hot-Deck방법과 Nearest Neighbor방법의 개념을 통합하여 응용된 방법으로 Hot-Deck방법에서 한 카테고리에 존재하는 무응답자수에 비하여 응답자수가 적을 때 응답자 선택을 비복원(without replacement)으로 수행한다면 카테고리외 카테고리의 통합과정이 필요하게 되는데 주로 카테고리의 통합은 한 class변수를 사전에 적절히 통합하여 하나의 새로운 class변수로 생성한

다음 이들의 순서를 적절히 배열하여 자동적으로 무응답 처리를 할 수 있도록 고안된 일종의 Hot-deck 대체방법이다. 미국 Census Bureau에서는 이 방법을 주로 Hot-deck이란 용어로 사용하고 있다(Donald B.Rubin, 1987). 어떤 연구 보고서에는 이 방법을 Flexible Matching Imputation 방법이라고도 한다.

Hot-deck 방법은 만약 한 정보제공자(donor) 선택에 비복원 방법을 사용한다면 대체군에 응답자수가 무응답자수보다 커야 하는 제한이 있으므로 class변수의 구조와 특성을 사전에 충분히 검토해야 하는데, 이 방법은 상관관계가 있다고 생각되는 무제한 class변수의 개수를 사용할 수 있다는 편리성이 있다.

성별과 같이 범주형 자료인 경우는 원자료를 얼마나 정확하게 대체 시키는가에 대한 적합률과 대체로 인한 분포의 왜곡문제가 발생하는지에 대해 살펴보고 방법에 따른 효율성을 비교해 본다.

나이와 같은 연속형 자료는 반복을 통한 표본평균의 차이 정도를 비교해 본다. 실제자료의 표본평균과 대체 후 자료의 표본평균이 얼마나 일치하는지 살펴보고 그 차이 값들의 평균을 ‘표본평균 차이’라고 명명한다. 이 표본평균 차이가 0에 가까울수록 좋은 무응답 처리방법이라고 할 수 있다. 그리고 대체 전후의 분포상의 왜곡문제를 검토하여 그 효율성을 비교해 보도록 한다.

다. 무응답 대체항목의 선정방법

인구주택총조사의 조사항목은 가구원에 관한 항목, 가구에 관한 항목, 주택에 관한 항목으로 구분할 수 있다. 무응답 처리대상 항목은 원칙적으로 모든 항목을 무응답 처리하는 것으로 간주한다. 성명, 산업 및 직업과 같이 주관식으로 응답하는 항목, 활동제약과 같이 왜곡될 가능성이 있는 항목, 국적 및 입국 연도와 같이 대체군 개발이 모호한 항목, 그리고 적합률 등을 검토하여 정확도가 낮은 항목은 무응답 대체항목에서 제외하도록 한다.

제 3 절 인구주택총조사 무응답 처리방법의 적용

1. 2008년 제2차 시험조사의 개요

2008년 4월에 2010년 인구주택총조사를 위한 제2차 시험조사를 실시하였다. 시험조사는 다양한 개선방안을 현장조사에 적용해 보고, 그 결과를 향후 총조사에 반영하고자 실시하는 것으로 조사대상은 2008년 4월 14일 0시 현재 조사지역 내에 거주하는 모든 사람과 이들이 살고 있는 주택이 조사대상이었다. 조사방법은 인터넷조사, 우편조사, 전화조사를 실시한 후 응답하지 않은 가구에 대해 14일간 가구방문조사를 실시하였다. 조사규모는 505조사구에 30,025가구이며 조사대상 지역은 부산, 경기도, 강원도 6개 시군구의 6개 읍면동이었다. 조사항목은 총 44개 항목을 조사하였으며 상세한 조사항목은 <표 4>와 같다.

<표 4> 제2차 시험조사 조사항목

구분	전수(18개 항목)	표본(26개 항목)
가구원 (24개)	①성명, ②성별, ③나이, ④가 구주와의 관계, ⑤교육정도, ⑥혼인상태, ⑦국적 및 입국 연도	①아동보육, ②출생지, ③1년 전 거주 지, ④5년 전 거주지, ⑤활동제약, ⑥통 근·통학 여부, ⑦통근·통학 장소, ⑧ 이용교통수단, ⑨통근·통학 소요시간, ⑩경제활동상태, ⑪종사상 지위, ⑫산 업, ⑬직업, ⑭근로장소, ⑮혼인년월, ⑯총 출생아수, ⑰고령자 생활비 원천
가구 (14개)	①가구 구분, ②사용 방 수, ③ 주거시설 형태, ④건물 및 거 주층, ⑤점유 형태, ⑥타지 주택 소유 및 주인가구 여부	①거주 기간, ②난방시설, ③식수 사용 형태, ④정보통신기기 보유현황, ⑤자 동차 보유 대수, ⑥주차시설, ⑦입차 료, ⑧가구소득
주택 (6개)	①거처의 종류, ②주거용 연 면적, ③층방수, ④건축년도, ⑤편익시설수	①대지면적

자료: 제2차 시험조사 조사지침서

주) 2005년 대비 신규항목은 밑줄로 표시된 6개 항목이며, 난방시설과 대지면적은 조사
표 종류를 변경(전수→표본)

2. 무응답 처리방법의 적용을 위한 자료처리

본 연구에서는 제2차 시험조사를 실시한 후 1차 내검을 거친 자료를 이용하여 항목 간 연관성 분석을 실시하고 가장 적절한 대체군을 개발하여, 그 결측치에 대한 가장 효율적인 무응답 대체방법을 찾아보고자 한다.

항목무응답 대체는 기존에 존재하는 자료 중 가장 유사한 자료로 그 항목을 대체하는 것이므로 전수조사의 무응답은 전수조사 자료에서, 표본조사의 무응답은 표본조사 자료에서 응답을 제공받아 대체된다. 그러기 위해서는 전수조사 자료와 표본조사 자료를 분리한다. 무응답 처리를 할 때, 표지항목에서 제공되는 조사구번호, 거처번호, 가구번호, 가구원수 등은 중요한 key가 되므로 표지항목과 가구원, 가구 및 주택 각각의 조사항목 테이블과 연계시키는 작업도 필요하다.

본 연구에서는 개별 항목의 무응답 대체를 하기 이전에 사전 자동내검 단계를 처리함으로써 실제적인 무응답 대체를 위한 원시입력자료를 작성하도록 한다. 그렇게 함으로써 실제적인 무응답 현황 분석도 용이할 것으로 판단된다. 조사를 실시하고 난 뒤 입력된 자료는 무응답과 '해당없음'에 해당하는 부분이 모두 공백으로 제공된다. 실제 무응답된 자료를 대체하기 위해서는 '해당없음'으로 인한 공백과 무응답으로 인한 공백의 구분이 필요하다. '해당없음'에 해당하여 공백인 경우는 '999(일부는 99999)'로 입력하고 실제적으로 무응답인 것은 공백으로 둔다. 그리고 상식적인 값의 범위를 벗어나는 이상치의 경우는 공백으로 처리하여 대체하도록 한다.

1년 전 거주지와 1년 전 거주지 행정구역코드 값은 일치성을 보이는 항목이므로 처리 시 유의해야 한다. 5년 전 거주지와 5년 전 거주지 행정구역코드, 통근·통학 장소와 통근·통학 장소 행정구역코드에는 일치성이 필요한 항목이므로 유의한다. 경제활동상태의 취업여부, 구직활동, 취업가능성의 항목과 종사상 지위항목은 논리적인 관계를 일치시켜야 하는 부분이다.

가구원, 가구 및 주택의 단위 무응답은 최소한의 가구와 주택정보는 입력이 될 수 있으므로 단위무응답으로 처리하지 않고 항목무응답으로

처리한다. 가구원과 관련된 항목은 확률적으로 선택된 가구의 구성원 정보를 그대로 가지고 오는 것이다. 대체할 가구를 선택하는 데 있어서 가구원수, 시·도, 시·군·구, 읍·면·동, 조사구번호, 가구 구분의 6가지 변수를 대체군으로 이용하여 Hierarchical Hot-deck 방법을 적용한다.

주택에 관한 항목은 대지면적을 제외하고는 모든 사항을 응답대상자가 응답해야 하는데 주택에 관한 항목 중 방수, 부엌수, 화장실수, 독립된 출입구수 및 주거용 연면적은 주택을 구성하는 데 필수 항목으로 간주되므로 반드시 무응답한 경우는 항목무응답 대체를 수행하여야 한다. 주택에 관한 사항의 항목무응답은 다음과 같은 순서로 처리한다.

첫째, 주택항목의 방수, 거실수, 식당수가 무응답이 발생했을 경우 가구정보로부터 방수, 거실수, 식당수를 이용하여 항목무응답 대체를 한다.

둘째, 앞뒤 주택의 형태가 동일하면 무응답 항목의 값을 앞뒤 주택의 응답항목의 값으로 대체하여도 상당한 정도가 의미있을 것으로 추정되어진다. 이것을 고려한 방법을 샌드위치(Sandwich)방법이라고 한다. 이 방법은 전후 주택의 동일한 구조를 확인하여 동일한 주택구조이면서 중간에 위치한 주택의 항목 가운데 무응답 항목이 있으면 이를 응답된 전후의 주택 가운데 응답항목의 값을 무응답한 항목의 값으로 이용하는 방법이다. 여기서 동일 주택구조를 확인하는 항목은 응답된 거처의 종류, 방수, 부엌수, 화장실수, 독립된 출입구수 항목들을 결합하여 형성된 합성변수를 기본으로 하여 앞뒤의 주택을 비교하도록 한다.

셋째, 주요 5개 변수 항목무응답 대체를 한다. 주요 5개 변수인 주거용 연면적, 방수, 부엌수, 화장실수, 독립된 출입구수 모두를 무응답한 경우는 변수 하나 하나를 독립하여 무응답 처리하기보다는 모든 변수를 한꺼번에 처리하도록 한다.

마지막으로 주택관련 개별 항목에 대하여 무응답 처리를 한다.

3. 항목 간 연관성 분석 실시

범주형 자료분석의 한 방법인 카이제곱 독립성 검정을 실시하여 1개의 항목에 대해 항목 간 연관성을 분석한다. 그 분석한 결과를 이용하여 본 연구에서는 ‘P값이 0.0001’이고, ‘크래머 V값’이 크며, 셀값이 ‘기대도수값 5’보다 큰 경우 중에서 3 ~ 4개 정도의 항목을 선택하여 대체군으로 선정한다. 향후, 이러한 대체군을 선택하는 조건에 대해서는 좀 더 많은 검토가 필요할 것으로 본다.

성별에 대한 응답률은 <표 5>와 같다. <표 5>를 보면 응답률의 경우 전수가 표본보다는 응답률이 높음을 알 수 있다.

<표 5> 성별에 대한 무응답 현황

2. 성별	전수		표본	
	명	%	명	%
Total	70,931	100.0	7,656	100.0
응답	70,641	99.59	7,603	99.31
무응답	290	0.41	53	0.69
단위무응답	232	0.33	9	0.12
항목무응답	58	0.08	44	0.57

카이제곱검정 결과 성별과 연관성이 높은 항목은 가구주와의 관계, 혼인상태, 교육정도이며 가장 연관성이 높은 항목은 가구주와의 관계로 나타났다. 항목별로 상기와 같이 무응답 현황을 분석하였으나 보고서에 수록하는 것은 생략하기로 한다.

<표 6> 성별에 대한 연관성 분석 결과

보조변수	카이제곱 값	유의확률	크래머의 V
가구주와의 관계	22842.6715	<math><0.0001</math>	0.5709
혼인상태	2154.8265	<math><0.0001</math>	0.1753
교육정도	806.9838	<math><0.0001</math>	0.1073

4. 연관성 분석 결과를 이용한 항목무응답 대체

일반적으로 손쉽게 사용할 수 있는 항목무응답 대체방법으로는 핫덱(hot-deck), 평균대체, 회귀대체 방법 등이 있으며, 김영원과 조선경(1996)은 실제 자료분석을 통해 이들 방법의 특성 및 효율성에 대한 비교·분석결과를 제시하고 있다. 하지만 이들 방법 중 평균대체와 회귀대체 방법은 그 특성상 관심변수가 연속형인 경우에 적용할 수 있는 방법이기 때문에 일반적인 사회조사에서 주로 관심대상이 되는 범주형 조사항목의 경우 핫덱이 가장 적합한 대체방법으로 알려져 있다(김영원·이주원, 2003).

본 연구에서는 앞에서 얻어진 카이제곱검정을 통한 연관성 분석 결과를 활용하여 다음 3가지 방법을 적용하고 그 효율성을 비교해 보고자 한다. 2005년에 처리한 항목무응답 대체방법은 Probability방법, Hot-deck방법, Hierarchical Hot-deck방법 3가지이다. 이 3가지 방법에 대한 이론적인 자세한 설명은 “2005 인구주택총조사 무응답 처리방법에 관한 자료(송순관, 2005)”를 참고하기 바란다.

- [방법1] 2005년 인구주택총조사에 적용한 대체군 및 대체방법을 활용하여 대체
- [방법2] 2005년 인구주택총조사에 적용한 대체군 및 2008년 제2차 시험조사에 적용하고자 하는 대체방법을 활용하여 대체
- [방법3] 2008년 제2차 시험조사에 적용하고자 하는 대체군 및 대체방법을 활용하여 대체

제시한 무응답 처리기법을 적용하여 각 항목들에 대한 모의실험을 실시한다. 전수자료에서 10%의 결측치를 생성하고 표본자료에서 5%의 결측치를 단순무작위(random sampling)로 발생시킨 후 상기의 3가지 방법으로 효율성을 비교해 보고자 한다. 본 연구에서는 항목무응답 대체 방법으로서 Hierarchical Hot-deck방법을 주로 이용하도록 하였다. 이 방법은 도너수가 적더라도 무응답을 결측치로 발생시키지 않고 상위 단계에서 자료를 찾아줄 수 있는 장점이 있기 때문이다.

