최종 보고서

2010 인구주택총조사 자료분석 및 활용제고방안 연구 (Ⅱ)

2010. 12.

통계청
STATISTICS KOREA

본 보고서에 수록된 내용의 일부 혹은 전부를 사전 승인 없이 전재, 역재, 복제할 수 없음

제 출 문

통계청장 귀하

본 보고서를 「2010 인구주택총조사 자료 분석 및 활용 제고방안 연구」의 최종보고서로 제출합니다.

2010년 12월

(사) 한국통계학회

목 차

제 Ⅰ 부 다중응답 항목의 통계적 처리 방안

제 Ⅱ 부 마이크로 데이터 제공 방안

제 Ⅰ 부

다중응답 항목의 통계적 처리 방안

Ⅰ-1. 연구개요

1. 연구 배경 및 목적

○ 자료의 무응답은 자료 분석에 어려움을 초래할 뿐 아니라 모수의 추정값에 편향이 발생하므로 무응답 자료에 대한 적절한 처리가 필요함. 현재 통계청에서는 인구주택총조사에서 발생한 무응답에 대하여 대체군(adjustment cell)을 사용한 핫덱대체(hotdeck imputation)를 실시하고 대체된 자료를 제공하고 있음.

○ 인구주택총조사 문항들 중에는 주된 2개의 항목을 선택하거나 해당되는 항목들을 모두 선택하는 다중응답 문항들이 존재함. 다중응답 항목은 한 개의 응답만을 요구하는 일반 문항과 달리 한 응답자로부터 복수의 항목에 대한 응답이 존재하고 한명의 응답자로부터 얻어진 복수의 응답들은 서로 연관성이 존재함.

○ 현재 인구주택총조사의 무응답에 대하여 핫덱대체를 실시하기 위해 대체군을 형성하는데, 대체군으로 사용할 변수들을 선택하기 위하여 연관성 분석 기법인 카이제곱 검정과 의사결정나무를 사용함. 이 방법은 근본적으로 한 개의 응답이 관측된 경우에 적용 가능하므로 현재 사용하는 대체군을 사용한 핫덱대체를 그대로 적용할 수 없음. 본 연구에서는 이 방법을 다중응답 문항의 처리를 위하여 확장하는 방법을 고려함.

○ 인구주택총조사 문항들 중 일부 문항은 주된 2개의 항목에 대한 응답을 요구하였으나 소수 응답자들에게서 2개 이상의 항목에 대하여 응답한 경우를 발견할 수 있음. 이와 같은 보기초과 응답에 대한 적절한 통계적 처리 방법에 관한 연구를 통해 고품질 자료의 생산 및 자료의 문제점에 해결 방안을 도출해 내는 것이 중요함.

○ 아동보육, 이용교통수단, 고령자 생활비 원천, 그리고 주차장소 문항의 경우 보기 항목들 중 한 가지 또는 2가지 이상의 항목이 응답자에게 적용되는 경우 주가 되는 2가지 항목에 대한 응답을 요구하였으나 일부 응답자들에게서 무응답이 발생함. 또한, 주된 것 2가지 항목에 대하여 3가지 이상의 응답을 제공하는 보기 초과 현상이 발생하여 응답 요구 조건에 대한 불일치를 해결하기 위한 통계적 처리 방안을 제시하고자 함.

○ 활동 제약, 사회활동, 그리고 정보통신기기 보유 및 이용 현황에 대한 문항의 경우 적용 가능한 항목을 모두 응답하도록 요구되었고 이 때 발생하는 무응답에 대한 핫덱대체 방법을 제안하고자 함.

○ 교통수단 보유 현황은 2010년에 신규로 추가된 문항들로서 가구 별로 여러 가지 교통수단의 보유대수를 파악함. 각 교통수단 별로 보유 여부에 관해 응답한 후 보유대수를 응답하므로 응답자의 응답값은 양의 정수로 측정됨. 이 문항에 대한 대체군을 형성하는 데 고려할 변수를 파악함.

2. 연구 범위

○ 인구주택총조사 자료에 대하여 무응답을 대체하여 무응답이 없는 완전한 자료를 제공해 왔는데 다중응답 변수의 응답은 단일응답 변수와 달리 한 응답자로부터 다중응답이 발생하고 이 값들 간에 연관성이 존재하므로 이를 고려한 대체 방법을 적용하는 것이 타당함.

○ 각 문항의 응답이 문항의 요구 사항에 일치해야 하지만 일부 다중응답 문항의 자료는 문항의 요구 사항에 부합하지 않음. 즉, 무응답으로 인하여 응답이 측정되지 않거나 주된 2가지 항목에 대한 응답이 요구되었으나 3개 이상의 응답값이 존재하는 경우가 발생하였으므로 자료의 완전성을 달성하고 제공된 자료에 대한 분석의 편의를 추구하기 위하여 다중응답 문항에서의 무응답에 대한 적절한 처리가 필요함.

○ 본 연구에서 고려하는 다중응답문항은 <표 1.1.1>에 나타난 바와 같이 주된 2개까지 응답하는 3개 문항과 해당되는 항목을 모두 선택하는 4개 문항임.

<표 1.1.1> 2010년 인구주택총조사의 다중응답 문항

주된 2개까지 응답 문항	해당되는 항목 모두 선택 문항
• 아동보육	• 활동제약
• 이용 교통수단	• 사회활동
• 고령자 생활비 원천	• 정보통신기기 보유 및 이용 현황
• 주차 장소	

○ 다중응답 문항에 대한 대체 방법 중 현재 사용되는 인구주택총조사 자료의 무응답 대체 프로그램인 대체군을 사용한 핫덱대체와 일치성을 가지면서 다중응답 문항에 대하여 정확한 대체를 실시할 수 있는 대체방법을 제안하기 위하여 본 연구가 진행됨.

○ 신규로 포함된 교통수단 보유대수에 대한 대체를 실시할 때 고려할 대체군을 선정하는 방법을 모색함.

3. 연구 내용

○ 무응답 대체를 위한 방법론에 관한 연구
  – 미국, 캐나다, 유럽, 호주 통계청에서 사용하는 무응답 대체를 위한 방법론을 연구함.

- 무응답에 대한 대체를 실시할 때 고려해야 하는 점을 기술함.
- 다중응답 문항에서 발생하는 무응답과 단일응답 문항의 무응답 간 차이점을 비교하고 다중응답 문항의 무응답에 대한 대체를 실시할 때 특별히 고려해야 할 점을 기술함.
- 다중응답 문항에서 발생하는 무응답에 대한 대체 방법을 제안함.

○ 보기초과 다중응답 문항에 대한 처리 방법 연구
- 다중응답 문항에서 발생하는 보기초과 문항에 대한 통계적 처리 방법을 제안함.

○ 모의실험
- 본 연구에서 제안된 대체 방법의 성능을 파악하기 위하여 모의실험을 실시함.
- 모의실험 자료는 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료를 사용하여 2010년 인구주택총조사 자료의 실제 상황과 가능한 한 비슷한 조건 하에서 모의실험을 실행함.

(1) 세계 각국의 통계청에서 사용하는 무응답 처리 방법에 대한 선행 연구 파악

○ 단일응답 항목에서 발생한 무응답에 대한 대체 방법들은 여러 가지 문헌을 통해 제안되어 왔음. 미국, 캐나다, 호주를 비롯한 여러 나라의 인구주택총조사에서는 무응답 대체를 위하여 다양한 핫덱대체 방법을 적용해 왔음.

○ 대부분의 무응답 대체 연구가 한 개의 해당 항목에 대하여 응답하는 방식으로 측정된 단일응답 문항에 대한 대체 방법에 중점을 두고 있고 다중응답 항목에서 발생하는 무응답에 대한 대체 방법에 대한 선행연구는 파악되지 않음.

○ 인구주택총조사 단일응답 문항에서 발생하는 무응답에 대하여 대체를 실시하는 방법인 대체군을 사용한 핫덱대체와 일치성을 유지하며 기존 대체 프로그램과 연동하여 적용 및 응용 가능성이 있는 대체 방법을 고려함.

○ 핫덱대체를 실시할 때 대체군의 형성에 따라 대체된 자료의 정확도가 결정됨. 대체군을 형성하기 위하여 고려해야 하는 변수들은 무응답이 발생한 변수와 연관성이 있는 변수 및 무응답 발생 여부와 연관성이 있는 변수들을 포함함.

○ Andridge and Little(2010)은 <표 1.1.2>에 나타난 바와 같이 대체군을 형성하기 위하여 무응답이 발생한 변수와 연관성이 높을 뿐 아니라 무응답 발생 원인과도 연관성이 높은 변수를 고려했을 때 평균의 추정량에 대한 편향이 감소할 뿐 아니라 추정량의 분산도 감소한다는 것을 보임. 또한, 대체군을 형성하기 위하여 무응답이 발생한 변수와 연관성이 높은 반면에 무응답 발생 여부와는 연관성이 낮은 변수를 고려한다면 추정량의 편향을 감소시키지는 않지만 추정량의 분산이 감소함을 보여줌. 한편, 대체군을 형성하기 위하여 무응답이 발생한 변수와 연관성이 낮은 반면에 무응답 발생 여부와는 연관성이 높은 변수를 고려한다면 추정량의 편향을 감소시키지는 않는데 반하여 추정량의 분산이 증가함을 보여줌. 즉, 평균 추정량의 편향을 감소시키고 동시에 추

정량의 분산을 줄이기 위하여 반응 변수 뿐 아니라 무응답 발생 여부와 연관된 변수들로 대체군을 형성하는 것이 가장 바람직함.

○ 무응답 대체를 실시할 때 대체군은 대체가 실시될 문항의 값이 대체군 내에서 비슷하도록 선택되어야 함. 더구나, 대체군을 형성하는 변수가 무응답 발생 확률과 연관되어 있다면 추정량의 편향을 줄일 수 있으므로 이를 함께 고려하는 것이 바람직함.

<표 1.1.2> 대체군에 사용된 변수들의 응답 여부 및 무응답 발생변수와의 연관성에 따라 평균의 추정에서 발생하는 편향과 분산의 보정효과 (Andridge and Little(2010)의 Table1에서 인용됨)

		무응답 발생 변수와의 연관성	
		낮음	높음
무응답 발생 여부와의 연관성	낮음	편향: 유지 분산: 유지	편향: 유지 분산: 감소
	높음	편향: 유지 분산: 증가	편향: 감소 분산: 감소

(2) 다중응답 항목에서 발생하는 무응답 처리 방법 연구

○ 단일응답 문항에서 발생하는 무응답에 대하여 대체를 실시하는 경우에 비하여 다중응답 문항에서 발생하는 무응답에 대해 핫덱대체를 실시할 때 어려운 점은 대체군의 형성이 복잡하다는 데 있음.

○ 단일응답 문항에서 대체군을 형성하기 위하여 카이제곱 검정이나 의사결정나무 기법 등이 사용되는 데 이 중 카이제곱 검정은 응답 변수와 대체군을 형성하는 데 고려할 변수 모두가 범주형인 경우에 적용이 가능함. 반면 의사결정 나무 기법은 변수의 타입이 연속형인 경우에도 적용이 가능하다는 장점을 지니지만 모형 선택 옵션에 따른 모형의 민감도가 존재함.

○ 다중응답 문항의 경우 한 응답자로부터 한 개 이상의 응답이 가능한 데 이 경우 여러 개의 응답들과 대체군을 형성하기 위하여 고려하는 변수들 사이의 연관성을 모형화해야 하므로 단일응답 항목보다 모형이 복잡해짐.

○ 무응답에 대한 대체를 실시할 때 다중응답 항목이 1가지 또는 주된 2가지 항목에 대한 응답만을 요구하는 아동보육, 이용교통수단, 고령자 생활비 원천, 그리고 2대 이상의 차를 소유한 가구의 주차장소에 대한 응답은 개인에 따라 1가지 또는 주된 2가지를 응답하는 것이 가능한데 이와 같이 응답자에 따라 변화하는 개수의 응답과 대체군을 형성하는 변수들 간의 연관성은 일반적인

카이제곱 검정 기법이나 의사결정 나무 기법을 이용하여 파악할 수 없음.

○ 무응답에 대한 대체를 실시할 때 다중응답 문항이 응답자에게 해당되는 모든 항목에 대한 응답들을 요구하는 경우 응답은 개인에 따라 1개부터 여러 개까지 다양하게 가능하고 이 경우 다양한 개수의 응답들과 대체군을 형성하는 변수들 간의 연관성을 카이제곱 검정이나 의사결정나무 기법으로 파악하기 어려움.

○ 여러 개의 응답들 중 첫 번째 응답 또는 임의로 선택된 한 개의 응답과 대체군을 형성하는 변수들 간의 연관성을 고려하여 대체군을 형성하는 방법이 적용될 수 있으나 이 경우 동일한 응답자의 여러 응답에서 나온 정보를 포함하지 못하는 단점을 지님.

○ 가능한 모든 항목에 대하여 응답이 가능한 다중응답 문항들의 경우 각 항목별로 대체군을 형성하는 방법이 있지만 이 방법은 한 개인으로부터의 다중 응답들 사이의 연관성을 고려하지 못하는 단점이 존재함.

○ 본 연구에서는 일반화선형모형(generalized linear model)을 고려하여 동일 응답자의 1개 또는 여러 개의 응답들과 대체군을 형성하기 위하여 변수들 간 연관성을 파악하고 이 방법을 통해 선택된 대체군을 사용한 대체를 실시하는 방법의 성능을 파악하고자 함.

○ 제안된 방법은 모의실험을 통해 응답 비율에 편향이 발생하는지 확인함. 모의실험은 2005년 인구주택총조사 10% 표본 자료와 2010년 인구주택총조사 시범조사의 원시자료를 사용하여 진행됨.

(3) 다중응답 문항에서 발생하는 보기 초과 응답에 대한 처리 방안 연구

○ 다중응답 문항이 주된 2개의 항목에 대한 응답만을 요구하지만 실제 수집된 자료에 3개 이상의 응답이 존재하는 경우 자료의 일치성을 만족시키기 위한 자료 처리 방안을 연구함.

○ 최대 2가지의 응답만 요구하였으나 3개 이상의 응답값을 제공한 경우 2개의 응답값만을 유지하고 나머지 응답들은 제외시키는 통계적 처리 방법을 고려함. 이 때 이 항목에 대한 초과 응답자는 3개 이상의 응답들 중 발생 가능한 2개의 항목의 조합에 응답한 사람들을 기증자로 사용하여 대체를 실시함으로써 선택된 대체값이 본인의 응답들 중에서 선정되는 것을 가능하게 함.

○ 예를 들어, 아동보육 문항은 아동의 보육 방법을 2가지까지 선택할 수 있지만 2010년 시범조사의 경우 3가지를 응답한 18명과 4가지를 응답한 3명, 총 0.4%에서 보기에서 요구한 것보다 초과한 개수의 응답을 한 경우가 존재함. 만약 한 응답자의 선택된 문항이 “부모”, “조부모”, “유치원”의 세 가지였다면 이 세 가지 중 2가지를 선택한 응답자들, 즉, (1) “부모”와 “조부모”를 응답한 사람들, (2) “부모”와 “유치원”을 선택한 응답자들, 그리고 (3) “조부모”와 “유치원”을 응답한 사람들만을 기증자로 고려하여 대체를 실시함. 이렇게 대체된 값들은 본인이 선택한 문항 중 2개의 문항을 고르는 효과를 지니므로 응답의 일치성을 유지할 수 있음.

○ 모의실험을 통해 제안된 방법의 성능을 평가하고자 하였으나 보기초과 응답자의 실제 숫자가 너무 적어 모의실험을 통해 평가가 어려움. 제한된 환경 하에서 실제자료와 비슷하게 추가 응답이 발생한 경우를 가정하고 모의실험을 실시함.

Ⅰ-2. 각 국 무응답 대체 실시 현황

1. 한국 인구주택총조사 무응답 대체

○ 인구주택총조사는 우리나라의 인구, 가구, 주택에 관한 정보를 파악하여 각종 정책 입안을 위한 기초 자료를 제공할 뿐 아니라 여러 가지 가구와 관련된 경상조사를 위한 표본틀(sampling frame)을 만드는 데 있어 기초자료로 활용하기 위하여 실시됨.

○ 0과 5로 끝나는 연도를 기준으로 매 5년마다 통계청에서 실시하는 이 조사는 대한민국 영토 중 행정권이 미치는 전 지역을 대상으로 조사기준 시점 현재 조사지역 내에 상주하는 내, 외국인 및 이들이 살고 있는 모든 거처에서 실시됨.

○ 2005년 인구주택총조사 자료의 조사표는 전수조사표와 표본조사표로 구분되어 있고 전수조사표는 기본적인 특성을 파악하기 위해 21개 문항으로 구성되어 있으며, 표본조사표는 전수조사 문항 이외에 보다 세부적인 특성을 파악하기 위한 20개 문항을 추가하여 총 41개 문항으로 구성되어 있음. 이 문항들 외에 추가로 16개 시, 도별로 각각 다른 조사 문항 3개가 포함되어 전체적으로는 44개 조사 문항으로 구성되어 있음.

○ 2005년 인구주택총조사 자료에 대하여 무응답에 대한 대체를 실시한 후 대체된 자료를 제공해 왔음. 사용된 대체 방법은 확률에 근거한 대체(probability imputation), 핫덱대체(hotdeck imputation), 계층적 핫덱대체(hierarchical hotdeck imputation)의 세 가지임. 문항에 따라 세 가지 대체 방법 중 한 가지 대체 방법이 적용됨(이현정, 2009).

(1) 확률에 근거한 대체

○ 확률에 근거한 대체는 무응답이 발생한 변수와 연관된 변수들로 대체군(또는 adjustment cell이라고도 부름)을 형성한 후 동일한 대체군 내에서 응답값들의 비율에 비례하게 무응답값을 대체하는 방법으로서 범주형 변수(categorical variable)의 대체에 적용됨.

(2) 핫덱대체

○ 무응답이 발생한 변수에 대한 응답자들을 모아 기증자 풀(donor pool)을 만들고 그 가운데 무응답

자의 숫자만큼의 기증자를 무작위로 추출하여 각 무응답자에게 한 명의 응답자를 무작위로 할당한 후 기증자의 값을 가지고 무응답을 대체하는 방법으로써 무응답이 발생한 변수들과 연관된 변수들로 구성된 대체군에 따라 자료를 나누어 각 대체군 내에서 대체를 실시함.

○ 가구관련 통계조사의 무응답 처리기법으로 흔히 사용되는 대체방법으로서 연속형 변수 및 범주형 변수 모두에 범용적으로 사용 가능함.

(3) 계층적 핫덱대체

○ 대체군을 이용한 핫덱대체 기법에서 대체군을 형성하는 변수들의 개수가 많아져 대체군의 일부에서 무응답의 숫자에 비하여 기증자의 숫자가 작거나 기증자가 없어 대체할 기증자를 구하기 어려운 경우에 흔히 사용되는 방법으로써 미국 통계청에서 Current Population Survey(CPS)의 소득영역 변수들을 대체하는데 사용했기 때문에 CPS 핫덱(David, et. al., 1986) 또는 융통성있는 짝짓기대체방법(flexible matching imputation method)이라고도 함.

○ 일반적으로 핫덱대체를 시행할 때 응답자 중에서 기증자를 비복원으로 추출하는데 이는 동일한 응답자가 기증자로 여러 번 사용되는 복원추출 시 동일한 값으로 대체된 자료값들이 모두 동일해 져서 추정량의 분산을 과소추정하는 것을 막기 위한 노력의 일환임. 하지만 기증자를 비복원으로 추출해 대체를 실시하려면 모든 대체군 내에서 응답자의 숫자가 무응답의 숫자보다 훨씬 커야 하는데 실제 자료의 경우 자료의 특성상 일부 대체군 내에서 무응답의 숫자가 응답자의 숫자보다 많아 무응답의 일부에 대하여 대체할 기증자를 발견할 수 없는 경우가 종종 발생하므로 이와 같은 경우에 모든 무응답에 대한 기증자를 얻기 위해 고안된 방법임.

○ 대체군을 이용한 핫덱대체 기법을 적용할 때 대체군을 만드는 변수들이 많아져 기증자를 찾기 어려운 경우에 흔히 사용되는 방법으로서 우선 대체군 내에서 기증자를 찾을 수 있는 응답값은 그 기증자의 값으로 대체하지만 기증자를 찾을 수 없는 일부 무응답값은 대체군을 형성하는 변수를 기증자를 찾을 때까지 하나씩 포기해 나가는 대체 방식을 의미함. 즉, 기증자를 찾을 때까지 대체군 형성에 사용되는 변수의 숫자를 하나씩 줄여나가는 핫덱대체 방식임.

○ 이 방법의 단점은 동일한 대체군 내에서도 일부는 기증자를 찾을 수 있는데 반하여 기증자를 못 찾은 일부 무응답값을 대체군을 형성하는 변수의 숫자를 줄여 기증자를 찾으므로 동일한 대체군 내의 무응답값에 대해서도 대체군이 계속적으로 변화하고 포기하는 변수의 순서에 따라 대체의 성능이 달라지므로 이 방법으로 대체된 자료의 경우 대체의 정확성에 대한 이론적 평가가 어려움.

○ 이 방법은 모든 대체군 내에서 기증자를 찾을 수 있는 경우 (2)에서 언급한 대체군에 근거한 핫덱

대체와 동일함.

○ 다중응답 문항의 무응답에 대하여 이 방법을 적용하는 방법에 대한 명확한 기술이 문헌에 나타나 있지 않음.

○ 2005년 인구주택총조사 자료의 대체를 실시할 때는 위의 세 가지 방법들 중 한 가지 방법이 변수에 따라 선택되어 적용되었지만 2010년 인구주택총조사 자료의 대체는 모든 변수에 대하여 계층적 핫덱대체를 실시하도록 통합될 예정임(통계청, 2010).

2. 미국 센서스 조사(Census Enumeration)에서의 무응답 대체

○ 1940년에 미국 인구통계국(U.S Census Bureau)은 응답자의 나이에서 발생하는 무응답에 대하여 대체를 실시하는 것을 시초로 대체가 실시되는 변수들의 숫자가 증가되어 옴(Cantwell, Hogan and Styples, 2005). 2000년 센서스에서는 가구에 대한 변수들 중 가족의 숫자, 거주여부, 그리고 거주상태 변수들에 대하여 대체를 실시하였고 나이, 성별, 인종, 히스패닉계(Hispanic origin) 여부와 같은 개인 변수들에 대한 대체도 실시함.

○ 1960년 센서스 이후 무응답 대체를 위해 핫덱방법이 사용됨. 순차적 핫덱 방법(sequential hotdeck method)이 사용되는데 이 방법은 일반적으로 동일한 특징을 지닌 거주지의 근처 가구(neighboring housing unit)나 동일가구 내 다른 가구원의 응답값을 가지고 무응답에 대한 대체를 실시함(Wetrogan and Cresce, 2001).

○ 대체는 전체 가구 대체(whole household imputation), 가구 내 대체(within household imputation), 그리고 개인 내 대체(within person imputation)로 구분됨(U.S. Census Bureau, 2004).

○ 전체 가구 대체는 가구 전체의 정보가 전혀 존재하지 않는 경우에 사용되는데 최근접이웃으로 구성된 기증자 풀(nearest neighbor donor poll)에서 동일한 가구 크기를 지닌 가구의 응답값을 가지고 대체를 실시하는 방법으로서 가구 대치(substitution)이라고도 부름.

○ 가구 내 대체와 개인 내 대체는 응답을 제공하지 않은 개인과 동일한 가구내 다른 가구원의 응답값에 근거하여 핫덱대체를 실시함.

○ Census 2000 Accuracy and Coverage Evaluation (A.C.E.)에서는 여러 가지 대체 방법을 혼용하여 사용하였는데 예를 들어 나이와 성별은 P sample의 분포를 고려하여 대체되었고 Tenure는 최근접 이웃 핫덱대체(nearest neighbor hotdeck imputation)를 사용하였으며 인종과 히스패닉 계 여부는

위 두 방법을 결합하여 대체를 실시하였음. Tenure, 인종과 히스패닉 계 여부의 대체 시 동일한 블록 군집(block cluster)에 사는 사람들로 대체하여 지리적 근접성(geographic proximility)을 고려한 대체를 실시함(U.S. Census Bureau, 2004).

3. 미국 인구통계국의 Survey and Income Program Participation(SIPP)에 대한 무응답 대체

○ 결측자료 메커니즘(missing data mechanism)이 임의결측(missing at random)(Little and Rubin, 2002)이라는 가정 하에서 대체를 실시함. 즉, 자료에서 무응답이 발생할 확률은 대체군 내에서는 완전히 임의라는 가정 하에서 대체가 실시되며 이는 대부분의 무응답 대체 방법이 가정하는 결측자료 메커니즘임.

○ 통계적 짝짓기(statistical matching)와 핫덱대체 방법(hotdeck procedure)이 무응답의 대체를 실시하기 위하여 사용됨(Westat and Mathematica Policy Research, 2001).

○ 이 자료는 패널자료(panel data)로서 1996년 자료에 대한 무응답 대체는 그 전 측정 시점(previous wave)의 응답값 정보에 의존하는 핫덱 방법을 사용함. 또한, 필요한 경우 핫덱대체와 콜드덱 대체(colddeck imputation)을 혼용하여 적용함.

○ 순차적 핫덱 방법(sequential hotdeck procedure)이 적용되었는데 이 방법은 우선 콜드덱이나 초기 기증값을 가지고 자료를 정렬한 후 대체군 내에서 핫덱대체를 실시하는 방법임. 추적조사(Wave 2 이후 자료)에서는 대체군을 형성할 때 지난 시점(previous wave)의 응답값을 포함시킴으로써 연관된 정보를 대체를 실시하기 위하여 포함함.

4. 캐나다 센서스(Census)에 대한 무응답 대체

○ 캐나다에서는 센서스 자료에 대한 편집 및 대체(edit and imputation)를 위하여 CANEDIT 프로그램을 사용함. C언어로 쓰인 CANEDIT은 최소변화 대체 시스템(minimum change imputation system)(Fellegi and Holt, 1976)의 원리를 사용하는 최근접이웃 대체방법을 사용함.

○ 2001년 센서스 자료의 인구학적 변수에 대한 편집 및 대체는 5단계로 진행됨(Mason, Bankier, and Poirier, 2002).

(1단계) 구조적 오류(systematic error)에 대한 결정적 대체(deterministic imputation) 실시.

(2단계) 편집을 위하여 커플, 자녀, 조부모 변수의 생성.

(3단계) 편집 실시.

(4단계) 최소변화 대체 시스템을 사용하여 최근접이웃 대체 방법으로 대체실시

- 최근접이웃은 유사성(similarity)에 근거하여 결정하는데 유사성은 지리학적 거리에 근거하여 결정됨.
- 대체되어야 하는 가구에 대하여 N개의 기증자를 선택하고 이 기증자의 값들 중 대체되어야 하는 변수에 대한 대체만 실시함.
- N개의 기증자로부터 가능한 대체들을 모두 고려하여 그 중 가장 최적인 대체를 선택함.

(5단계) 전수조사 항목에 대하여 보조 변수를 사용함.

- 특정 연령대에 대하여 질의된 항목은 그 정보를 사용하여 대체를 실시함.

○ 2006년 Census에서는 모든 변수에 대하여 이 방법이 적용됨.

○ 대체는 개인 단위가 아니라 가구 단위로 실시됨.

5. 호주통계국의 주제별 조사지를 사용하는 센서스(Thematic Form Census)에 대한 무응답 대체

○ 호주 통계청에서 전수/표본 조사를 실시하는 대신 몇 가지 다른 형식의 조사지를 사용하여 센서스를 실시한 후 발생하는 무응답에 대한 대체 방법으로 균형 대체 기법(balanced imputation approach)을 고려함(Bell and Whiting, 2007).

○ 균형대체란 대체된 자료가 특정한 기준(criteria)을 만족시키도록 대체 자료들 중에서 선택하는 기법을 의미함. 이 방법은 관심 있는 여러 가지 표에 대한 균형 잡힌 대체 자료를 생성하는 것을 목적으로 하므로 표의 칸(cell) 중에서 큰 표본오차(sampling error)를 가지는 칸에 의해 상대적으로 덜 영향 받고 추정량을 안정적으로 구할 수 있는 기준을 선택하여 대체를 실시하는 기법을 의미함. 이 때 사용된 대체 기법은 핫덱대체임.

○ 균형대체는 각 개인 단위가 아니라 표의 통합된 수준(aggregate level)에서 우수한 대체를 실시하는 대체 자료를 선택함으로써 개인 단위의 대체에 사용될 분포를 보완하고자 함.

○ 지리적 위치(geography)가 균형 잡힌 표를 생성하는 데 주요한 역할을 함. 즉, 알고리즘은 각 지역(geographical area)에 대하여 다음과 같이 4단계로 진행됨

(1단계) 예측된 대체(predicted imputation)의 균형을 평가.