2005년 인구주택총조사에 실시한 방법과 2008년 제2차 시험조사에 적용한 방법이 동일할 경우는 [방법2]에 의한 효율성 비교는 생략하고자 한다.

다음은 성별 항목에 대하여 무응답 처리를 한 후 대체 후의 적합률을 비교해 보았다.

[방법1]의 무응답 처리방법에서는 성별에 대한 대체군으로 시도, 시군구, 읍면동, 조사구번호를 사용하였고, 무응답 대체방법으로는 Probability방법을 적용하였다. 그 결과 전수자료의 적합률은 50.5%이고, 표본자료의 적합률은 54.8%이다.

<표 7> 성별에 대한 [방법1]의 대체군 및 대체결과

대체군	대체방법
시도, 시군구, 읍면동, 조사구번호	Probability

전수자료 적합률(50.5%)

대체결과 전수조사	남	여	합계	구성비
남(1)	1768	1712	3480	49.9
여(2)	1739	1751	3490	50.1
합계	3507	3463	6970	100
구성비(%)	50.3	49.7	100	

표본자료 적합률(54.8%)

대체결과 표본조사	남	여	합계	구성비(%)
남(1)	83	70	153	46.4
여(2)	79	98	177	53.6
합계	162	168	330	100
구성비(%)	49.1	50.9	100	

상기와 같은 방법으로 대부분의 항목에 대하여 전수자료와 표본자료를 분리하여 적합률을 구한다.

제 4 절 무응답 처리방법의 효율성 비교

1. 항목무응답 대체결과 비교

가. 가구원에 관한 사항

1) 성별

〈표 8〉 성별에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	시도, 시군구, 읍면동, 조사구번호	Probability	50.5	54.8
방법2	가구주와의 관계, 혼인상태, 교육정도	상동	68.9	72.7
방법3	상동	Hierarchical Hot-deck	68.6	70.0

대체 전후의 분포변화(표본)		
실제인원	남(153명)	여(177명)
방법1	162	168
방법2	161	169
방법3	156	174

성별에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 68.6%, 표본자료 70.0%의 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 18.4%, 표본 17.9% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.3%, 표본 2.7% 감소를 보이고 있지만 감소 효과는 미미하다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 남자 153명, 여자 177명에서 남자 156명, 여자 174명으로 대체가 되어 분포의 왜곡이 가장 적은 편이라고 볼 수 있다.

2) 나이

<표 9> 만 나이에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	교육정도, 혼인상태, 시도, 시군구, 읍면동, 조사구번호, 성별	Hierarchical Hot-deck	0.344	1.142
방법3	혼인상태, 교육상태, 교육정도, 가구주와의 관계	상동	0.128	1.148

대체 전후의 분포변화(표본)

구성비(%)	0-9 (43)	10-19 (49)	20-29 (37)	30-39 (58)	40-49 (51)	50-59 (38)	60-69 (36)	70-79 (12)	80-89 (6)	90- (-)
방법1	35	47	47	62	52	45	27	8	6	1
방법3	37	41	43	64	63	46	28	6	1	1

[방법1]의 표본평균 차이는 전수자료 0.344, 표본자료 1.142이고, [방법3]의 표본평균 차이는 전수자료 0.128, 표본자료 1.148이다.

표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수 0.216 감소하고 표본 0.006 증가한다. 두 가지 방법 모두 전수자료인 경우는 실제 표본평균과 대체한 후의 표본평균의 차이가 거의 없는 것으로 판단할 수도 있으나 표본자료인 경우는 1보다 큰 값을 보이고 있어 판단이 곤란하다.

만 나이는 연속형 항목이므로 범주화를 하여 대체 전후의 분포변화를 보고자 한다. 대체 전후의 항목분포의 변화를 살펴보면 [방법3]이 실제 인원수가 0-9세 43명, 10-19세 49명, 20-29세 37명 등에서 0-9세 37명, 10-19세 41명, 20-29세 43명 등으로 대체가 되어 분포의 변화도 약간 발생한다고 볼 수 있다. [방법1]도 [방법3]과 비슷한 분포를 가지며 두 방법의 효율성 비교는 어렵다.

3) 가구주와의 관계

<표 10> 가구주와의 관계에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	gacount, 시도, 시군구, 읍면동, 조사구번호	Set 1, 14 Hierarchical Hot-deck	36.3	37.3
방법2	성별, 혼인상태, 교육상태, 만 나이	상동	36.3	37.3
방법3	상동	Hierarchical Hot-deck	76.6	79.4

주) gacount는 가구원수를 의미하며 가구원전체가 무응답인지의 여부를 비교하기 위한 변수임

대체 전후의 분포변화(표본)

실제인원	가구주 (123)	가구주의 배우자 (71)	자녀 (113)	자녀의 배우자 (1)	가구주의 부모 (11)	배우자의 부모 (3)	손자녀, 그 배우자 (3)
방법1	133	0	0	0	0	0	0
방법2	133	0	0	0	0	0	0
방법3	126	73	106	2	4	1	9

대체 전후의 분포변화(계속)

실제인원	조부모 (1)	형제자매, 그 배우자 (3)	형제자매의 자녀, 그 배우자 (0)	부모의형제자매, 그 배우자 (0)	기타 친·인척 (1)	그외 같이 사는 사람 (0)
방법1	0	0	0	0	0	197
방법2	0	0	0	0	0	197
방법3	0	4	2	1	0	0

가구주와의 관계에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 76.6%, 표본자료 79.4%의 높은 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경시켜도 적합률이 증가하지 않는다. 그것은 무응답 대체를 함에 있어서 가구원의 일부가 가구주 관계를 무응답했을 경우는 첫 번째 가구원이면 '1'로, 그 외 가구원은 '14'로 부여하기 때문에 대체군의 영향은 거의 없다고 볼 수 있다. [방법3]처럼 무응답 처리방법을 변경할 경우는 전수 40.3%, 표본 42.1%의 현저한 증

가가 나타난다. 이는 대체군의 변경보다는 항목대체방법의 변경이 효율성 측면에서 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 가구주 123명, 가구의 배우자 71명 등에서 가구주 126명, 가구의 배우자 73명 등으로 대체되어 분포의 왜곡이 가장 적은 편이라고 할 수 있다. 그러나, [방법3]의 경우는 사후내검의 단계가 필요하다. 예를 들면 동일한 성별인데도 가구주 및 가구의 배우자 관계가 될 수 있기 때문이다.

4) 교육정도

〈표 11〉 교육정도에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	시도, 만 나이	Probability	42.1 (36개 결측값)	38.9 (9개 결측값)
방법2	혼인상태, 만 나이, 교육상태, 시도, 시군구, 읍면동, 조사구번호, 성별, 가구주와의 관계	상동	45.0 (4개 결측값)	38.7 (2개 결측값)
방법3	상동	Hierarchical Hot-deck	42.3	36.1

대체 전후의 분포변화(표본)

실제인원	안받았음(49)	초등학교(59)	중학교(31)	고등학교(87)	대학(38)	대학교(62)	대학원석사과정(3)	대학원박사과정(1)
방법1	2	85	35	114	30	51	4	0
방법2	0	85	42	103	28	67	3	0
방법3	6	85	44	111	26	55	5	1

교육정도에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 42.3%, 표본자료 36.1%의 적합률을 보인다.

적합률은 [방법1], [방법2], [방법3] 어떤 경우라도 50%를 넘지 못하고 있다. 특히, [방법2]와 [방법3]의 경우는 대체군으로 가구원관련 전수

조사 항목을 모두 사용하였으나 적합률을 향상시키지 못한다. 이는 기존의 조사항목으로서는 교육정도 항목대체에 있어서 정도를 높이기가 어렵다는 것을 보여준다.

대체 전후의 분포 왜곡문제가 발생하고 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 안받았음 49명, 초등학교 59명, 중학교 31명 등에서 안받았음 6명, 초등학교 85명, 중학교 44명 등으로 대체가 되어 분포의 왜곡을 가지고 온다. 다른 방법도 동일하므로 결국 값이 발생할 수 있는 Probability방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

<표 12> 교육상태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	시도, 교육정도	Probability	65.7	64.5
방법2	만 나이, 가구주와의 관계, 혼인상태, 교육정도	상동	86.3	85.2
방법3	상동	Hierarchical Hot-deck	85.7	83.3

대체 전후의 분포변화(표본)

실제인원	졸업(196)	재학(72)	수료(1)	휴학(1)	중퇴(15)	해당없음(45)
방법1	200	65	1	6	12	46
방법2	193	72	2	4	15	44
방법3	195	71	2	4	13	45

교육상태에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 85.7%, 표본자료 83.3%의 높은 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 20.6%, 표본 20.7% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.6%, 표본 1.9% 감소를 보이고 있지만 감소 효과는 미미하다. 이는 항목대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 졸업 196명, 재학 72명 등에서 졸업 195명 재학 71명 등으로 대체가 되어 분포의 왜곡이 가장 적은 편이라고 볼 수 있다.

5) 아동보육

〈표 13〉 아동보육에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	시도, 시군구, 읍면동, 만 나이	Hierarchical Hot-deck	89.1
방법3	교육정도, 출생지 행정구역코드, 만 나이, 5년 전 거주지	상동	89.1

대체 전후의 분포변화(표본)

실제인원	자녀의 부모 (22)	조부모 (7)	기타 가족 (2)	유치원 (1)	어린이집 (3)	기타보육시설 (0)	방과후 학교 (2)	학원 (14)	혼자 (3)	해당없음 (276)
방법1	32	6	0	4	4	1	1	5	0	227
방법3	27	10	0	1	6	0	1	8	1	276

아동보육에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 89.1%의 높은 적합률을 보인다.

아동보육은 [방법1]과 [방법3]의 대체방법이 동일하므로 [방법2]의 과정은 거치지 않도록 한다. 이 두 가지 방법의 적합률은 89.1%로 동일하고 이 정도이면 잘 대체시킨다고 볼 수 있을 것이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 자녀의 부모 22명, 유치원 1명, 해당없음 276명 등에서 자녀의 부모 27명, 유치원 1명, 해당없음 276명 등으로 대체되어 분포의 왜곡이 가장 적은 편이라고 할 수 있다.

6) 출생지

〈표 14〉 출생지에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	5년 전 거주지 장소, 만 나이, 시도, 시군구, 읍면동, 가구주와의 관계	Hierarchical Hot-deck	79.4

대체 전후의 분포변화(표본)

실제인원	현재 살고 있는 집(21)	같은 시군구내 다른집(91)	다른 시군구(218)
방법3	9	52	269

출생지는 제2차 시험조사에서 신규로 추가된 항목이기 때문에 [방법 3]만 적용하도록 한다. [방법3]은 표본자료 79.4%의 적합률을 보인다.

대체 전후의 분포 왜곡문제가 발생한다고는 볼 수 없지만 만족할 만한 수준은 되지 못한다고 볼 수 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 현재 살고 있는 집 21명, 같은 시군구내 다른집 91명, 다른 시군구 218명에서 현재 살고 있는 집 9명, 같은 시군구내 다른집 52명, 다른 시군구 269명으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 없다.

〈표 15〉 출생지 행정구역코드에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	1년 전 거주지 행정구역코드, 아동보육, 시도, 시군구, 읍면동, 5년 전 거주지 행정구역코드	Hierarchical Hot-deck	17.3

대체 전후의 분포변화(표본)

실제인원	서울(30)	부산(15)	대구(3)	인천(6)	광주(3)	대전(4)	울산(3)	경기(40)	강원(19)	충북(13)	충남(11)	전북(12)	전남(13)	경북(24)	경남(41)	해당 없음(93)
방법3	44	21	6	6	4	4	7	46	28	18	17	13	17	32	53	14

출생지 행정구역코드는 시도까지만 일치하는지 여부를 판단해 보았다. 이 때 적합률은 17.3%로써 낮다고 할 수 있다. 이것은 16개 시도에서 대체를 시킨다는 것은 무리가 있다고 볼 수 있다.

대체 전후의 분포 왜곡문제가 발생하고 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 서울 30명, 경기 40명, 해당없음 93명 등에서 서울 44명, 경기 46명, 해당없음 14명 등으로 대체되어 분포의 왜곡을 가지고 온다. 그러므로 이 항목을 대체시킬 것인가에 대해서는 고민이 필요하다.

7) 1년 전 거주지

<표 16> 1년 전 거주지에 대한 대체결과

표본자료	대체군		대체방법	적합률(%)	
방법3	5년 전 거주지, 출생지, 만 나이, 교육정도		Hierarchical Hot-deck	79.4	
대체 전후의 분포변화(표본)					
실제인원	태어나지 않았음(6)	현재 살고 있는 집(263)	같은 시군구내 다른집(33)	다른 시군구(27)	북한 또는 외국(1)
방법3	6	250	34	40	0

1년 전 거주지는 신규로 추가된 항목이기 때문에 [방법3]만 적용하도록 한다. [방법3]은 표본자료 79.4%의 적합률을 보인다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 태어나지 않았음 6명, 현재 살고 있는 집 263명 등에서 태어나지 않았음 6명, 현재 살고 있는 집 250명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

<표 17> 1년 전 거주지 행정구역코드에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	5년 전 거주지 행정구역코드, 시도, 시군구, 읍면동, 남자 출생이수, 혼인년도	Hierarchical Hot-deck	11.8

대체 전후의 분포변화(표본)

실제인원	서울 (3)	부산 (0)	대구 (0)	인천 (1)	광주 (1)	대전 (0)	울산 (0)	경기 (8)	강원 (1)	충북 (0)	충남 (1)	전북 (1)	전남 (2)	경북 (0)	경남 (1)	해당 없음 (311)
방법3	44	25	2	12	1	8	7	155	9	5	5	6	5	6	10	30

1년 전 거주지 행정구역코드는 시도까지만 일치하는지 여부를 판단해 보았다. 이 때 적합률은 11.8%로써 낮다고 할 수 있다. 따라서 16개 시도에서 대체를 시킨다는 것은 무리가 있다고 볼 수 있다.