(2단계) 임의난수를 생성하여 대체의 순서를 결정.

(3단계) 차례로 대체 실시.

- 3.1단계: 새로운 지역을 읽음.
- 3.2 단계: 개별 표에 대한 적합 변화 척도(fit change measures)를 계산.
- 3.3 단계: 각 대체에 대한 전체적 적합도(overall measures of fit)를 계산.
- 3.4 단계: 대체 자료의 선택.
- 3.5 단계: 각 개체의 대체를 위한 척도를 개정함.
- 3.6 단계: 지역의 마지막 단위를 저장함.

(4단계) 전체적 적합도를 저장하고 모수를 조정함.

6. 유럽 EUROSTAT의 European Community Household Panel(ECHP)의 무응답 대체

○ ECHP 설문에서 소득은 가구 단위와 개인 단위로 측정됨. 조사된 소득은 실제 국가 전체의 소득과 차이가 존재하므로 이 차이를 제거하기 위하여 대체가 실시됨(European Commission, 2002).

○ 순차회귀 다중대체 방법(sequential regression multiple imputation)을 사용하는 SAS 매크로 프로그램 IVEWare를 사용하여 소득 액수에 대한 대체를 실시함. 이 방법은 연속형 변수 뿐 아니라 가산형 또는 범주형 변수 등 다양한 분포를 따르는 변수들에 대하여 동시에 대체를 실시하는 것이 가능함.

○ 특이점을 제외시킨 후 대체를 실시하였음.

Ⅰ-3. 다중응답 문항에 대한 무응답 대체

1. 무응답 대체 기법

○ 자료를 수집하는 과정에서 일부 문항의 응답이 측정되지 않으면 그 문항에 대한 응답이 발생하지 않았다는 의미로 무응답(nonresponse)이 발생했다고 함. 무응답은 거의 대부분의 자료에서 발생하지만 설문 조사 자료에서는 더 흔히 발생함. 무응답을 무시한 채 완전히 응답된 자료 만에 근거하여 분석을 실시하는 경우 추정값에 편향(bias)이 발생할 수 있으므로 무응답을 포함한 자료에 대한 적절한 분석이 필요함.

○ Little and Rubin(2002)은 무응답 자료의 메커니즘을 완전임의결측(Missing Completely At Random, MCAR), 임의결측(Missing At Random, MAR) 그리고 비임의결측(Not Missing At Random, NMAR)의 세 가지로 분류함.

○ 완전임의결측(MCAR)은 무응답이 발생할 확률이 자료의 값에 상관없이 완전임의인 경우를 의미함. 임의결측(MAR)은 무응답이 발생할 확률은 자료의 관측된 부분과는 연관되지만 자료의 관측되지 않은 부분과는 연관이 없는 경우를 의미하고 비임의결측은 무응답이 발생할 확률이 자료의 관측되지 않은 부분과 연관된 경우를 의미함. 결측자료 메커니즘이 임의결측을 따르고 자료 모형의 모수와 응답 지시행렬(response indicator matrix)과 관련된 모수가 서로 별개(distinct)일 때 결측자료 메커니즘은 무시할 수 있는 결측자료 메커니즘(ignorable missing data mechanism)을 따른다고 함. 대부분의 무응답 자료 분석 방법은 자료가 무시할 수 있는 결측자료 메커니즘을 따른다고 가정하고 있음.

○ 무응답을 포함한 자료의 결측값을 그럴듯한 값으로 채우면 대체된 자료는 무응답이 없이 완전하게 응답된 형태를 지니므로 여러 가지 분석이 용이해 짐(Rubin, 1987). 명시적 모형(explicit model)에 근거한 대체방법은 무응답의 예측분포를 다변량 정규분포와 같은 통계적 모형에 근거하여 설정한 후 대체를 실시하는 방법이고 내재적 모형(implicit model)에 근거한 대체 방법은 자료의 분포에 대한 모형화 없이 무응답에 대하여 그럴 듯한 값을 대체하는데 중점을 둔 대체방법으로서 대표적인 방법으로 핫덱대체 방법이나 콜드덱 대체 방법 등이 존재함.

○ 핫덱대체는 자료의 응답값들 중에서 한 개의 값을 선택하여 무응답을 대체하는 방법을 의미하는데

이 때 무응답을 대체하는데 사용되는 응답값을 기증자(donor)라 함. 응답값들 중에서 완전히 임의로 기증자를 선택하여 대체를 실시하는 단순임의대체보다는 그럴 듯한 값을 가지고 대체하기 위하여 연관된 변수들로 대체군을 형성한 후 무응답값과 동일한 대체군에 속하는 응답값들 중에서 대체를 실시하는 대체군에 근거한 핫덱대체가 주로 사용됨.

○ 핫덱대체는 여러 가지 분포를 따르는 자료에 두루 적용될 수 있고 대체된 값들이 관찰된 값들이므로 자료의 분포를 왜곡하지 않는 장점을 지님. 이 방법은 세계 여러 나라의 센서스 및 대규모 사회조사의 대체 방법으로 널리 사용되고 있음.

○ 명시적 모형에 근거한 대체는 자료에 대하여 특정 분포를 따른다고 가정하므로 상대적으로 분포의 성질에 근거하여 추정량의 성질을 평가하기 쉬운 반면에 핫덱대체는 자료에 대한 모형을 가정하지 않고 그럴듯한 값으로 대체를 실시하기 위하여 대체군을 형성하거나 계층적 핫덱대체 등 상대적으로 복잡한 알고리즘을 사용하여 대체를 실시하는 것이 일반적이므로 대체된 자료로부터 구한 추정량의 성질을 이론적으로 규명하기 어려움.

○ 핫덱대체의 이론적 특성을 파악하기 어려우므로 핫덱대체의 성능은 모의실험을 통해 평가하는 것이 일반적임.

2. 무응답 대체를 결정할 때 고려해야 하는 사항

○ 자료에 대한 대체를 실시해야 할지 결정하기 위하여 여러 가지 사항들이 고려되어야 함. 대체를 실시하는 작업은 상당한 시간 및 전문성을 필요로 하므로 전문 인력의 투입 및 비용이 뒷받침되어야 함. 이와 같은 원인으로 인하여 대규모 자료에 대한 무응답 대체는 개인 연구자가 실행하기 부담스럽고 자료를 수집한 기관에서 대체를 실시한 후 대체된 자료를 제공하는 것이 바람직함.

○ 통계청과 같이 자료를 수집한 기관은 일반 연구자들보다 접근할 수 있는 정보를 많이 보유하는 경우가 흔함. 예를 들어, 응답자의 자세한 주소 정보나 개인 소득에 관한 정보는 개인정보 보호를 위하여 일반인들에게 열람되지 않는 경우가 존재하지만 이 정보는 무응답값의 정확한 대체를 위해 유용하게 사용될 수 있음. 따라서 자료수집 기관에서 일반연구자에게 공개할 수 없는 유용한 정보를 포함하여 대체를 실시하고 대체된 자료를 제공하는 것이 바람직함. 이 경우 대체를 위해 고려한 정보의 양이 연구자가 분석에 사용하는 정보의 양보다 많게 되는데 이 때 연구자 모형의 모수 추정에 편향이 발생하지 않음(Meng, 1995).

○ 무응답에 대한 대체를 실시할지 여부를 결정할 때 많은 경우 결측 비율을 고려하는데 정확하게는 “무응답으로 인하여 손실된 모수에 대한 정보량(fraction of missing information about parameters due to nonresponse)”을 고려해야 함. 무응답으로 인하여 손실된 모수에 대한 정보량이란 무응답이 발생하지 않은 완전한 자료(complete data)와 비교하여 무응답으로 인해서 발생한 모수의 정밀도(precision)의 감소분(reduction)을 의미함.

○ 무응답으로 인하여 손실된 모수의 정보량은 무응답의 비율과 연관되어 있지만 정확하게 무응답의 비율과 일치하지는 않음. 일반적으로 무응답이 발생한 변수와 연관성이 높은 변수를 포함하여 대체를 실시한다면 무응답으로 인하여 손실된 모수의 정보량은 무응답의 비율보다 낮게 되며 무응답값에 대한 예측이 정확하게 가능할수록 무응답으로 인하여 손실된 모수에 대한 정보량은 크게 감소함. 따라서 무응답 대체의 성능은 무응답이 발생한 변수와 높은 연관성을 지니는 변수들을 대체모형에 포함하여 무응답값에 대한 정확한 예측이 가능한 정도에 크게 의존함.

○ 무응답으로 인하여 손실된 모수의 정보량이 어느 정도 되면 무응답에 대한 대체를 실시해야 한다는 이론은 존재하지 않음. 또한, 무응답으로 인하여 손실된 모수의 정보량은 관심 있는 모수에 따라 다르게 나타남. 이는 또한 연구자가 모수의 분산을 어느 정도 정확하게 추정해야 하는가에 따라 달라지는데, 필요한 모수추정의 정확도는 연구의 목적에 따라 달라질 수 있음. 일반적으로 무응답을 인하여 손실된 모수의 정보량이 20% 미만인 경우 상대적으로 무응답으로 인한 손실이 크지 않을 것으로 기대함 (Schafer, 1997).

○ 무응답으로 인하여 손실된 모수의 정보량이 매우 크다면 대체의 효율성이 낮아짐. 충분한 숫자의 응답값이 존재하여야 무응답값을 잘 예측할 수 있으므로 자료의 결측 비율이 매우 높다면 대체를 실시하는데 어려움이 발생함.

○ 핫덱대체를 실시하는 경우 응답자의 비율이 무응답자의 비율보다 높아야 함. 즉, 무응답의 비율이 50% 이상인 경우 핫덱대체를 실시할 수 없음. 핫덱대체의 경우 각각의 무응답값을 응답자의 값으로 대체하는데 한 개의 응답값을 한 번 이상 기증자로 사용하지 않는 것이 일반적이기 때문임. 특히, 핫덱대체는 대체군을 형성하여 대체를 실시하는 경우가 대부분인데 이 경우 각 대체군 내에서 응답자의 비율이 무응답자의 비율보다 높아야 하므로 응답자의 비율이 무응답자의 비율보다 훨씬 많은 것이 바람직함.

○ 핫덱대체를 사용하여 무응답 대체를 실시할 때 대체군은 대체가 실시될 문항의 값이 대체군 내에서 비슷하도록 선택되어야 함. <표 1.1.2>에 나타난 바와 같이 무응답이 발생한 변수 뿐 아니라 무응답 발생 확률과 연관되어 있는 변수들로 대체군을 형성한다면 추정량의 편향 뿐 아니라 분산도 줄

일 수 있으므로 대체군을 형성하는 변수를 잘 선택하는 것이 중요함.

○ 대규모 조사 자료의 경우 무응답의 비율이 높지 않더라도 여러 연구자가 다양한 연구를 진행할 가능성이 높고 많은 연구자들은 무응답을 포함한 자료에 대한 적절한 분석을 실시하는데 어려움을 느끼므로 무응답의 비율이 높지 않아도 대체를 실시하고 대체된 자료를 제공하는 경향이 높음.

○ 주의할 점은 무응답값을 그럴 듯한 값으로 대체하였다고 하더라도 자료가 가진 정보의 양이 늘어나지는 않음. 예를 들어 100개의 자료 중 10개가 무응답인 경우 자료는 총 90개 분량의 정보를 지니며 대체를 실시하였다고 정보의 분량이 100개로 늘어나는 것은 아님. 무응답값을 한 개의 그럴듯한 값으로 대체하는 경우 분석자는 어느 자료가 관측된 값이고 어느 자료가 대체된 값인지 판별하기 어려워 100개 분량의 정보를 가진 것으로 간주하고 분석을 시행하게 됨. 이 경우 모수의 추정값의 분산이 과소추정되는 문제가 발생함을 유의해야 함. 이를 보정하는 여러 가지 방법들이 제안되어 왔는데 분산보정 방식이나 다중대체(multiple imputation)이 주로 사용되어짐 (Little and Rubin, 2002).

3. 다중응답 문항의 대체

○ 설문 문항이 여러 개의 항목 중 해당되는 하나만을 고르라고 요구하는 대신 한 개 이상의 항목에 대한 응답을 허락하는 경우 다중응답(multiple response)이 발생하게 됨. 한 명의 응답자가 여러 개의 응답을 하는 것이 가능하므로 이를 대체모형에 고려하여야 함. 한 명의 응답자의 다중응답은 다른 응답자의 응답과 달리 연관성이 존재할 가능성이 높음. 다중응답 자료에서 무응답이 발생하는 경우 한 명의 응답자의 다중응답 항목 간 연관성을 고려하여 대체를 실시해야 함.

○ 다중응답 문항은 여러 개의 항목 중 한 개 또는 여러 개를 선택하도록 요구하므로 선택된 응답값은 응답 항목이 되고 따라서 일종의 범주형 변수에 해당됨. 이와 같은 다중응답 자료에 대하여 특화된 대체 방법에 대하여 기술된 문헌을 찾을 수 없었음.

○ 통계청 인구주택총조사의 단일응답 문항은 대체군을 형성하여 대체군 내에서 대체를 실시하므로 다중응답 문항에 대한 대체도 대체군을 형성하여 대체군 내에서 대체를 실시한다면 대체 방법 간 일관성을 유지할 수 있으므로 바람직함.

○ 다중응답 문항의 경우 대체군 내에서 기증자를 찾은 다음에 기증자의 다중 응답값들을 가지고 무응답자의 다중응답 문항 전체 항목을 대체함으로써 다중 응답 항목 간 연관성을 유지하는 일종의

n-partition 대체 기법(Marker 등, 2002)을 사용하여 대체를 실시한다면 동일한 응답자의 다중응답들을 세트로 대체함으로써 한 개인의 다중응답들간의 일치성을 만족시키고 다중응답값들 사이의 연관관계를 유지할 수 있음.

○ 대체군을 형성하는 변수가 선택된다면 다중응답 문항의 대체는 n-partition 기법을 사용하여 기존은 대체 프로그램을 사용하여 시행할 수 있음. 통계청에서 대체를 위해 사용하는 SAS macro %hhdeck은 기증자 정보를 저장하는 것을 가능하게 하므로 우선 대체군에 근거하여 기증자를 선정한 후 저장된 기증자 정보에 근거하여 다중응답 항목들에 대한 무응답값을 기증자의 다중응답값으로 대체하면 됨.

○ 다중응답 문항에 대한 대체를 실시할 때 어려운 부분은 대체군을 형성하는 문제임. 본 연구에서는 다중응답 문항에 대한 대체군을 선정하기 위하여 (1) 일반화다항로짓모형과 (2) 로짓모형을 함께 사용할 것을 권장함.

○ 인구주택총조사에서는 대체군을 형성하는 변수를 선택하기 위하여 연관성 분석인 카이제곱 검정 기법과 의사결정나무 기법을 이용함. 연관성 분석 중 한 범주의 응답값이 2개인 경우 카이제곱 검정의 결과는 로짓모형에 근거한 검정과 동일해 짐. 범주의 응답값이 2개 초과인 경우 일반화로짓모형(generalized logit model)을 사용하여 검정을 실시할 수 있음.

○ 본 연구에서는 <표 1.1.2>에 나타난 바와 같이 무응답이 발생한 변수 뿐 아니라 무응답 발생 확률과 연관되어 있는 변수들로 대체군을 형성한다면 추정량의 편향 뿐 아니라 분산도 줄일 수 있으므로 대체군을 형성할 때 (1) 무응답이 발생한 변수와 연관되어 있는 변수의 선택 및 (2) 무응답 발생 확률과 연관된 변수를 선택하여 대체를 실시하는 것을 권장함.

○ 무응답이 발생한 변수와 연관되어 있는 변수가 대체군 형성 시 제외된다면 추정값에 편향이 발생할 수 있으므로 대체군 형성시 가능한 한 많은 변수를 고려하는 경향이 있음. 한편, 대체군을 형성하는 변수의 개수가 많아지면 기증자를 찾지 못하는 경우가 발생하고 대체군 형성 변수의 일부를 포기해야 하므로 포기하는 변수들의 순서를 조심스럽게 결정해야 함.

(1) 일반화다항로짓모형을 사용하여 무응답이 발생한 변수와 연관되어 있는 변수의 선택

○ 본 연구에서는 일반화선형 모형을 사용하여 여러 개의 응답들과 대체군으로 고려하는 변수들 간의 연관성을 고려하고자 함. 가장 간단한 형태의 연속형 자료에 대한 일반화선형모형은

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_{ij}$$

으로 표현할 수 있음. 이 때, y_{ij}는 i번째 응답자의 다중응답 문항에 대한 j번째 응답을 의미하고 $x_{i1}, x_{i2}, \cdots, x_{ip}$는 대체군으로 고려하는 연관된 변수들을 의미하며 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$은 이 모형에서의 회귀계수를 의미함. 또한, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{ik})'$는 i번째 응답자의 k개의 오차항들의 벡터로서 이 오차항들 사이에 연관성을 고려하여 동일한 응답자의 응답들 간의 연관성을 나타냄. 이 모형은 범주형 변수를 고려하도록 확장이 가능함. 이 모형에서 유의한 회귀계수에 해당되는 변수들은 응답과 연관된 변수라 할 수 있음.

○ 인구주택총조사의 다중응답 문항들은 범주형 변수들로 이루어져 있으므로 연관된 변수들에 대한 다항로짓모형(multinomial logit model)을 적용할 수 있음. 한편, 다중응답의 경우 한 개인 또는 가구로부터의 다중응답들은 서로 연관되어 있으므로 이를 모수에 포함시켜 검정을 실시하고 유의한 변수들에 근거하여 대체군을 형성하여야 함. 다항로짓모형은 동일한 응답자의 여러 개의 응답들 간 연관성을 고려하는 일반화 다항 로짓모형(generalized multinomial logit model)으로 확장될 수 있음.

○ k개의 다중응답 문항에 대한 일반화 다항로짓모형의 우도함수는

$$L(\mu_{ij}|y_{ij}) = \prod_{j=1}^{k} \mu_{ij}^{y_{ij}}$$

으로 나타낼 수 있음. 여기서, y_{ij}는 i번째 응답자의 j번째 다중응답 문항에 대한 응답 여부를 의미하는 0(응답하지 않음) 또는 1(응답)의 값을 가지는 지시변수이며 μ_{ij}는 i번째 응답자의 다중응답 문항에 대한 j번째 응답 평균을 나타내는 모수임.

○ 일반화 다항로짓모형에서 i번째 응답자의 다중응답 문항에 대한 j번째 응답의 평균 μ_{ij}는

$$E(y_{ij}|\beta, \gamma) = \mu_{ij} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma z_i$$

로 표현할 수 있음. 이 때, $x_{i1}, x_{i2}, \cdots, x_{ip}$는 대체군으로 고려하는 연관된 변수들을 의미하며 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$은 이 모형에서의 회귀계수를 의미함. 또한, z_i는 i번째 응답자의 랜덤효과(random effect)에 대한 디자인 변수(design variable)를 나타내고 γ는 랜덤효과의 모수를 의미하며 일반적으로 정규분포를 따른다고 가정함. 이 랜덤효과를 사용하여 i번째 응답자의 다중응답 문항들 간의 연관성을 모형에 포함할 수 있음. 우도를 최대화하는 모수 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$을 추정하고 이 모수들의 유의성을 검정하여 유의하게 나타난 계수에 해당되는 변수를 대체군으로 선정하는 것

이 바람직함.

○ SAS를 사용하는 경우 GLIMMIX procedure나 NLMIX procedure를 사용하여 일반화 다항로짓모형을 적합할 수 있음. 모형을 적합한 후 유의하게 나타나는 변수들을 대체군 형성에 고려할 수 있음.

(2) 로짓모형을 사용하여 무응답 발생 확률과 연관되어 있는 변수의 선택

○ 다중응답 문항에 대한 응답여부는 실제로 문항이 다중응답 문항인지 단일응답문항인지와 상관이 없음. 무응답 발생 확률을 모형화하기 위하여 다음과 같은 응답 여부 변수를 생성함.

$$R_i = \begin{cases} 1 & i\text{번째 응답자가 다중응답 문항에 응답하지 않은 경우} \\ 0 & i\text{번째 응답자가 다중응답 문항에 응답한 경우} \end{cases}$$

응답 여부는 지시변수로 나타낼 수 있으므로 이 변수와 대체군으로 고려하고자 하는 변수들 간의 연관성은 로짓모형을 사용하여 적합할 수 있음.

○ i번째 응답자의 대체군으로 고려하고자 하는 변수들 $x_1,\cdots,x_p$가 주어졌을 때 무응답 확률을 $\pi_i = P(R_i = 1|x_1,\cdots,x_p)$로 나타내면 로짓모형은

$$\log\ddot{y}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

으로 나타낼 수 있음.

○ SAS를 사용한다면 LOGISTIC Procedure를 사용하여 모형을 적합한 후 유의하게 나타난 계수에 해당되는 변수들을 대체군 형성에 고려할 수 있음.

(3) 대체군을 형성할 변수의 선택

○ (1)에 나타난 일반화다항로짓모형 및 (2)에 나타난 로짓모형에서 공통적으로 유의하게 나타난 변수들을 포함하여 대체군을 형성하는 것이 바람직함.

○ 유의하게 나타난 변수가 여러 개인 경우 대체군의 순서는 가장 중요한 의미를 가지는 변수를 선택해야 함. 하지만 대체군을 형성하는 특정 변수로 인하여 기증자를 찾지 못할 가능성이 있으므로 유의해야 함. 통계청에서 사용하는 SAS macro는 대체군을 형성하는 변수의 숫자를 줄여감에 따라 기증자를 발견하는 단계를 결과(OUTPUT)로 제공하므로 이 결과를 주의 깊게 살펴보고 특정 변수로 인하여 대체군을 형성하는데 문제가 발생한다면 이 변수를 우선적으로 포기하는 방식으로 순위를 조정하는 것이 바람직함.

4. 교통수단 보유 문항의 대체

○ 교통수단 보유 문항은 우선 교통수단 보유 여부를 질문한 후 보유한 항목에 대하여 보유대수를 측정함. 일부 가구는 보유 여부는 응답하였으나 보유대수를 응답하지 않아 이 문항이 무응답으로 남게 됨. 한편, 일부 가구는 보유 여부에 응답하지 않았으나 보유대수를 응답하였고 이 가구의 보유 여부는 "보유"로 간주하였음.

○ 이 문항은 보유가구에 대하여 응답을 요구하였으므로 교통수단 보유대수는 양의 정수(positive integer)로 측정됨. 특히 대부분의 가구에서 보유대수는 1-2대로 측정되지만 일부 가구의 보유대수는 이보다 많게 나타남. 예를 들어, 2010년 인구주택총조사 시범자료의 승용차 보유대수는 최대 9대로 보고되었음.

○ 양의 정수로 측정되고 2대 보유한 경우는 1대 보유한 경우보다 많은 형태이므로 이 자료는 순서형 변수(ordinal variable)로 측정되었음. 따라서 순서형 변수에 대한 로짓모형(ordinal logit model)을 고려하여 연관성이 있는 변수들을 선택하여 대체군을 형성하는데 사용하는 것이 바람직함. 이 변수의 경우 각 가구별로 한 개의 응답만을 제공하므로 단일응답 문항으로 간주할 수 있음.

○ K개의 순서형 응답이 가능한 순서형 로짓모형에서 i번째 응답자의 문항에 대한 응답값 y_i는

$$\log\left(P(y_i \le k|x_{i1},\cdots,x_{ip})\right)=\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip},\quad k=1,2,...,K$$

로 표현할 수 있음. 이 때, $x_{i1},x_{i2},\cdots,x_{ip}$는 대체군으로 고려하는 연관된 변수들을 의미하며 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$은 이 모형에서의 회귀계수를 의미함. 이 모형의 모수 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$을 추정하고 이 모수들의 유의성을 검정하여 유의하게 나타난 계수에 해당되는 변수를 대체군으로 선정하는 것이 바람직함.

○ 로짓모형을 사용하여 무응답 발생 확률과 연관되어 있는 변수를 선택하는 방법은 이 장의 3절 (2)에서와 동일하게 적용할 수 있음. 위의 순서형 로짓모형에서 유의하게 나타난 변수들과 로짓모형에서도 공통적으로 유의하게 나타난 변수들을 포함하여 대체군을 형성하는 것이 바람직함.

Ⅰ-4. 인구주택총조사 자료

○ 본 연구에서는 2010년 인구주택총조사의 다중응답 문항에서 발생하는 무응답에 대하여 파악하기 위하여 현재 이용 가능한 두 자료인 (1) 2005년 인구주택총조사 10% 표본 자료와 (2) 2010년 인구주택총조사 시범자료(rehearsal data)의 다중응답 문항에서 발생하는 무응답에 대하여 고찰하고 이 자료를 이용하여 모의실험을 실행함.

○ 본 연구에서 고려한 2005년 인구주택총조사 10% 표본 자료의 대상은 2010년 인구주택총조사 모집단과 상당수 겹칠 것으로 예상되므로 유사한 특성을 보일 것으로 예상되며 자료는 시도, 시군구, 읍면동 및 조사구 등의 정보를 포함하므로 실제 2010년 인구주택총조사 자료에 대한 무응답 대체를 실시할 때와 유사한 대체 모형을 가질 것으로 기대됨.

○ 2010년 인구주택총조사 시범자료는 2010년 인구주택총조사 문항으로 조사되었으므로 실제 2010년 인구주택총조사에서 측정되는 모든 문항에 대한 정보를 제공함. 하지만, 시범자료의 특성상 2개의 시도에서 각각 1개의 시군구에서만 조사되어 시도, 시군구, 읍면동의 행정구역 정보를 모형에 포함시켜 분석할 수 없음. 또한, 관찰값의 숫자가 상대적으로 적어 전체자료에 적절한 복잡한 모형을 적합하는 데 어려움이 발생하기도 함.

1. 설문 문항

○ 2005년 인구주택총조사 표본 설문 문항은 총 41개 문항으로써 가구원을 대상으로 한 인구 사항에 관한 문항 24개와 가구를 대상으로 하는 가구 관련 문항 11개 및 주택에 관한 문항 6개로 이루어짐.

○ 2010년 인구주택총조사 표본 설문 문항은 총 2005년에 비하여 5개 문항이 증가되어 46개 문항으로써 가구원을 대상으로 한 인구 사항에 관한 문항 28개와 가구를 대상으로 하는 가구 관한 문항 12개 및 주택에 관한 문항 6개로 이루어짐.

○ 2005년과 2010년 인구주택총조사 설문조사 문항 간에 일부 변동이 있었으며 (예를 들면, 활동제약은 2005년에는 2개의 문항으로 질문하였으나 2010년에 한 개의 문항으로 통합됨) 문항 내에서도 항목에 변화가 생기기도 함. <표 1.4.1>은 2005년과 2010년 인구주택총조사 다중응답 설문 문항 간 차이를 표로 나타낸 것임.

<표 1.4.1> 2005년과 2010년 인구주택총조사 다중응답 설문 문항의 변화

	2005년 인구주택총조사	2010년 인구주택총조사
아동보육	10개 응답항목	11개 응답항목 ① 항목 추가(“방과 후 학교” 추가)
활동제약	2개 문항 (6 + 5 항목) ① 각 문항에 “없음” 항목 포함 ② 항목명 변경(“걷기, 계단 오르기, 들고 운반하기 등 육체적 제약”) ③ 항목명 변경(“학습의 어려움, 정신적 질환 등 정신적 제약”) ④ 항목명 변경(“옷입기, 목용하기, 밥먹기, 집안 돌아다니기”) ⑤ 항목명 변경(“쇼핑, 병원가기, 집 밖 돌아다니기”)	1개 문항으로 통합 (8개 항목) ① “치매,” “중풍” 항목 추가 ② 항목명 변경(“걷기, 계단 오르기 등 이동 제약”으로 육체적 제약에서 이동 제약으로 변경) ③ 항목명 변경(“정신적 질환 등 정신적 제약”으로 “학습의 어려움” 제외) ④ 항목명 변경(“옷입기, 목용하기, 밥먹기”로 “집안 돌아다니기” 제외) ⑤ 항목명 변경(“장보기, 병원가기”로 “쇼핑”이 “장보기”로 변경되고 “집 밖 돌아다니기” 제외)
이용교통수단	항목명 변경(“기타(오토바이, 트럭 등”)	항목명 변경(“기타(오토바이, 화물차 등”으로 “트럭”에서 “화물차”로 변경)
사회활동		신규
고령자 생활비 원천	① 항목명 변경(“국민, 공무원, 교직원 연금”) ② 항목명 변경(“주식, 채권, 증권”)	① 항목명 변경(“국민, 공무원, 사학 ·군인 연금”으로 “교직원 연금”이 “사학 ·군인 연금”으로 변경) ② 항목명 변경(“주식, 펀드, 채권”으로 “증권”에서 “펀드”로 변경)
정보통신기기 보유 및 이용 현황		10년 주기 문항
주차장소	① 항목명 변경(“공터(공휴지)”)	① 항목명 변경(“공터”로 “공휴지” 제거)

○ 2005년 인구주택총조사에서 조사되지 측정되지 않았으나 2010년 조사에서 추가된 다중응답 문항들이 존재함. 10년 주기로 측정되는 “정보통신기기 보유 및 이용현황”과 신규로 조사되는 “사회활동” 문항은 2005년 조사에는 조사되지 않았음.