대체 전후의 분포 왜곡문제가 발생하고 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 서울 3명, 경기 8명, 해당없음 311명 등에서 서울 44명, 경기 155명, 해당없음 30명 등으로 대체되어 분포의 왜곡을 가지고 온다. 그러므로 이 항목을 대체시킬 것인지에 대해서는 고민이 필요하다.

8) 5년 전 거주지

<표 18> 5년 전 거주지에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	1년 전 거주지, 만 나이, 교육정도, 출생지	Hierarchical Hot-deck	59.1

대체 전후의 분포변화(표본)

실제인원	태어나지 않았음(15)	현재 살고 있는 집(146)	같은 시군구내 다른집(85)	다른 시군구(70)	북한 또는 외국(1)	해당없음 (13)
방법3	28	117	71	113	1	0

5년 전 거주지는 신규로 추가된 항목이기 때문에 [방법3]만 적용하도록 한다. [방법3]은 표본자료 59.1%의 적합률을 보인다. 좀 더 적합률을 향상시킬 수 있는 대체군을 개발해야 할 필요가 있다.

대체 전후의 분포 왜곡문제가 발생한다고는 할 수 없어도 만족할 만한 수준은 되지 못한다고 볼 수 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 태어나지 않았음 15명, 현재 살고 있는 집 146명 등에서 태어나지 않았음 28명, 현재 살고 있는 집 117명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 없다.

〈표 19〉 5년 전 거주지 행정구역코드에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	시도, 시군구, 읍면동, 만 나이, 가구주와의 관계, 아동보육	Hierarchical Hot-deck	22.7

대체 전후의 분포변화(표본)

실제인원	서울 (17)	부산 (8)	대구 (1)	인천 (2)	광주 (1)	대전 (0)	울산 (0)	경기 (36)	강원 (1)	충북 (0)	충남 (0)	전북 (3)	전남 (0)	경북 (1)	경남 (4)	해당 없음 (256)
방법3	59	33	6	7	2	7	7	99	25	2	1	5	2	11	19	45

5년 전 거주지 행정구역코드는 시도까지만 일치하는지 여부를 판단해 보았다. 이 때 적합률은 22.7%로써 대체방법을 적용시키기에는 낮다고 할 수 있다. 이것은 16개 시도에서 대체를 시킨다는 것은 무리가 있다고 볼 수 있다.

대체 전후의 분포 왜곡문제가 발생하고 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 서울 17명, 경기 36명, 해당없음 256명 등에서 서울 59명, 경기 99명, 해당없음 45명 등으로 대체되어 분포의 왜곡을 가지고 온다. 그러므로 이 항목을 대체시킬 것인지에 대해서는 고민이 필요하다.

9) 통근·통학 여부

〈표 20〉 통근·통학 여부에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	읍면동, GOSchool, 성별, 만 나이, 교육정도	Hierarchical Hot-deck	76.4
방법3	만 나이, 교육상태, 취업여부, 가구주와의 관계	상동	89.4

주) GOSchool은 재학여부를 의미하는 변수임

대체 전후의 분포변화(표본)

실제인원	통근함(148)	통학함(29)	안함(110)	해당없음(43)
방법1	190	51	48	41
방법3	138	31	120	41

통근·통학 여부에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 89.4%의 높은 적합률을 보인다. 적합률은 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 표본 13% 증가한다. 이를 통해 대체군의 변경이 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 통근함 148명, 통학함 29명 등에서 통근함 138명, 통학함 31명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

10) 통근·통학 장소

〈표 21〉 통근·통학 장소에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	통근·통학 여부, 읍면동, GOSchool, 성별, 만 나이, 교육정도	Hierarchical Hot-deck	76.4
방법3	시도, 시군구, 읍면동, 조사구번호, 이용교통수단, 통근·통학 소요시간, 교육정도	상동	78.2

주) GOSchool은 재학여부를 의미하는 변수임

대체 전후의 분포변화(표본)

실제인원	현재 살고 있는 읍면동(55)	같은 시군구내 다른 읍면동(65)	다른 시군구(62)	해당없음(148)
방법1	45	69	69	147
방법3	45	80	59	146

통근·통학 장소에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 78.2%의 적합률을 보인다.

적합률은 [방법1]에서 [방법3]으로 대체군을 변화시키더라도 큰 향상을 가져오지는 않는다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 현재 살고 있는 읍면동 55명, 같은 시군구내 다른 읍면동 65명 등에서 현재 살고 있는 읍면동 45명, 같은 시군구내 다른 읍면동 80명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

<표 22> 통근·통학 장소 행정구역코드에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	읍면동, 통근·통학 여부, GOschool, 이용교통수단, 통근·통학 소요시간	Hierarchical Hot-deck	79.7
방법3	시도, 시군구, 읍면동, 1년 전 거주지 행정구역코드, 통근·통학 소요시간, 5년 전 거주지 행정구역코드	상동	79.7

주) GOschool은 재학여부를 의미하는 변수임

대체 전후의 분포변화(표본)

실제인원	서울 (14)	부산 (11)	대구 (0)	인천 (0)	울산 (1)	경기 (22)	강원 (0)	충북 (1)	충남 (1)	경북 (1)	경남 (2)	해당 없음 (277)
방법1	15	18	0	1	5	43	5	1	2	0	8	232
방법3	14	22	1	0	2	43	9	1	0	0	7	231

통근·통학 장소 행정구역코드에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 79.7%의 적합률을 보인다.

통근·통학 장소 행정구역코드는 출생지 행정구역코드 등 다른 행정 구역코드를 대체하는 것에 비해 적합률이 79.7%로 상대적으로 높게 나타나고 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 서울 14명, 경기 22명, 해당없음 277명 등에서 서울 14명, 경기 43명, 해당없음 231명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

11) 이용교통수단

<표 23> 이용교통수단에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	읍면동, 통근·통학 여부, 통근·통학 장소, GOschool, 성별, 만 나이, 교육정도	Hierarchical Hot-deck	73.3
방법3	통근·통학 여부, 통근·통학 장소, 성별, 취업여부	상동	71.5

주) GOschool은 재학여부를 의미하는 변수임

대체 전후의 분포변화(표본)

실제인원	걸어서 (48)	승용차 (70)	사내버스 (35)	통근버스 (4)	고속버스 (1)	전철 (3)	택시 (1)	자전 거(2)	기타 (5)	해당없음 (161)
방법1	55	60	33	5	0	8	2	2	6	159
방법3	42	70	46	3	1	5	0	0	3	160

이용교통수단에 대한 무응답 처리를 할 때 [방법1]이 효율적인 것으로 나타났다. [방법1]은 표본자료 73.3%의 높은 적합률을 보인다.

적합률에 있어서는 [방법1]이 [방법3]보다 높다. 대체군을 선택하는데 있어서 서로 연관성이 있는 변수 중 상위 일부만을 가지고 오게 되므로 같이 [방법1]과 [방법3]의 대체군을 결합하여 새로운 대체군을 개발하는 것이 좋을 것 같다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 걸어서 48명, 승용차 70명 등에서 걸어서 42명, 승용차 70명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

12) 통근·통학 소요시간

〈표 24〉 통근·통학 소요시간에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1	읍면동, 통근·통학 여부, 통근·통학 장소, 이용교통수단, GOschool, 성별, 만 나이, 교육정도	Hierarchical Hot-deck	1.512
방법3	통근·통학 장소, 시도, 시군구, 읍면동, 교육정도	상동	1.922

주) GOschool은 재학여부를 의미하는 변수임

대체 전후의 분포변화(표본)

실제인원	0-29분(99)	30-59분(50)	60-89분(18)	90-119분(5)	120-149분(2)	해당없음(156)
방법1	101	50	15	5	2	157
방법3	98	48	19	5	4	156

통근·통학 소요시간에 대한 무응답 처리를 할 때 [방법1]이 효율적인 것으로 나타났다. [방법1]의 표본평균 차이는 표본자료 1.512이다.

표본평균의 차이는 [방법1]이 [방법3]보다 0.41 적다. 표본평균의 차이 1.512는 차이가 적다고 할 수 없다. 대체군을 선택하는 데 있어서 서로 연관성이 있는 변수 중 상위 일부만을 가지고 오게 되므로 [방법1]과 [방법3]의 대체군을 결합하여 새로운 대체군을 개발하는 것도 좋을 것 같다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법1]이 실제인원수인 0-29분 99명, 30-59분 50명 등에서 0-29분 101명, 30-59분 50명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

13) 경제활동상태

<표 25> 취업여부에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	시군구, 읍면동, 성별, 만 나이, 교육정도	Hierarchical Hot-deck	67.9
방법3	통근·통학 여부, 성별, 만 나이, 이용교통수단	상동	89.1

대체 전후의 분포변화(표본)

실제인원	주로 일하였음(143)	틈틈이 일하였음(8)	잠시 쉬고 있음(8)	없었음(107)	해당없음(64)
방법1	144	48	4	70	64
방법3	141	15	7	103	64

취업여부에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 89.1%의 높은 적합률을 보인다.

[방법3]처럼 대체군을 변경하면 [방법1]보다 적합률이 21.2% 증가한다. 이는 대체군의 변경이 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 주로 일하였음 143명, 틈틈이 일하였음 8명 등에서 주로 일하였음 141명, 틈틈이 일하였음 15명 등으로 대체되어 분포의 왜곡이 가장 적은 편이라고 할 수 있다.

<표 26> 구직여부에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	성별, 만 나이, 교육정도, 시군구, 읍면동	Hierarchical Hot-deck	88.2
방법3	혼인년도, 시도, 시군구, 읍면동, 조사구번호, 교육정도, 교육상태	상동	90.6

대체 전후의 분포변화(표본)

실제인원	찾아보지 않았음(104)	찾아보았음(28)	해당없음(198)
방법1	113	16	201
방법3	119	14	197

구직여부에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 90.6%의 높은 적합률을 보인다.

적합률은 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 2.4% 증가한다. 이는 대체군의 변경이 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 찾아보지 않았음 104명, 찾아보았음 28명 등에서 찾아보지 않았음 119명, 찾아보았음 14명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 27〉 취업가능성에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	성별, 만 나이, 교육정도, 시군구, 읍면동	Hierarchical Hot-deck	97.6
방법3	시도, 시군구, 읍면동, 조사구번호	상동	98.5

대체 전후의 분포변화(표본)			
실제인원	일할 수 있었음(5)	일할 수 없었음(9)	해당없음(316)
방법1	6	7	317
방법3	7	6	317

취업가능성에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 98.5%의 높은 적합률을 보인다.

적합률에 있어서는 [방법1]과 [방법3]이 비슷하다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 일할 수 있었음 5명, 일할 수 없었음 9명, 해당없음 316명에서 일할 수 있었음 7명, 일할 수 없었음 6명, 해당없음 317명으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

14) 종사상 지위

<표 28> 종사상 지위에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	성별, 만 나이, 교육정도, 시군구, 읍면동	Hierarchical Hot-deck	74.5
방법3	성별, 통근·통학 여부, 근로장소, 만 나이	상동	76.4

대체 전후의 분포변화(표본)

실제인원	임금근로자 (93)	고용원이없는 자영자(35)	고용원일둔 사업주(13)	무급가족 종사자(13)	해당없음 (176)
방법1	91	39	8	11	181
방법3	91	35	12	14	178

종사상 지위에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 76.4%의 높은 적합률을 보인다.

적합률에 있어서는 [방법1]과 [방법3]이 비슷하다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 임금근로자 93명, 고용원이 없는 자영자 35명 등에서 임금근로자 91명, 고용원이 없는 자영자 35명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

15) 근로장소

<표 29> 근로장소에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	성별, 만 나이, 교육정도, 시군구, 읍면동	Hierarchical Hot-deck	84.8
방법3	통근·통학 여부, 성별, 종사상 지위, 시도, 시군구, 읍면동	상동	85.8

대체 전후의 분포변화(표본)

실제인원	사업장 (115)	자기집 (5)	남의집 (1)	거리 (5)	야외 작업현장(13)	운송수단 (4)	기타 (3)	해당없음 (184)
방법1	115	3	4	7	8	7	1	185
방법3	116	9	2	5	7	5	1	185

근로장소에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 85.8%의 높은 적합률을 보인다.

적합률에 있어서는 [방법1]과 [방법3]이 비슷하다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 사업장 115명, 자기집 5명 등에서 사업장 116명, 자기집 9명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

16) 혼인상태

<표 30> 혼인상태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	시도, 시군구, 성별, 만 나이	Probability	75.6	80.3
방법2	가구주와의 관계, 만 나이, 교육정도, 교육상태	상동	81.1	84.8
방법3	상동	Hierarchical Hot-deck	80.9	83.6

대체 전후의 분포변화(표본)

실제인원	미혼(74)	배우자있음(169)	사별(17)	이혼(8)	해당없음(62)
방법1	80	165	12	7	66
방법2	77	163	18	11	61
방법3	74	158	26	10	62

혼인상태에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 80.9%, 표본자료 83.6%의 높은 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 5.5%, 표본 4.5% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.2%, 표본 1.2%의 감소를 보이고 있지만 효과는 미미하다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 미혼 74명, 배우자 있음 169명 등에서 미혼 74명, 배우자 있음 158명 등으로 대체가 되어 분포의 왜곡이 가장 적은 편이라고 볼 수 있다.

17) 혼인년월

<표 31> 혼인년도에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1	성별, 만 나이	Probability	0.163
방법2	만 나이, 교육정도, 혼인상태, 통근·통학 여부	상동	0.127
방법3	상동	Hierarchical Hot-deck	0.067

대체 전후의 분포변화(표본)

실제인원	1940-1949 (3)	1950-1959 (8)	1960-1969 (27)	1970-1979 (30)	1980-1989 (49)	1990-1999 (59)	2000- (22)	해당없음 (132)
방법1	2	8	22	42	52	37	30	137
방법2	3	10	21	38	48	47	28	134
방법3	3	9	18	39	55	47	26	133

혼인년도에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 표본자료인 경우 표본평균 차이가 0.067이다.