○ 본 연구의 주요 관심사인 다중응답 문항 외에 다른 연관 변수들도 추가되거나 변경되는 경우가 발생함. 예를 들어, 2010년 조사에서는 아동보육과 연관된 이전 거주지 항목인 5년 전 거주지 외에 1년 전 거주지 및 출생지가 포함됨.

○ 2010년 인구주택총조사 시범자료는 2005년 자료와 달리 2010년 조사 문항에 대하여 실시되었으므로 신규로 추가된 문항(사회활동, 정보통신기기 보유 및 이용현황)에 대한 정보를 제공하고 있음. 이 자료의 제약은 2개의 시도에서 각각 1개의 시군구를 선택하여 조사를 실시하였으므로 대체군을 형성하는 데 주요 변수로 고려될 수 있는 행정구역 정보를 활용하기 어려움.

○ 교통수단 보유 및 이용현황에 관한 문항은 2010년 조사에서 추가되었음.

2. 다중응답 문항의 무응답

○ <표 1.4.2>는 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료의 다중응답 문항들 각각에 대한 유효표본의 숫자 및 무응답 비율을 나타냄. 2010년 시범자료에서는 무응답 비율이 상대적으로 높게 나타나지만 정보통신기기 보유 및 이용현황에서 10% 정도로 나타나고 나머지 문항의 경우 10%보다 작게 나타나고 있으며 2005년 인구주택총조사 자료에서는 무응답 비율이 모두 2% 미만으로 나타남.

○ 2005년 인구주택총조사 10% 표본 자료의 경우 최대 2개까지 응답하는 문항에 대하여 2개 이상의 응답을 한 정보가 존재하지 않으므로 보기초과 응답 문항에 대한 처리 방법에 대한 연구가 불가능함.

○ 2010년 인구주택총조사 시범자료의 경우 최대 2개까지 응답하는 문항에 대하여 2개 이상의 응답을 한 정보가 존재하지만 관찰값의 수가 너무 작아 적절한 모의실험을 실시하는 데 심각한 제약이 있음.

<표 1.4.2> 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료의 다중응답 문항의 유효표본의 숫자 및 무응답 비율

	2005년 인구주택총조사 자료			2010년 인구주택총조사 시범자료		
	유효표본수	응답자수	무응답 비율 (%)	유효표본수	응답자수	무응답 비율 (%)
아동보육	736,201	723,933	1.67%	5,395	5,056	6.28%
활동제약 1a)	4,820,648	4,786,634	0.71%	29,030	27,416	5.56%
활동제약 2	4,820,648	4,786,657	0.71%			
이용교통수단d)	2,555,968	2,540,203	0.62%	16,997	16,753	1.44%
사회활동b)				24,347	23,540	3.31%
고령자 생활비 원천	760,590	752,064	1.12%	4,506	4,349	3.48%
정보통신기기 보유 및 이용 현황c)				12,158	10,921	10.17%
주차장소e)	477,767	473,642	0.86%	6,623	6,513	1.66%

a) 2005년에 2개 문항으로 질의된 활동제약이 2010년에 1개 문항으로 통합됨
b) 2010년 신규 문항
c) 10년 주기로 측정되므로 2005년에는 측정되지 않음
d) 13세 이상으로 통근, 통학여부에 대하여 응답자 중에서 계산함.
e) 주차장소는 승용차, 승합차, 화물·승합차에 대한 응답자 중에서 계산함

○ 주된 두 곳까지 응답하는 아동보육문항에서 2개 이상의 항목에 대하여 응답한 인원은 시범 자료의 응답자수 5,056명 중 21명에 불과하였음. <표 1.4.3>은 2010년 인구주택총조사 시범자료의 아동보육문항에 대한 응답자들의 응답분포를 나타내는데 이 중 보기초과 응답의 경우 회색으로 표시하였음. 대부분의 응답조합에서 보기초과응답을 보인 인원수는 1명이며 부모-학원-혼자로 응답한 인원이 5명, 그리고 부모-방과후학교-학원으로 응답한 인원이 5명이었음. 한편, 5,056명 중 3,824명인 75.6%가 1개의 응답만을 제공함.

<표 1.4.3> 2010년 인구주택총조사 시범자료의 아동보육문항에 대한 응답자들의 응답분포

부모	조부모	기타가족	가사도우미	유치원	어린이집	기타보육시설	방과후학교	학원	혼자	기타	빈도	백분율(%)	누적빈도	누적백분율(%)
										1	12	0.24%	12	0.24%
									1		58	1.15%	70	1.38%
									1	1	6	0.12%	76	1.50%
								1			570	11.27%	646	12.78%
								1		1	12	0.24%	658	13.01%
								1	1		83	1.64%	741	14.66%
							1				136	2.69%	877	17.35%
							1			1	4	0.08%	881	17.42%
							1		1		8	0.16%	889	17.58%
							1	1			101	2.00%	990	19.58%
						1					12	0.24%	1002	19.82%
						1		1			3	0.06%	1005	19.88%
						1	1				1	0.02%	1006	19.90%
					1						519	10.27%	1525	30.16%
					1					1	1	0.02%	1526	30.18%
					1				1		2	0.04%	1528	30.22%
					1			1			11	0.22%	1539	30.44%
					1			1	1		1	0.02%	1540	30.46%
					1	1					3	0.06%	1543	30.52%
				1							227	4.49%	1770	35.01%
				1					1		4	0.08%	1774	35.09%
				1				1			26	0.51%	1800	35.60%
				1			1				1	0.02%	1801	35.62%
				1			1			1	1	0.02%	1802	35.64%
				1		1					1	0.02%	1803	35.66%
				1	1						8	0.16%	1811	35.82%
				1	1			1			1	0.02%	1812	35.84%
			1								52	1.03%	1864	36.87%
			1						1		1	0.02%	1865	36.89%
			1					1			3	0.06%	1868	36.95%
			1				1				1	0.02%	1869	36.97%
			1		1						5	0.10%	1874	37.06%
			1	1							8	0.16%	1882	37.22%
		1									27	0.53%	1909	37.76%

		1								1	1	0.02%	1910	37.78%
		1						1			4	0.08%	1914	37.86%
		1					1				4	0.08%	1918	37.94%
		1			1						5	0.10%	1923	38.03%
		1		1							3	0.06%	1926	38.09%
	1										208	4.11%	2134	42.21%
	1								1		4	0.08%	2138	42.29%
	1							1			37	0.73%	2175	43.02%
	1						1				4	0.08%	2179	43.10%
	1					1					1	0.02%	2180	43.12%
	1				1						19	0.38%	2199	43.49%
	1			1							18	0.36%	2217	43.85%
	1		1								1	0.02%	2218	43.87%
	1	1									3	0.06%	2221	43.93%
1											2003	39.62%	4224	83.54%
1										1	11	0.22%	4235	83.76%
1									1		14	0.28%	4249	84.04%
1								1			346	6.84%	4595	90.88%
1								1	1		5	0.10%	4600	90.98%
1							1				67	1.33%	4667	92.31%
1							1			1	1	0.02%	4668	92.33%
1							1	1			5	0.10%	4673	92.42%
1						1					5	0.10%	4678	92.52%
1					1						171	3.38%	4849	95.91%
1				1							159	3.14%	5008	99.05%
1				1				1			1	0.02%	5009	99.07%
1				1	1						1	0.02%	5010	99.09%
1			1								4	0.08%	5014	99.17%
1		1									2	0.04%	5016	99.21%
1		1			1						1	0.02%	5017	99.23%
1	1										35	0.69%	5052	99.92%
1	1							1			1	0.02%	5053	99.94%
1	1						1	1			3	0.06%	5056	100.00%

○ 주된 두 곳까지 응답하는 이용 교통수단 문항에서 2개 이상의 항목에 대하여 응답한 인원은 시범자료의 응답자수 16,753명 중 4명에 불과하였음. <표 1.4.4>는 2010년 인구주택총조사 시범자료의 이용 교통수단에 대한 응답자들의 응답분포를 나타내는데 이 중 보기초과 응답의 경우 회색으로 표시하였음. 승용차-버스-전철로 응답한 인원이 2명이었으며 나머지 응답조합에서 보기초과응답

을 보인 인원수는 각 1명이었음. 한편, 16,753명 중 16,113명은 1개의 응답을 제공하여 대부분의 응답자가 1개의 교통수단만을 이용하는 것으로 나타남.

<표 1.4.4> 2010년 인구주택총조사 시범자료의 이용 교통수단 문항에 대한 응답자들의 응답분포

걸어서	승용차	시내, 좌석, 마을버스	통근통학버스	고속, 시외버스	전철, 지하철	기차	택시	자전거	기타	빈도	백분율 (%)	누적빈도	누적 백분율 (%)
.	.	.	.	.	.	.	.	.	1	512	3.06%	512	3.06%
.	.	.	.	.	.	.	.	1	.	381	2.27%	893	5.33%
.	.	.	.	.	.	.	.	1	1	16	0.10%	909	5.43%
.	.	.	.	.	.	.	1	.	.	38	0.23%	947	5.65%
.	.	.	.	.	.	1	.	.	.	50	0.30%	997	5.95%
.	.	.	.	.	.	1	.	.	1	1	0.01%	998	5.96%
.	.	.	.	.	1	.	.	.	.	248	1.48%	1246	7.44%
.	.	.	.	.	1	.	.	.	1	1	0.01%	1247	7.44%
.	.	.	.	.	1	.	.	1	.	1	0.01%	1248	7.45%
.	.	.	.	.	1	.	1	.	.	3	0.02%	1251	7.47%
.	.	.	.	.	1	1	.	.	.	8	0.05%	1259	7.52%
.	.	.	.	1	.	.	.	.	.	92	0.55%	1351	8.06%
.	.	.	.	1	.	.	1	.	.	1	0.01%	1352	8.07%
.	.	.	.	1	.	1	.	.	.	3	0.02%	1355	8.09%
.	.	.	.	1	1	.	.	.	.	2	0.01%	1357	8.10%
.	.	.	1	.	.	.	.	.	.	635	3.79%	1992	11.89%
.	.	.	1	.	.	.	.	1	.	2	0.01%	1994	11.90%
.	.	.	1	.	.	.	1	.	.	2	0.01%	1996	11.91%
.	.	.	1	.	1	.	.	.	.	12	0.07%	2008	11.99%
.	.	.	1	1	.	.	.	.	.	3	0.02%	2011	12.00%
.	.	1	.	.	.	.	.	.	.	1522	9.08%	3533	21.09%
.	.	1	.	.	.	.	.	.	1	2	0.01%	3535	21.10%
.	.	1	.	.	.	.	.	1	.	15	0.09%	3550	21.19%
.	.	1	.	.	.	.	1	.	.	14	0.08%	3564	21.27%
.	.	1	.	.	.	1	.	.	.	6	0.04%	3570	21.31%
.	.	1	.	.	1	.	.	.	.	92	0.55%	3662	21.86%
.	.	1	.	1	.	.	.	.	.	9	0.05%	3671	21.91%
.	.	1	1	.	.	.	.	.	.	44	0.26%	3715	22.18%

.	1	.	.	.	.	.	.	.	.	7434	44.37%	11149	66.55%
.	1	.	.	.	.	.	.	.	1	30	0.18%	11179	66.73%
.	1	.	.	.	.	.	.	1	.	87	0.52%	11266	67.25%
.	1	.	.	.	.	.	1	.	.	4	0.02%	11270	67.27%
.	1	.	.	.	.	1	.	.	.	15	0.09%	11285	67.36%
.	1	.	.	.	1	.	.	.	.	55	0.33%	11340	67.69%
.	1	.	.	.	1	1	.	.	.	1	0.01%	11341	67.70%
.	1	.	.	1	.	.	.	.	.	11	0.07%	11352	67.76%
.	1	.	1	.	.	.	.	.	.	59	0.35%	11411	68.11%
.	1	1	.	.	.	.	.	.	.	138	0.82%	11549	68.94%
.	1	1	.	.	.	1	.	.	.	1	0.01%	11550	68.94%
.	1	1	.	.	1	.	.	.	.	2	0.01%	11552	68.95%
1	.	.	.	.	.	.	.	.	.	5201	31.05%	16753	100.00%

○ 주된 두 곳까지 응답하는 고령자 생활비 원천 문항에서 2개 이상의 항목에 대하여 응답한 인원은 시범 자료의 응답자수 4,349명 중 50명이었음. <표 1,4.5>는 2010년 인구주택총조사 시범자료의 고령자 생활비 원천 문항에 대한 응답자들의 응답분포를 나타내는데 이 중 보기초과 응답의 경우 회색으로 표시하였음. 2문항 이상에 대하여 응답한 응답자 대부분은 3개의 응답을 제공하였으며 4개의 응답을 한 경우가 1명 존재함. 대부분의 응답조합에서 보기초과응답을 보인 인원수는 1-2명이었으나 본인·배우자의 일, 직업-따로 사는 자녀-국가·지방자치단체 보조로 응답한 인원이 18명이었고 예금,적금-개인연금-따로 사는 자녀로 응답한 인원이 4명이었음. 한편, 4,349명 중 3,006명인 69.1%가 1가지 생활비 원천에 근거하여 생활비를 마련하고 있는 것으로 나타남.

<표 1.4.5> 2010년 인구주택총조사 시범자료의 고령자 생활비 원천 문항에 대한 응답자들의 응답분포

본인, 배우자 직업	예금, 적금	국민/공무원/교직연금	개인연금	부동산	주식, 채권, 증권	함께 사는 자녀	따로 사는 자녀	친인척	국가지자체보조	이웃종교보조	기타	빈도	백분율 (%)	누적빈도	누적 백분율 (%)
.	.	.	.	.	.	.	.	.	.	.	1	96	2.21%	96	2.21%
.	.	.	.	.	.	.	.	.	.	1	.	21	0.48%	117	2.69%

.	.	.	.	.	.	.	.	.	1	.	.	100	2.30%	217	4.99%
.	.	.	.	.	.	.	.	.	1	.	1	5	0.11%	222	5.10%
.	.	.	.	.	.	.	.	.	1	1	.	3	0.07%	225	5.17%
.	.	.	.	.	.	.	.	1	.	.	.	20	0.46%	245	5.63%
.	.	.	.	.	.	.	.	1	1	.	.	1	0.02%	246	5.66%
.	.	.	.	.	.	.	1	.	.	.	.	546	12.55%	792	18.21%
.	.	.	.	.	.	.	1	.	.	.	1	15	0.34%	807	18.56%
.	.	.	.	.	.	.	1	.	.	1	.	1	0.02%	808	18.58%
.	.	.	.	.	.	.	1	.	1	.	.	110	2.53%	918	21.11%
.	.	.	.	.	.	.	1	1	.	.	.	3	0.07%	921	21.18%
.	.	.	.	.	.	1	.	.	.	.	.	568	13.06%	1489	34.24%
.	.	.	.	.	.	1	.	.	.	.	1	9	0.21%	1498	34.44%
.	.	.	.	.	.	1	.	.	.	1	.	1	0.02%	1499	34.47%
.	.	.	.	.	.	1	.	.	1	.	.	43	0.99%	1542	35.46%
.	.	.	.	.	.	1	1	.	.	.	.	88	2.02%	1630	37.48%
.	.	.	.	.	.	1	1	.	1	.	.	1	0.02%	1631	37.50%
.	.	.	.	.	1	.	.	.	.	.	.	2	0.05%	1633	37.55%
.	.	.	.	1	.	.	.	.	.	.	.	55	1.26%	1688	38.81%
.	.	.	.	1	.	.	.	.	.	.	1	2	0.05%	1690	38.86%
.	.	.	.	1	.	.	1	.	.	.	.	17	0.39%	1707	39.25%
.	.	.	.	1	.	1	.	.	.	.	.	6	0.14%	1713	39.39%
.	.	.	.	1	.	1	1	.	.	.	.	2	0.05%	1715	39.43%
.	.	.	1	.	.	.	.	.	.	.	.	45	1.03%	1760	40.47%
.	.	.	1	.	.	.	.	.	.	.	1	2	0.05%	1762	40.52%
.	.	.	1	.	.	.	.	.	.	1	.	1	0.02%	1763	40.54%
.	.	.	1	.	.	.	.	.	1	.	.	5	0.11%	1768	40.65%
.	.	.	1	.	.	.	1	.	.	.	.	8	0.18%	1776	40.84%
.	.	.	1	.	.	1	.	.	.	.	.	8	0.18%	1784	41.02%
.	.	.	1	.	.	1	1	.	.	.	.	1	0.02%	1785	41.04%
.	.	.	1	1	.	.	.	.	.	.	.	6	0.14%	1791	41.18%
.	.	1	.	.	.	.	.	.	.	.	.	238	5.47%	2029	46.65%
.	.	1	.	.	.	.	.	.	.	.	1	10	0.23%	2039	46.88%
.	.	1	.	.	.	.	.	.	.	1	.	2	0.05%	2041	46.93%
.	.	1	.	.	.	.	.	.	1	.	.	14	0.32%	2055	47.25%
.	.	1	.	.	.	.	1	.	.	.	.	88	2.02%	2143	49.28%
.	.	1	.	.	.	.	1	.	1	.	.	1	0.02%	2144	49.30%
.	.	1	.	.	.	1	.	.	.	.	.	28	0.64%	2172	49.94%
.	.	1	.	.	.	1	1	.	.	.	.	2	0.05%	2174	49.99%
.	.	1	.	.	1	.	.	.	.	.	.	1	0.02%	2175	50.01%
.	.	1	.	1	.	.	.	.	.	.	.	20	0.46%	2195	50.47%
.	.	1	.	1	.	.	1	.	.	.	.	2	0.05%	2197	50.52%

.	.	1	1	.	.	.	.	.	.	.	.	5	0.11%	2202	50.63%
.	1	.	.	.	.	.	.	.	.	.	.	147	3.38%	2349	54.01%
.	1	.	.	.	.	.	.	.	.	.	1	7	0.16%	2356	54.17%
.	1	.	.	.	.	.	.	.	1	.	.	3	0.07%	2359	54.24%
.	1	.	.	.	.	.	1	.	.	.	.	70	1.61%	2429	55.85%
.	1	.	.	.	.	.	1	.	1	.	.	2	0.05%	2431	55.90%
.	1	.	.	.	.	1	.	.	.	.	.	22	0.51%	2453	56.40%
.	1	.	.	.	.	1	1	.	.	.	.	1	0.02%	2454	56.43%
.	1	.	.	.	1	.	.	.	.	.	.	1	0.02%	2455	56.45%
.	1	.	.	1	.	.	.	.	.	.	.	15	0.34%	2470	56.79%
.	1	.	.	1	.	.	.	.	.	.	1	1	0.02%	2471	56.82%
.	1	.	1	.	.	.	.	.	.	.	.	13	0.30%	2484	57.12%
.	1	.	1	.	.	.	1	.	.	.	.	4	0.09%	2488	57.21%
.	1	.	1	.	.	1	.	.	.	.	.	2	0.05%	2490	57.25%
.	1	.	1	.	1	1	.	.	.	.	.	1	0.02%	2491	57.28%
.	1	1	.	.	.	.	.	.	.	.	.	31	0.71%	2522	57.99%
.	1	1	.	.	.	.	1	.	.	.	.	2	0.05%	2524	58.04%
.	1	1	.	.	.	1	.	.	.	.	.	1	0.02%	2525	58.06%
.	1	1	.	.	1	.	.	.	.	.	.	2	0.05%	2527	58.11%
.	1	1	.	1	.	.	.	.	.	.	.	1	0.02%	2528	58.13%
1	.	.	.	.	.	.	.	.	.	.	.	1168	26.86%	3696	84.99%
1	.	.	.	.	.	.	.	.	.	.	1	19	0.44%	3715	85.42%
1	.	.	.	.	.	.	.	.	1	.	.	84	1.93%	3799	87.35%
1	.	.	.	.	.	.	.	1	.	.	.	1	0.02%	3800	87.38%
1	.	.	.	.	.	.	1	.	.	.	.	218	5.01%	4018	92.39%
1	.	.	.	.	.	.	1	.	1	.	.	18	0.41%	4036	92.80%
1	.	.	.	.	.	1	.	.	.	.	.	61	1.40%	4097	94.21%
1	.	.	.	.	.	1	.	.	1	.	.	1	0.02%	4098	94.23%
1	.	.	.	1	.	.	.	.	.	.	.	25	0.57%	4123	94.80%
1	.	.	1	.	.	.	.	.	.	.	.	23	0.53%	4146	95.33%
1	.	.	1	.	.	.	1	.	.	.	.	1	0.02%	4147	95.36%
1	.	1	.	.	.	.	.	.	.	.	.	119	2.74%	4266	98.09%
1	.	1	.	.	.	.	1	.	.	.	.	2	0.05%	4268	98.14%
1	.	1	.	1	.	.	.	.	.	.	.	1	0.02%	4269	98.16%
1	1	.	.	.	.	.	.	.	.	.	.	79	1.82%	4348	99.98%
1	1	.	.	.	.	1	.	.	.	.	.	1	0.02%	4349	100.00%

○ 주된 두 곳까지 응답하는 주차 장소 문항에서 2개 이상의 항목에 대하여 응답한 경우는 시범 자료의 응답가구 6,513가구 중에는 존재하지 않았음. 한편, 6,513가구 중 6,382가구가 주로 주차하는

장소로 1곳을 응답하여 대부분의 응답자가 1곳에만 주차하는 것으로 나타남.

3. 교통수단 보유 문항의 무응답

○ <표 1.4.6>은 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료의 교통수단 보유 문항들 각각에 대한 유효표본의 숫자 및 무응답 비율을 나타냄. 2010년 시범자료에서 무응답 비율은 5% 미만으로 나타나고 있으며 승용차를 제외한 다른 교통수단의 무응답의 숫자가 20가구 미만이어서 적절한 대체 모형을 고려하는데 한계가 있음. 더구나, 2005년 인구주택총조사 자료에서는 이 문항에 대하여 무응답이 없었음.

<표 1.4.6> 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료의 교통수단 보유 문항의 유효표본의 숫자 및 무응답 비율

	2005년 인구주택총조사 자료			2010년 인구주택총조사 시범자료		
	유효표본수	응답자수	무응답 비율 (%)	유효표본수	응답자수	무응답 비율 (%)
승용차 (경차)a)	709,054	709,054	0.00%	2,619	2,587	1.22%
승용차 (경차 외)				5,129	5,095	0.66%
승합차	32,467	32,467	0.00%	261	248	4.98%
화물·특수자동차	113,080	113,080	0.00%	817	807	1.22%
오토바이b)				470	457	2.77%
자전거b)				1,786	1,768	1.01%

a) 2005년 인구주택총조사에서는 승용차를 “경차”와 “경차 외”로 구분하지 않고 한 문항으로 조사함
b) 2010년 신규 문항

Ⅰ-5. 모의실험

1. 모의실험 자료

○ 본 연구에서 고려한 다중응답 자료의 무응답 및 응답 불일치 문항에 대하여 제안된 (1) 무응답 발생확률과 대체군 형성 변수들 간의 연관성에 대한 로짓모형 및 (2) 무응답이 발생한 변수와 대체군 형성 변수들 간의 연관성에 대한 일반화다항로짓모형 또는 순서형 변수에 대한 로짓모형을 이용하여 대체를 실시하는 방법의 성능을 평가하기 위하여 모의실험이 실시됨.

○ 실제 인구주택총조사 자료 중 응답된 자료에 근거하여 모의실험을 실시함으로써 2010년 실제 인구주택총조사의 상황과 가능한 한 비슷한 조건 하에서 모의실험을 진행하고 그 결과를 2010년 인구주택총조사의 무응답 대체를 위해 적용할 수 있도록 돕고자 함.

○ 모의실험은 (1) 2005년 인구주택총조사 10% 표본 자료와 (2) 2010년 인구주택총조사 시범자료(rehearsal data)에 대하여 각 다중응답 또는 교통수단 보유 문항에 대하여 완전하게 응답된 값들을 모집단으로 가정하고 이 모집단에서 임의로 표본을 뽑은 후 선택된 표본에 무응답을 발생시켜 시행함. 즉, 각 문항에 따라 모집단이 변동되는데 이는 문항에 따라 응답 대상자가 바뀌므로 관심 모집단이 달라지는 실제 자료의 상황을 재현한 것임.

○ 모의실험에서 고려한 2005년 인구주택총조사 10% 표본 자료는 2010년 인구주택총조사 모집단과 대부분 겹치므로 유사한 특성을 보일 것으로 예상되며 시도, 시군구, 읍면동 및 조사구 등의 정보를 포함하므로 실제 2010년 인구주택총조사 자료에 대한 무응답 대체를 실시할 때와 유사한 모형을 가질 것으로 기대됨.

○ 2010년 인구주택총조사 시범자료는 2010년 인구주택총조사 문항으로 조사되었으므로 실제 2010년 인구주택총조사에서 측정되는 모든 문항에 대한 정보를 제공함. 하지만, 시범자료의 특성상 2개의 시군구 만에서 조사되어 시도, 시군구, 읍면동의 행정구역 정보가 대체군을 형성할 때 고려해야 하는 변수인지 모의실험에서 확인할 수 없음. 또한, 시범자료의 숫자가 상대적으로 작아 모형의 적합 시 과다적합으로 인한 문제점이 발생함.

○ 2005년 인구주택총조사 10% 표본과 2010년 인구주택총조사 자료는 모의실험을 실시할 때 각각 다른 특성을 보일 수 있으므로 이 두 자료에 대하여 각각 모의실험을 실시하고 결과를 비교함으로써 실제 2010년 인구주택자료의 다중응답 및 교통수단 보유 문항에 대한 대체를 실시할 때 고려할

사항들을 도출하였음.

○ 가능한 한 실제 인구주택총조사 자료와 비슷한 조건하에서 모의실험을 실시하기 위하여 이 두 자료의 다중응답 및 교통수단 보유 문항들 각각에 대하여 완전하게 응답한 자료를 모집단으로 가정하여 모의실험을 실시함.

2. 대체군의 선정

○ 각 다중응답 문항에 대하여 대체군을 형성하기 위하여 고려한 변수들이 <표 1.5.1>에 나타남. 기본변수는 다중응답 문항(또는 교통수단 보유대수 문항)과 연관성이 있을 것으로 예상되는 인구학적 특성 변수들 중에서 선택된 변수들 및 거주 시도 정보를 포함함. 정보통신기기 보유여부와 주차장소는 가구별로 응답하였고 응답자는 가구주(35.7%), 가구주의 자녀(33.8%), 가구주의 배우자(23.0%) 등으로 다양하게 나타나 응답자의 나이, 교육정도, 성별의 이 변수와의 연관성을 가정하기 어려워 고려하지 않음. 실제 대체를 시행할 때에는 본 연구에서 고려한 변수 외에 다른 연관된 것으로 추측되는 변수들을 추가하거나 본 연구에서 고려한 변수들을 제외하고 진행할 수 있음.

<표 1.5.1> 다중응답 및 교통수단 보유 문항별 대체군 형성을 위해 고려한 변수들

문항	기본변수	추가 변수
아동보육	시도 성별 만나이 교육정도	출생지 시도 출생지 시군구 5년전 거주지 1년전 거주지 가구주 경제활동상태a) 가구주 교육정도a) 가구주 배우자 경제활동상태a) 가구주 배우자 교육정도a)
활동제약		통근/통학여부 경제활동상태 경제활동가능여부
이용교통수단		통근/통학여부 통근/통학장소 재학여부 취업여부 활동제약 유무

		사회활동 유무 경제활동상태 종사상 지위 근로장소 혼인상태 자동차 보유대수 자동차 이용횟수 주거용 연면적 점유형태(자기집/전세/월세(사글세)/무상)
사회활동		경제활동상태 혼인상태 종사상 지위 산업코드 가구구분 거처의 종류 주거용연면적 자녀수 점유형태(자기집/전세/월세(사글세)/무상)
고령자 생활비 원천		통근/통학여부 경제활동상태
정보통신기기 보유 여부	시도	가구구분 점유형태 거처종류 난방시설
주차장소		거처의 종류 주인가구여부 난방시설 자동차 보유댓수
교통수단 보유대수		침실수 기타방수 종사상지위 가구구분 거처종류 주인가구여부 주거용연면적 난방시설 주차장소(영업용,건물부설주차장) 식사방수 주차장소(자가) 주차장소(도로변) 주차장소(노상주차장)

		주차장소(공터) 점유형태 이용횟수b)

a) 가구주나 가구주의 배우자가 없는 응답자의 경우 이 문항들을 별도의 범주로 분류하여 포함함.
b) 2005년 인구주택총조사에서는 측정되지 않아 2010년 자료에 대한 모형에만 포함됨. 각 항목별 이용횟수가 포함됨.