표본평균의 차이는 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 표본자료 0.036만큼 감소한다. [방법3]처럼 대체방법을 변경할 경우는 0.06 더 감소하고 있다. 이는 대체군의 변경보다도 대체방법의 변경이 더 효과적이라고 할 수 있다. 세 가지 방법 모두 실제 표본평균과 대체 후의 표본평균의 차이는 거의 없는 것으로 판단된다.

혼인년도는 연속형 항목으로 범주화를 하여 대체 전후의 분포변화를 보고자 한다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인

원수가 1940-1949년 3명, 1980-1989년 49명, 1990-1999년 59명 등에서 1940-1949년 3명, 1980-1989년 55명, 1990-1999년 47명 등으로 대체가 되어 분포의 변화도 약간 발생한다고 볼 수 있다. [방법2]도 비슷한 분포를 가진다.

〈표 32〉 혼인월에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	성별, 만 나이	Probability	50.6
방법2	시도, 시군구, 읍면동, 혼인년도, 교육정도, 5년 전 거주지 장소	상동	58.5
방법3	상동	Hierarchical Hot-deck	59.7

대체 전후의 분포변화(표본)

실제인원	1월 (14)	2월 (12)	3월 (12)	4월 (26)	5월 (21)	6월 (8)	7월 (5)	8월 (2)	9월 (8)	10월 (25)	11월 (30)	12월 (20)	해당 없음 (147)
방법1	10	11	25	28	25	6	3	4	7	27	18	21	145
방법2	12	9	16	23	21	3	5	4	6	24	36	23	148
방법3	10	9	14	23	19	5	5	5	7	31	32	23	147

혼인월에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 59.7%의 적합률을 보인다.

적합률에 있어서는 [방법1], [방법2], [방법3]이 비슷한 적합률을 보이고 있다. 혼인월은 12개월의 월 중에 하나를 선택해야 하므로 무응답 대체를 할 것인지 고민해야 한다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 1월 14명, 11월 30명, 해당없음 147명 등에서 1월 10명, 11월 32명, 해당없음 147명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

18) 총 출생아 수

<표 33> 남자 출생아수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1 방법3	대체군 없음	무응답은 0으로 처리	1.173

표본자료 대체정확도

구성비(%)	일치	1명 차이	2명 차이	3명 차이	4명 차이	기타 차이
방법1,3	69.1	17.0	8.5	2.7	0.6	2.1

주) 기타 차이는 해당없음으로 대체해야 하는데 일정한 값으로 대체하거나 일정한 값으로 대체해야 하는데 해당없음으로 대체한 경우임

공백으로 조사된 것은 0으로 처리하여 무응답 처리를 하도록 한다. 그러면 표본평균의 차이는 1.173명이 된다. 이것은 1명 이상 차이가 나므로 실제 자료와 대체 후의 차이가 적다고 할 수 없다.

대체정확도 검토결과 이 방법은 정확한 남자 출생아수 69.1%, 1명 차이 17.0%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

<표 34> 여자 출생아수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1 방법3	대체군 없음	무응답은 0으로 처리	1.095

표본자료 대체정확도

구성비(%)	일치	1명 차이	2명 차이	3명 차이	4명 차이	기타 차이
방법1,3	78.8	11.2	5.2	2.4	0.6	1.8

주) 기타 차이는 해당없음으로 대체해야 하는데 일정한 값으로 대체하거나 일정한 값으로 대체해야 하는데 해당없음으로 대체한 경우임

공백으로 조사된 것은 0으로 처리하여 무응답 처리를 하도록 한다. 그러면 표본평균의 차이는 1.095명이 된다. 이것은 1명 이상의 차이가 나므로 실제 자료와 대체 후의 차이가 적다고 할 수 없다.

대체정확도 검토결과 이 방법은 정확한 여자 출생아수 78.8%, 1명 차이 11.2%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

〈표 35〉 남자 사망아수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1	대체군 없음	무응답은 0으로 처리	0.032
방법3	여자 사망아수, 남자 출생아수, 여자 출생아수, 혼인년도	무응답은 0으로 처리 Hierarchical Hot-deck	0.032

표본자료 대체정확도

구성비(%)	일치	1명 차이	2명 차이	3명 차이	4명 차이	기타 차이
방법1,3	98.8	0.3	0.6	-	-	0.3

주) 기타 차이는 해당없음으로 대체해야 하는데 일정한 값으로 대체하거나 일정한 값으로 대체해야 하는데 해당없음으로 대체한 경우임

공백으로 조사된 것은 0으로 처리하여 무응답 처리를 하도록 하고, 남자 출생아수보다 많이 조사된 경우만 Hierarchical Hot-deck방법을 적용하도록 한다. 그러나 두 방법에 있어서 표본평균의 차이는 동일하고 0.032는 차이가 거의 없다고 볼 수 있다.

대체정확도 검토결과 이 방법은 정확한 남자 사망아수 98.8%, 1명 차이 0.3%를 추정하여 정확도가 매우 높은 방법이라는 것을 알 수 있다.

〈표 36〉 여자 사망아수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1	대체군 없음	무응답은 0으로 처리	0.045
방법3	남자 사망아수, 여자 출생아수, 혼인년도, 만 나이	무응답은 0으로 처리 Hierarchical Hot-deck	0.045

표본자료 대체정확도

구성비(%)	일치	1명 차이	2명 차이	3명 차이	4명 차이	5명 이상	기타 차이
방법1,3	97.6	0.3	0.3	-	-	0.3	1.5

주) 기타 차이는 해당없음으로 대체해야 하는데 일정한 값으로 대체하거나 일정한 값으로 대체해야 하는데 해당없음으로 대체한 경우임

공백으로 조사된 것은 0으로 처리하여 무응답 처리를 하도록 하고, 여자 출생아수보다 많이 조사된 경우만 Hierarchical Hot-deck방법을 적용하도록 한다. 그러나 두 방법에 있어서 표본평균의 차이는 동일하고 0.045는 차이가 거의 없다고 볼 수 있다.

여자 사망아수에 대한 정확도 검토결과 이 방법은 정확한 가구원수 97.6%, 1명 차이 0.3%를 추정하여 정확도가 매우 높은 방법이라는 것을 알 수 있다.

19) 고령자 생활비 원천

<표 37> 고령자 생활비 원천에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	시도, 시군구, 성별, 만 나이	Hierarchical Hot-deck	90.0
방법3	통근·통학 여부, 성별, 시도, 시군구, 읍면동, 조사구번호	상동	91.5

대체 전후의 분포변화(표본)

실제인원	본인 직업 (17)	예금, 적금 (8)	국민 연금 (4)	개인 연금 (0)	부동산 (0)	함께 사는 자녀 (7)	따로 사는 자녀 (8)	국가 보조 (1)	이웃 보조 (0)	기타 (0)	해당없음 (285)
방법1	15	3	5	0	1	9	4	3	0	3	287
방법3	17	6	6	1	0	6	4	2	1	1	286

고령자 생활비 원천에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 91.5%의 적합률을 보인다.

적합률에 있어서는 [방법1], [방법3]이 비슷한 적합률을 보이고 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제인원수인 본인직업 17명, 예금·적금 8명 등에서 본인직업 17명, 예금·적금 6명 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

나. 가구에 관한 사항

1) 가구 구분

〈표 38〉 가구 구분에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	시도, 시군구, 읍면동	Probability	69.2 (27개 결측값)	70.1
방법2	ga_wonsu_group, 거실수, 주인가구 여부, 점유형태	상동	93.3 (20개 결측값)	95.5
방법3	상동	Hierarchical Hot-deck	92.9	97.8

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수임

대체 전후의 분포변화(표본)

실제가구	가족으로 이루어진 가구(105)	가족과 가족이외의 사람이 함께 사는 가구(1)	1인 가구(27)	가족이 아닌 남남끼리 함께 사는 5인 이하의 가구(1)
방법1	111	0	19	4
방법2	102	2	27	3
방법3	106	0	27	1

가구 구분에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 92.9%, 표본자료 97.8%의 매우 높은 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 24.1%, 표본 25.4% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.4% 감소, 표본 2.3% 증가를 보이고 있지만 감소 효과는 미미하다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 가족으로 이루어진 가구 105가구, 1인가구 27가구 등에서 가족으로 이루어진 가구 106

가구, 1인가구 27가구 등으로 대체가 되어 분포의 왜곡이 가장 적은 편이라고 볼 수 있다.

〈표 39〉 가구원수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	시도, 시군구, 읍면동, 가구 구분	Hot-deck	0.117 (30개 결측값)	0.104
방법2	가구 구분, 거실수, 주인가구 여부, 점유형태	상동	0.079 (29개 결측값)	0.110 (6개 결측값)
방법3	상동	Hierarchical Hot-deck	0.065	0.040

표본자료 대체정확도

구성비(%)	일치	1명 차이	2명 차이	3명 차이	4명 차이
방법1	34.3	38.1	20.9	6.7	-
방법2	33.6	42.2	16.4	7.0	0.8
방법3	40.3	32.1	17.9	8.2	1.5

가구원수에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]의 표본평균 차이는 전수자료 0.065, 표본자료 0.040이다.

표본평균의 차이는 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 0.038 감소하고 표본 0.006 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.014, 표본 0.06만큼 더 감소하고 있다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다. 세 가지 방법 모두 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다.

대체정확도 검토결과 [방법3]이 정확한 가구원수 40.3%, 1명 차이 32.1%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

2) 거주 기간

〈표 40〉 거주 기간에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	주인가구 여부, 점유형태, 건물층수, 거주층수, 시도, 시군구, 읍면동, 조사구번호	Hierarchical Hot-deck	27.6

대체 전후의 분포변화(표본)

실제가구	1년 미만(24)	1-2년 미만(21)	2-3년 미만(15)	3-5년 미만(11)	5-10년 미만(25)	10-15년 미만(17)	15-20년 미만(3)	20-25년 미만(6)	25년 이상(12)
방법3	29	23	12	12	24	12	10	5	7

거주 기간의 적합률은 27.6%로 낮은 편이다. 이 적합률로는 무응답 대체가 타당하다고 설명할 수 없으므로 적합률을 높이기 위한 대체군을 개발하는 것이 필요하다.

대체 전후의 분포 왜곡문제가 발생하고 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 1년 미만 24가구, 5-10년 미만 25가구, 15-20년 미만 3가구 등에서 1년 미만 29가구, 5-10년 미만 24가구, 15-20년 미만 10가구 등으로 대체되어 분포의 왜곡을 가지고 온다. 그러므로 이 항목을 대체시킬 것인지에 대해서는 좀 더 고민이 필요하다.

3) 사용 방 수

〈표 41〉 침실수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	ga_wonsu_group	Hot-deck	0.016	0.070
방법2	거실수, ga_wonsu_group, 가구 구분, 주인가구 여부	상동	0.009 (4개 결측값)	0.138
방법3	상동	Hierarchical Hot-deck	0.020	0.043

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이	3개 차이
방법1	48.5	43.3	6.7	1.5
방법2	56.7	34.3	9.0	0
방법3	50.0	38.1	9.7	2.2

침실수에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]의 표본평균 차이는 전수자료 0.020, 표본자료 0.043이다.

표본평균의 차이는 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 0.007 감소, 표본 0.068 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.011 증가, 표본 0.095 감소한다. 이는 대체군과 대체방법 중 어느 것이 더 효과적인 것인지 알 수 없다. 세 가지 방법 모두 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다.

대체정확도 검토 결과 [방법2]가 정확한 침실수 56.7%, 1개 차이 34.3%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다. 하지만 [방법3]도 비슷한 정확도를 가지고 있으므로 결측값이 발생할 수 있는 Hot-deck방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

<표 42> 침실이외의 방수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	ga_wonsu_group	Hot-deck	0.147	0.102
방법2	침실수, 거실수, ga_wonsu_group, 시도, 시군구, 읍면동	상동	0.064 (18개 결측값)	0.053 (5개 결측값)
방법3	상동	Hierarchical Hot-deck	0.024	0.030

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이	3개 차이
방법1	42.6	44.0	13.4	-
방법2	62.8	34.1	3.1	-
방법3	60.4	32.1	6.7	0.8

침실이외의 방수에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]의 표본평균 차이는 전수자료 0.024, 표본자료 0.030이다.

표본평균의 차이는 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 0.083, 표본 0.049 감소한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.04, 표본 0.023만큼 더 감소하고 있다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다. 세 가지 방법 모두 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다.

대체정확도 검토 결과 [방법2]가 정확한 침실이외의 방수 62.8%, 1개 차이 34.1%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다. 하지만 [방법3]도 비슷한 정확도를 가지고 있으므로 결측값이 발생할 수 있는 Hot-deck방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

〈표 43〉 거실수(대청마루)에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	ga_wonsu_group	Hot-deck	0.035	0.037
방법2	침실수, 거처의 종류, 시도, 시군구, 읍면동, ga_wonsu_group	상동	0.018 (24개 결측값)	0.022 (4개 결측값)
방법3	상동	Hierarchical Hot-deck	0.004	0.030

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이
방법1	85.8	14.2
방법2	86.9	13.1
방법3	83.6	16.4

거실수에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]의 표본평균 차이는 전수자료 0.004, 표본자료 0.030이다.

표본평균의 차이는 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 0.017, 표본 0.015 감소한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.014 감소, 표본 0.008 증가한다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다. 세 가지 방법 모두 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다.

대체정확도 검토 결과 [방법2]가 정확한 거실수 86.9%, 1개 차이 13.1%를 추정하여 정확도가 매우 높은 방법이라는 것을 알 수 있다. 하지만 [방법3]도 비슷한 정확도를 가지고 있으므로 결측값이 발생할 수 있는 Hot-deck방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

<표 44> 식당수(부엌이 딸린 식당 포함)에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	ga_wonsu_group	Hot-deck	0.112	0.094
방법2	거처의 종류, 거실수, 시도, 시군구, 읍면동, 건물층수	상동	0.027 (17개 결측값)	0.033 (6개 결측값)
방법3	상동	Hierarchical Hot-deck	0.003	0.022

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이
방법1	71.6	28.4
방법2	80.5	19.5
방법3	83.4	16.6

식당수에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]의 표본평균 차이는 전수자료 0.003, 표본자료 0.022이다.

표본평균의 차이는 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 0.085, 표본 0.061 감소한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.024, 표본 0.011만큼 더 감소하고 있다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다. 세 가지 방법 모두 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단한다.

대체정확도 검토 결과 [방법3]이 정확한 식당수 83.4%, 1개 차이 16.6%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

4) 건물 및 거주층

〈표 45〉 건물층수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법3	거주층수, 거처의 종류, 시도, 시군구, 읍면동, 주인가구 여부	Hierarchical Hot-deck	0.125	0.213

대체 전후의 분포변화(표본)

실제가구	1-4층(87)	5-9층(13)	10-14층(1)	15-19층(21)	20-24층(5)	25-29층(6)	30-34층(1)
방법3	88	13	1	22	6	4	0

건물층수에 대한 실제값과 대체값의 차이의 평균은 전수자료 0.125, 표본자료 0.213으로 거의 차이가 없다고 볼 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 1-4층 87가구, 5-9층 13가구 등에서 1-4층 88가구, 5-9층 13가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

<표 46> 거주층 구분에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법3	화장실 사용여부, 점유형태, 주인가구 여부	Hierarchical Hot-deck	100.0	100.0
대체 전후의 분포변화(표본)				
실제가구	지상(134)			
방법3	134			

거주층 구분에 대한 대체는 전수자료와 표본자료 모두 100% 적합시키고 있다. 이것은 거주층 구분이 ‘①지하, ②지상, ③옥상(옥탑)’인 경우 중 ‘②지상’에 거주하는 것이 대부분이기 때문이다.

<표 47> 거주층수에 대한 대체결과

	대체군	대체방법	표본평균 차이			
			전수	표본		
방법3	건물층수, 거처의 종류, 시도, 시군구, 읍면동, 주인가구 여부	Hierarchical Hot-deck	0.095	0.130		
대체 전후의 분포변화(표본)						
실제가구	1-4층(95)	5-9층(18)	10-14층(11)	15-19층(7)	20-24층(3)	30-34층(0)
방법3	93	19	14	6	1	1

거주층수에 대한 실제값과 대체한 후의 값 차이의 평균은 전수자료 0.095, 표본자료 0.130으로 거의 차이가 없다고 볼 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 1-4층 95가구, 5-9층 18가구 등에서 1-4층 93가구, 5-9층 19가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

5) 주거 시설 형태

주거 시설 형태와 관련한 항목에 대한 무응답 대체는 높은 적합률을 보여주고 있다. 또한 상이한 대체군을 사용하더라도 적합률에는 큰 변화를 보이지 않는다. 주거 시설 형태에서 무응답이 발생했을 경우는 무응답 처리를 하더라도 큰 무리는 없을 것으로 판단된다.

〈표 48〉 부엌 형태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck	95.8	96.3
방법3	목욕시설 형태, 목욕시설 사용여부, 화장실 형태, 화장실 사용여부	상동	96.1	96.3

대체 전후의 분포변화(표본)

실제가구	현대식(131)	재래식(3)	없음(0)
방법1	132	2	0
방법3	130	3	1

부엌 형태에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 전수자료 96.1%, 표본자료 96.3%의 매우 높은 적합률을 보인다.

적합률은 어떤 방법이더라도 높게 나타나고 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후

항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 현대식 131가구, 재래식 3가구, 없음 0가구에서 현대식 130가구, 재래식 3가구, 없음 1가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 49〉 부엌 사용여부에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동, 부엌 형태	Hierarchical Hot-deck	99.9	100.0
방법3	목욕시설 사용여부, 화장실 사용여부, 점유형태, 주인가구 여부	상동	99.6	100.0

대체 전후의 분포변화(표본)

실제가구	단독사용(134)
방법1	134
방법3	134

부엌 사용여부에 대한 무응답 처리를 할 때 [방법1]과 [방법3]이 모두 효율적인 것으로 나타났다. 두 방법 모두 99.6% 이상의 매우 높은 적합률을 보인다. 적합률은 어떤 방법이더라도 높게 나타나고 있다. [방법1]의 대체군에 [방법3]의 대체군을 추가하여 새로운 대체군을 개발한다.

대체 전후의 분포 왜곡문제도 단독사용 134가구가 그대로 단독사용 134가구로 대체됨으로써 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 50〉 수도 형태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck	91.4	89.6
방법3	시도, 시군구, 읍면동, 거처의 종류, 화장실 형태	상동	91.4	89.6

대체 전후의 분포변화(표본)

실제가구	상수도(125)	마을상수도(3)	자가수도(2)	없음(4)
방법1	129	1	2	3
방법3	126	3	3	2

수도 형태에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 전수자료 91.4%, 표본자료 89.6%의 매우 높은 적합률을 보인다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 상수도 125가구, 마을상수도 3가구 등에서 상수도 126가구, 마을상수도 3가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 51〉 화장실 형태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck	93.1	94.8
방법3	목욕시설 형태, 부엌 형태, 시도, 시군구, 읍면동, 거처의 종류	상동	94.7	97.0

대체 전후의 분포변화(표본)

실제가구	수세식(129)	재래식(5)	없음(0)
방법1	128	5	1
방법3	129	5	0

화장실 형태에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 전수자료 94.7%, 표본자료 97.0%의 매우 높은 적합률을 보인다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 수세식 129가구, 재래식 5가구 등에서 수세식 129가구, 재래식 5가구 등으로 대체되어 분포의 왜곡은 없다.

<표 52> 화장실 사용여부에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동, 화장실 형태	Hierarchical Hot-deck	96.8	98.5
방법3	목욕시설 사용여부, 목욕시설 형태, 주인가구 여부, 점유형태	상동	96.7	96.3

대체 전후의 분포변화(표본)

실제가구	단독사용(132)	공동사용(1)	해당없음(1)
방법1	130	3	1
방법3	127	6	1

화장실 사용여부에 대한 무응답 처리를 할 때 [방법1]이 효율적인 것으로 나타났다. [방법1]은 전수자료 96.8%, 표본자료 98.5%의 매우 높은 적합률을 보인다.

적합률은 어떤 방법이더라도 높게 나타나고 있다. [방법1]의 대체군에 [방법3]의 대체군을 추가하여 새로운 대체군을 개발하는 것도 고려해 볼 수 있을 것이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법1]이 실제가구수인 단독사용 132가구, 공동사용 1가구 등에서 단독사용 130가구, 공동사용 3가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

<표 53> 목욕시설 형태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck	90.5	93.3
방법3	화장실 형태, 화장실 사용여부, 부엌 형태, 거처의 종류	상동	91.9	95.5

대체 전후의 분포변화(표본)

실제가구	온수시설(128)	비온수시설(0)	없음(6)
방법1	129	1	4
방법3	128	1	5

목욕시설 형태에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 전수자료 91.9%, 표본자료 95.5%의 매우 높은 적합률을 보인다. [방법1]과 [방법3]의 대체군도 상이한 특성을 지니고 있으나, 자료를 대체하는 데 있어서는 비슷한 적합률을 보이고 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 온수시설 128가구, 비온수시설 0가구, 없음 6가구에서 온수시설 128가구, 비온수시설 1가구, 없음 5가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 54〉 목욕시설 사용여부에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동, 목욕시설 형태	Hierarchical Hot-deck	98.5	99.3
방법3	부엌 사용여부, 화장실 사용여부, 부엌 형태, 점유형태	상동	96.2	99.3

대체 전후의 분포변화(표본)

실제가구	단독사용(124)	해당없음(10)
방법1	123	11
방법3	123	11

목욕시설 사용여부에 대한 무응답 처리를 할 때 [방법1]이 효율적인 것으로 나타났다. [방법1]은 전수자료 98.5%, 표본자료 99.3%의 매우 높은 적합률을 보인다.

적합률은 어떤 방법이더라도 높게 나타나고 있다. [방법1]과 [방법3]의 대체군도 상이한 특성을 지니고 있으나, 자료를 대체하는 데 있어서는 비슷한 적합률을 보이고 있다. [방법1]의 대체군에 [방법3]의 대체군

을 추가하여 새로운 대체군을 개발하는 것도 고려해 볼 수 있을 것이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법1]이 실체가구수인 단독사용 124가구, 해당없음 10가구에서 단독사용 123가구, 해당없음 11가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

6) 난방시설

<표 55> 난방시설에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1	거처의 종류, 시도, 시군구, 읍면동	Hierarchical Hot-deck	73.9
방법3	시도, 시군구, 읍면동, 주인가구 여부, 목욕시설 형태, 건물층수	상동	83.6

대체 전후의 분포변화(표본)

실체가구	지역난방 (5)	도시가스 보일러(64)	기름 보일러(44)	프로판가스 보일러(7)	전기 보일러(11)	연탄 보일러(2)	기타 (1)
방법1	9	67	36	10	8	2	2
방법3	5	70	38	6	11	4	0

난방시설에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 표본자료 83.6%의 적합률을 보인다.

적합률은 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 9.7% 증가한다. 이는 대체군의 변경이 효율성 측면에서는 영향력이 크다는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실체가구수인 지역난방 5가구, 도시가스보일러 64가구, 기타 1가구 등에서 지역난방 5가구, 도시가스보일러 70가구, 기타 0가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

7) 식수 사용 형태

〈표 56〉 식수 사용 형태에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	수도 형태, 시도, 시군구, 읍면동, 조사구번호, 주인가구 여부, 가구 구분, 거처의 종류	Hierarchical Hot-deck	41.8

대체 전후의 분포변화(표본)

실제가구	수돗물을 그대로 먹음(2)	수돗물을 끓여 먹음(63)	수돗물을 정수해 먹음(47)	생수를 사서 먹음(10)	약수를 떠서 먹음(7)	지하수를 먹음(5)	기타 (0)
방법3	4	76	34	12	5	2	1

식수 사용 형태에 대한 적합률은 41.8%로 적합률이 높다고 할 수 없다. 적합률을 높일 수 있는 다른 대체군을 좀 더 고민해야만 한다.

대체 전후의 분포 왜곡문제가 발생한다고는 볼 수 없지만 만족할 만한 수준은 되지 못한다고 볼 수 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 수돗물을 그대로 먹음 2가구, 수돗물을 끓여 먹음 63가구, 기타 0가구 등에서 수돗물을 그대로 먹음 4가구, 수돗물을 끓여 먹음 76가구, 기타 1가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

8) 정보통신기기 보유현황

〈표 57〉 개인용 컴퓨터 보유에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	대체군 없음. 컴퓨터 보유대수가 1대 이상이면 '1(있음)'으로 변경	무응답 0과 1로 처리	95.5

대체 전후의 분포변화(표본)

실제가구	없음(27)	있음(107)
방법3	33	101

개인용 컴퓨터 보유에 대한 적합률은 95.5%로 적합률이 높다. 무응답인 경우는 모두 0으로 변경할 때 개인용 컴퓨터를 보유한 가구가 많을수록 높은 적합률을 얻을 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 27가구, 있음 107가구에서 없음 33가구, 있음 101가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

<표 58> 인터넷 회선 보유에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	대체군 없음. '1'은 인터넷 회선이 있는 경우임	무응답은 1로 처리	69.4

대체 전후의 분포변화(표본)

실제가구	없음(41)	있음(93)
방법3	0	134

인터넷 회선 보유에 대한 적합률은 69.4%로 적합률이 높다. 이것은 인터넷 회선을 보유한 가구가 많기 때문이다.

대체 전후의 분포 왜곡문제가 발생한다고는 볼 수 없지만 만족할 만한 수준은 되지 못한다고 볼 수 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 41가구, 있음 93가구에서 없음 0가구, 있음 134가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 없다.

〈표 59〉 케이블TV 보유에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	대체군 없음. '1'은 케이블TV가 있는 경우임	무응답은 1로 처리	76.1
대체 전후의 분포변화(표본)			
실제가구	없음(32)	있음(102)	
방법3	0	134	

케이블TV 보유에 대한 적합률은 76.1%로 적합률이 높다. 이것은 케이블TV를 보유한 가구가 많기 때문이다.

대체 전후의 분포 왜곡문제가 발생한다고는 볼 수 없지만 만족할 만한 수준은 되지 못한다고 볼 수 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 32가구, 있음 102가구에서 없음 0가구, 있음 134가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 없다.

〈표 60〉 위성방송 수신기 보유에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	대체군 없음. '0'은 위성방송수신기가 없는 경우임	무응답은 0으로 처리	94.0
대체 전후의 분포변화(표본)			
실제가구	없음(126)	있음(8)	
방법3	134	0	

위성방송 수신기 보유에 대한 적합률은 94.0%로 적합률이 높다. 이것은 위성방송 수신기를 보유하지 않은 가구가 많기 때문이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 126가구, 있음 8가구에서 없음 134가구, 있음 0가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 61〉 팩스 보유에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	대체군 없음. '0'은 팩스가 없는 경우임	무응답은 0으로 처리	97.0
대체 전후의 분포변화(표본)			
실제가구	없음(130)	있음(4)	
방법3	134	0	

팩스 보유에 대한 적합률은 97.0%로 적합률이 높다. 이것은 팩스를 보유하지 않은 가구가 많기 때문이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 130가구, 있음 4가구에서 없음 134가구, 있음 0가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 62〉 정보통신기기 보유여부에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	대체군 없음. 정보통신기기가 모두 '없음'이면 '1(없음)'으로 처리하고 1종류라도 있으면 '0(있음)'으로 처리	무응답은 0과 1로 처리	97.8
대체 전후의 분포변화(표본)			
실제가구	없음(9)	있음(125)	
방법3	12	122	

정보통신기기 보유 여부에 대한 적합률은 97.8%로 적합률이 높다. 이것은 정보통신기기를 보유하지 않은 경우는 모두 1로 지정하여 없음으로 하기 때문이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 9가구, 있음 125가구에서 없음 12가구, 있음 122가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 63〉 개인용 컴퓨터 보유대수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법3	대체군 없음. 개인용 컴퓨터를 보유하고 개인용 컴퓨터 보유대수가 0대인 경우는 모두 '1'로 변경하고, 개인용 컴퓨터를 보유하지 않으면 '0'으로 변경	무응답 0과 1로 처리	0.164
표본자료 대체정확도			
구성비(%)	일치	1대 차이	
방법3	86.6	13.4	

개인용 컴퓨터 보유대수에 대한 표본평균의 차이는 0.164로 실제값과 대체한 후의 값의 차이는 거의 없다고 볼 수 있다.