(1) 무응답 발생 확률과 연관되어 있는 변수의 선택

○ 무응답 발생확률과 연관되어 있는 변수를 선택하기 위하여 응답여부를 반응변수로 하고 <표 1.5.1>의 변수들을 설명변수로 포함하여 로짓모형을 적합한 후 유의한 변수를 선택함. 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료에 대하여 다중응답 및 교통수단 보유 문항별로 로짓모형을 실시하여 유의수준 5% 정도에서 유의하게 나타난 변수들이 <표 5.2>에 나타남. 표에서 진하게 표시한 변수들은 2005년 및 2010년 시범 자료에서 공통으로 유의하게 나타난 변수를 의미함.

<표 1.5.2> 로짓모형에서 유의하게 나타난 변수들

문항	2005년 자료	2010년 시범자료
아동보육	교육정도 만 나이 성별 5년전 거주지 가구주 배우자 경제활동상태 가구주 배우자 교육정도	교육정도 가구주 경제활동상태 가구주 교육정도
활동제약	시도 만나이 통근/통학여부 경제활동상태 경제활동가능여부	시도
이용교통수단		교육정도 활동제약유무 근로장소 혼인상태 점유형태
사회활동a)		시도 만나이 교육정도 주거용 연면적

고령자 생활비 원천	시도 만나이	시도 만나이
정보통신기기 보유 여부b)		시도 점유형태
주차장소c)		만나이 거처의종류 자동차 보유댓수
승용차 (경차)d),e)		주차장소(자가)
승용차 (경차 외)		
승합차e)		주차장소(자가)
화물·특수자동차e)		
오토바이f)		주거용연면적 기타방수 침실수 식사방수 주차장소(자가) 시도
자전거f)		주인가구여부 주차장소(자가) 시도

a) 2010년 신규 문항이므로 2005년 인구주택총조사에서는 측정되지 않음.
b) 10년 주기로 측정되므로 2005년 인구주택총조사에서는 측정되지 않음
c) 2005년 자료의 경우 주차장소 응답에 결측이 없어 모형 적합이 불가능함.
d) 2005년 인구주택총조사에서는 승용차를 “경차”와 “경차 외”로 구분하지 않고 한 문항으로 조사함
e) 2005년 인구주택총조사에서는 이 문항에 결측이 없어 모형 적합 불가
f) 2010년 신규 문항이므로 2005년 인구주택총조사에서는 측정되지 않음.

(2) 무응답이 발생한 변수와 연관되어 있는 변수의 선택

○ 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료에 대하여 다중응답 문항별로 일반화다항로짓모형을 실시하여 5% 유의수준 하에서 유의하게 나타난 변수는 다음의 <표 1.5.3>에 나타남. 또한, 이 표는 교통수단 보유대수에 관한 문항별로 순서형 변수에 대한 로짓모형을 실시하여 5% 유의수준 하에서 유의하게 나타난 변수들도 포함함.

○ 다중응답 문항들의 응답 항목의 숫자는 적게는 6개(주차장소)로부터 많게는 12개(고령자 생활비

원천)까지 상당히 많은데 반하여 응답자의 분포를 살펴보면 주로 2-4 개 항목에 집중되어 있었고 나머지 항목들의 응답비율은 10%보다도 낮거나 일부 항목은 1% 미만으로 매우 낮게 나타남. 이와 같이 응답항목의 측정 비율이 낮은 경우 일반화다항로짓모형을 적합하기 위한 유효한 관찰값의 숫자가 작아 모형 적합에 어려움이 발생하는 경우가 있었음.

○ 이 조사와 같이 대규모 자료에 대하여 SAS의 NLMixed나 GLIMMIX procedure가 "메모리 부족"으로 실행되지 않는 경우가 종종 발생하였음. 이 경우, 응답자의 일부를 랜덤하게 뽑거나 응답비율이 적은 항목을 통합하여 모형 적합을 시도하였음. 또한, 랜덤효과를 제외한 다항로짓모형을 함께 실시하여 유의할 것으로 생각되는 변수들을 먼저 추려낸 후 일반화다항로짓모형을 적합하는 등의 모형 적합을 위한 시도를 하였음. 이와 같은 문제는 분석이 실행되는 컴퓨터의 메모리 용량에 따라 발생하지 않을 수도 있음.

<표 1.5.3> 일반화다항로짓모형에서 유의하게 나타난 변수들

문항	2005년 자료	2010년 시범자료
아동보육	시도 성별 5년전 거주지 만나이 교육정도 5년전 거주지 가구주 배우자 경제활동상태 가구주 배우자 교육정도 가구주 경제활동상태 가구주 교육정도	시도 성별 5년전 거주지 만나이 교육정도 가구주 배우자 경제활동상태 가구주 배우자 교육정도 가구주 경제활동상태 가구주 교육정도 출생지 시도 출생지 시군구 1년전 거주지
활동제약	시도 만나이 교육정도 경제활동상태 경제활동가능여부 성별 통근/통학여부	시도 만나이 교육정도 경제활동상태 경제활동가능여부
이용교통수단	시도 만나이 교육정도 성별 통근/통학장소	시도 만나이 교육정도 성별 통근/통학장소

	종사상지위 주거용 연면적 재학여부 점유형태 자동차 보유댓수 통근/통학여부 경제활동상태 근로장소 혼인상태	종사상지위 주거용연면적 재학여부 점유형태 자동차 보유댓수 사회활동유무 자동차 이용횟수
사회활동a)		시도 만나이 성별 경제활동상태 혼인상태 자녀수 가구구분 거처의 종류
고령자 생활비 원천	시도 만나이 성별 교육정도 경제활동상태 통근/통학여부	시도 만나이 성별 교육정도 경제활동상태
정보통신기기 보유 여부b)		시도 가구구분 점유형태 거처종류 난방시설
주차장소	시도 거처의 종류 자동차보유댓수 난방시설 주인가구여부	시도 거처의 종류 자동차보유댓수 난방시설
승용차 (경차) c),	시도 침실수 기타방수 종사상지위 가구구분 거처종류 주인가구여부 주거용연면적 난방시설	시도 침실수 기타방수 종사상지위 가구구분 경차이용횟수 식사방수 주차장소(자가) 주차장소(공터)
승용차 (경차 외)	주차장소(영업용,건물부설주차장)	침실수

	식사방수 주차장소(자가) 주차장소(도로변) 주차장소(노상주차장) 주차장소(공터) 점유형태	기타방수 종사상지위 가구구분 거처종류 주인가구여부 주거용연면적 난방시설 주차장소(영업용,건물부설주차장) 주차장소(자가) 주차장소(도로변) 주차장소(공터) 이용횟수
승합차	시도 침실수 기타방수 종사상지위 가구구분 거처종류 주거용연면적 난방시설 주차장소(영업용,건물부설주차장) 주차장소(자가) 주차장소(도로변) 주차장소(노상주차장) 주차장소(공터) 점유형태	주차장소(노상주차장)
화물·특수자동차	시도 침실수 종사상지위 가구구분 거처종류 주인가구여부 주거용연면적 난방시설 주차장소(영업용,건물부설주차장) 주차장소(자가) 주차장소(도로변) 주차장소(노상주차장) 주차장소(공터)	주차장소(영업용,건물부설주차장) 침실수
오토바이d)		주차장소(노상주차장) 오토바이 이용횟수
자전거d)		침실수

a) 2010년 신규 문항이므로 2005년 인구주택총조사에서는 측정되지 않음.
b) 10년 주기로 측정되므로 2005년 인구주택총조사에서는 측정되지 않음
c) 2005년 인구주택총조사에서는 승용차를 "경차"와 "경차 외"로 구분하지 않고 한 문항으로 조사함
d) 2010년 신규 문항이므로 2005년 인구주택총조사에서는 측정되지 않음.

(3) 대체군의 선택

○ 2005년 인구주택총조사 자료 및 2010년 인구주택총조사 시범자료에 대하여 일반화다항로짓분석(또는 순서형 변수에 대한 로짓분석) 및 로짓분석을 실시하여 유의하게 나타난 변수들에 기초하여 <표 1.5.4>에 나타난 변수들을 대체군을 형성하는데 사용하기 위한 변수들로 선택함.

○ 본 연구의 모의실험 자료가 실제 자료의 일부분인 10% 표본과 시범자료이므로 전체 인구주택총조사에 대한 대체를 실시할 때 고려할 수 있는 변수들을 가능한 한 포함하기 위하여 가능한 한 여러 개의 연관된 변수들을 대체군 형성을 위해 선택함.

○ 무응답 발생확률과 연관되어 있는 변수를 선택하는 로짓모형에 근거한 변수선택 결과인 <표 1.5.2>보다 무응답이 발생한 변수와 연관되어 있는 변수를 선택하는 일반화다항로짓모형(또는 순서형 변수에 대한 로짓모형)이 더 많은 유의한 변수를 포함하고 <표 1.5.2>에서 유의하게 나타난 변수를 포함하면 추정량의 분산을 줄일 수 있으므로 이에 근거하여 충분한 숫자의 대체군 형성 변수를 선택함.

○ 교통수단 보유여부에서는 주차장소에 대한 변수들이 유의한 경우가 많아 일부만 유의하면 기타를 전체 항목 변수들을 모형에 포함시켰음.

○ 2005년 인구주택총조사 자료에서는 교통수단 이용횟수가 측정되지 않아 이 자료에 근거한 모형에서는 이용횟수가 대체군 형성을 위해 고려될 수 없었음. 한편, 2010년 시범조사에서는 승용차(경차 및 경차 외 모두)에서 이용횟수가 유의하게 나타나 모형에 포함되었음. 2010년 시범자료에서 승합차 및 화물·특수자동차에 관한 응답자가 적어 유의하게 나타난 변수들이 많지 않았고 따라서 대체군을 형성하는 변수는 2005년 10% 자료에 크게 의존하여 결정됨. 2010년 인구주택총조사 본조사 자료에서 이 항목에 대한 대체군을 형성하기 위해 사용할 변수를 선택할 때 이용횟수가 포함되어야 할지 고려할 가치가 있는 것으로 예상됨.

<표 1.5.4> 대체군을 형성하기 위하여 선택된 변수들

문항	선택변수 리스트 (대체군 형성시 우선순위에 따라)
아동보육	교육정도, 가구주배우자 경제활동상태, 가구주 교육정도, 만나이, 5년전 거주지, 시도, 성별, 가구주 배우자 교육정도, 가구주 경제활동상태, 1년전 거주지, 출생지 시도, 출생지 시군구
활동제약	만나이, 경제활동상태, 교육정도, 성별, 통근/통학여부,시도, 경제활동가능여부
이용교통수단	종사상 지위, 혼인상태, 교육정도, 자동차 보유댓수, 성별, 시도, 근로장소, 통근/통학장소, 만나이, 주거용 연면적, 재학여부, 점유형태, 자동차 이용 횟수, 경제활동상태
사회활동a)	시도, 가구구분, 주거용 연면적, 거처의 종류, 교육정도, 혼인상태, 점유형태, 성별, 자녀수, 경제활동상태, 만나이
고령자 생활비 원천	경제활동상태, 교육정도, 성별, 만나이, 시도
정보통신기기 보유 여부b)	난방시설, 가구구분, 거처종류, 시도, 점유형태
주차장소	거처의 종류, 시도, 난방시설, 자동차 보유댓수, 주인가구 여부
승용차 (경차)	침실수, 기타방수, 경차이용횟수, 가구구분, 종사상지위, 주차장소(자가), 시도, 주차장소(공터), 식사방수, 주차장소(노상주차장), 주차장소(영업용,건물부설주차장), 주차장소(도로변)
승용차 (경차 외)	침실수, 기타방수, 승용차 이용횟수, 가구구분, 종사상지위, 주거용연면적, 난방시설, 주인가구여부, 주차장소(영업용,건물부설주차장), 주차장소(자가), 주차장소(도로변), 주차장소(공터), 거처종류, 주차장소(노상주차장), 시도
승용차 (경차 및 경차 외)	침실수, 기타방수, 종사상지위, 가구구분, 시도, 거처종류, 주인가구여부, 주거용연면적, 난방시설, 주차장소(영업용,건물부설주차장), 식사방수, 주차장소(자가), 주차장소(도로변), 주차장소(노상주차장), 주차장소(공터), 점유형태
승합차	종사상지위, 주차장소(영업용,건물부설주차장), 점유형태, 침실수, 기타방수, 주차장소(자가), 주차장소(도로변), 주차장소(노상주차장), 주차장소(공터), 시도, 난방시설, 가구구분, 주거용연면적, 거처종류, 주인가구여부, 식사방수
화물·특수자동차	종사상지위, 주차장소(영업용,건물부설주차장), 난방시설, 침실수, 주차장소(자가), 가구구분, 주차장소(공터), 주차장소(도로변), 주차장소(노상주차장), 시도, 거처종류, 주인가구여부, 주거용연면적,

	기타방수, 식사방수, 점유형태
자전거	침실수, 가구구분, 주차장소(자가), 시도

○ 인구주택총조사에서 사용하는 계층적 핫덱대체를 실시하기 위해서는 대체군을 형성하는 변수의 순서를 정해야 함. 일반화다항로짓모형은 랜덤효과를 포함하므로 변수선택(variable selection)을 위한 옵션을 제공하지 않는 것이 일반적임. 따라서 본 모의실험에서는 <표 1,5.4>에서 선택된 변수들을 사용하여 다항로짓모형하에서 변수선택을 실시하고 선택된 변수의 순서에 따라 대체를 실시함. <표 1.5.4>의 변수들은 대체군을 형성할 때 우선순위로 놓는, 즉 제일 나중에 포기하는 변수들의 순서로 정렬됨.

○ 실제자료에 대하여 대체를 실시하는 경우 특정 대체군 형성 변수로 인하여 대체군을 형성하는 많은 숫자의 변수를 포기해야 하는 문제가 발생할 수 있으므로 이를 검토하여 순서를 조정할 수 있을 것으로 기대됨. 본 연구에서 진행한 모의실험의 경우 자료를 여러 번 생성하여 진행하므로 이와 같은 조정 없이 진행되었음.

3. 모의실험 자료의 무응답 생성

○ 각 문항별로 완전하게 응답한 자료를 모집단으로 가정한 후 완전 임의로 표본을 추출함. 다중응답 문항들에 대하여 2005년 인구주택총조사 10% 표본 자료는 50,000개의 완전하게 응답한 자료를 임의로 추출하고 2010년 인구주택총조사 시범조사 자료는 500개의 완전하게 응답된 자료를 임의로 추출하였음. 표본의 숫자는 모집단의 10% 미만이 되어 지나치게 높은 비율의 자료를 추출하는 것을 지양함. 응답자의 5% 이상의 자료가 추출되는 경우 유한모집단수정(finite population correction factor)을 사용하여 분산을 보정함(Kish, 1965).

○ 교통수단 보유대수 문항들에 대하여 2005년 인구주택총조사 10% 표본 자료는 5,000개의 완전하게 응답한 자료를 임의로 추출하고 2010년 인구주택총조사 시범조사 자료는 200개의 완전하게 응답된 자료를 임의로 추출하였음. 이 문항들은 다중응답 문항에 비해 응답자가 상대적으로 적어 표본의 개수가 적게 선택됨. 2010년 인구주택총조사 시범조사 자료에서 승합차, 화물·특수자동차, 오토바이 항목은 응답자의 숫자가 800명 정도 또는 그보다 훨씬 작아 모의실험을 실시할 수 없었음.

○ 무응답을 발생시킬 때 무응답 자료 메커니즘이 임의결측(MAR)이 되도록 하여야 함. 이는 무응답

자료 메커니즘이 완전임의결측(MCAR)인 경우 적절한 어느 대체 방법을 사용하더라도 추정량의 편향이 발생하지 않으므로 대체 방법의 비교가 불가능하고 실제 자료가 완전임의결측인 무응답 자료 메커니즘을 만족하는 경우는 매우 드물기 때문임. 인구주택총조사의 경우도 본 연구의 관심사인 표본항목에 대한 응답자들의 무응답이 완전히 임의로 발생했다고 가정하기는 어려움.

○ 무응답 자료 메커니즘이 가능하면 실제 자료와 비슷하게 임의결측이 되도록 하기 위하여 응답여부와 관련된 변수들을 찾아내고 이 변수들의 연관관계에 따라 무응답 자료를 생성함.

○ 2010년 시범자료의 경우 다중응답 문항에서 무응답의 비율이 10% 정도 또는 그 미만으로 나타나고 있으며 2005년 인구주택총조사 10% 표본에서는 대부분의 다중응답 문항에서 무응답의 비율이 2% 미만으로 매우 적게 나타나고 있음. 2% 무응답은 무응답의 효과를 파악하기에 너무 낮은 비율이므로 본 연구에서는 모의실험의 무응답 비율을 2010년 시범자료의 최대 무응답 비율이면서 상대적으로 크지 않은 10%로 설정함.

○ 다중응답의 발생하는 응답 불일치(2개 초과 응답)를 재연하기 위하여 현재 발생하는 자료의 응답 불일치 패턴에 따라 임의로 응답 불일치 자료를 생성한 후 모의실험을 시행함.

○ 무응답을 생성하기 위하여 고려한 변수들은 <표 1.5.4>에 나타난 대체군을 형성하는데 사용된 변수들 중 일부로서 선택하여 대체군을 사용한 대체가 임의결측(MAR)을 만족하도록 생성하였음. 무응답을 생성하기 위하여 고려한 변수들은 <표 1.5.5>에 나타남.

○ 무응답을 포함한 전체 자료에 대하여 <표 1.5.5>에 나타난 변수들을 설명변수로 포함하고 무응답 발생여부를 반응변수로 하여 로짓모형을 적합한 후 무응답이 발생할 확률을 계산함. 완전히 응답된 자료에 대하여 이 확률에 근거하여 무응답 여부를 임의로 추출한 후 절편을 조정하여 무응답 자료의 의 10%가 무응답이 되도록 무응답 자료를 생성함. 이와 같이 생성된 무응답 자료는 원래 자료의 무응답과 유사한 특색을 가질 것으로 기대되며 무응답 발생이 대체군을 형성하는 데 포함된변수들에 의존하므로 임의결측 조건을 만족할 것으로 기대됨.

○ 반복적으로 분석이 시행될 수 있도록 동일한 상황하에서 독립적으로 1,000개의 무응답 자료를 생성함.

<표 1.5.5> 무응답을 생성하기 위한 모형에 포함된 변수들

문항	선택변수 리스트
아동보육	시도, 만나이, 교육정도, 성별, 5년전 거주지

활동제약	시도, 만나이, 교육정도, 경제활동상태, 경제활동가능여부
이용교통수단	통근통학장소, 종사상 지위, 시도, 교육정도, 만나이, 성별, 재학여부, 점유형태, 자동차보유댓수
사회활동a)	시도, 만나이, 성별, 경제활동상태, 혼인상태, 자녀수, 가구구분, 거처의 종류
고령자 생활비 원천	시도, 만나이, 교육정도, 성별, 경제활동상태
정보통신기기 보유 여부b)	시도, 가구구분, 점유형태, 거처종류, 난방시설
주차장소	시도, 거처의 종류, 난방시설, 자동차 보유대수
승용차 (경차)	시도, 주인가구여부, 거처종류, 주차장소(자가)
승용차 (경차 외)	시도, 주인가구여부, 거처종류, 주차장소(자가)
승합차	시도, 주차장소(자가), 주차장소(도로변)
화물·특수자동차	시도, 주차장소(자가)
자전거	시도, 주인가구여부, 주차장소(자가)

4. 모의실험 결과

○ 생성된 1,000개의 무응답을 포함한 자료 각각에 대하여 <표 1.5.4>에서 선택된 대체군을 이용하여 본 연구에서 제안한 방법으로 대체를 실시한 후 결과를 무응답이 발생하기 전 완전 자료 및 무응답을 무시한 채 응답자만에 근거한 자료에서 얻어진 비율과 비교함.

(1) 아동보육

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 아동보육 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.6>과 <표 1.5.7>에 나타남.

○ “완전자료 응답비율”이란 무응답이 생성되기 전 완전하게 응답된 자료에서의 각 항목별 응답비율을 의미함. “대체자료 응답비율”은 본 연구에서 제안된 방법에 의하여 대체된 자료에서 관찰된 각 항목별 응답비율을 의미함. “응답자 응답 비율”은 무응답을 제외하고 응답자만에 근거하여 각 항목별 응답 비율을 계산한 결과를 의미함. <표 1.5.6>의 2005년 자료의 경우 대체된 자료의 응답 비율이

완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. <표 1.5.7>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

<표 1.5.6> 2005년 10% 표본자료의 아동보육 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (자녀의 부모)	0.640	0.641	0.647	0.001	0.007	0.881	0.122
2 (조부모)	0.088	0.088	0.089	0.000	0.001	0.999	0.999
3 (기타가족, 친인척)	0.010	0.010	0.010	0.000	0.000	1.000	1.000
4 (가사도우미, 이웃)	0.007	0.007	0.007	0.000	0.000	1.000	1.000
5 (유치원)	0.076	0.073	0.078	−0.002	0.003	0.962	0.959
6 (어린이집, 놀이방)	0.092	0.091	0.097	−0.001	0.005	0.995	0.249
7 (기타보육시설)	0.018	0.018	0.018	0.000	0.000	1.000	1.000
8 (학원)	0.316	0.315	0.299	−0.001	−0.017	0.937	0.000
9 (혼자, 아동끼리)	0.053	0.053	0.050	0.000	−0.003	1.000	0.904
10 (기타)	0.011	0.011	0.011	0.000	0.000	1.000	1.000

<표 1.5.7> 2010년 시범조사의 아동보육 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (자녀의 부모)	0.542	0.541	0.542	−0.001	0.000	0.924	0.955
2 (조부모)	0.050	0.051	0.051	0.001	0.000	1.000	1.000
3 (기타가족, 친인척)	0.008	0.008	0.008	0.000	0.000	1.000	1.000
4 (가사도우미, 이웃)	0.012	0.012	0.012	0.000	0.000	1.000	1.000
5 (유치원)	0.098	0.098	0.109	0.000	0.011	1.000	0.999
6 (어린이집, 놀이방)	0.061	0.061	0.067	0.000	0.006	1.000	1.000
7 (기타보육시설)	0.006	0.007	0.007	0.000	0.000	1.000	1.000
8 (학원)	0.103	0.103	0.099	0.000	−0.004	0.997	0.999
9 (혼자, 아동끼리)	0.358	0.359	0.348	0.001	−0.010	0.913	0.917
10 (기타)	0.054	0.054	0.052	0.000	−0.002	1.000	1.000
기타	0.014	0.013	0.013	0.000	0.000	1.000	1.000

○ “대체자료 편향”은 1,000개의 대체된 자료와 완전히 응답한 자료 사이의 응답비율의 차이의 평균을 나타내고 “응답자료 편향”은 1,000개의 무응답 자료에서 무응답을 제외한 후 응답된 자료 만에 근거한 응답비율과 완전히 응답한 자료 사이의 응답비율의 차이의 평균을 나타냄. <표 1.5.6>의 2005년 자료는 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적음을 알 수 있음. <표 1.5.7>의 2010년 시범자료에서도 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 많은 항목에서 적게 나타남을 알 수 있음,

○ “대체자료 Coverage"는 대체자료의 모비율에 관한 95% 신뢰구간이 참값(완전자료의 비율)을 포함하는 비율을 나타내고 “응답자료 Coverage"는 응답자만에 근거한 95% 신뢰구간이 참값을 포함하는 비율을 나타냄. 각 행이 응답 비율을 나타내고 비율이 매우 낮거나 높은 경우가 대부분이므로 이 경우 arcsin 변환을 통해 분산 안정화를 실시하고 신뢰구간을 계산하였음. <표 1.5.6>의 2005년 자료의 경우 대체자료의 Coverage가 응답자만에 근거한 Coverage보다 높아 대체가 잘 시행되었음을 나타냄. <표 1.5.7>의 2010년 시범자료의 경우 대체자료에 대한 “혼자, 아동끼리” 항목에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 나머지 항목에서는 Coverage가 우수하여 대체가 잘 시행되었음을 나타냄.

(2) 활동제약

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 활동제약 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.8>과 <표 1.5.9>에 나타남.

○ <표 1.5.8>의 2005년 자료의 경우 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. <표 1.5.9>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

○ <표 1.5.8>의 2005년 자료는 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적음을 알 수 있음. 비슷하게 <표 1.5.9>의 2010년 시범자료에서도 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적게 나타남을 알 수 있음.

○ <표 1.5.8>의 2005년 자료의 경우 대체자료의 Coverage가 응답자만에 근거한 Coverage보다 높아 대체가 잘 시행되었음을 나타냄. <표 1.5.9>의 2010년 시범자료의 경우 대체자료에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 “없음” 항목에서 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 나머지 항목에서는 Coverage가 우수하여 대체가 잘 시행되었음을 나타냄.

<표 1.5.8> 2005년 10% 표본자료의 활동제약 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (시각,청각,언어장애)	0.017	0.017	0.017	0.000	0.000	1.000	1.000
2 (치매)	0.006	0.006	0.006	0.000	0.000	1.000	1.000
3 (중풍)	0.007	0.007	0.008	0.000	0.000	1.000	1.000
4 (걷기등 육체적제약)	0.059	0.059	0.061	0.000	0.001	1.000	0.998
5 (학습어려움 등 정신적 제약)	0.028	0.028	0.029	0.000	0.001	1.000	1.000
6 (없음)	0.910	0.910	0.907	0.000	−0.002	0.827	0.361

a) 이 문항은 2010년에 변경되어 항목이 달라짐.

<표 1.5.9> 2010년 시범조사의 활동제약 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (시각,청각,언어장애)	0.005	0.006	0.006	0.001	0.001	1.000	1.000
2 (걷기 등 이동제약)	0.017	0.017	0.018	0.000	0.001	1.000	1.000
3 (정신적 제약)	0.002	0.003	0.003	0.001	0.001	1.000	1.000
4 (배우기. 기억하기)	0.004	0.005	0.005	0.001	0.001	1.000	1.000
5 (옷입기, 목욕하기)	0.002	0.003	0.003	0.001	0.001	1.000	1.000
6 (장보기, 병원가기)	0.003	0.004	0.005	0.001	0.001	1.000	1.000
7 (취업활동)	0.002	0.004	0.004	0.001	0.001	1.000	1.000
8 (없음)	0.971	0.971	0.970	0.000	−0.001	0.791	0.835

(3) 이용교통수단

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 이용교통수단 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.10>과 <표 1.5.11>에 나타남.

○ <표 1.5.10>의 2005년 자료의 경우 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. <표 1.5.11>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

<표 1.5.10> 2005년 10% 표본자료의 이용교통수단 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (걸어서)	0.308	0.308	0.314	0.000	0.006	0.944	0.278
2 (승용차,승합차)	0.320	0.320	0.321	0.000	0.000	0.943	0.949
3 (버스)	0.187	0.187	0.183	0.000	−0.004	0.974	0.569
4 (통근,통학버스)	0.050	0.050	0.049	0.000	−0.001	1.000	1.000
5 (고속,시외버스)	0.008	0.008	0.007	0.000	0.000	1.000	1.000
6 (전철,지하철)	0.091	0.091	0.086	0.000	−0.005	0.997	0.341
7 (기차)	0.002	0.002	0.002	0.000	0.000	1.000	1.000
8 (택시)	0.005	0.005	0.005	0.000	0.000	1.000	1.000
9 (자전거)	0.013	0.013	0.013	0.000	0.000	1.000	1.000
10(기타)	0.045	0.045	0.048	0.000	0.003	1.000	0.981

<표 1.5.11> 2010년 시범조사의 이용교통수단 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (걸어서)	0.254	0.255	0.247	0.001	−0.008	0.946	0.930
2 (승용차,승합차)	0.557	0.556	0.567	−0.001	0.010	0.941	0.931
3 (버스)	0.085	0.084	0.082	0.000	−0.003	0.998	0.999
4 (통근,통학버스)	0.038	0.038	0.036	0.000	−0.002	1.000	1.000
5 (고속,시외버스)	0.007	0.007	0.007	0.000	0.000	1.000	1.000
6 (전철,지하철)	0.023	0.023	0.023	0.000	0.000	1.000	1.000
7 (기차)	0.006	0.007	0.007	0.000	0.000	1.000	1.000
8 (택시)	0.002	0.004	0.004	0.001	0.001	1.000	1.000
9 (자전거)	0.028	0.028	0.028	0.000	0.000	1.000	1.000
10(기타)	0.040	0.041	0.043	0.001	0.002	1.000	1.000

○ <표 1.5.10>의 2005년 자료는 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적음을 알 수 있음. <표 1.5.11>의 2010년 시범자료에서도 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적게 나타남을 알 수 있음.

○ <표 1.5.10>의 2005년 자료의 경우 대체자료의 Coverage가 응답자만에 근거한 Coverage보다 높아 대체가 잘 시행되었음을 나타냄. <표 1.5.11>의 2010년 시범자료의 경우도 Coverage가 우수함.

(4) 사회활동

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 사회활동 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.12>에 나타남. 이 문항은 2010년 신규개설 문항이므로 2010년 시범조사 결과만 평가가 가능함.

○ <표 1.5.12>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. 또한, 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적게 나타남을 알 수 있음.

<표 1.5.12> 2010년 시범조사의 사회활동 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (사회분야단체)	0.040	0.040	0.040	0.000	0.000	1.000	1.000
2 (경제분야단체)	0.026	0.026	0.027	0.000	0.001	1.000	1.000
3 (문화분야단체)	0.089	0.089	0.089	0.000	0.000	0.999	1.000
4 (종교분야단체)	0.005	0.006	0.006	0.001	0.001	1.000	1.000
5 (지역단체)	0.099	0.099	0.099	0.000	−0.001	0.998	0.998
6 (친목단체)	0.018	0.017	0.018	0.000	0.000	1.000	1.000
7 (교육단체)	0.232	0.234	0.239	0.002	0.007	0.926	0.933
8 (기타)	0.021	0.021	0.021	0.000	0.000	1.000	1.000
9 (없음)	0.005	0.006	0.006	0.001	0.001	1.000	1.000

○ <표 1.5.12>의 2010년 시범자료의 경우 대체자료에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 "교육단체" 항목에서 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 나머지 항목에서는 Coverage가 우수하여 대체가 잘 시행되었음을 나타냄.

(5) 고령자 생활비 원천

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 고령자 생활비 원천 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.13>과 <표 1.5.14>에 나타남.

○ <표 1.5.13>의 2005년 자료의 경우 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. <표 1.5.14>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

<표 1.5.13> 2005년 10% 표본자료의 고령자 생활비원천 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (본인, 배우자의 일, 직업)	0.351	0.350	0.340	−0.001	−0.011	0.938	0.000
2 (예금,적금)	0.088	0.088	0.087	0.000	−0.002	1.000	0.980
3 (연금)	0.076	0.076	0.074	0.000	−0.002	0.999	0.964
4 (개인연금)	0.020	0.019	0.019	0.000	−0.001	1.000	1.000
5 (부동산)	0.053	0.053	0.052	0.000	−0.001	1.000	1.000
6 (주식,채권,증권)	0.001	0.001	0.001	0.000	0.000	1.000	1.000
7 (함께 사는 자녀)	0.213	0.213	0.218	0.000	0.005	0.948	0.251
8 (따로 사는 자녀)	0.301	0.300	0.306	−0.001	0.006	0.927	0.238
9 (친,인척)	0.007	0.007	0.007	0.000	0.000	1.000	1.000
10 (국가보조)	0.104	0.104	0.108	0.000	0.004	0.998	0.653
11 (이웃,사회보조)	0.017	0.017	0.018	0.000	0.001	1.000	1.000
12 (기타)	0.025	0.025	0.026	0.000	0.000	1.000	1.000

<표 1.5.14> 2010년 시범조사의 고령자 생활비원천 문항에 대한 모의실험결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (본인, 배우자의 일, 직업)	0.908	0.908	0.908	0.001	0.000	0.950	0.964
2 (예금,적금)	0.052	0.052	0.052	0.000	0.000	1.000	1.000
3 (연금)	0.085	0.085	0.084	0.001	0.000	1.000	1.000
4 (개인연금)	0.018	0.018	0.017	−0.001	−0.001	1.000	1.000
5 (부동산)	0.012	0.012	0.011	0.000	0.000	1.000	1.000
6(주식,채권,증권)a)	.	.	.	.	.	.	.
7 (함께 사는 자녀)	0.046	0.046	0.047	0.000	0.001	1.000	1.000
8 (따로 사는 자녀)	0.162	0.164	0.169	0.001	0.007	1.000	1.000
9 (친,인척)	0.001	0.003	0.003	0.001	0.001	1.000	1.000
10 (국가보조)	0.072	0.072	0.075	0.001	0.004	1.000	1.000
11 (이웃,사회보조)	0.001	0.002	0.003	0.001	0.001	1.000	1.000
12 (기타)	0.014	0.014	0.015	0.000	0.000	1.000	1.000

a) 이 항목에 대한 응답자는 없었음.

○ <표 1.5.13>의 2005년 자료는 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적음을 알 수 있음. <표 1.5.14>의 2010년 시범자료에서도 거의 대부분의 항목에서 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적게 나타남을 알 수 있음.

○ <표 1.5.13>의 2005년 자료의 경우 대체자료의 Coverage가 응답자만에 근거한 Coverage보다 높아 대체가 잘 시행되었음을 나타냄. <표 1.5.14>의 2010년 시범자료의 경우도 Coverage가 우수함.

(6) 정보통신기기 보유 여부

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 정보통신기기 보유 여부 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.15>에 나타남. 이 문항은 10년마다 측정되는 문항이므로 2005년 인구주택총조사 자료에는 존재하지 않음.

○ <표 1.5.15>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율

과 매우 유사하게 나타나고 대부분의 항목에서 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. 또한, 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 비슷하거나 적게 나타남을 알 수 있음.

○ <표 1.5.15>의 2010년 시범자료의 경우 대체자료에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 많은 항목에서 Coverage가 우수하여 대체가 잘 시행되었음을 나타냄.

<표 1.5.15> 2010년 시범조사의 정보통신기기 보유 여부 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (개인용 컴퓨터)	0.679	0.679	0.678	0.000	−0.001	0.924	0.943
2 (인터넷회선)	0.588	0.588	0.587	0.000	−0.001	0.920	0.948
3 (디지털티비)	0.284	0.285	0.283	0.001	−0.001	0.925	0.944
4 (인터넷티비)	0.071	0.071	0.071	0.000	−0.001	1.000	1.000
5 (케이블티비)	0.525	0.524	0.525	−0.001	−0.001	0.919	0.949
6 (위성방송수신기)	0.132	0.132	0.133	0.000	0.001	0.996	0.997
7 (팩스)	0.082	0.083	0.082	0.000	−0.001	1.000	1.000
8 (없음)	0.046	0.046	0.046	0.000	0.000	1.000	1.000

(7) 주차장소

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 주차장소 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.16>과 <표 1.5.17>에 나타남.

○ <표 1.5.16>의 2005년 자료의 경우 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄. <표 1.5.17>의 2010년 시범 자료도 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

<표 1.5.16> 2005년 10% 표본자료의 주차장소 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (자가주차장)	0.775	0.775	0.788	0.000	0.014	0.934	0.000
2 (영업용 또는 건물주차장)	0.030	0.030	0.030	0.000	0.000	1.000	1.000
3 (노상 주차장)	0.060	0.060	0.057	0.000	−0.003	1.000	0.959
4 (도로변,골목길)	0.135	0.135	0.126	0.000	−0.009	0.992	0.003
5 (공터)	0.017	0.017	0.016	0.000	−0.001	1.000	1.000
6 (기타)	0.004	0.004	0.004	0.000	0.000	1.000	1.000

<표 1.5.17> 2010년 시범조사의 주차장소 여부 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1 (자가주차장)	0.909	0.908	0.913	0.000	0.004	0.831	0.839
2 (영업용 또는 건물주차장)	0.025	0.026	0.026	0.000	0.001	1.000	1.000
3 (노상 주차장)	0.025	0.025	0.024	0.000	−0.001	1.000	1.000
4 (도로변,골목길)	0.043	0.043	0.041	0.000	−0.002	1.000	1.000
5 (공터)	0.015	0.015	0.014	0.000	−0.001	1.000	1.000
6 (기타)	0.004	0.005	0.005	0.001	0.001	1.000	1.000

○ <표 1.5.16>의 2005년 자료는 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적음을 알 수 있음. <표 1.5.17>의 2010년 시범자료에서도 거의 대부분의 항목에서 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 적게 나타남을 알 수 있음.

○ <표 1.5.16>의 2005년 자료의 경우 대체자료의 Coverage가 응답자만에 근거한 Coverage보다 높아 대체가 잘 시행되었음을 나타냄. <표 1.5.17>의 2010년 시범자료의 경우 대체자료에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 “자가주차장” 항목에서 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 나머지 항목에서는 Coverage가 우수하여 대체가 잘 시행되었음을 나타냄.

5. 응답초과 문항에 대한 모의실험

○ 주된 2가지를 선택하는 문항에 대한 응답이 2개를 초과하는 경우 제안된 대체방법의 성능을 평가하는 모의실험은 2개 이상 응답자의 숫자가 너무 적어 시행에 어려움이 존재함. 응답자의 숫자가 충분한 2005년 인구주택총조사 10% 표본 자료는 응답초과 정보를 포함하지 않고 있으며 2010년 시범자료의 경우 전체적으로 표본의 숫자가 작은데다 응답초과는 설문지 응답자에게서만 발생하므로(인터넷 설문조사는 구조적으로 응답초과를 방지함) 주된 2가지를 초과해 응답한 경우가 거의 없었음.

○ 본 연구에서는 제한적으로나마 모의실험을 실시하기 위하여 아동보육 문항에서 2개 이상의 항목에 대하여 응답한 사람들 중 5명의 빈도를 보인 부모-방과후학교-학원 조합에 응답한 경우에 제한하여 응답초과가 발생하였다고 가정하고 모의실험을 진행함. 이 모의실험의 결과는 제한적일 수밖에 없으며 2010년 인구주택총조사 원시자료를 가지고 다시 모의실험을 수행할 것을 적극 추천함.

○ 모의실험에서는 부모-방과후학교-학원 조합에 대하여 응답한 응답초과 아동의 성별과 나이를 확인한 후 본인의 조합 중 2개를 응답한 응답자 중 성별과 나이가 동일한 아동이 표본으로 선택되면 1/2의 확률로 그 아동에게 응답초과를 생성하였음. 예를 들어, 10세 남아의 아동보육 문항 응답이 부모-방과후학교-학원이었으면 표본으로 선택된 10세 남아 중 이 문항의 응답이 부모-방과후학교 또는 부모-학원 또는 방과후학교-학원인 경우 1/2의 확률로 응답이 부모-방과후학교-학원이라고 응답했다고 가정하여 보기초과 응답을 생성함. 이렇게 생성된 응답초과 부모-방과후학교-학원 조합과 부모-방과후학교-혼자로 응답한 응답초과 아동에 대한 모의실험 자료에 제안된 방법으로 대체(실제로는 세 가지 중 두 가지 항목을 고르는 효과)를 실시하였음.

○ 무응답 자료에 대한 모의실험과 마찬가지로 500개의 표본을 임의로 뽑아 보기초과 자료를 만들면 대략 20-40명의 부모-방과후학교-학원 조합의 보기초과 응답자가 생성됨. 이 대체자료에 대하여 본 연구에서 제안한 대체방법을 적용하고 이를 1,000번 반복한 결과가 <표 1.5.18>에 나타남.

<표 1.5.18> 보기응답 초과 문항에 대한 모의실험 결과

	부모	방과후	학원
모집단	54.23%	10.31%	35.87%
제안된 방법에 의한 대체	54.40%	10.12%	35.98%
보기초과 응답자 제외시	53.65%	9.60%	34.52%
보기초과 응답자 자료 그대로 포함	54.89%	12.02%	36.27%

○ “모집단” 행은 보기초과 응답을 생성하기 전 주된 2곳까지 응답한 사람들의 부모-방과후-학원의 응답 분포를 나타냄. “제안된 방법에 의한 대체” 행은 본 연구에서 제안한 대체 방법에 근거하여 보기초과 응답을 제거한 후 부모-방과후-학원의 응답 분포를 나타냄. “보기초과 응답자 제외시” 행은 보기초과응답을 한 아동의 자료를 제외한 후 부모-방과후-학원의 응답 분포를 나타내고 “보기초과 응답자 자료 그대로 포함” 행은 보기초과응답을 한 아동의 자료를 보기초과 상태로 포함시킨 후 부모-방과후-학원의 응답 분포를 나타냄. 제안된 방법에 의한 대체를 실시한 경우 부모-방과후-학원의 응답 분포가 모집단의 응답분포와 가장 근접하게 나타나 제안된 방법의 타당성을 뒷받침함.

6. 교통수단 보유 문항에 대한 모의실험

○ 생성된 무응답 자료에 대하여 <표 1.5.4>의 보유교통수단 행에 나타난 변수들을 사용하여 대체군을 형성하여 대체를 실시한 후 결과가 <표 1.5.19>부터 <표 1.5.24>에 나타남.

○ <표 1.5.19>부터 <표 1.5.21>은 2005년 10% 표본 자료에서 측정된 승용차(경차와 경차외 구분 없음), 승합차 및 화물차에 대한 대체 결과로서 대체된 자료의 응답 비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

<표 1.5.19> 2005년 10% 표본자료의 승용차 보유 대수 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1대	0.840	0.840	0.841	0.000	0.001	0.825	0.845
2대	0.151	0.150	0.149	0.000	-0.001	0.991	0.994
3대	0.009	0.009	0.009	0.000	0.000	1.000	1.000
4대	0.001	0.001	0.001	0.000	0.000	1.000	1.000
5대	0.000	0.000	0.000	0.000	0.000	1.000	1.000
6대	0.000	0.000	0.000	0.000	0.000	1.000	1.000
7대	0.000	0.000	0.000	0.000	0.000	1.000	1.000
8대	0.000	0.000	0.000	0.000	0.000	1.000	1.000
9대	0.000	0.000	0.000	0.000	0.000	1.000	1.000

<표 1.5.20> 2005년 10% 표본자료의 승합차 보유 대수 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1대	0.924	0.924	0.924	0.000	0.000	0.832	0.875
2대	0.069	0.069	0.069	0.000	0.000	1.000	1.000
3대	0.007	0.007	0.007	0.000	0.000	1.000	1.000

<표 1.5.21> 2005년 10% 표본자료의 화물차 보유 대수 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1대	0.921	0.920	0.921	−0.001	−0.001	0.804	0.844
2대	0.063	0.062	0.062	0.000	−0.001	0.999	0.999
3대	0.016	0.016	0.016	0.000	0.000	0.999	0.999
4대	0.000	0.000	0.000	0.000	0.000	1.000	1.000

○ <표 1.5.19>부터 <표 1.5.21>의 2005년 10% 표본 자료에서 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 대부분 작게 나타남을 알 수 있음. 일부 항목에서 대체자료에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 나머지 항목에서는 Coverage가 우수하여 대체가 잘 시행되었음을 나타냄.

○ <표 1.5.22>부터 <표 1.5.24>는 2010년 시범 자료 중 모의실험이 가능하도록 충분한 자료가 존재하는 승용차(경차), 승용차(경차 외), 그리고 자전거에 대한 대체 결과로서 대체된 자료의 응답비율이 완전하게 응답된 자료의 응답비율과 매우 유사하게 나타나고 응답자만의 응답비율보다 완전자료의 응답비율에 가까워 대체가 우수하게 이루어졌음을 나타냄.

○ <표 1.5.22>부터 <표 1.5.24>의 2010년 자료는 대체자료의 편향이 응답자만에 근거한 자료의 편향보다 대부분 작게 나타남을 알 수 있음. 일부 항목에서 대체자료에 대한 Coverage가 응답된 자료 만에 근거한 Coverage보다 상대적으로 낮게 나타났는데 이는 단일대체의 문제점인 분산의 과소추정 때문일 것으로 추측됨. 나머지 항목에서는 Coverage가 우수하여 대체가 잘 시행되었음을 나타

냄.

<표 1.5.22> 2010년 시범자료의 승용차(경차) 보유 대수 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1대	0.861	0.868	0.861	0.007	0.000	0.823	0.876
2대	0.130	0.130	0.130	−0.001	−0.001	0.998	1.000
3대	0.007	0.007	0.007	0.000	0.000	1.000	1.000
4대	0.001	0.001	0.001	0.000	0.000	1.000	1.000
5대	0.001	0.000	0.001	0.000	0.000	1.000	1.000

<표 1.5.23> 2010년 시범자료의 승용차(경차 외) 보유 대수 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1대	0.735	0.735	0.734	0.000	−0.001	0.933	0.952
2대	0.253	0.253	0.254	0.000	0.001	0.941	0.946
3대	0.011	0.010	0.010	0.000	0.000	1.000	1.000
4대	0.001	0.001	0.001	0.000	0.000	1.000	1.000
5대	0.000	0.000	0.000	0.000	0.000	1.000	1.000

<표 1.5.24> 2010년 시범자료의 자전거 보유 대수 문항에 대한 모의실험 결과

항목	완전자료 응답비율	대체자료 응답비율	응답자 응답비율	대체자료 편향	응답자 편향	대체자료 Coverage	응답자 Coverage
1대	0.570	0.567	0.566	−0.003	−0.004	0.917	0.950
2대	0.299	0.301	0.302	0.002	0.003	0.938	0.948
3대	0.101	0.101	0.101	0.000	0.000	1.000	1.000
4대	0.025	0.025	0.025	0.000	0.000	1.000	1.000
5대	0.005	0.005	0.005	0.000	0.000	1.000	1.000
7대	0.001	0.001	0.001	0.000	0.000	1.000	1.000

Ⅰ-6. 토의

○ 본 연구에서는 다중응답 문항에 대한 무응답을 대체하기 위하여 (1) 무응답 발생확률과 대체군 형성 변수들 간의 연관성에 대한 로짓모형 및 (2) 무응답이 발생한 변수와 대체군 형성 변수들 간의 연관성에 대한 일반화다항로짓모형을 이용하여 대체를 실시하는 방법을 제안하고 제안된 방법의 성능을 평가하기 위하여 모의실험이 실시하였음. 모의실험은 제안된 방법에 근거한 대체는 (1) 2005년 인구주택총조사 5% 표본자료와 (2) 2010년 인구주택총조사 시범자료 모두에서 편향 없이 각 항목의 비율을 추정하는 것이 가능함을 나타냄.

○ 교통수단 보유 문항에 대한 무응답을 대체하기 위하여 (1) 무응답 발생확률과 대체군 형성 변수들 간의 연관성에 대한 로짓모형 및 (2) 무응답이 발생한 변수와 대체군 형성 변수들 간의 연관성에 대한 순서형 변수에 대한 로짓모형을 이용하여 대체를 실시하는 방법을 제안하고 제안된 방법의 성능을 평가하기 위하여 모의실험이 실시하였음. 모의실험은 제안된 방법에 근거한 대체는 (1) 2005년 인구주택총조사 5% 표본자료와 (2) 2010년 인구주택총조사 시범자료 모두에서 편향 없이 각 항목의 비율을 추정하는 것이 가능함을 나타냄.

○ 인구주택총조사의 경우 자료의 숫자가 크므로 본 연구에서는 대체군을 형성하는 변수들을 충분히 많이 고려하여 대체를 실시하였음. 대체군의 숫자가 늘어나면 대체군을 형성하는 특정 변수로 인하여 기증자를 찾지 못할 가능성이 있으므로 유의해야 함. 통계청에서 사용하는 SAS macro는 대체군을 형성하는 변수의 숫자를 줄여감에 따라 기증자를 발견하는 단계를 결과(OUTPUT)로 제공하므로 이 결과를 주의 깊게 살펴보고 문제가 있는 변수가 있다면 순위를 조정하는 것이 바람직함.

○ 본 연구에서 제안한 일반화다항로짓모형은 각 항목들을 비교하므로 일부 항목의 응답자가 매우 드문 경우 모형 적합에 문제가 생길 수 있음. 이 경우 이 항목을 다른 항목과 통합한 후 분석하는 방법도 가능할 것임. 이 모형을 통해 예측하는 것이 아니라 어느 변수들이 무응답이 발생한 변수와 연관되어 있는지 여부만을 확인하는데 사용하므로 제안된 방법은 모형 적합의 민감성에 상당히 강건할(robust) 것으로 기대됨.

○ 2005년 인구주택총조사 10% 표본 자료는 응답초과 정보를 포함하지 않고 있으며 2010년 인구주택총조사 시범조사 자료의 경우 다중응답 문항의 최대 2가지 응답 문항에서 발생하는 응답 불일치(2개 이상의 무응답)가 발생하는 비율이 낮아 실제 자료의 불일치 문항과 비슷한 모의실험 자료를 생성하기 어려움. 또한, 이 모의실험의 결과는 실제 자료의 불일치 발생이 모의실험의 불일치 발생과

비슷하다는 가정 하에서 정확도가 결정될 수 있음.

○ 응답초과를 보인 자료의 숫자가 작아 제한된 응답초과 상황에서 본 연구에서 제안한 대체를 통한 응답불일치 처리방법에 대한 모의실험을 진행하였고 제안된 방법에 의한 대체를 실시한 경우 응답분포가 모집단의 응답분포와 가장 근접하게 나타나 제안된 방법의 유용성을 뒷받침함.

○ 모의실험에서 나타난 바와 같이 단일대체를 시행하는 경우 대체된 자료로 인하여 실제자료보다 정보가 과대 추정되어 추정량의 분산이 과소추정될 수 있음을 유념하여야 함.

○ 2010년 인구주택총조사 시범자료는 실제 인구주택총조사보다 표본의 숫자가 작아 이 자료에 근거한 모형이 실제 전체 인구주택총조사 자료의 모형과 다를 가능성이 존재함.

○ 2005년 인구주택총조사 자료에 비교하여 2010년 조사는 문항이 변경되거나 추가되었으므로 2005년 자료에 근거한 결과를 100% 신뢰하여 그대로 시행하는 것 보다는 이를 참조하여 2010년 자료에 바람직한 모형을 선택하기 위한 가이드라인으로 사용할 것을 권장함.

참 고 문 헌

이현정 (2009), 인구주택총조사 무응답 처리기법 연구, 통계청.

통계청 (2010), 2010년 인구주택총조사 무응답 자료처리 기본계획, 통계청 조사관리국 인구총조사과.

Andridge, R. R. and Little, R. J. A. (2010), "A Reivew of Hot Deck Imputation for Survey Non-response," International Statistical Review, 78, 40-64.

Bell, P. A. and Whiting, J. P. (2007), "Imputation and Estimation for a Thematic Form Census," Research Paper, Austrailian Bureau of Statistics.

David, M., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1986), "Alternative Methods for CPS Income Imputation," Journal of the American Statistical Association, 81, 29-41.

Cantwell, P. J., Hogan, H. and Styles, K. M. (2005), "Imputation, Apportionment, and Statistical Methods in the U.S. Census: Issues Surrounding Utah v. Evans," Research Technical Series (Statistics 2005-01), U.S. Census Bureau.

European Commission (2002), "Imputation of Income in the ECHP," DOC. PAN 164, Eurostat.

Fellegi, I. P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, 71, 17-35.

Kish, L. (1965), Survey Sampling, Wiley: New York.

Little, R. J. A. and Rubin, D. B. (2002), Statistical Analysis with Missing Data, Wiley: New York.

Marker, D. A., Judkins, D. R., and Winglee, M. (2002), "Large-Scale Imputation for Complex Surveys," Survey Nonresponse, Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. A. J. (eds.), 329-341.

Mason, P., Bankier, M. and Poirier, P. (2002), "Imputation of Demographic Variables from the 2001 Canadian Census of Population," Proceeding of the Joint Statistical Meetings - Section on Survey Research Methods, The American Statistical Association.

Meng, X.-L. (1995), "Multiple Imputation with Uncongenial Sources of Input (with discussion)," Statistical Science, 10, 538-573.

Rubin D. B. (1987), Multiple Imputation for Nonresponse in Surveys, John Wiley: New York.

Schafer, J. L. (1997), Analysis of Incomplete Multivariate Data, Chapman & Hall: Londonl.

U. S. Census Bureau (2004), "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology," Technical Document, U.S. Census Bureau.

Westat and Mathematica Policy Research (2001), "Survey of Income and Program Participation Users' Guide," Supplement to the Technical Documentation, U.S. Census Bureau.

Wetrogan, S. I. and Cresce, A. R. (2001), " ESCAP II: Characteristics of Census Imputations," Technical Report, No. 22, U.S. Census Bureau.

제 Ⅱ 부

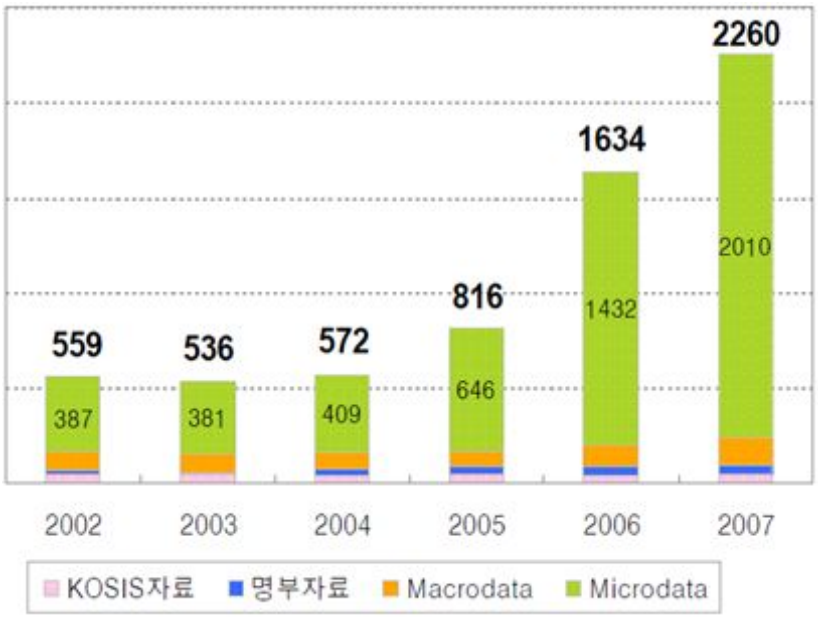
마이크로 데이터 제공 방안

Ⅱ-1. 연구개요

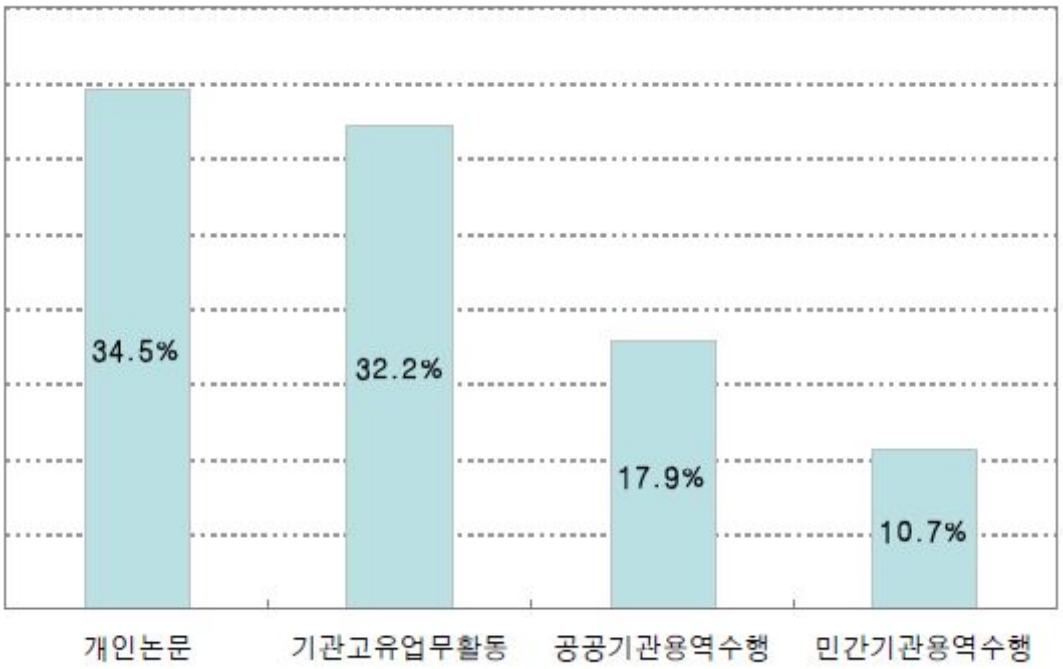
1. 연구 배경

○ 사회구조가 복잡해짐에 따라 의사결정의 기초자료로서 통계수요가 급격히 증가하고 있음. 특히, 중앙통계기관인 통계청의 자료는 각종 국가정책 수립 및 평가, 기업경영, 사회현상 탐구 등의 기초자료로 활용되는데, <그림 1.1>에서도 살펴볼 수 있듯이 2002년 559건의 자료 이용건수가 5년 후인 2007년에는 2,260건으로 그 수요가 급격히 증가함을 알 수 있다.