대체정확도 검토결과 [방법3]이 정확한 개인용 컴퓨터 보유대수 86.6%, 1대의 차이 13.4%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

9) 자동차 보유대수

〈표 64〉 승용차 보유대수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1 방법3	대체군 없음	무응답은 0으로 처리	0.595
표본자료 대체정확도			
구성비(%)	일치	1대 차이	2대 차이
방법1, 3	54.5	36.6	8.9

승용차 보유대수에 대한 표본평균의 차이는 0.595로 실제 자료와 대체한 자료의 표본평균의 차이가 적다고 할 수는 없다. 그러나 이 방법을 받아들이는 데 있어서는 문제가 없다고 볼 수 있다.

대체정확도 검토결과 [방법3]이 정확한 승용차 보유대수 54.5%, 1대의 차이 36.6%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

〈표 65〉 승합차 보유대수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1 방법3	대체군 없음	무응답은 0으로 처리	0.020
표본자료 대체정확도			
구성비(%)	일치	1대 차이	
방법1, 3	97.8	2.2	

승합차 보유대수에 대한 표본평균의 차이는 0.020으로 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다. 그러므로 [방법3]을 선택할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 승합차 보유대수 97.8%, 1대 차이 2.2%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

〈표 66〉 화물 및 기타자동차 보유대수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1 방법3	대체군 없음	무응답은 0으로 처리	0.122
표본자료 대체정확도			
구성비(%)	일치	1대 차이	2대 차이
방법1, 3	90.3	8.2	1.5

화물 및 기타자동차 보유대수에 대한 표본평균의 차이는 0.122로 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다. 그러므로 [방법3]을 선택할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 화물 및 기타자동차 보유대수 90.3%, 1대 차이 8.2%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

〈표 67〉 자동차 보유여부에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법1 방법3	대체군 없음. 승용차, 승합차, 화물 및 기타자동차 보유대수가 모두 0이면 자동차보유여부를 '1(없음)'으로 변경하고 보유대수가 1대라도 있으면 무응답은 '0(있음)'으로 변경	무응답은 1로 처리	97.8
대체 전후의 분포변화(표본)			
실제가구	없음(39)	있음(95)	
방법1, 3	42	92	

자동차 보유여부에 대한 적합률은 97.8%로 적합률이 높다. 이것은 승용차, 승합차, 화물 및 기타자동차 보유대수가 모두 0대이면 자동차 보유여부를 없음으로 지정함에 따라 적합률이 높아졌다고 볼 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 없음 39가구, 있음 95가구에서 없음 42가구, 있음 92가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

10) 주차시설

〈표 68〉 주차시설에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)			
방법3	거처의 종류, 주인가구 여부, 난방시설, 시도, 시군구, 읍면동	Hierarchical Hot-deck	85.1			
대체 전후의 분포변화(표본)						
실제가구	자가 주차장(70)	영업용 주차장(4)	노상 주차장(5)	도로변, 골목길(4)	공터(1)	해당없음(50)
방법3	70	4	2	6	2	50

주차시설에 대한 적합률은 85.1%로 잘 대체한다고 볼 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 자가주차장 70가구, 영업용주차장 4가구에서 자가주차장 70가구, 영업용주차장 4가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

11) 점유 형태

<표 69> 주거영업구분에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류	Probability	94.3	91.8
방법2	거처의 종류, 건물층수, 거주층수, 시도, 시군구, 읍면동	상동	93.1	89.6
방법3	상동	Hierarchical Hot-deck	93.0	93.3

대체 전후의 분포변화(표본)

실제가구	주거전용(126)	영업겸용(8)
방법1	129	5
방법2	126	8
방법3	131	3

주거영업구분에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 93.0%, 표본자료 93.3%의 적합률을 보인다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법2]가 실제가구수인 주거전용 126가구, 영업겸용 8가구에서 주거전용 126가구, 영업겸용 8가구로 대체되어 분포의 왜곡은 없다.

그러나 적합률은 어떠한 방법이든 비슷하므로 결측값이 발생할 수 있는 Probability방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

〈표 70〉 점유형태에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	주거영업구분	Probability	60.6	58.2
방법2	주인가구 여부, 거실수, 화장실 사용여부, 시도, 시군구, 읍면동	상동	72.9	76.9
방법3	상동	Hierarchical Hot-deck	73.7	75.4

대체 전후의 분포변화(표본)

실제가구	자기집 (67)	전세 (16)	보증금 있는 월세(41)	보증금 없는 월세(2)	사글세 (0)	무상 (8)
방법1	98	17	14	1	0	4
방법2	62	25	40	1	0	6
방법3	62	25	38	1	1	7

점유형태에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 73.7%, 표본자료 75.4%의 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 12.3%, 표본 18.7% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0.8% 증가, 표본 1.5%의 감소를 보이고 있지만 효과는 미미하다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법2]가 실제가구수인 자기집 67가구, 전세 16가구 등에서 자기집 62가구, 전세 25가구 등으로 대체가 되어 분포의 왜곡이 가장 적은 편이라고 볼 수 있다. 하지만 [방법3]도 비슷한 분포를 가지고 있으므로 결측값이 발생할 수 있는 Probability방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

<표 71> 사글세월수에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법3	가구소득, 침실이외의 방수, 건물층수, 거주층수	Hierarchical Hot-deck	-

표본자료 대체정확도

구성비(%)	일치	기타 차이
방법3	99.3	0.7

주) 기타차이는 일정한 값으로 대체해야 하는데 해당없음으로 대체하거나 해당없음으로 대체해야 하는 데 일정한 값으로 대체한 경우

사글세월수에 대한 표본평균의 차이는 0으로 실제 표본평균과 대체한 후의 표본평균의 차이는 없다. 이것은 사글세월수가 12개월이라는 값만 응답되어 있기 때문이다.

대체정확도 검토결과 [방법3]이 정확한 사글세월수를 99.3%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

12) 임차료

<표 72> 전세보증금에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법3	주인가구 여부, 점유형태, 거실수, 거주 기간, 시도, 시군구, 읍면동	Hierarchical Hot-deck	158.565

대체 전후의 분포변화(표본)

실제가구	0-1999만원(30)	2000-3999만원(10)	4000-5999만원(6)	6000-7999만원(0)	8000-9999만원(1)	10000만원 이상(4)	해당없음(83)
방법3	28	15	3	2	3	1	82

전세보증금에 대한 표본평균의 차이는 158.565만원으로 실제 표본평균과 대체한 후의 표본평균의 차이가 적다고 할 수 없다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 0-1999만원 30가구, 2000-3999만원 10가구, 해당없음 83가구 등에서 0-1999만원 28가구, 2000-3999만원 15가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 73〉 월세에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법3	점유형태, 주인가구 여부, 전세보증금, 거처의 종류	Hierarchical Hot-deck	0.360

대체 전후의 분포변화(표본)

실제가구	0-39만원(32)	40-79만원(4)	80-119만원(1)	해당없음(97)
방법3	33	2	1	98

월세에 대한 표본평균의 차이는 0.360으로 실제 표본평균과 대체한 후의 표본평균의 차이는 거의 없는 것으로 판단된다. 그러므로 [방법3]을 선택할 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 0-39만원 32가구, 해당없음 97가구 등에서 0-39만원 33가구, 해당없음 98가구 등으로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

13) 가구소득

〈표 74〉 가구소득에 대한 대체결과

표본자료	대체군	대체방법	적합률(%)
방법3	시도, 시군구, 읍면동, 조사구번호, 타지주택소유, 전세보증금, 가구원수, 가구 구분	Hierarchical Hot-deck	23.1

대체 전후의 분포변화(표본)

실제 가구	1000만원 미만(33)	1000-2000 만원미만(30)	2000-3000 만원미만(26)	3000-4000 만원미만(21)	4000-6000 만원미만(14)	6000-8000 만원미만(8)	8000-1억 원미만(2)
방법3	39	35	23	16	19	1	1

가구소득에 대한 적합률은 23.1%로 낮다. 현재의 상태로는 이 방법을 적용하기 힘들다. 현재 조사된 항목으로 적합률을 높이는 것이 어려울 수도 있다.

대체 전후의 분포 왜곡문제가 발생하고 있다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 1000만원 미만 33가구, 3000-4000만원 미만 21가구, 4000-6000만원 미만 14가구 등에서 1000만원 미만 39가구, 3000-4000만원 미만 16가구, 4000-6000만원 미만 19가구 등으로 대체되어 분포의 왜곡을 가지고 온다. 그러므로 이 항목을 대체시킬 것인지에 대해서는 좀 더 고민이 필요하다.

14) 타지 주택 소유 및 주인가구 여부

<표 75> 타지 주택 소유에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	주인가구 여부	Probability	76.6	76.9
방법2	점유형태, 시도, 시군구, 읍면동, 건물층수, 거주층수	상동	78.4	82.8
방법3	상동	Hierarchical Hot-deck	77.4	79.1

대체 전후의 분포변화(표본)

실제가구	다른 곳에 주택 소유(14)	다른 곳에 주택 미소유(120)
방법1	19	115
방법2	17	117
방법3	16	118

주거영업구분에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 77.4%, 표본자료 79.1%의 적합률을 보인다.

적합률은 어떠한 방법이든 비슷하므로 결측값이 발생할 수 있는 Probability방법보다 Hierarchical Hot-deck방법을 선택하는 것이 더 타당하다고 본다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 다른 곳에 주택 소유 14가구, 다른 곳에 주택 미소유 120가구에서 다른 곳에 주택 소유

16가구, 다른 곳에 주택 미소유 118가구로 대체되어 분포의 왜곡이 적은 편이라고 할 수 있다.

〈표 76〉 주인가구 여부에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	점유형태	Probability	79.4	75.4
방법2	점유형태, 시도, 시군구, 읍면동, 조사구번호, 화장실 사용여부	상동	86.5	82.8
방법3	상동	Hierarchical Hot-deck	86.5	84.3

대체 전후의 분포변화(표본)

실제가구	주인가구(73)	대표가구(35)	기타 세들어 살고 있는 가구(26)
방법1	72	43	19
방법2	67	41	26
방법3	67	41	26

주인가구 여부에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 86.5%, 표본자료 84.3%의 높은 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 7.1%, 표본 7.4% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 0%, 표본 1.5% 증가한다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제가구수인 주인가구 73가구, 대표가구 35가구, 기타 세들어 살고 있는 가구 26가구에서 주인가구 67가구, 대표가구 41가구, 기타 세들어 살고 있는 가구 26가스로 대체가 되어 분포의 왜곡이 적은 편이라고 볼 수 있다. [방법2]의 경우도 동일한 분포를 가진다.

다. 주택에 관한 사항

1) 거처의 종류

<표 77> 거처의 종류에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	시도, 시군구, 읍면동, Type	Probability	42.6	47.0
방법2	방수, 시도, 시군구, 읍면동, 건축년도, 주거용연면적	상동	81.3	90.4
방법3	상동	Hierarchical Hot-deck	80.3	91.3

주) Type은 주택관련 항목이 모두 무응답인지 아닌지를 판단하는 변수임

대체 전후의 분포변화(표본)

실제 주택	단독주택 (35)	아파트 (59)	연립주택 (4)	다세대 주택(9)	비주거용 건물(0)	오피스텔 (7)	판잣집, 비닐하우스(0)	기타(1)
방법1	37	51	3	16	4	4	0	0
방법2	28	61	7	8	3	7	0	1
방법3	29	64	4	9	2	6	1	0

거처의 종류에 대한 무응답 처리를 할 때 [방법3]이 가장 효율적인 것으로 나타났다. [방법3]은 전수자료 80.3%, 표본자료 91.3%의 적합률을 보인다.

적합률은 [방법2]처럼 대체군을 변경함으로써 [방법1]보다 전수 38.7%, 표본 43.4% 증가한다. [방법3]처럼 대체방법을 변경할 경우는 전수 1.0% 감소, 표본 0.9% 증가를 보이고 있지만 효과는 미미하다. 이는 대체방법의 변경보다는 대체군의 변경이 효율성 측면에서는 더 효과적이라는 것을 알 수 있다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제주택수 단독주택 35주택, 아파트 59주택, 연립주택 4주택 등에서 단독주택 29주택, 아파트 64주택, 연립주택 4주택 등으로 대체가 되어 분포의 왜곡이 적은 편이라고 볼 수 있다.

〈표 78〉 단독주택의 종류에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	Size_m, 주거용 연면적, 방수, 대지면적, 시도, 시군구, 읍면동, 조사구번호	Hierarchical Hot-deck	87.9	93.9
방법3	부속수, 독립된 출입구수, 화장실수, 방수	상동	89.7	94.8

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

대체 전후의 분포변화(표본)

실제주택	일반 단독주택(21)	다가구 단독주택(4)	영업겸용단독주택(2)	해당없음(88)
방법1	21	2	4	88
방법3	19	5	3	88

단독주택의 종류에 대한 적합률은 [방법3]처럼 대체군을 변경시켜도 적합률은 크게 향상시키지는 못했다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 일반 단독주택 21주택, 다가구 단독주택 4주택 등에서 일반 단독주택 19주택, 다가구 단독주택 5주택 등으로 대체가 되어 분포의 왜곡이 적은 편이라고 볼 수 있다.