○ 통계청에서 실시하는 인구주택총조사(census: 이하 인총조사)는 매 5년마다 전국을 대상으로 실시되고 있으며, 이를 통해 수집된 결과에는 우리나라의 인구· 사회현상 파악에 도움이 되는 여러 기본정보들을 포함하고 있다. 이러한 데이터는 민간 혹은 학계에서는 직접 수집이 거의 불가능함은 당연한 사실이다. 따라서 통계청에서는 그간 수집된 데이터를 집계(aggregation)하여 제공하여 왔다. 그러나 최근에는 원천 자료에 대한 직접적 분석에 대한 욕구가 증가하였고 따라서, 학계나 연구기관의 통계이용자들이 마이크로자료의 제공 대상 확대를 요구하고 있는 실정이다.



<그림 2.1.1> 통계청 자료 제공수



<그림 2.1.2> 마이크로 데이터 이용목적

○ 또한 <그림 1.2>에서 볼 수 있듯이 최근에는 민간기관의 용역수행, 특히 기업의 마케팅이나 광고 효과 분석에서 수요가 증가하고 있는데 이는 주로 데이터를 이용한 기업고객 분석의 환경이 발전하여 나타나는 결과로 유추할 수 있다.(통계청, 2006)

○ 현재 우리나라에서 마이크로 데이터는 크게 세 가지 방법으로 이용자에게 제공된다. 먼저 개인식별 데이터를 제거한 마이크로데이터를 CD- ROM에 저장하여 제공하는 형식인데, 통계쇼핑몰(http://www.nso.go.kr /shop)을 통해 제공되고 있다. 다음으로 통신망을 이용한 방법인데, 이용자가 인터넷을 통해 시스템에 접속하여 원하는 자료를 직접 가공하거나 다운받을 수 있도록 지원하는 원-스톱서비스 시스템인 MDSS (MicroData Service System)를 이용하는 방법이다.

○ MDSS는 2006년 1월 1일부터 서비스를 시작하였고 홈페이지를 통해 제공 가능한 자료의 목록과 범위, 이용방법 및 비용 등을 안내하고 있으며, MDSS 사이트(http://mdss. nso.go.kr)에 접속하여 신청한 후 자료를 다운로드 받을 수 있다.

○ 또한 요청기관에 따라 중앙정부 및 지방자치단체에서 신청할 경우에는 통계청에서 작성하여 제공하고 있으며, 기타 기관 즉, 대학 및 연구기관에서 신청할 경우에는 통계진흥원으로 이첩하여 제공하는 방식을 취하고 있다.

<표 2.1.1> 데이터 제공 형식

형식		자료의 특징 및 내용	통계청 적용형태
통계표 형태 서비스		심층 이용자 요구에 따라 통계표 작성. 지나치게 세분화되면 비밀노출 위험	간행물, KOSIS
익명자료 파일서비스	공공이용 파일	외부 이용자를 위해 사전에 만들어진 마이크로데이터 파일로 CD- ROM등에 저장	마이크로 데이터 CD-ROM
	인가파일	특정한 심층 이용자에게 제공되며, 제공 전에 법적인 조치가 필요	개별자료 제공
원격접속 파일 서비스		통신망을 통하여 심층 이용자들이 마이크로 데이터를 활용하는 제도	MDSS
데이터 실험실을 통한 서비스		통계기관의 엄격한 감독 하에 기관 내에서 마이크로 데이터를 이용하는 제도	On-site

○ 이러한 내용을 다시 표로 정리하면 <표 1.1>과 같다.(2008, 정동명 외)

○ 인구주택 총조사 데이터의 경우 2005년 조사 기준으로 개인 식별이 가능한 이름 및 성씨 /본관 항목을 제외한 1, 2, 5% 마이크로 데이터(Micro data)가 제공되고 있다.

○ 2% 마이크로 데이터의 경우 가중값의 유무에 따라 달리 제공되고 있는데, 이전까지는 시/군/구 단위의 표본조사 결과를 가중값 없이 제공하였으나, 2007년부터는 노출제한방법이 적용된 2% 마이크로 데이터를 제공함으로써, 개인정보노출을 방지함과 동시에 원천 자료의 이용이라는 두 가지의 목적을 달성하고자 노력 중이다.

2. 연구의 필요성

○ 현재 제공되는 인총조사에 대한 마이크로 데이터는 개인 정보가 식별 가능한 민감한 항목들은 일부 제거한 후 제공하고 있기 때문에 활용 면에서 일부 제한적인 것이 사실이다.

○ 이에 따라 최근 통계청을 중심으로 개인정보의 비밀보장을 유지하면서 마이크로 데이터의 제공 폭을 확대하고자 여러 통계적 기법에 대한 연구가 진행되고 있다.

○ 기존 연구는 주로 인총조사 및 기타 통계청에서 제공되는 데이터에 대한 마이크로 데이터에서 개인정보 보호와 관련된 노출위험에 관한 연구가 중심이라고 할 수 있다. 노출위험이란 기존 문헌 연구

에서 데이터의 유일성으로 언급되는데, 데이터의 유일성이란 어떤 조사단위가 데이터 내에서 식별될 가능성을 나타내는 측도라고 할 수 있다. 데이터의 유일성을 통해 노출위험이 증가함을 억제하기 위해 통계적 확률모델에 대한 연구가 진행되어 왔다.

○ 상기한 여러 기법을 통해 마이크로 데이터가 제공되고 있으나, 노출위험과 관련된 여러 세부적 연구는 추가적으로 필요한 실정다. 특히, 노출을 막기 위해 제한된 여러 항목에서 어떠한 변수를 추가 제공할 지는 본 연구의 핵심이라고 할 수 있다.

○ 인구센서스 마이크로 데이터에 대한 개별 식별단위가 전국 단위라면 노출위험과 유일성에 대한 문제가 상대적으로 덜 하겠으나, 실제 연구자들은 소지역 단위 마이크로 데이터 제공을 원하는 것이 현실이다. 이러 요구에 부응하기 위해 상대적으로 노출위험이 적으며 이전보다 상세한 단위의 데이터 수준은 어디까지인지에 대한 연구가 필요하다고 하겠다.

○ 기존 제공 마이크로 데이터는 인총조사의 경우 1%인 경우에는 전국, 5%인 경우에는 시/군/구, 2%인 경우에는 가중치 유무에 따라 전국과 시/군/구 단위로 나뉘어 제공되고 있으나, 실제 읍/면/동 단위나 이하 세부 소지역에서 제공되는 경우 어느 정도 노출위험이 존재하는 지에 대한 연구가 절실하다고 할 수 있다.

○ 소지역로 제공되는 데이터는 필수적으로 노출위험을 증가시키는 바 이를 위해 소지역 단위의 요약에 대한 연구도 수반되어야 하겠다. 여기서 언급되는 소지역 단위 데이터 요약이란 자료 파일 내에 많은 레코드들은 선정된 하나의 기준변수에 따라 정렬하고, 이를 이용해 관측값을 그룹화한 뒤 그룹 내의 각 자료값을 그룹평균으로 대체하는 방법을 의미한다.

○ 또한 소지역 단위로 제공되는 마이크로 데이터 중 원 자료의 표본 집단을 제공하는 경우에는 모집단의 모수 추정과 관련되는 '가중치'와 '상대표준오차'의 문제가 발생할 수 있다. 표집의 추정치의 상대표준오차가 낮은 경우 가중치를 추정치에 곱하더라도 모수 추정을 위한 값의 정확도가 담보되지 않으며, 이러한 현상은 특히 표본을 소지역 단위로 제공하여 소지역의 추정치를 계산하는 경우 심각하게 발생할 수 있다고 하겠다.

○ 결과적으로 소지역 단위의 마이크로 데이터 제공은 정보의 노출위험도의 증가와 동시에 정보의 정확도에서 많은 문제점을 야기시킬 가능성이 높지만, 소지역 단위 마이크로 데이터 요청이 사회적으로 증가하고 있는 현 시점에서는 반드시 해결해야 하는 과제이기도 하다.

○ 노출방지를 유지하며 데이터 이용자의 만족도를 높이기 위해서는 두 가지 목적을 달성하는 최적의 조합을 찾아내고 이를 위한 통계적 추출 기법을 밝히는 필수적이라 사료된다.

3. 연구 결과 활용방안

○ 마이크로 데이터에 대한 상세 제공 변수 및 소지역 단위 요약 방법에 대한 통계적 절차 수립 및 적정 제공 표본 비율 파악을 위한 통계적 방법 도출을 통해서 실제 마이크로 데이터 제공의 실무적 지침으로 활용하도록 한다.

○ 본 연구 대상인 인총조사 외의 다른 조사 데이터의 마이크로 데이터 제공시 이에 대한 제공 변수 범위 결정 위한 참고 자료로 활용하도록 한다.

Ⅱ-2. 인구주택총조사 마이크로 데이터 제공 현황

○ 먼저 마이크로 데이터의 적절한 제공을 위한 연구에 앞서, 현재 통계청에서 제공되고 있는 인총 조사의 마이크로 데이터 제공 현황을 살펴보고자 한다.

○ 앞에서 언급하였듯이 현재 인총 10% 표본조사에 대해 이를 각각 10, 20, 50% 재추출(resample)한 전국 대비 1, 2, 5%의 마이크로 데이터가 제공되고 있다.

1. 인총 5%의 개요

○ 5% 마이크로 데이터는 2005년 인총조사 10% 표본조사 자료에서 추출하였으며 인구파일과 가구파일로 나누어져 있다.

○ 데이터의 형태는 ASCII 코드로 구성되어 있고 두 파일은 가구키를 통해 연결하여 이용할 수 있다. 또한 각각의 파일에는 인구 가중값과 가구 가중값을 포함하고 있으므로 5% 자료만으로 전체 모집단을 추정할 수 있도록 하였다.

2. 표본추출

○ 추출틀은 조사구 특성별로 아파트 조사구, 보통조사구 및 섬조사구로 구성되었으며, 특수조사구인 기숙시설조사구, 특수하회시설조사구나 외국인 거주 지역조사구는 제외되었다. 또한 가구구분에서 일반가구만을 대상으로 추출틀을 구성되어 제공 중이다. .

○ 표본 설계는 복합표본설계(complex survey design)로서 먼저, 행정구역과 거처종류(주택 유형으로써, 단독, 아파트, 비거주용 등의 10개의 범주로 구성됨) 및 농가구분(농가/비농가) 변수를 이용하여 층화하고, 각 층 내에서는 계통추출에 의해 센서스의 '가구구분' 변수에서 '일반가구'에 해당하는 대상을 추출하여 전국 대비 5%(표본조사 내의 50%)로 구성되어 있다.

3. 가중값

○ 가중값은 인구관련 변수와 가구관련 변수 각각에 달리하여 가중값 테이블을 제공되고 있다. 가중값은 기본적으로 추출률의 역수를 제공하는 통계적 기본원리에 따르고 있으나, 표본값을 통한 모집단의 보다 정확한 추정을 위해 가중값의 집단별로 서로 다른 가중값을 제공하였다.

○ 인구부문의 가중값은 총 4가지의 변수를 통해 서로 다른 집단을 구성하고 각 셀에 대해 서로 다른 가중값을 제공하고 있다. 4가지 변수는 시군구 단위를 구분하는 변수와 해당 지역의 동(洞) 부, 읍(邑) 부, 면(面)부의 구분 및 성별과 연령 그룹 (17개, 5세 단위)이다. 이를 통해 총 15,020개의 가중치를 제공하고 있다.

○ 시/군/구 단위는 우리가 흔히 알고있는 기초단체 행정구역(행정자치부 기준)과 달리 2005년 당시 행정구역 234개와 26개의 비자치구 즉, 광역시가 아닌 '시(市)' 단위에 소속한 '구(區)'까지 포함한 단위로써, 전국을 동일한 '급(級)'의 단위로 상호 배반적(exclusive)지역으로 나누어 가중치의 정확도를 높이고자 하였다.

○ 가구부문도 인구부문과 유사하게 시, 군, 구 260개와 동, 읍, 면부인 3개의 지역구분을 공통적으로 채택하여 가중치를 부여하고 있으나, 인구부문과는 달리 주택유형(아파트, 단독, 기타주택)과 가구원수를 가중치를 제공하는 기준 변수로 채택하였다. 이를 통해 총 5,342개의 가중치를 제공하고 있다.

4. 적용된 비밀보호 방법

○ 제공 마이크로 데이터는 2005년 인총조사 표본조사표의 41개 항목을 비교·분석한 후, 특성에 따라 제공범위에서 항목을 제외하거나 자료를 축소, 세부항목을 통합·제거하는 방법을 적용한 후 제공되고 있다. 즉, 자료의 제공범위를 제한하는 방법으로 지역관련 변수를 이용하거나 가구 관련 변수의 일부를 제한하였으며, 항목의 범주값을 통합하여 범주수 자체를 줄여 제공하고 있다.

○ 제공범위를 지역 관련 변수로 제한하기 위해 제공되는 지역을 시군구 단위까지만 제공하고, 인구 5만 이하 시·군·구 지역은 인접지역을 묶어 그룹핑한 후 기타지역으로 표시하여 제공하고 있다. 또한 제공되는 가구를 6인 이상의 가구를 제한함으로써 분석의 편의는 제공하되 정보노출을 줄이는 방법을 채택하고 있다.

○ 범주형 변수에 대한 자료 제공시 각 범주값이 유일함으로써 정보노출이 되는 경우를 막고자 정보노출의 대표적인 기법인 'Grouping' 방법과 'top-coding/bottom -coding'방법을 적용하여 제공한다.

전자의 방법에 적용된 변수로는 '행정구역 코드 (시/도, 시/군/구, 읍/면/동)' 관련된 변수와 가구원에 관련된 '나이, 가구주와의 관계, 교육정도, 아동보육, 5년전 거주지, 이용교통수단, 통근·통학소요시간, 경제활동상태, 종사상 지위, 혼인연령, 고령자 생활비' 등의 변수와, 가구에 관한 사항으로써 '가구구분, 거주기간, 주차시설, 소유형태, 주인가구, 세대구성' 및 주택에 관한 사항으로써 '거처종류, 연건평, 건축년도'가 그에 해당된다.

○ 또한 Top-coding/Bottom-coding방법을 적용한 항목으로써, 가구원에 관한 '나이, 통근·통학소요시간, 혼인연령, 자녀수'와 가구와 관련된 변수인 ' 총방수, 자동차보유대수' 및 주택에 관련 변수로써 '연건평'에 그러한 기법이 적용되어 제공되고 있다.

○ 또한 아래의 변수들에서 각 세부 범주를 제외하여 데이터를 제공함으로써 정보노출을 막고 있다.

- 가구원에 관한 사항
  - 종교 종류, 활동제약, 산업, 직업, 근로장소, 추가계획 자녀수
  - "경제활동상태" 항목의 '구직여부', '일의여부' 하위항목 제외
  - "자녀수" 항목의 '사망한 자녀수' 하위항목 제외
- 가구에 관한 사항
  - 거주시설, 거주층, 난방시설
  - "점유형태" 항목의 '주거전용'과 '영업겸용'을 제외
- 주택에 관한 사항
  - "거처종류" 항목의 '단독주택' 하위항목 제외
  - 총 방수, 편익시설 수

Ⅱ-3. 노출제한을 위한 문헌 연구

1. 비밀보호 방법에 대한 연구 추세

○ 외국의 통계작성기관에서는 1970년대부터 마이크로 데이터의 제공에 대한 요구와 개인정보의 비밀보호라는 서로 상반된 요인을 동시에 만족시키고 적정한 수준의 자료를 제공할 수 있는 방법을 찾기 위해 노력해 왔으며, 현재까지도 이에 대한 다양한 연구논문들이 꾸준히 발표되고 있다.

○ 예를 들어, 미국 상무부 센서스국에서는 마이크로 데이터 제공시 통계적 비밀보호방법을 적용하고 있으며, 네덜란드나 스웨덴 등 유럽의 여러 나라에서도 개인사생활 보호에 대해 각별한 노력을 기울이고 있다. 대학교이나 연구기관 등에서도 비밀보호방법에 대한 연구가 활발하게 진행되고 있다.

○ Bethelehem et al (1990)은 자료파일에서 개인의 식별(identification)과 노출(disclose)의 문제에 대해 언급하고, 이를 해결하기 위한 방법으로 모집단 유일성 추정을 위한 통계적 모형을 설정한 후 예제를 통해 제시하였다. Marsh et al (1991)은 영국의 센서스에서 개인비밀보호를 위해 익명화된 마이크로 데이터 파일 작성의 필요성과 효과 등에 대해 언급하였다. 이외에도 Dalenius (1977)나 Kim(1986), Fuller (1993) 등 여러 학자들이 노출의 위험성과 비밀보호에 대한 다양한 방법들을 연구하였다.

○ 우리나라의 경우 개인정보 노출의 위험성에 대한 인식부족 등으로 인해 노출방지 방법에 대한 연구가 상당히 미흡하였으나, 최근 들어 통계청을 중심으로 이에 대한 연구가 활발히 진행되고 있다.

○ 특히 통계청에서는 개인의 비밀을 보호하면서 이전 보다 자세한 수준에서 마이크로 데이터를 제공할 수 있는 방법을 개발하고자 수년 전부터 연구를 수행하고 있으며, 정동명(2007) 등은 통계자료의 비밀보호에 대한 개념과 실제 자료에 적용한 사례를 지속적으로 연구하였다.

○ 주어진 자료에 대한 노출제한은 변수의 성격에 따라 크게는 두 종류로 나눌 수 있다. 즉, 범주형과 연속형에 변수에 대한 노출제한이 그것으로써, 주로 자료의 속성값을 최대한 자료 이용자에게 전달하면서 그와 동시에 개인의 정보는 보호하는 기법들이다.

2. 범주형 변수에 대한 노출제한 기법

○ 자료교환(data swapping): Dalenius가 1979년에 제안한 이산형 변수에 대한 비밀보호 방법이다. 마

이크로 자료 내의 민감한 항목의 자료노출을 방지하기 위하여, 동일한 키(key)변수 조합을 갖는 레코드(record)간에 자료값을 상호 교환하는 방식을 제안하였다. 이 방법의 특징은 정보가 비밀보호된 후에도 각 key 변수별로 민감한 항목의 빈도수는 교환하기 이전의 빈도수와 동일하게 되어 자료의 분석이 용이하고, 이진변수(dichotomous variable)와 다지분리 변수(polychotomous variable)인 경우에도 모두 적용이 가능하다는 특징을 지니고 있다고 할 수 있다. 이를 예를 들어 설명하면 아래와 같다.

<표 2.3.1> 자료교환 예시 데이터

(a) 원데이터

관찰치#	X	Y	Z
1	0	1	0
2	0	1	0
3	0	0	1
4	0	0	1
5	1	1	1
6	1	0	0
7	1	0	0

(b) 자료교환 후 데이터

관찰치#	X	Y	Z
1	1	1	0
2	0	1	0
3	0	0	1
4	1	0	1
5	0	1	1
6	1	0	0
7	0	0	0

먼저 원데이터가 <표 2.3.1>의 (a)와 같이 관찰치 7개와 3개의 변수 X, Y, Z로 구성되어 있다고 하자. 특히, 이 예에서 변수 X는 노출이 될 경우 민감한 문제가 발생하는 변수라고 하자. 자료교환은 이러한 민감한 변수에 대해 관찰치의 위치를 바꿈으로써 데이터의 부분적인 결과에 영향을 미치지 않는 동일한 결과를 제공하는데 주안점을 둔다는 것이다. 즉, 변수 X의 관찰치 속성값 1을 가지는 번호 '5,6 및 7'번째의 관찰치 위치를 <표 2.3.1> (b)와 같이 번호 '1, 4, 6'번으로 위치 변경한다. 이러한 방법은 <표 2.3.2>를 통해 볼 수 있는 세 개의 모든 변수에 대한 분할표를 제외한 각 변수에 대한 빈도표나 세 변수의 어떤 조합의 2차원 분할표도 동일한 결과를 제공한다. 이러한 방법은 이후 기술한 다른 연구와 유사하게 자료 자체에 인위적인 조작을 가함으로써, 이를 통해 정보노출을 제한한다는 방법적 특성을 지니고 있다. 또한 제공하여야 하는 변수가 범주형이어야 하며 연속형인 경우에는 범주형 변수로 변환하는 과정에서 정보의 손실을 감수해야한다는 불편한 점이 존재한다.

<표 2.3.2> 3차원 분할표

(a) 원데이터

Z=0			*Z=1*		
	Y			*Y*	
X	*0*	*1*	X	*0*	*1*
0	0	2	*0*	2	0
1	2	0	*1*	0	2

(b) 자료교환 후

Z=0			*Z=1*		
	Y			*Y*	
X	0	1	*X*	0	1
0	1	1	0	1	1
1	1	1	1	1	0

○ 코딩 접근법(Coding Approach) : 이 방법은 원 자료에 적절한 값(noise)을 추가하여 원래 관측치에 대한 정보를 보호하고자 다른 자료값으로 변형하는 방법의 일종으로 볼 수 있다. 즉, 원 자료 값에 약간의 무시할 만한 값을 가감하기는 하나 원래의 값의 성질과 합이나 평균값은 유지할 수 있게 데이터를 조작하는 방법의 일종이라 할 수 있다.

○ 그룹화(Grouping) : 자료파일에서 어떤 변수들은 특성상 노출되기가 쉬운 범주로 구성된 경우, 인근의 다른 범주들과 통합함으로써 쉽게 식별이 되지 않도록 할 수 있는 기법으로 현재 2005년 기준 센서스 마이크로 데이터에서 적용된 기법이다.

3. 연속형 변수에 대한 노출제한 기법

○ 반올림(Rounding) : 주어진 자료를 적당한 몫과 나머지의 형태로 나타내어 원자료의 값을 식별할 수 없게 보호하는 기법이다.

○ 구간 그룹화(Grouping into Intervals) : 연속형 자료를 적당한 구간으로 그룹화하여 각 구간을 대표하는 코드(code) 값으로 원 자료를 대체하는 방법이다. 대체하는 통계량에 따라 여러 기법이 만들어 질 수 있다는 특징이 있다. 예를 들어, 분포 성격에 따라 평균 혹은 중위수를 달리 사용할 수 있다.

○ 마이크로 요약(Micro Aggregation) : 자료 파일 내에 많은 레코드들은 선정된 하나의 기준변수에 따라 정렬될 수 있으며, 이를 이용하여 보통 3~4개의 관측값을 그룹화하고 그룹 내의 각 자료값은 그룹평균으로 대체하는 기법이다. 현재 국내 연구에 적용된 방법으로 새로운 기법 적용시 비교, 연구되어야 하는 기법이라고 사료된다.

4. CTA(Controlled Tabular Adjustment) 기법

○ 원자료에서 요약된 결과표를 공표하는 경우에 정보노출을 제한하기 위한 기법으로 최근 많이 연구되고 있는 CTA기법을 소개하고자 한다.

○ 이 기법은 모든 매크로 데이터에 대한 요약 기법은 아니다. 기존 알려진 재입력(recoding)이나 정보제한(suppression)기법이 정보 손실이 발생한다는 단점을 존재하는 바, 이를 극복하여 손실을 최소화하며 정보노출을 막자는 취지 하에서 연구된 기법이다.

○ CTA를 설명하기 위해 아래와 같은 식을 우선 가정하자.

- m개의 선형결합 $Aa = b$를 가지는 일반적인 분할표 $a_i,\ i = 1,\cdots,n$가 있다고 가정하자.
- 특정 셀의 값에 대해 $l \le a \le u$의 조건을 만족하는 상한 u와 하한 l이 있다고 가정하자.
- 셀 연동 가중치인 w_i는 음이 아닌 값을 가지며, $i = 1, \cdots, n$ 이라고 하자.
- 정보노출이 우려되는 민감한 셀의 집합은 $P \subseteq \{1, \cdots n\}$이라고 하자.
- 각각의 초기 셀에 대해 보호하고자 하는 상한 및 하한을 각각 lpl_p과 upl_p로 표기하고 이때 p는 $p \in P$라고 하자.

○ 위에 가정된 표기에 의해, CTA는 임의의 거리인 L_w을 이용하여 a에 가장 가깝고 정보노출이 되지 않는 '안전한 결과 테이블'인 x를 다음의 (식 1)과 같은 로직에 의해 찾는 것이다.

(식 1)

$$\min_x \| x - a \|_{L(w)}$$
$$\text{subject to } Ax = b,\ l_x \le x \le u_x,$$
$$x_p \le a_p - lpl_p \text{ 이거나 } x_p \ge a_p - upl_p,\ p \in P$$

만약 $z = x - a$, $l_z = l_x - a$이고 $u_z = u_x - a$라고 한다면, 위의 식 (1)은 아래와 같은 식(2)로 재정의할 수 있다.

(식 2)

$$\min_x \| z \|_{L(w)}$$
$$\text{subject to } Az = 0,\ l_z \le z \le u_z,$$
$$z_p \le -lpl_p \text{ 이거나 } z_p \ge upl_p,\ p \in P$$

○ CTA는 기존의 반올림이나 정보제한과 같이 셀 값의 주변정보(marginal values)를 변화시키지 않고 특정 셀들의 값 만으로 정보가 노출되지 않는 테이블을 생성하는 '섭동(perturbative)'기법의 대표적인 예이다. 섭동 기법은 CTA외에도 CR(controlled rounding)기법이 있다. 재입력이나 정보제한 기법은 섭동 기법에 대별하여 '비섭동(non-perturbative)'기법이라고 한다.

○ 실제 수치의 예를 통해 섭동 기법인 CTA와 이를 비교하기 위해 비섭동 기법인 재입력 방법을 살펴보기로 하자. <표 2.3.3>은 원 자료와 재입력 방법에 의해 수정된 매크로 정보가 두 개의 표로 제시되어 있다. 또한 <표 2.3.4>는 CTA에 의해 생성된 2개의 테이블과 원자료가 함께 제시되어 있다.

○ 먼저 <표 2.3.4>를 살펴 보면 원래 생성된 좌측의 테이블의 (P_3, M_2)번째 셀 값인 '40'이 정보노출을 막아야 하는 민감한 값이라고 하자. 이 경우 CTA기법에서는 민감한 값 '40'에 대한 정보노출을 막기 위해 주어진 규칙에 의해 일정한 값을 가감한다. <표 2.3.4>의 예에서는 이러한 가감값을 '5'로 설정하여 가운데와 우측에 있는 테이블을 생성한 것이다. 즉, 민감한 값 '40'을 막기 위해 해당 셀에 '5'를 빼거나 더한 결과인 두 개의 테이블을 재생성한다. 이러한 경우 해당 셀에 일정한 값을 더하거나 빼기를 했기 때문에 다른 셀의 값도 영향을 받을 수 밖에 없다. <표 2.3.4>의 가운데 테이블을 살펴보면 (P_3, M_2)셀의 값 '40'에 '5'를 빼 '35'의 값을 얻었고 아울러 셀 (P_3, M_1)은 '28'에서 '33'으로 '5'가 늘어난 것을 볼 수 있을 것이다. 또한 셀 (P_1, M_1)은 '20'에서 '15'로 변동되었다. 이러한 결과로 특정 셀의 값이 변동이 되더라도 테이블의 주변 정보인 합계정보는 변하지 않았다. <표 2.3.4>의 우측 테이블은 정해진 가감치인 '5'를 더하고 이상에서 설명한 내용과 동일하게 각 셀의 값들을 조정한 결과이다.

○ 앞에서 설명한 CTA기법을 <표 2.3.3>의 재입력 기법과 비교하여 보자. 예에서 볼 수 있는 재입력 기법의 특징은 민감한 셀이 (P_3, M_2)라고 가정하는 경우, 이에 대한 정보를 노출하지 않기 위해 X축의 범주 P_i 중에서 P_2와 P_3를 합쳐서 새로운 범주를 생성하는 기법이다. 따라서, P_i에 대한 주변 정보도 (130, 73, 46, 90, 31)에서 (130, 119, 90, 31)로 변동하게 되어 정보 왜곡이 발생하였

다.

<표 2.3.3> 재입력(recoding)기법의 적용의 예

Original table

	P_1	P_2	P_3	P_4	P_5	TOTAL
M_1	20	15	30	20	10	95
M_2	72	20	**1**	30	10	133
M_3	38	38	15	40	11	142
TOTAL	130	73	46	90	31	370

Recoded table

	P_1	P_2+P_3	P_4	P_5	TOTAL
M_1	20	45	20	10	95
M_2	72	**21**	30	10	133
M_3	38	53	40	11	142
TOTAL	130	119	90	31	370

<표 2.3.4> CTA기법의 적용의 예

Original table

	P_1	P_2	P_3	TOTAL
M_1	20	24	28	72
M_2	38	38	**40**	116
M_3	40	39	42	121
TOTAL	98	101	110	309

Adjusted table, lower protection sense

	P_1	P_2	P_3	TOTAL
M_1	15	24	33	72
M_2	43	38	**35**	116
M_3	40	39	42	121
TOTAL	98	101	110	309

Adjusted table, upper protection sense

	P_1	P_2	P_3	TOTAL
M_1	25	24	23	72
M_2	33	38	**45**	116
M_3	40	39	42	121
TOTAL	98	101	110	309

Ⅱ-4. 마이크로 데이터 제공을 위한 통계적 방법 연구

1. 데이터 구축의 목적 및 방법

○ 2005년 인총조사 데이터 중 일부를 표본 데이터로 추출하여 효과적인 MD 구축의 방법을 서로 비교, 연구하고자 한다.

○ 또한 분석결과 및 데이터 구축에 관련성이 낮은 일부 지역 제외하고 분석에 의미있는 대상 만을 표본추출하여 데이터를 구축하였다.

○ 현재 5% 마이크로 데이터에 적용된 가구대상으로 계통추출하는 방법 외에도 방법상의 효과를 상호 비교하기 위해 집락추출(clustering sampling) 방법을 이용하였다. 두 가지의 추출 방법을 인총 10% 자료의 10~100%, 즉 전체 기준 1~10% 표본 데이터로 다시 세분화하여 정보노출의 정도, 정보노출 기법의 용이성을 비교하였다.

2. 표본 대상 선정 및 % 데이터 구축

○ MD의 외부 제공시 주요 이슈 중에 하나는 제공 정보의 노출정도이며, 정보 노출 위험은 '키(key) 변수 선택 과정', '표본추출 기법' 및 '표본 크기'와 밀접한 연관이 있다고 할 수 있다. 이에 대한 연구를 인총 10% 표본 자료를 대상으로 수행하기 위해 표본기법별, %구간별로 데이터를 세분화하였다.

○ 우선 인총 10%표본에 대해 각 '%구간' 별(1%~10%)로 데이터를 구축하였다. 구간 별 % 데이터를 여러 번 반복하여 결과를 종합함으로써 결과에 대한 신뢰도를 제고하고자 하였다.

○ 추출기법은 복합표본추출(complex survey sampling)의 방법 중에 하나인 층화 후 집락추출과 계통추출을 이용하였다. 예를 들어 인구 관련 조사에서는 조사특성에 따른 변수를 통해 층화한 후 이를 다시 가구를 집락추출하거나 계통추출하는 방식이다.

○ 또한 본 연구와 관련성이 낮은 집단 조사구, 예컨대 기숙시설, 군부대와 같은 특수 사회시설 및 관광호텔 조사구 등은 제외하는 것이 바람직하다고 사료되어 표본 대상에서 제외하였다. 이러한 조사구를 제외한 아파트 조사구, 보통조사구, 섬조사구 등 일반 주민에 대한 조사구 만을 대상으로 표

본추출을 시도하였다.

○ 위에서 적용된 원칙에 의해 최종 추출 표본 데이터를 생성하고, 모집단과 비교를 통해 표본의 특성을 고찰하였다. 추출된 표본은 조사구 선정과 표본 크기 등 가급적 모집단의 특성을 반영하도록 설계 및 추출하고자 하였다.

3. 데이터 변환

○ 본 연구에서 사용한 자료는 지난 2005년 11월에 통계청에서 실시한 2005 인총조사의 10% 표본조사 결과이다. 인총조사의 조사표는 크게 전수조사표와 표본조사표로 구분하여 가구원과 가구, 그리고 주택에 관한 사항들을 조사하도록 구성되어 있다.

○ 표본조사표는 전수항목을 포함하여 보다 세부적인 특성을 파악코자 가구원에 대한 사항 24개, 가구에 관한 사항 11개, 주택에 관한 사항 6개 등 총 41개 항목으로 구성되어 있으며, 이외에 추가로 16개 시도별로 각각 서로 다른 조사항목 3개가 포함되어 전체적으로 44개 조사항목으로 구성되어 있다.

○ 2005 인총 조사의 대상가구는 평균 60가구 내외로 구성된 조사구(enu- meration district)내에 포함되어있으며, 집락추출시 이러한 조사구를 하나의 조사단위로 간주한다. 조사구는 출입의 통제여부에 따라 일반조사구와 특별조사구로 나누어진다. 10% 표본조사에서는 군대, 교도소 등으로 이루어진 특별조사구를 제외하고, 아파트조사구(A), 보통조사구(1), 섬조사구(2), 기숙시설조사구(3), 특수사회시설조사구(4), 관광호텔 및 외국인조사구(5)로 이루어진 일반조사구 중에서도 관광호텔 및 외국인조사구를 제외한 일반조사구만을 대상으로 되어 있다.

○ 10% 표본조사 대상 추출은 모집단 조사구 자료를 행정구역에 따라 읍면동별로 1차 층화한 후 각 읍면동 내에서 조사구 특성에 따라 다음과 같이 2차 층화한다.

- 제1층 : 아파트조사구
- 제2층 : 보통조사구, 섬조사구
- 제3층 : 기숙시설조사구, 특수사회시설조사구

○ 2차 층화 후 층별로 조사구 번호순으로 자료를 나열하고, 제1층과 제2층에서는 각 읍면동에 해당되는 조사구수 만큼의 조사구를 계통추출방법으로 추출한다. 제3층은 모집단조사구의 약 2.7%에 불과하여 모두 표본조사구가 되었다.

○ 2005 인총 10% 표본 조사 대상은 <표 4-1>에 나타난 바와 같이, 10% 표본의 경우 조사구는 32,241개이고 가구원은 5,042,490명, 가구는 1,591,631가구, 주택은 1,303,668호로 구성되어 있

다.

<표 2.4.1> 2005 인총조사 10% 표본

조사	가구원수	가구수	주택수
전수	47,278,951	15,988,274	13,222,641
10%표본	5,042,490 (10.7%)	1,591,631 (10.0%)	1,303,668 (9.9%)

○ 또한 10% 표본 자료를 인구와 가구주택별 소지역 분석이 가능하도록 읍면동 행정구역코드를 확보하여 정리하였다. 이를 인구/가구별 및 시도별로 정리하면 각각 <표 2.4.2>와 <표 2.4.3>과 같다.

<표 2.4.2> 2005 인총조사 10% 표본 자료

조사	관측치	식별변수	분석변수
인구	1,591,631	14	104
주택가구	5,042,490	15	114

<표 2.4.3> 2005 인총 10% 표본 자료 시도별 집계

시도명	가구원수	가구수	주택수
서울특별시	956,545	307,646	216,704
부산광역시	349,654	111,648	88,244
대구광역시	236,907	76,096	56,250
인천광역시	244,420	79,811	68,124
광주광역시	150,333	44,258	36,076
대전광역시	147,343	44,563	34,494
울산광역시	103,785	30,783	24,338
경기도	1,038,155	310,923	251,651
강원도	182,459	58,684	51,453
충청북도	177,548	55,761	48,923
충청남도	245,616	71,668	65,945
전라북도	229,554	75,146	69,284
전라남도	239,000	84,251	80,413
경상북도	346,452	109,728	98,699
경상남도	339,728	113,118	98,698

제주도	54,991	17,547	14,372
합계	5,042,490	1,591,631	1,303,668

○ 인수 데이터에 대한 소지역 단위 분석을 원활하게 하기 위해 지역코드에 대한 식별을 수행하였다. 인수 데이터의 지역코드는 2005년 인총조사 시점의 지역 기준 코드로서 행자부의 행정구역 코드와는 다른 바, 인수 데이터의 지역코드에 대한 실제 지명 테이블을 입수하여 데이터를 병합(merge)하고, 이를 통해 실제 추정량을 계산하는 과정을 밟았다.

4. 키변수 선정을 위한 다른 기관의 데이터 조사 내용

○ 개인이나 가구의 정보노출에서 키변수는 핵심적인 역할을 한다. 키변수란 정보추적의 대상이 되는 개체의 식별자로써 대표적인 예가 개인인 경우 주민등록번호라고 할 수 있다. 한 개인의 주민등록번호를 파악하는 경우 주민등록번호가 포함된 다른 데이터 세트의 여러 변수에 대한 값을 식별할 수 있게 되어 정보추적이 매우 용이해 진다. 이렇듯 하나의 변수를 통해 개체를 식별하고 이를 통해 다른 정보까지 파악할 수 있는 변수를 키변수라고 할 수 있다.

○ 키변수는 단 하나의 변수가 아니더라도 여러 변수가 결합되어 한 개체를 식별 혹은 노출시킬 수 있다면 이 또한 광의의 개념에서 키변수로 간주할 수 있다. 즉, 특정 개인이 거주하는 동네이름과 그 사람의 이름을 알 수 있고, 같은 동네 내에서 이름이 서로 같을 확률이 극히 드물다면, 두 개의 변수를 결합하여 개체를 식별 가능하다.

○ 특히 최근에 대용량의 데이터를 여러 기관에서 다양한 목적으로 수집하는 경우가 많은바, 특정 기관에서 수집하는 데이터에서 키변수가 될 수 있는 정보는 날로 증가하고 있다. 따라서, 실제 데이터 수집 기관의 데이터를 살펴보고, 이를 통해 키변수로 간주할 수 있는 변수를 살펴보고자 한다.

○ 본 연구에서는 32개 기관(금융기관 11개 및 일반기업 21개)의 데이터를 수집하여, 키변수로 활용될 수 있는 변수를 검토하였다. 대부분의 수집된 데이터가 개인 식별이 가능한 주민번호와 주소, 이름 및 전화번호 등을 수집하고 있었으며, MD에도 포함되어 있는 직업 및 가구관련 정보 및 주택관련 정보를 포함하고 있었다.

○ 이 중 하나의 금융기관에 대한 사례를 기술하고자 한다. 아래의 <표 4-4>는 한 금융기관에서 수집된 개인 및 가구정보 데이터세트의 사례이다. 그곳에 보유중인 전체 고객 데이터 중 1,000개의 표본을 단순임의추출하여, 각 변수에 대한 키변수 가능성 여부를 결측치 빈도와 결부하여 살펴

보았다. 실제 결측이 많이 발생하는 변수는 해당 기관에서 보유하고 있더라도 활용의 가능성이 낮기 때문에 정보노출이 어렵다는 이유에서이다.

<표 2.4.4> S캐피탈의 고객신용정보

정보분류	변수명	설명	결측치 빈도(%)
개인신상 정보	주민번호	거래자 주민번호	0.0
	성명	거래자 성명	0.0
	실거주지주소	현재 실거주지	0.0
	주민등록지주소	주민등록상 기재주소	0.0
	직장주소	현재 근무 직장주소	45.5
소득정보	근로소득	거래자 근로소득	36.6
	배우자소득	배우자 근로소득	81.6
	임대소득	임대시 임대소득	81.6
	이자소득	이자수입	81.6
	연금소득	연금수입	81.6
	기타소득	비경상소득	81.6
	연간가구소득	가구 경상소득	24.1
	재산세	전년 재산세 증빙자료	68.8
	재산세지역구분코드	전년 재산세 납부지역	87.2
직업정보	직업군코드	현재 근무 직장의 직업군	0.0
	근속년수	현재 직장에서 근속년	0.0
	직위코드	해당 직장의 직위	57.8
	직무코드	해당 직장의 직무	50.1
가구정보	맞벌이여부	배우자 맞벌이 여부	66.8
	가족수	가구내 구성 가족수	14.0
	결혼유무	결혼여부	14.0
	기취학 아동수	가구내 취학 아동수	100.0
	세대주와의 관계	가구주와 관계	0.0
주택정보	주택임대형태구분코드	주택 임대여부	16.4
	주거형태	거주 주택형태	0.0
	주거년수	현 주택에서 주거년수	0.0
	평수	거주 실평수	87.6
	보유주택 시가	해당연월 거래시세	92.7
	거주사항코드	거주 여부	0.2
거래정보	고객구분코드	내부고객 구분 코드	75.4
	총연체횟수	타 금융기관 포함 연체횟수	0.0
	총거래금액	총 금융거래 금액	0.0
	최초거래일	해당기관 최초거래일	0.6
	최종거래일	해당기관 최종거래일	0.0
	전산입력일자	고객 등록일	4.4
	카드소지코드	신용카드 총 거래수	58.0
	론 서류코드	대출상품 서류코드	0.0
	증빙서류코드	증빙서류 종류코드	18.3
	현금서비스한도	해당기관 현금서비스한도	13.6
	통장잔고	해당기관 거래통장잔고	0.0

○ 앞의 예에서 알 수 있듯이 대부분의 금융기관에서는 거래자의 신상을 파악할 수 있는 주민번호는 물론, 거주 및 행정상의 주거지 주소, 직업 및 가구 정보를 수집하고 있었다. 특히 해당 거래자의 직업군, 근속년수 등의 직업정보, 주거형태와 주거년수 등의 거주 정보 및 세대주 관계와 같은 민감한 정보를 빠짐없이 수집하고 있었다.

○ 이상에서 살펴본 내용에 의하면 주소, 직업 및 주거관련 정보가 센서스에서도 동일하게 수집되기 때문에 동일 변수를 MD로 제공시 신중을 기해야 한다는 것을 보여준다. 또한, 그러한 변수를 키변수로 이용하여 정보노출이 어느 정도 가능한 지에 대한 연구가 필요하다는 것을 역설하기도 한다.

5. MD자료에서 키변수 설정

○ 인총 10%자료에서 정보보호의 대상이 되는, 키변수로 상정가능한 변수는 아래와 같다.

- 행정구역 코드에 관한 사항 : 시·도, 시/군/구, 읍/면/동
- 가구원에 관한 사항 : 가구주와의 관계, 교육정도, 아동보육, 5년 전 거주지, 이용교통수단, 통근/통학소요시간, 경제활동상태, 종사상 지위, 혼인연령, 고령자 생활비
- 가구에 관한 사항 : 가구구분, 거주기간, 주차시설, 소유형태, 주인가구 및 세대구성, 총 방수, 자동차 보유대수
- 주택에 관한 사항 : 거처종류, 연건평, 건축년도

○ 통계청의 인총 데이터와 외부기관의 데이터에 공통으로 존재하여 개인정보 노출의 키변수가 될 가능성이 높은 변수 추출한 후, 정보노출이 어느 정도 가능한 가를 우선 파악하여야 한다.

○ 이를 위해 통계청 외의 타 기관에서 보유가능성이 높은 항목인 인구관련 항목 11개(거주시도, 성별, 나이, 가구주와의 관계, 교육정도, 종교유무, 5년 전 거주지 1과 2, 종사상 지위, 혼인상태)와 가구관련 항목 11개(가구구분, 거주기간, 방수, 차량보유, 가구소유형태, 주인가구 유무, 거처종류, 연건평, 건축년도, 가구주의 만나이, 가구주의 혼인상태) 총 11개 항목을 키변수로 선택하였다. 이를 통해 정보노출의 정도에 대한 기준을 제시하였다.

○ 키변수의 선정은 현재 일반에 공개하고 있는 5% MD 자료에 존재하는 변수를 대상으로 선정하였다. 이는 본 연구와 현재 제공 MD자료와 동일한 기준에서 키변수를 선택한 후, 두 경우에 대한 노출위험을 비교연구하기 위함이다.

○ 키변수 선정에서 주요 이슈 중에 하나는 지역코드 관련 변수이다. 지역관련 변수에서 특히 기초지자체 단위의 제공은 각 지자체의 센서스 활용도를 높힌다는 점에서 긍정적이라고 사료된다. 그러나, 지역코드 내의 시/군/구는 260여개의 범주값으로 구성되어 이 변수가 키변수로 포함되는 경우, 범

주수가 적은 다른 변수에 비해 정보노출의 위험을 획기적으로 높인다는 문제가 존재한다.

○ 즉, 범주형 변수로 이루어지진 키변수에서 정보 노출 건수는 각 변수에 대한 범주의 곱으로 생성되는 셀에서 유일하게 존재하는 개체수의 합이다. 따라서, 시/군/구와 같이 하나의 범주형 변수가 많은 범주를 가지게 되면 다른 변수의 범주수와 관계없이 정보노출의 위험은 커지기 때문이다. 예를 들어 키변수가 3개이며, 변수 모두가 범주형이라고 가정하고 각 키변수의 범주값의 개수가 5개씩이라고 가정하더라도 키변수의 조합에 의한 셀은 총 5^3=125개가 된다. 이 경우 각 셀에서 유일한 관측치는 정보 추적 혹은 노출이 가능한 유일성의 개체가 되면 이러한 관측치의 총합이 증가하면 증가할수록 해당 데이터의 노출위험은 증가하게 된다.

○ 본 연구에서 상정한 키변수는 인구와 가구에서 모두 11개의 변수이며 인구 항목만을 놓고 살펴보더라도 원데이터에서 발생하는 셀의 개수는 총 84,367,360개이다. 물론 나이 항목이 1세 단위의 연속형 변수라는 속성을 띄어 많은 값(108개)를 가지기 때문이기는 하나, 이 변수를 제외하고 셀 개수를 계산하더라도 78만 1천여개의 셀이 존재한다. 따라서, 시/군/구 단위 데이터를 제공하기 위해 260개의 범주가 추가되는 경우 총 셀의 개수는 기하급수적으로 발생하게 된다. 이런 이유로 본 연구에서는 시/군/구를 식별할 있는 변수는 연구에서 제외되었다.

6. 노출위험의 계산

○ 노출 위험은 기존의 연구 결과, 예컨대 Kim, J. (1986) 등에서 제시된 것과 동일하게 정의되었다. 먼저 유일성과 관련된 관찰치와 파일을 아래와 같이 정의하자.

- A : 마이크로 데이터 속한 임의의 사람
- S_1 : 통계작성기관의 마이크로데이터로 구성된 파일
- S_2 : 외부인(intruder)에 의해 구성된 파일
- U_P : 모집단의 유일성 집단
- U_S : 표본의 유일성 집단

○ 만약 금융기관이나 이웃 주민 등과 같은 외부인들이 스스로가 작성한 파일인 S_2 내에 정보를 추적하고자 관심을 두고 있는 어떤 사람 A가 포함되어 있다는 것을 모르고 있다면, 특정 사람의 노출위험 $DR(A)$은 다음과 같이 정의할 수 있다.

$$DR(A) = \Pr[(A \in S_1) \cap (A \in S_2) \cap (A \in U_P)]$$

그러나, 외부인이 자신의 파일에 A가 포함되어 있다는 건을 이미 파악한 후 마이크로 데이터서 그와 관련된 내용을 추적하고자 한다면 $\Pr(A \in S_2) = 1$ 이 될 것이며, 이때 노출위험은 다음과 아래

와 같이 된다.

$$DR(A) = \Pr(A \in S_1) \cdot \Pr(A \in U_P)$$

○ 현재 통계청에서는 1, 2, 5% 마이크로 데이터를 센서스 표본 조사에 대해 적용하고 있으나, 본 연구에서는 추출비율 별 자료의 제공 타당성 파악을 위해 센서스 표본 조사에 대한 36개 내외의 표본추출을 시도하고 이를 통해 표본 간의 유일성 파악 및 정보보호를 위한 자료 비교를 수행하고자 하였다.

○ 아래의 <표 2.4.6>은 <표 2.4.5>의 인총 10% 데이터를 추출틀로 하여 표본크기와 추출기법별로 2번 반복하여 뽑은 시도별 표본의 크기이다. 즉, 전국 대비 1%~9% 내에서 %구간별로 2번씩 집락과 계통추출에 의해 추출된 총 36개의 표본의 표본크기를 정리한 것이다.

<표 2.4.5> 추출틀 내의 가구 및 주택 현황

시도	가구원수	가구수	주택수
서울특별시	956,545	307,646	216,704
부산광역시	349,654	111,648	88,244
대구광역시	236,907	76,096	56,250
인천광역시	244,420	79,811	68,124
광주광역시	150,333	44,258	36,076
대전광역시	147,343	44,563	34,494
울산광역시	103,785	30,783	24,338
경기도	1,038,155	310,923	251,651
강원도	182,459	58,684	51,453
충청북도	177,548	55,761	48,923
충청남도	245,616	71,668	65,945
전라북도	229,554	75,146	69,284
전라남도	239,000	84,251	80,413
경상북도	346,452	109,728	98,699
경상남도	339,728	113,118	98,698
제주도	54,991	17,547	14,372
합계	5,042,490	1,591,631	1,303,668

<표 2.4.6> 추출표본의 시도별 크기

표본비율	집락		계통	
	표본1	표본2	표본1	표본2
1%	445,554	445,382	444,797	446,745
2%	890,692	890,767	891,174	891,113
3%	1,335,799	1,336,570	1,335,380	1,337,664
4%	1,781,084	1,781,153	1,782,287	1,781,502
5%	2,228,092	2,228,265	2,229,053	2,229,053
6%	2,675,073	2,673,498	2,672,676	2,674,025
7%	3,119,081	3,117,307	3,119,951	3,118,231
8%	3,564,458	3,565,289	3,563,789	3,564,414
9%	4,010,450	4,010,143	4,010,534	4,455,527

○ 추출의 단위는 앞서 설명한 바와 같이 표본조사구 내의 가구를 추출 대상으로 하였다. 이를 통해 해당 가구 내의 모든 인구에 대한 값을 산출하였다.

○ 추출 데이터 내에 노출위험이 존재하는 대상 U_s를 계산한 결과가 <표 2.4.7>과 <표 2.4.8>에 제시되어 있다. 두 표는 10% 각각 인구와 가구/주택 부문의 원데이터에서 50%를 추출한 전국 대비 5% 집락추출의 결과로써, 노출제한기법 적용 전의 유일건수, 구성비가 전국 및 시도별로 제시되었다.

○ 인구부문에 있어 앞에서 상정한 키변수 조합에 대한 유일성 건수는 시도별로는 대략 2.35~4.97%였으며, 전국 대비 3.12%가 노출위험에 있는 것으로 나타났다. 또한 가구/주택부문에서 유일성 건수는 인구부문에 비해 높은 26.06~58.53%로 나타났다.

<표 2.4.7> 인구부문, 노출기법 적용 전의 유일성, 집락추출 결과

시도	인구수	노출기법 적용 전	
		유일건수 (Us)	구성비 (%)
전국	2,228,092	69,468	3.12
서울특별시	439,283	11,311	2.57
부산광역시	159,892	3,764	2.35
대구광역시	110,473	3,148	2.85

인천광역시	117,975	4,582	3.88
광주광역시	65,098	2,389	3.67
대전광역시	64,134	3,189	4.97
울산광역시	45,559	1,931	4.24
경기도	467,283	12,948	2.77
강원도	76,241	3,153	4.14
충청북도	74,716	2,985	4.00
충청남도	94,296	4,392	4.66
전라북도	97,980	2,556	2.61
전라남도	104,763	2,974	2.84
경상북도	136,938	4,393	3.21
경상남도	149,192	4,656	3.12
제주도	24,269	1,097	4.52

<표 2.4.8> 가구부문, 노출기법 적용 전의 유일성, 집락추출 결과

시도	가구수	노출기법 적용 전	
		유일건수 (Us)	구성비 (%)
전국	791,341	296,472	37.46
서울특별시	152,667	61,700	40.41
부산광역시	55,578	24,663	44.38
대구광역시	37,819	18,222	48.18
인천광역시	39,980	15,595	39.01
광주광역시	21,976	8,923	40.60
대전광역시	22,133	10,456	47.24
울산광역시	15,261	7,064	46.29
경기도	154,820	50,465	32.60
강원도	29,125	12,104	41.56
충청북도	27,892	10,885	39.03
충청남도	35,468	11,438	32.25
전라북도	37,139	11,649	31.37
전라남도	41,923	10,924	26.06
경상북도	54,584	17,223	31.55
경상남도	56,428	20,158	35.72
제주도	8,548	5,003	58.53

7. 노출제한을 위한 부분 카이제곱 통계량 그룹화 방법 적용

○ 앞에서 언급한 기존 연구방법들은 임의의 데이터 값에 부분적으로 값을 가감하거나(Coding Approach, Rounding), 구간 그룹화 같은 방법처럼 연속형 값을 중위수나 평균으로 대체(replace)하는 방법이 주로 연구되어 왔다.

<표 2.4.9> 그룹화 이해를 위한 예제

(a) 그룹화 이전 2×2 분할표

x	y				
	가	나	다	라	합계
a	46	2	1	1	50
b	2	46	1	1	50
c	1	1	47	1	50
d	1	1	1	47	50
합계	50	50	50	50	200

(b) y변수 기준의 그룹화

x	y			
	가	나	다/라	합계
a	46	2	2	50
b	2	46	2	50
c	1	1	48	50
d	1	1	48	50
합계	50	50	100	200

(c) y변수 그룹핑 후 x변수 그룹화

x	y			
	가	나	다/라	합계
a	46	2	2	50
b	2	46	2	50
c/d	2	2	96	100
합계	50	50	100	200

(d) 단변량 기준으로 그룹화가 필요한 경우

x	y			
	가	나	다/라	합계
a	48	0	0	48
b	0	1	0	1
c	0	0	1	1
합계	48	1	1	50

○ 그러나, 센서스 데이터처럼 키변수로 상정되는 변수의 속성들이 대부분 범주형 변수이며, 고려해야 하는 키변수가 기존의 연구보다 많고, 키변수조합에 의해 생성되는 셀의 개수가 많은 경우에는 기존의 연구방법을 직접 적용하기 많은 무리가 따른다고 할 수 있다. 따라서, 센서스와 같이 많은 후보 키변수를 가진 경우에 대해서는 우선 합리적인 기준에 의해 셀의 개수를 줄여 특정 개체의 노출위험을 줄이는 것이 우선적이라 사료된다.

○ 셀 개수를 줄이는 개념을 설명하기 위해 먼저, <표 2.4.9>에서 제시된 여러 형태의 2×2 분할표 예제를 살펴보자. 위의 예제는 각각 4개의 범주로 이루어진 변수 x(a, b, c, d)와 변수 y(가, 나, 다, 라)에 대해 50개의 관찰치로 구성된 자료의 2×2분할표 결과이다. 원천 데이터에 대한 분할표 결과가 만약 (a)와 같다면, 총 10개의 관측치가 데이터 세트에서 유일하게 되어 50개의 관측치 중 전체 20%가 노출위험에 빠지게 된다. 이를 해결하기 위해 그룹화를 통해 노출위험을 줄이고자 변수 y의 범주 '다'와 '라'를 합치는 것은 합리적인 선택이라고 할 수 있다. 이런 후 변수 x의 범주 'c'와 'd'를 합쳐 (c)의 분할표와 같은 결과를 얻는다면, 50개의 관측치 중 노출이 되는 개체는 사라지게 될 것이다.

○ 이러한 논리는 2×2 분할표의 일부분에 대한 카이제곱 통계량으로도 설명이 가능하다고 할 수 있다. 즉, (a)에서 y변수의 특정 범주를 결합하기 위해서는 변수 x와 변수 y의 (가, 나)로 이루어진 분할표와 (다, 라)로 이루어진 분할표 중 카이제곱 통계량 값이 더 큰 분할표의 결과에 의해 범주를 통합하면 될 것이다.

○ 다만 위의 예에서 사전에 고려하여야 할 점이 존재한다. (a)~(c)의 분할표는 원래 데이터의 x, y의 분포가 균일하다는 점을 전제로 한다. 만약 분할표 (d)와 같이 두 변수가 주변 분포(marginal distribution)에서부터 한쪽으로 치우친 결과를 보인다면, 어떠한 방법을 적용하더라도 노출을 막을 수는 없다고 할 수 있다.

○ 두 변수에 대한 분할표 내에 부분적인 분할표를 만들었을 때 이들이 서로 다른 패턴을 보이는 경우는 쉽게 찾을 수 있다. <표 2.4.10>에서 행 변수인 '직업코드'의 '기능원 및 관련 기능 종사자'와 '장치기계조작 및 조립종사자' 및 '단순노무종사자'에 대해 열 변수 '교육정도1'의 '초등학교'이전의 빈도패턴은 '고등학교' 이상의 학력에 대한 그것과 큰 차이를 보이고 있음을 알 수 있다.

○ 앞에서 언급한 내용을 통해 하나의 알고리즘으로 정리하고자 한다. 이러한 알고리즘은 먼저 각 키변수의 주변분포를 균일한 형태로 병합(merge)한다. 각 키변수를 균일분포에 가깝게 병합하는 경우 많은 셀에서 발생하는 유일건수를 줄이는데 도움이 된다. 그런 후, 변환된 키변수들의 임의의 두 개 변수에 대해 <표 4-10>의 분할표에서 볼 수 있는 '부분적인 빈도 패턴의 차이'를 감지할 수 있는 통계적 모형을 세운다. 이러한 통계적 모형으로 고려될 수 있는 것이 '로그선형모형

(log-linear model)'이다. 로그선형모형의 카이제곱 통계량이 유의한 두 변수의 짝에서 부분적으로 발생하는 모든 분할표에 대한 '두 변수의 독립성 검정'을 실시하고 독립성 검정의 결과가 '유의하게' 줄어드는 상태에서 범주를 병합을 결정한다.