2) 주거용 연면적

〈표 79〉 주거용 연면적에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, 방수, 부속수, 화장실수, 독립된 출입구수	Hierarchical Hot-deck	1.435	4.887
방법3	방수, 거처의 종류, 시도, 시군구, 읍면동, 단독주택의 종류	상동	2.034	9.909

대체 전후의 분포변화(표본)

실제주택	16-32m ² (5)	33-48m ² (9)	49-65m ² (32)	66-98m ² (46)	99-131m ² (7)	132-164m ² (7)	165-198m ² (5)	198-230m ² (3)	231-329m ² (1)
방법1	7	13	28	46	12	5	3	1	0
방법3	3	14	33	49	8	5	3	0	0

주거용 연면적에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료는 0.599, 표본자료는 5.022만큼 증가하고 있다. 이것은 [방법1]이 더 좋은 대체방법이라는 것을 의미한다. [방법3]의 대체군에 [방법1]의 대체군을 고려해야만 할 것이다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제주택수 16-32m² 5주택, 66-98m² 46주택 등에서 16-32m² 3주택, 66-98m² 49주택 등으로 대체가 되어 분포의 왜곡이 적은 편이라고 볼 수 있다.

3) 대지 면적

<표 80> 대지 면적에 대한 대체결과

표본자료	대체군	대체방법	표본평균 차이
방법1	Size_m, 주거용 연면적, 방수, 시도, 시군구, 읍면동, 조사구번호	Hierarchical Hot-deck	62.209
방법3	시도, 시군구, 읍면동, 단독주택의 종류, 주거용 연면적, 화장실수	상동	43.817

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

대체 전후의 분포변화(표본)

실제주택	33-48m ² m ² (1)	49-65m ² m ² (2)	66-98m ² m ² (6)	99-131m ² m ² (4)	132-164m ² m ² (7)	165-198m ² m ² (5)	198-230m ² m ² (1)	231-329m ² m ² (5)	330m ² 이 상(7)	해당없음 (77)
방법1	0	2	7	6	5	5	0	6	6	78
방법3	1	1	3	7	9	7	0	4	6	77

대지면적에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 18.392만큼 감소하고 있다. 이것은 [방법1]보다는 [방법3]이 좋다는 것을 의미하지만 43.317m²의 차이를 실제값과 대체한 후의 값의 차이가 적다고 말할 수 없다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제주택수 33-48m² 1주택, 132-164m² 7주택 등에서 33-48m² 1주택, 132-164m² 7주택 등으로 대체가 되어 분포의 왜곡이 적은 편이라고 볼 수 있다.

4) 총 방 수

〈표 81〉 방수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, Size_m, 주거용 연면적, 부엌수, 화장실수, 독립된 출입구수	Hierarchical Hot-deck	0.096	0.105
방법3	단독주택의 종류, 주거용 연면적, 거처의 종류, 시도, 시군구, 읍면동	상동	0.082	0.090

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이	3개 차이	4개 차이
방법1	89.6	6.1	2.6	-	1.7
방법3	90.4	7.8	-	-	1.8

방수에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료는 0.014, 표본자료는 0.015 만큼 감소하고 있다. 두 가지 방법 모두 표본평균의 차이는 거의 없는 것으로 판단할 수 있다. 그러므로 좀 더 적은 평균의 차이가 있는 [방법3]이 더 타당하다고 할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 방수 90.4%, 1개의 차이 7.8%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

<표 82> 거실수(대청마루)에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 화장실수, 독립된 출입구수	Hierarchical Hot-deck	0.024	0.057
방법3	부엌수, 독립된 출입구수, 화장실수, 단독주택의 종류	상동	0.028	0.057

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이
방법1	96.5	3.5
방법3	96.5	3.5

거실수에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료 0.004만큼 증가하고 있다. 두 가지 방법 모두 표본평균의 차이는 거의 없는 것으로 판단할 수 있다. [방법1]과 [방법3]의 차이는 크다고 할 수 없으므로 [방법3]을 받아들이는 것에 문제가 없다.

대체정확도 검토결과 [방법3]이 정확한 거실수 96.5%, 1개의 차이 3.5%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

<표 83> 식당수(부엌이 딸린 식당 포함)에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 화장실수, 독립된 출입구수	Hierarchical Hot-deck	0.057	0.059
방법3	독립된 출입구수, 부엌수, 거실수, 거처의 종류	상동	0.042	0.038

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이
방법1	93.0	7.0	-
방법3	94.8	4.3	0.9

식당수에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료 0.015, 표본자료 0.021 만큼 감소하고 있다. 두 가지 방법 모두 표본평균의 차이는 거의 없는 것으로 판단할 수 있다. 그러므로 좀 더 적은 평균의 차이가 있는 [방법3]이 더 타당하다고 할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 방수 94.8%, 1개의 차이 4.3%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

5) 건축년도

〈표 84〉 건축년도에 대한 대체결과

	대체군	대체방법	적합률(%)	
			전수	표본
방법1	거처의 종류, 시도, 시군구, 읍면동, 조사구번호	Hierarchical Hot-deck	74.7	73.0
방법3	시도, 시군구, 읍면동, 조사구번호, 거처의 종류, 단독주택의 종류, 방수	상동	76.2	77.4

대체 전후의 분포변화(표본)

실제주택	2008년 (0)	2007년 (0)	2006년 (13)	2000-20 05년(39)	1995-19 99년(28)	1990-19 94년(10)	1980-19 89년(12)	1970-19 79년(9)	1960-19 69년(2)	1959년 이전(2)
방법1	0	0	16	33	33	3	15	10	3	2
방법3	1	3	14	38	25	11	13	6	2	2

건축년도에 대한 무응답 처리를 할 때 [방법3]이 효율적인 것으로 나타났다. [방법3]은 전수자료 76.2%, 표본자료 77.4%의 적합률을 보인다.

적합률은 어떠한 방법이든 비슷하다. 현재의 대체군으로는 적합률을 현저히 향상시키는 것은 어려운 것 같다.

대체 전후의 분포 왜곡문제는 거의 없는 것으로 판단된다. 대체 전후 항목분포의 변화를 살펴보면 [방법3]이 실제주택수 2008년 0주택, 2000-2005년 39주택 등에서 2008년 1주택, 2000-2005년 38주택 등으로 대체가 되어 분포의 왜곡이 적은 편이라고 볼 수 있다.

6) 편익시설 수

<표 85> 부엌수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, Size_m, 주거용 연면적, 방수, 화장실수, 독립된 출입구수	Hierarchical Hot-deck	0.050	0.120
방법3	독립된 출입구수, 화장실수, 거실수, 단독주택의 종류	상동	0.012	0.032

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이	3개 차이	4개 차이	5개 이상
방법1	88.7	6.9	2.6	-	-	0.9
방법3	95.6	2.6	0.9	-	-	0.9

부엌수에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료 0.038, 표본자료 0.088만큼 감소하고 있다. 두 가지 방법 모두 표본평균의 차이는 거의 없는 것으로 판단할 수 있다. 그러므로 좀 더 적은 평균의 차이가 있는 [방법3]이 더 타당하다고 할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 부엌수 95.6%, 1개의 차이 2.6%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

〈표 86〉 화장실수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 독립된 출입구수	Hierarchical Hot-deck	0.040	0.087
방법3	부엌수, 독립된 출입구수, 거실수, 단독주택의 종류	상동	0.015	0.052

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이	3개 차이	4개 차이
방법1	85.2	11.3	2.6	-	0.9
방법3	75.6	22.6	0.9	0.9	-

화장실수에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료 0.025, 표본자료 0.035만큼 감소하고 있다. 두 가지 방법 모두 표본평균의 차이는 거의 없는 것으로 판단할 수 있다. 그러므로 좀 더 적은 평균의 차이가 있는 [방법3]이 더 타당하다고 할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 화장실수 75.6%, 1개의 차이 22.6%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

〈표 87〉 독립된 출입구수에 대한 대체결과

	대체군	대체방법	표본평균 차이	
			전수	표본
방법1	거처의 종류, Size_m, 주거용 연면적, 방수, 부엌수, 화장실수	Hierarchical Hot-deck	0.044	0.091
방법3	부엌수, 화장실수, 거실수, 단독주택의 종류	상동	0.004	0.044

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

표본자료 대체정확도

구성비(%)	일치	1개 차이	2개 차이	3개 차이	4개 차이	5개 이상
방법1	87.8	7.8	2.6	-	-	0.9
방법3	94.8	3.4	-	0.9	-	0.9

독립된 출입구수에 대한 표본평균의 차이는 [방법3]처럼 대체군을 변경함으로써 [방법1]보다 전수자료 0.040, 표본자료 0.047만큼 감소하고 있다. 두 가지 방법 모두 표본평균의 차이는 거의 없는 것으로 판단할 수 있다. 그러므로 좀 더 적은 평균의 차이가 있는 [방법3]이 더 타당하다고 할 수 있다.

대체정확도 검토결과 [방법3]이 정확한 독립된 출입구수 94.8%, 1개의 차이 3.4%를 추정하여 정확도가 높은 방법이라는 것을 알 수 있다.

2. 무응답 처리방법에 대한 제안

가. 가구원에 관한 사항

가구원관련 항목을 무응답 처리하여 효율성을 비교해 본 결과 <표 88>과 같은 대체군 및 대체방법 결과를 얻을 수 있다.

<표 88> 가구원에 관한 사항의 무응답 처리

조사항목	무응답 처리여부	대체군	대체방법
성명	X	주관식으로 응답	-
성별	○	가구주와의 관계, 혼인상태, 교육정도	Hierarchical Hot-deck
나이	○	혼인상태, 교육상태, 교육정도, 가구주와의 관계	"
가구주와의 관계	○	성별, 혼인상태, 교육상태, age_group, 만 나이	"
교육 정도	교육정도	○ 혼인상태, age_group, 만 나이, 교육상태, 시도, 시군구, 읍면동, 조사구번호, 성별, 가구주와의 관계	"
	교육상태	○ age_group, 만 나이, 가구주와의 관계, 혼인상태, 교육정도	"
아동보육	○	교육정도, 출생지 행정구역코드, age_group, 만 나이, 5년 전 거주지	"
출생 지	출생지	○ 5년 전 거주지, age_group, 만 나이, 시도, 시군구, 읍면동, 가구주와의 관계	"
	출생지 행정구역코드	X	정확도 낮음

1년 전 거주지	1년 전 거주지	○	5년 전 거주지, 출생지, age_group, 만 나이, 교육정도	Hierarchical Hot-deck
	1년 전 거주지 행정구역코드	X	정확도 낮음	-
5년 전 거주지	5년 전 거주지	○	1년 전 거주지, age_group, 만 나이, 교육정도, 출생지	Hierarchical Hot-deck
	5년 전 거주지 행정구역코드	X	정확도 낮음	-
활동제약		X	민감함 항목임. 왜곡될 가능성 존재	-
통근·통학 여부		○	age_group, 만 나이, 교육상태, 취업여부, 가구주와의 관계	Hierarchical Hot-deck
통근·통학 장소	통근·통학 장소	○	시도, 시군구, 읍면동, 조사구번호, 이용교통수단, 통근·통학 소요시간, 교육정도	"
	통근·통학장소 행정구역코드	○	시도, 시군구, 읍면동, 1년 전 거주지 행정구역코드, 통근·통학 소요시간, 5년 전 거주지 행정구역코드	"
이용교통수단		○	읍면동, 통근·통학 여부, 통근·통학 장소, GOschool, 성별, age_group, 만 나이, 교육정도, 취업여부	"
통근·통학 소요시간		○	시도, 시군구, 읍면동, 통근·통학 여부, 통근·통학 장소, 이용교통수단, GOschool, 성별, age_group, 만 나이, 교육정도	"
경제활동상태	취업여부	○	통근·통학 여부, 성별, age_group, 만 나이, 이용교통수단	"
	구직여부	○	marry_group, 혼인년도, 시도, 시군구, 읍면동, 조사구번호, 교육정도, 교육상태	"
	취업가능성	○	시도, 시군구, 읍면동, 조사구번호	"
종사상 지위		○	성별, 통근·통학 여부, 근로장소, age_group, 만 나이	"
산업		X	주관식으로 응답	-
직업		X	주관식으로 응답	-
근로장소		○	통근·통학 여부, 성별, 종사상 지위, 시도, 시군구, 읍면동	Hierarchical Hot-deck
혼인상태		○	가구주와의 관계, age_group, 만 나이, 교육정도, 교육상태	"
혼인년월	혼인년도	○	age_group, 만 나이, 교육정도, 혼인상태, 통근·통학 여부	"
	혼인월	○	시도, 시군구, 읍면동, marry_group, 혼인년도, 교육정도, 5년 전 거주지	"

총 출생 아 수	남자 출생아수	○	대체군 없음	무응답은 0으로 처리
	여자 출생아수	○	대체군 없음	"
	남자 사망아수	○	여자 사망아수, 남자 출생아수, 여자 출생아수, marry_group, 혼인년도	무응답은 0으로 처리 Hierarchical Hot-deck
	여자 사망아수	○	남자 사망아수, 여자 출생아수, marry_group, 혼인년도, age_group, 만 나이	"
고령자 생활비 원천	○	통근·통학 여부, 성별, 시도, 시군구, 읍면동, 조 사구번호	Hierarchical Hot-deck	
국적 및 입국연도	X	무응답 처리가 필요한 범위를 한정하기 어려움. 대체군 개발이 모호	-	

주) age_group은 만 나이를 5세별로 그룹화한 변수, marry_group은 혼인년도를 10년별로 그룹화한 변수, GOschool은 재학여부를 의미하는 변수임

나. 가구에 관한 사항

가구관련 항목을 무응답 처리하여 효율성을 비교해 본 결과 <표 89>와 같은 대체군 및 항목대체방법 결과를 얻을 수 있다.

<표 89> 가구에 관한 사항의 무응답 처리

조사항목	무응답 처리여부	대체군	대체방법
가구 구분	가구 구분	ga_wonsu_group, 가구원수, 거실수, 주인가구 여부, 점유형태	Hierarchical Hot-deck
	가구원수	가구 구분, 거실수, 주인가구 여부, 점유형태	"
거주 기간	X	정확도 낮음	-
사용 방 수	침실수	거실수, ga_wonsu_group, 가구원수, 가구 구분, 주인가구 여부	Hierarchical Hot-deck
	침실이외의 방수	침실수, 거실수, ga_wonsu_group, 가구원수, 시도, 시군구, 읍면동	"
	거실수	침실수, 거처의 종류, 시도, 시군구, 읍면동, ga_wonsu_group, 가구원수	"
	식당수	거처의 종류, 거실수, 시도, 시군구, 읍면동, floor_group, 건물층수	"
건물 및 거주 층	건물층수	floor2_group, 거주층수, 거처의 종류, 시도, 시군구, 읍면동, 주인가구 여부	"
	거주층 구분	화장실 사용여부, 점유형태, 주인가구 여부	"
	거주층수	floor_group, 건물층수, 거처의 종류, 시도, 시군구, 읍면동, 주인가구 여부	"

주거 시설 형태	부엌 형태	○	목욕시설 형태, 목욕시설 사용여부, 화장실 형태, 화장실 사용여부	Hierarchical Hot-deck
	부엌 사용여부	○	거처의 종류, 시도, 시군구, 읍면동, 부엌 형태, 목욕시설 사용여부, 화장실 사용여부, 점유형태, 주인가구 여부	"
	수도 형태	○	시도, 시군구, 읍면동, 거처의 종류, 화장실 형태	"
	화장실 형태	○	목욕시설 형태, 부엌형태, 시도, 시군구, 읍면동, 거처의 종류	"
	화장실 사용여부	○	거처의 종류, 시도, 시군구, 읍면동, 화장실 형태, 목욕시설 사용여부, 목욕시설 형태, 주인가구 여부, 점유형태	"
	목욕시설 형태	○	화장실 형태, 화장실 사용여부, 부엌 형태, 거처의 종류	"
	목욕시설 사용여부	○	거처의 종류, 시도, 시군구, 읍면동, 목욕시설 형태, 부엌 사용여부, 화장실 사용여부, 부엌 형태, 점유 형태	"
난방시설	○	시도, 시군구, 읍면동, 주인가구 여부, 목욕시설 형태, floor_group, 건물층수	"	
식수 사용 형태	X	정확도 낮음	-	
정보통신기기 보유현 황	개인용 컴퓨터	○	대체군 없음. 개인용 컴퓨터 보유대수가 1대 이상 이면 '1(있음)'로 처리	무응답은 0과 1로 처리
	인터넷 회선	○	대체군 없음. '1'은 인터넷 회선이 있는 경우임	무응답은 1로 처리
	케이블 TV	○	대체군 없음. '1'은 케이블TV가 있는 경우임	"
	위성방송 수신기	○	대체군 없음. '0'은 위성방송수신기가 없는 경우임	무응답은 0으로 처리
	팩스	○	대체군 없음. '0'은 팩스가 없는 경우임	"
	정보통신 기기보유 여부	○	대체군 없음. 정보통신기기가 모두 '없음'이면 '1(없음)'으로 처리하고 1종류라도 있으면 '0(있음)'으로 처리	무응답은 0과 1로 처리
	개인용컴퓨터보유 대수	○	대체군 없음. 개인용 컴퓨터를 보유하고 개인용 컴퓨터 보유대수가 0대인 경우는 모두 '1'로 변경하고 개인용 컴퓨터를 보유하고 있지 않으면 '0'으로 변경	"
자동차 보유 대수	승용차 보유대수	○	대체군 없음	무응답은 0으로 처리
	승합차 보유대수	○	대체군 없음	"
	화물 및 기타자동차 보유대수	○	대체군 없음	"
	자동차 보유여부	○	대체군 없음. 승용차, 승합차, 화물 및 기타자동차 보유대수가 모두 0대이면 자동차보유여부를 '1(없음)'으로 변경하고 보유대수가 1대라도 있으면 무응답은 '0(있음)'으로 변경	무응답은 0과 1로 처리
주차시설	○	거처의 종류, 주인가구 여부, 난방시설, 시도, 시군구, 읍면동	Hierarchical Hot-deck	

점유 형태	주거영업 구분	○	거처의 종류, floor_group, 건물층수, floor2_group, 거주층수, 시도, 시군구, 읍면동	Hierarchical Hot-deck
	점유형태	○	주인가구 여부, 거실수, 화장실사용여부, 시도, 시군구, 읍면동	"
	사글세월수	○	가구소득, 침실이외의 방수, floor_group, 건물층수, floor2_group, 거주층수	"
임차료	전세보증금	○	주인가구 여부, 점유형태, 거실수, 거주 기간, 시도, 시군구, 읍면동	"
	월세 (사글세)	○	점유형태, 주인가구 여부, rent1_group, 전세보증금, 거처의 종류	"
	가구소득	X	정확도 낮음	-
타지 주택 소유 및 주인가구 여부	타지주택 소유	○	점유형태, 시도, 시군구, 읍면동, floor_group, 건물층수, floor2_group, 거주층수	Hierarchical Hot-deck
	주인가구 여부	○	점유형태, 시도, 시군구, 읍면동, 조사구번호, 화장실 사용여부	"

주) ga_wonsu_group은 6인 이상을 동일한 값으로 준 가구원수의 그룹을 의미하는 변수, floor_group은 건물층수를 5층별로 그룹화한 변수, floor2_group은 거주층수를 5층별로 그룹화한 변수, rent1_group은 전세보증금을 2000만원별로 그룹화한 변수임

다. 주택에 관한 사항

주택관련 항목을 무응답 처리하여 효율성을 비교해 본 결과 <표 90>과 같은 대체군 및 대체방법 결과를 얻을 수 있다.

<표 90> 주택에 관한 사항의 무응답 처리

조사항목		무응답 처리여부	대체군	대체방법
거처의 종류	거처의 종류	○	방수, 시도, 시군구, 읍면동, 건축년도, Size_m, 주거용 연면적	Hierarchical Hot-deck
	단독주택의 종류	○	부속수, 독립된 출입구수, 화장실수, 방수	"
주거용 연면적		○	거처의 종류, 방수, 부속수, 화장실수, 독립된 출입구수	"
대지 면적		○	시도, 시군구, 읍면동, 단독주택의 종류, Size_m, 주거용 연면적, 화장실수	"
총 방수	방수	○	단독주택의 종류, Size_m, 주거용 연면적, 거처의 종류, 시도, 시군구, 읍면동	"
	거실수	○	부속수, 화장실수, 독립된 출입구수, 단독주택의 종류	"

	식당수	○	독립된 출입구수, 부엌수, 거실수, 거처의 종류	Hierarchical Hot-deck
	건축년도	○	시도, 시군구, 읍면동, 조사구번호, 거처의 종류, 단독주택의 종류, 방수	"
편의 시설 수	부엌수	○	독립된 출입구수, 화장실수, 거실수, 단독주택의 종류	"
	화장실수	○	부엌수, 독립된 출입구수, 거실수, 단독주택의 종류	"
	독립된 출입구수	○	부엌수, 화장실수, 거실수, 단독주택의 종류	"

주) Size_m은 면적(m²)을 5평과 10평별로 그룹화한 변수임

제5 절 결론

본 연구에서는 2008년에 실시한 제2차 시험조사 자료에 대하여 무응답을 대체할 수 있는 3가지 방법을 고려하고 있다. [방법1]은 2005년 인구주택총조사 자료 무응답 처리 시 사용한 방법이고, [방법2]는 2008년 대체군과 2005년 대체방법을 적용한 것이고 [방법3]은 2008년 대체군과 2008년 대체방법을 적용한 것이다.

실제 결측치를 발생시킨 후 실시한 모의실험 결과 대체군은 카이제곱 독립성 검정을 통하여 얻어진 새로운 대체군을 사용하는 것이 효율적이라는 결론을 얻을 수 있었다. 또한 대체방법을 Probability방법이나 Hot-deck방법을 변경하여 Hierarchical Hot-deck방법으로 적용하는 것이 꼭 필요하다는 것은 아니라는 사실을 알 수 있었다. 그러나 Hierarchical Hot-deck방법을 이용하면 무응답 처리를 할 때 결측값이 발생하지 않는다는 장점이 있다. 대체의 정확도를 향상시키는 것은 대체방법의 변경보다는 가장 적절한 대체군을 발견하여 적용하는 것이 훨씬 더 큰 영향력이 있다는 결론을 얻을 수 있다.

본 연구결과의 적합률은 주로 70% 이상으로 무응답 대체를 처리해도 좋다는 결론을 얻을 수 있다. 그러나 교육정도, 거주 기간, 가구 소득 등은 기존의 조사문항으로는 대체군을 형성하는 데는 한계가 있

다는 것을 보여준다. 이 항목들을 대체함에 있어서 정확도를 높이기 위해서는 연관성이 높은 새로운 문항의 조사가 필요할 것으로 판단된다. 무응답 대체가 이루어진 후에는 대체된 값이 다른 변수와의 연관 관계에 있어서 논리적으로 문제가 없는지 반드시 확인하는 사후내검 단계를 수행해야 한다. 효율성을 상세히 비교해 보면 무응답 처리가 가능한 총 80개의 세부항목 중 11개 항목(13.75%)은 2005년에 비해 효율성이 향상되었다고 볼 수 있고, 40개 항목(50%)은 효율성이 비슷하다고 할 수 있다. 2008년 제2차 시험조사에서 추가된 항목 등 15개 항목(18.75%)은 새로운 방법을 개발하여 적용하여도 무난하다고 할 수 있다. 그러나 14개 항목(17.5%)은 효율성이 낮은 항목으로 나타났다. 여기서 효율성이 낮다고 보는 것은 적합률이 50% 미만이거나 표본평균의 차이가 1보다 큰 것으로 간주했다. 본 연구에서는 효율성이 낮은 14개 세부항목을 다음과 같이 처리하도록 결론을 내렸다. 나이, 교육정도, 주거용 연면적, 대지 면적은 전수조사와 표본조사의 공통 항목이므로 무응답 처리를 하고 남자 출생아수, 여자 출생아수, 전세보증금과 같이 조사항목의 세부항목인 경우는 무응답 처리를 한다. 제2차 시험조사의 신규항목 및 2005년도에 무응답 처리를 하지 않은 거주 기간, 식수 사용 형태, 가구소득은 무응답 처리를 하지 않았다. 조사항목의 세부항목이지만 출생지 행정구역코드, 1년 전 거주지 행정구역코드, 5년 전 거주지 행정구역코드는 무응답 처리를 하지 않고 2005년도에 무응답 처리를 한 통근통학 소요시간 표본조사 항목은 무응답 처리를 하는 것으로 하였다.

인구주택총조사의 현장조사가 완료되면 자료처리를 한다. 통계자료의 가치는 정확성과 신속성에 의하여 수준을 평가하는 만큼 공표시기의 단축은 의미가 있다. 일련의 자료처리 과정이 공표 이전에 선행 작업으로 이루어지는 만큼 그 과정을 줄여야 신속한 조사결과를 제공할 수 있으므로 무응답 및 오류자료의 처리가 중요한 역할을 차지하게 되었다.

현재 인구주택총조사의 무응답 처리기법 연구는 선행연구가 잘 되어 있고 실제 적용도 하고 있기 때문에 커다란 변경은 불필요할 수도 있다. 그렇지만 이론적인 배경 및 원칙을 정립하는 부분에 있어서는

지속적인 연구가 필요하다. 본 연구는 인구주택총조사의 전체 항목에 대해서 일부 이론적으로 정립을 시키고 효율성도 비교해 보았지만 많은 자료의 양으로 대체군을 개발하지 못한 것에 그 한계점이 있다. 소규모 자료와 대규모 자료에서의 대체군은 다를 수 있기 때문이다. 향후에는 2005년에 실시한 인구주택총조사의 자료를 이용하여 본 연구의 결과에 대한 타당성을 검증해 보고 인구주택총조사의 항목에 더욱 적합한 대체군 및 대체방법을 제시해 보고자 한다.

참고문헌

- 김영원 · 이주원(2003), “CART를 활용한 결측값 대체방법: 인구주택총조사 혼인상태 항목을 중심으로”, 「조사연구」, 제4권 제2호, pp.1-21.
- 김영원 · 조선경(1996), “표본조사에서 항목 무응답 대체 방법”, 「한국통계학회논문집」, 제3권 제3호, pp. 145-159.
- 김재광 · 한근식 · 윤연옥(2004), “가계조사 무응답 처리기법 연구”, 통계청, 「통계연구」, 제9권 제1호, pp.79-102.
- 김진(2004), “농가경제조사에 대한 대체법 비교”, 통계청, 「통계연구」, 제9권 제2호, pp.133-145.
- 송순관(2005), “무응답 처리방법 연구 및 읍·면·동 통계작성 가능성 검토”.
- 이기성 · 한성실 · 정기문(2007), “한글SPSS통계자료분석”, 자유아카데미.
- 이내성(2008), “회귀분석을 이용한 Imputation기법활용 연구”, 통계청, 「국가통계의 품질제고를 위한 방법론 연구」, 2008년 제2호, pp.44-105.
- 이재원(2000), “무응답 및 오류자료의 Imputation 적용 결과”, 「무응답오차」, pp.131-145. 조사통계연구회, 자유아카데미.
- 조사통계연구회(2000), “무응답 오차”, 자유아카데미.
- 통계청(2006), “2005 인구주택총조사 종합평가보고서”. 내부자료.
- 통계청(2008), “제2차 시험조사 조사지침서”. 내부자료.