<표 2.4.10> 표본 10%자료에 대한 직업코드와 교육정도의 분할표

교육정도1 (빈도, %, 행%)	직업코드												
	의회 의원 고위 임직원/ 관리자	전문가	기술공 및 준 전문가	사무 종사자	서비스 종사자	판매 종사자	농업, 임업, 어업 숙련 종사자	기능원 및 관련 기능 종사자	장치기계 조작 및 조립 종사자	단순 노무 종사자	군인	분류 불능	합계
안 받았음	54	97	107	194	1,121	1,281	30,115	**707**	**518**	**3,093**	2	242,963	280,252
	0	0	0	0.01	0.05	0.06	1.35	0.03	0.02	0.14	0	10.9	12.57
	0.02	0.03	0.04	0.07	0.4	0.46	10.75	0.25	0.18	1.1	0	86.69	
초등 학교	497	211	584	1,440	8,147	6,430	65,662	**6,921**	**6,420**	**13,403**	3	303,303	413,021
	0.02	0.01	0.03	0.06	0.37	0.29	2.95	0.31	0.29	0.6	0	13.61	18.53
	0.12	0.05	0.14	0.35	1.97	1.56	15.9	1.68	1.55	3.25	0	73.44	
중학교	1,174	353	1,461	3,661	13,394	9,377	24,806	13,510	14,878	14,550	34	175,718	272,916
	0.05	0.02	0.07	0.16	0.6	0.42	1.11	0.61	0.67	0.65	0	7.88	12.24
	0.43	0.13	0.54	1.34	4.91	3.44	9.09	4.95	5.45	5.33	0.01	64.39	
고등 학교	9,264	5,316	19,400	50,178	44,335	50,546	24,838	**48,079**	**61,615**	**29,895**	1,209	312,429	657,104
	0.42	0.24	0.87	2.25	1.99	2.27	1.11	2.16	2.76	1.34	0.05	14.02	29.48
	1.41	0.81	2.95	7.64	6.75	7.69	3.78	7.32	9.38	4.55	0.18	47.55	
대학 (4년제 미만)	3,705	13,553	16,518	28,828	9,294	12,586	2,787	**10,431**	**9,812**	**3,963**	865	78,513	190,855
	0.17	0.61	0.74	1.29	0.42	0.56	0.13	0.47	0.44	0.18	0.04	3.52	8.56
	1.94	7.1	8.65	15.1	4.87	6.59	1.46	5.47	5.14	2.08	0.45	41.14	
대학교 (4년제 이상)	14,058	44,964	35,926	54,541	11,479	18,020	3,044	**9,617**	**7,785**	**4,181**	1,107	157,287	362,009
	0.63	2.02	1.61	2.45	0.51	0.81	0.14	0.43	0.35	0.19	0.05	7.06	16.24
	3.88	12.42	9.92	15.07	3.17	4.98	0.84	2.66	2.15	1.15	0.31	43.45	
대학원 (석사 과정)	2,846	16,600	4,417	5,119	562	918	215	**452**	**317**	**164**	568	10,777	42,955
	0.13	0.74	0.2	0.23	0.03	0.04	0.01	0.02	0.01	0.01	0.03	0.48	1.93
	6.63	38.65	10.28	11.92	1.31	2.14	0.5	1.05	0.74	0.38	1.32	25.09	
대학원 (박사 과정)	505	6,867	414	417	43	60	23	**22**	**28**	**8**	40	1,514	9,941
	0.02	0.31	0.02	0.02	0	0	0	0	0	0	0	0.07	0.45
	5.08	69.08	4.16	4.19	0.43	0.6	0.23	0.22	0.28	0.08	0.4	15.23	
합계	32,103	87,961	78,827	144,378	88,375	99,218	151,490	89,739	101,373	69,257	3,828	1,282,504	2,229,053
	1.44	3.95	3.54	6.48	3.96	4.45	6.8	4.03	4.55	3.11	0.17	57.54	100

○ 범주의 병합은 변수의 속성에 의해 두 방향으로 고려되어야 했다. 즉, 명목형(nominal)변수인 경우와 순서형(ordinal)변수가 그것이다. 순서형 변수, 예컨대 '교육정도 1'과 같이 범주값이 '무학, 초등학교, 중학교,..,대학원졸(박사)'로 구성된 경우에는 서로 인접한 범주 간의 결합 만을 고려해야 하고, '직업코드'와 같이 범주간의 순서를 고려할 수 없는 경우에는 범주의 총 개수에 대한 임의의 2개 짝에 대한 모든 병합을 고려하여야 한다. 또한 나이, 주택 내 방수, 자동차 대수와 같이 연속형으로 간주하는 것이 합리적인 변수들도 카이제곱 검정을 위해 순서형 변수로 간주하였다. 본 절에 언급된 내용을 정리하여 하나의 알고리즘을 제시하면 다음과 같다.

<카이제곱 통계량을 이용한 부분 그룹화 알고리즘>

- step 1. 각 키변수에 대한 동일성 검정(homogeneity test)을 수행하여 카이제곱 검정 통계량의 값이 줄어드는 방향으로 범주를 통합한다.

- step 2. step 1을 거친 키변수 집단에 대해 로그선형분석을 실시한다. 만약 키변수가 n개라면 총 nC2개의 로그선형분석을 수행하게 된다. 여기서 로그선형 모형은 아래와 같다.

  • 임의의 두 변수 x, y가 각각 $i = 1,\cdots,a$, $j = 1,\cdots,b$ 개의 범주로 이루어진 범주형 변수라고 하자.

  $$\log\ m_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij}$$

  여기서, m_{ij}는 x의 i번째, y의 j번째 범주에 대한 관측빈도. 즉, 키변수 중 임의의 2개 변수에 대한 교호작용(interaction effect)λ_{ij}을 고려한 로그선형모형이다.

- step 3. step 2.의 각 모델 중 교호작용이 유의한 모델에 대해 아래와 같은 규칙에 의해 범주를 부분 통합한다.

  • 명목형 변수의 범주값의 개수가 j개인 경우 총 jC2개의 2×2분할표에 대한 카이제곱 검정을 수행하여 가장 낮은 값을 가진 분할표의 범주 통합의 결과를 선택한다.

  • 순서형 변수의 범주값의 개수가 j개인 경우 총 $j-1$개의 2×2분할표에 대한 카이제곱 검정을 수행하여 가장 낮은 값을 가진 분할표의 범주 통합의 결과를 선택한다.

- step 4. step 1.~step 3.의 과정을 반복하여 '충분히 낮은 노출비율'을 얻을 때까지 반복한다.

○ 위의 알고리즘을 적용하여 <표 2.4.11>과 <표 2.4.12> 같은 결과를 도출하였다. '충분히 낮은 노출비율'에 대한 기준은 이후에 언급할 현 제공 5% MD에서 계산되는 유일성 건수 비율과 비교하여 조정하였다.

○ 또한, 본 계산시 특정 키변수에 결측치가 존재하는 경우에는 유일성 계산에서 제외하였다. 이는 결측치가 존재하는 특정 범주가 다른 키변수의 범주와 결합되어 유일한 관측치가 되더라도 정보의 노출이 일어날 수 없기 때문이다.

<표 2.4.11> 노출제한을 위해 그룹화 조정 결과 : 인구 부문

변수	원 데이터	그룹화 후
성별	① 남자	① 남자
	② 여자	② 여자
만 나이	① 000세	① 10세미만
		② 20세미만
		③ 30세미만
		④ 40세미만
		⑤ 50세미만
		⑥ 60세미만
		⑦ 70세미만
		⑨ 70세이상
가구주와의 관계	① 가구주	① 가구주
	② 가구주의 배우자	② 가구주의 배우자
	③ 자녀	③ 자녀 및 자녀의 배우자
	④ 자녀의 배우자	
	⑤ 가구주의 부모	④ 기타
	⑥ 배우자의 부모	
	⑦ 조부모	
	⑧ 손자녀, 그 배우자	
	⑨ 증손자녀, 그 배우자	
	⑩ 형제자매, 그 배우자	
	⑪ 형제자매의 자녀, 그 배우자	
	⑫ 부모의 형제자매, 그 배우자	
	⑬ 기타 친인척	
	⑭ 기타 동거인	
교육정도1	① 안 받았음(미취학 포함)	① 안 받았음(미취학 포함)
	② 초등학교	② 중학교 이하
	③ 중학교	
	④ 고등학교	③ 고등학교
	⑤ 대학(4년제 미만)	④ 대학교 이상
	⑥ 대학교(4년제 이상)	
	⑦ 대학원(석사 과정)	
	⑧ 대학원(박사 과정)	

종교1	① 있다	① 있다
	② 없다	② 없다
5년 전 거주지	② 현재 살고 있는 집	① 현재 살고 있는 집
	③ 같은 시군구 내 다른 집	② 같은 시군구
	① 태어나지 않았음	③ 다른 시군구/기타 (태어나지 않았음+북한 또는 외국)
	④ 다른 시군구	
	⑤ 기타	
5년 전 거주지 시도 코드	① 서울	① 서울
	② 부산	② 부산, 경남, 울산
	⑦ 울산	
	⑮ 경남	
	③ 대구	③ 대구, 경북
	⑭ 경북	
	⑤ 광주	④ 광주, 전남, 전북
	⑬ 전남	
	⑫ 전북	
	⑥ 대전	⑤ 대전, 충남, 충북
	⑩ 충북	
	⑪ 충남	
	⑧ 경기	⑥ 인천, 경기
	④ 인천	
	⑨ 강원	⑦ 제주, 강원
	⑯ 제주	
	⑰ 아시아주	⑧ 기타 외국
	⑱ 아메리카주	
	⑲ 유럽주	
	⑳ 오세아니아주	
	㉑ 아프리카주	
종사상 지위	① 임금 근로자	① 임근근로자 + 무급가족종사자
	④ 무급 가족 종사자	
	② 고용원이 없는 자영자	② 사업주
	③ 고용원을 둔 사업주	
혼인 상태	② 배우자 있음	② 배우자 있음
	① 미혼	① 배우자 없음
	③ 사별	
	④ 이혼	

<표 2.4.12> 노출제한을 위해 그룹화 조정 결과 : 가구주택 부문

변수	원 데이터	그룹화 후
가구구분	① 가족으로 이루어진 가구	① 가족가구
	③ 1인 가구	② 1인 가구
	② 가족과 가족이외의 함께 사는 가구	③ 기타(가족과 가족이외의 함께 사는 가구 + 가족이 아닌 남남끼리 5인 이하의 가구)
	④ 가족이 아닌 남남끼리 5인 이하의 가구	
거주기간	① 1년 미만	① 1년 미만
	② 1년~2년 미만	② 1년~5년 미만
	③ 2년~3년 미만	
	④ 3년~5년 미만	
	⑤ 5년~10년 미만	③ 5년 이상
	⑥ 10년~15년 미만	
	⑦ 15년~20년 미만	
	⑧ 20년~25년 미만	
	⑨ 25년 이상	
방수	① OO개	① 2개 이하
		② 3개
		③ 4개 이상
자동차 대수	① 0	① 1대
	② 1	
	③ 2	② 2대 이상
	④ 3	
	⑤ 4	
	⑥ 5	
	⑦ 6	
가구소유 형태	① 자기 집	① 자기 집
	② 전제(월세 없음)	② 전세, 월세 (보증금 있는 월세+ 보증금 없는 월세 + 사글세)
	③ 보증금 있는 월세	
	④ 보증금 없는 월세	
	⑤ 사글세	
	⑥ 무산(관사,사택,친척집 등)	
주인가구	① 주인가구	① 주인가구
	② 대표가구	② 주인 아닌 가구 (대표가구+기타 세 들어 사는 가구)
	③ 기타 세들어 살고 있는 가구	
주택소유 여부	① 다른 곳에 주택 소유	① 다른 곳에 주택 소유
	② 다른 곳에 주택 미소유	② 다른 곳에 주택 미소유
거처종류1	① 단독주택	① 단독주택

	② 아파트	② 공동주택 (아파트+연립 + 다세대 주택)
	③ 연립주택	
	④ 다세대주택	
	⑤ 비주거용 건물 내 주택	③ 기타 주거 형태
	⑥ 오피스텔	
	⑦ 숙박업소(호텔, 여관)의 객실	
	⑧ 기숙사 및 특수 사회시설	
	⑨ 판잣집, 비닐하우스, 움막	
	⑩ 기타	
연건평	① OOO평	① 25평 미만
		② 35평 미만
		③ 45평 미만
		④ 45평 이상
건축년도	① 2005년	① 2000년~2005년
	② 2004년	
	③ 2003년	
	④ 2002년	
	⑤ 2001년	
	⑥ 2000년	
	⑦ 1995년~1999년	② 1990년~1999년
	⑧ 1990년~1994년	
	⑨ 1985년~1989년	③ 1989년 이전
	⑩ 1980년~1984년	
	⑪ 1970년~1979년	
	⑫ 1960년~1969년	
	⑬ 1959년 이전	

8. 노출제한 기법 적용 후 적용 전후의 결과 비교

○ 앞에서 제시된 변수별 노출제한 기법을 통해 결과를 산출하였다. <표 2.4.13>은 인총10% 표본에 대해 가구의 집락추출을 통해 얻어진 5% 데이터의 결과로서, 앞에서 제안된 노출제한 기법이 어느 정도 효과가 있는 지를 보여주고 있다. <표 2.4.7>에 볼 수 있는 노출제한 기법 적용 전 정보노출이 가능한 유일성 개체가 69,468개인 반면, 기법 적용 후에는 9,561개로 감소하였으며, 이는 전국을 기준으로 1/13.76 정도 줄어들은 것으로 나타났다. 특히 개체수(인구)가 많은 서울/경기에서 각각 1/8.18, 1/7.17 정도 감소함으로서 다른 시도에 비해 높은 감소를 보였다.

○ <표 2.4.14>에서는 가구/주택부문에 대한 노출제한 기법 적용 후의 결과를 볼 수 있다. 기법 적용 전에 시도별로 26.06~58.53%, 전국적으로는 37.46%의 노출을 보인 반면, 적용 후에는 1.54%만이 유일한 결과를 얻었다.

<표 2.4.13> 노출기법 적용 전후의 결과 비교: 인구부문, 집락추출, 5%

시도	인구수	노출기법 적용 후		노출제한 전후대비 (후/전)
		유일건수(Us)	구성비(%)	
전국	2,228,092	9,561	0.43	13.76
서울특별시	439,283	925	0.21	8.18
부산광역시	159,892	540	0.34	14.35
대구광역시	110,473	448	0.41	14.23
인천광역시	117,975	578	0.49	12.61
광주광역시	65,098	369	0.57	15.45
대전광역시	64,134	500	0.78	15.68
울산광역시	45,559	377	0.83	19.52
경기도	467,283	929	0.20	7.17
강원도	76,241	688	0.90	21.82
충청북도	74,716	605	0.81	20.27
충청남도	94,296	681	0.72	15.51
전라북도	97,980	600	0.61	23.47
전라남도	104,763	614	0.59	20.65
경상북도	136,938	684	0.50	15.57
경상남도	149,192	611	0.41	13.12
제주도	24,269	412	1.70	37.56

<표 2.4.14> 노출기법 적용 전후의 결과 비교: 가구부문, 집락추출, 5%

시도	가구수	노출기법 적용 후		노출제한 전후대비 (후/전)
		유일건수 (Us)	구성비 (%)	
전국	791,341	12,198	1.54	4.1
서울특별시	152,667	989	0.65	1.6
부산광역시	55,578	772	1.39	3.1
대구광역시	37,819	685	1.81	3.8
인천광역시	39,980	791	1.98	5.1
광주광역시	21,976	554	2.52	6.2
대전광역시	22,133	687	3.1	6.6
울산광역시	15,261	568	3.72	8.0
경기도	154,820	1,033	0.67	2.0
강원도	29,125	842	2.89	7.0
충청북도	27,892	713	2.56	6.6
충청남도	35,468	786	2.22	6.9
전라북도	37,139	770	2.07	6.6
전라남도	41,923	766	1.83	7.0
경상북도	54,584	870	1.59	5.1
경상남도	56,428	791	1.4	3.9
제주도	8,548	581	6.8	11.6

9. 노출 기법 적용 후 현 제공 MD와 결과 비교

○ 앞에서는 카이제곱 및 로그선형 모형을 기본으로 한 그룹화 기법에 의해 노출위험이 어느 정도 감소되었는지를 제시하였다. 본 절에서는 현재 통계청에서 제공되는 마이크로 데이터와 비교함으로써, 제시된 알고리즘이 얼마나 효과가 있는지를 밝히고자 한다.

○ 현재 통계청에서는 전국 단위 1, 2, 5%의 MD를 제공하고 있으며, 본 연구와 유사한 그룹화 방법을 적용하고 있다. 이러한 기존 제공 방법과 본 연구 방법을 비교하기 위해 다음과 같은 조건을 동일하게 조정하였다. 우선 비교를 위한 표본크기는 전국 단위 5%(인총 표본 10% 데이터의 50%)로, 표본추출 방법은 층화 후 집락추출(추출 단위는 가구)로 하는 비교자료를 만든 뒤, 서로 그룹화 방법만을 달리하여 두 방법의 유일건수를 계산하였다. 현 제공 5% MD에 대한 그룹화 기준은 'http://mdss. kostat.go.kr/'의 자료 제공 범위에서 밝힌 기준에 의거하였다.

○ 인구 및 가구주택부분에 대한 각각의 결과가 <표 2.4.15>와 <2.4.16>에 제시되었다. 인구는 약 1/2.8(=0.43/1.20), 가구주택은 약1/3.4(5.24/ 1.54)정도 감소한 것으로 나타났다.

<표 2.4.15> MD 5%와 비교, 인구부문

시도	인구수	현 제공기준		변수 추가 전	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,228,092	26,678	1.20	9,561	0.43
서울특별시	439,283	3,206	0.73	925	0.21
부산광역시	159,892	1,517	0.95	540	0.34
대구광역시	110,473	1,234	1.12	448	0.41
인천광역시	117,975	1,767	1.50	578	0.49
광주광역시	65,098	984	1.51	369	0.57
대전광역시	64,134	1,275	1.99	500	0.78
울산광역시	45,559	900	1.98	377	0.83
경기도	467,283	3,570	0.76	929	0.20
강원도	76,241	1,499	1.97	688	0.90
충청북도	74,716	1,488	1.99	605	0.81
충청남도	94,296	1,910	2.03	681	0.72
전라북도	97,980	1,371	1.40	600	0.61
전라남도	104,763	1,479	1.41	614	0.59
경상북도	136,938	1,934	1.41	684	0.50
경상남도	149,192	1,880	1.26	611	0.41
제주도	24,269	664	2.74	412	1.70

<표 2.4.16> MD 5%와 비교, 가구/주택부문

시도	가구수	현 제공기준		변수 추가 전	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	791,341	41,491	5.24	12,198	1.54
서울특별시	152,667	4,530	2.97	989	0.65
부산광역시	55,578	3,204	5.76	772	1.39
대구광역시	37,819	2,404	6.36	685	1.81
인천광역시	39,980	2,517	6.30	791	1.98
광주광역시	21,976	1,779	8.10	554	2.52
대전광역시	22,133	1,949	8.81	687	3.10
울산광역시	15,261	1,599	10.48	568	3.72
경기도	154,820	4,056	2.62	1,033	0.67
강원도	29,125	2,558	8.78	842	2.89
충청북도	27,892	2,207	7.91	713	2.56
충청남도	35,468	2,395	6.75	786	2.22
전라북도	37,139	2,323	6.25	770	2.07
전라남도	41,923	2,410	5.75	766	1.83
경상북도	54,584	3,040	5.57	870	1.59
경상남도	56,428	3,038	5.38	791	1.40
제주도	8,548	1,482	17.34	581	6.80

10. 키변수의 추가

○ 현 제공 MD에서는 인구와 가구주택부문에서 분석에 충분한 변수를 제공하고 있으나, 정보노출을 포함한 제공시 심각한 문제를 야기할 수 있는 경우에 대해서는 정보 제공을 제한하고 있다. 허나 반대로 분석자의 입장에서 보면 추가적인 정보제공은 분석이나 결과활용에서 당연히 반기는 입장이다.

○ 앞의 결과에서 현 제공기준 보다 본 보고서에서 제시하는 정보노출의 결과가 낮은 이유로 현재 제공기준에 변수의 개수보다 많은 변수를 제공하는 방안을 고려하고자 한다.

○ 변수추가를 위해 두 가지 면을 고려하였다. 먼저 사회통념상 분석자 입장에서 추가로 원하는 변수가 어떤 변수인지에 대한 것이다. 인구부문에서 현재 제공되지 않는 변수는 장애여부를 파악하는 '활동제약' 관련 변수들과 '어머니의 교육정도 및 출산' 관련 변수 및 '노후방법준비'관련 변수 및 '자원봉사 참여' 등과 '직업' 및 '종교' 관련 변수이다. '종교'의 경우 종교를 믿는지에 대한 여부는 제공되고 있으나, '실제 믿는 종교가 어떤 종교'인지에 대해서는 제공되지 않고 있다. 따라서 본 보고서에서는 분석에 활용 범위가 사회적으로 넓은 '직업'과 '종교'에 대한 변수를 키변수에 추가하여 노출비율을 도출하고자 하였다.

○ 비슷한 과정을 통해 가구주택의 경우 '총가구원수'와 '가구주의 교육정도'를 추가 변수로 선택하였다. 이 두 변수가 제공되지 않는 다른 변수에 활용도가 높다고 판단하여 추가 변수로 선택되었다. 물론 '가구주의 교육정도'의 경우 현 제공 MD의 '가구일련번호'를 통해 두 데이터를 결합한 후 파악이 가능하나 연구 편의상 가구주택의 제공하지 않는 데이터로 간주하였다.

○ 앞서 언급한 인구부문의 직업과 종교, 가구주택 부문의 가구원수와 가구주 교육정도는 분석에 활용도가 높아 제공시 분석자의 만족도를 높일 수 있다는 장점은 있다. 그러나 이러한 변수는 <표 2.4.4>의 외부기관 데이터 사례에도 볼 수 있듯이 정보노출의 가능성이 높은 변수이기도 하다. 실제 <표 2.4.4>에서 직업과 가구원수(표에서는 '가족수')는 데이터 조사비율(1-결측치 비율)이 각각 100%, 86%로서 데이터 충실도도 높은 편이다. 따라서, 이러한 변수를 추가한 후에도 충분한 노출제한이 유지되도록 결과를 살펴야 하겠다.

○ 인구 및 가구주택의 추가 2개의 변수에 대해 앞서 제시한 그룹화 방법을 적용한 결과가 <표 2.4.17>과 <2.4.18>이다.

<표 2.4.17> 변수의 추가, 인구부문

변수	원 데이터	그룹화 후
종교2	① 불교	① 불교
	② 기독교(개신교)	② 개신교
	③ 기독교(천주교)	③ 기타 종교
	④ 유교	
	⑤ 원불교	
	⑥ 증산교	
	⑦ 천도교	
	⑧ 대종교	
	⑨ 기타	
	⑩ 미륵대도	
	⑪ 이슬람교	
	⑫ 천리교	
	⑬ 한국SGI(창가학회)	
	⑭ 기타(무속 등 미신)	
	⑮ 미상	
자동직업코드_대	① 의회의원	④ 의회의원,전문가
	② 전문가	
	④ 사무직	② 사무직, 서비스직, 판매직
	⑤ 서비스직	
	⑥ 판매직	
	③ 기수공	③ 기수공, 기능원, 장치기계조작, 단순노무
	⑧ 장치, 기계조작	
	⑨ 단순노무	
	⑦ 농림어업	① 농림어업
	⑩ 군인	⑤ 기타
	⑪ 분류불능	

<표 2.4.18> 변수의 추가, 가구/주택부문

변수	원 데이터	그룹화 후
가구내 총가구원수	① OO명	① 2명 이하
		② 3명
		③ 4명 이상
가구주의 교육정도	① 안 받았음	① 안받았음/초중학교
	② 초등학교	
	③ 중학교	② 고등학교
	④ 고등학교	
	⑤ 2년제 대학교	③ 대학교이상
	⑥ 4년제 대학교	
	⑦ 석사	
	⑧ 박사	

11. 변수 추가 후 결과 비교

○ 변수추가 후, 현재 제공되는 MD 방법과 두 개의 변수가 추가된 이후의 유일건수를 비교하고자 한다. 변수추가 전에는 두 방법이 11개의 키변수로 동일하였으나, 추가 후에는 MD는 그대로이나 본 보고서에서 제안한 방법이 적용된 데이터는 13개의 변수로 바뀌었다. 이를 서로 비교하면 다음과 같다.

○ 변수를 추가하여 노출위험이 증가한다면, 추가 정보를 제공하는 것이 무의미하겠으나, <표 2.4.19>와 <표 2.4.20>에서 결과와 같이 큰 변화가 없다면 얼마든지 추가변수를 고려할 수 있을 것이다. 즉, 인구부문에서 변수 추가 전후의 유일비율의 차이는 0.24%(=0.67−0.43)로서 현 MD제공 방법 보다 낮은 유일비율을 나타내고 있다. 따라서, 현 제공 기준에 의한 유일비율을 허용한다면, 인구부문 MD에 대해서는 추가적인 변수를 제공하는 것도 고려할 수 있을 것이다.

○ 가구주택 부문에서는 2개의 변수 추가 및 알고리즘 적용 결과 전국 단위로 6.19%의 유일비율이 발생하였으며, 이는 현재 제공 MD 기준 5.12%보다 약 1.07%가 높은 수준이다. 물론 변수 추가 전의 비율은 낮았다고 하나, 추가 후 유일비율이 급격히 증가하는 경우에는 반대로 변수제거를 시도해야 할 것으로 사료된다.

<표 2.4.19> 인구부문 변수추가 후 노출비율, 계통추출 전국 5%

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,229,053	26,729	1.20	9,644	0.43	14,831	0.67
서울특별시	439,727	3,142	0.71	878	0.20	1,831	0.42
부산광역시	160,134	1,551	0.97	542	0.34	875	0.55
대구광역시	111,136	1,256	1.13	432	0.39	671	0.60
인천광역시	117,703	1,791	1.52	632	0.54	917	0.78
광주광역시	64,815	965	1.49	369	0.57	494	0.76
대전광역시	64,544	1,297	2.01	515	0.80	717	1.11
울산광역시	45,707	915	2.00	364	0.80	469	1.03
경기도	466,264	3,564	0.76	942	0.20	2,072	0.44
강원도	76,261	1,558	2.04	720	0.94	895	1.17
충청북도	73,989	1,460	1.97	607	0.82	823	1.11
충청남도	94,548	1,884	1.99	687	0.73	1,088	1.15
전라북도	98,719	1,334	1.35	607	0.61	801	0.81
전라남도	104,421	1,454	1.39	601	0.58	700	0.67
경상북도	136,878	1,939	1.42	721	0.53	1,027	0.75
경상남도	149,259	1,864	1.25	586	0.39	966	0.65
제주도	24,948	755	3.03	441	1.77	485	1.94

<표 2.4.20> 가구주택부문 변수추가 후 노출비율, 계통추출 전국 5%

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	791,340	41,258	5.21	12,287	1.55	48,946	6.19
서울특별시	152,748	4,358	2.85	928	0.61	4,348	2.85
부산광역시	55,654	3,140	5.64	772	1.39	3,299	5.93
대구광역시	37,904	2,413	6.37	712	1.88	2,863	7.55
인천광역시	39,799	2,538	6.38	761	1.91	3,070	7.71
광주광역시	21,998	1,779	8.09	575	2.61	2,274	10.34
대전광역시	22,185	1,890	8.52	667	3.01	2,399	10.81

울산광역시	15,341	1,643	10.71	599	3.90	1,991	12.98
경기도	154,593	3,922	2.54	1,020	0.66	4,711	3.05
강원도	29,159	2,520	8.64	839	2.88	3,194	10.95
충청북도	27,687	2,156	7.79	727	2.63	2,798	10.11
충청남도	35,515	2,482	6.99	792	2.23	3,025	8.52
전라북도	37,351	2,320	6.21	729	1.95	2,950	7.90
전라남도	41,915	2,463	5.88	841	2.01	3,023	7.21
경상북도	54,467	3,011	5.53	860	1.58	3,469	6.37
경상남도	56,292	3,145	5.59	867	1.54	3,581	6.36
제주도	8,732	1,478	16.93	598	6.85	1,951	22.34

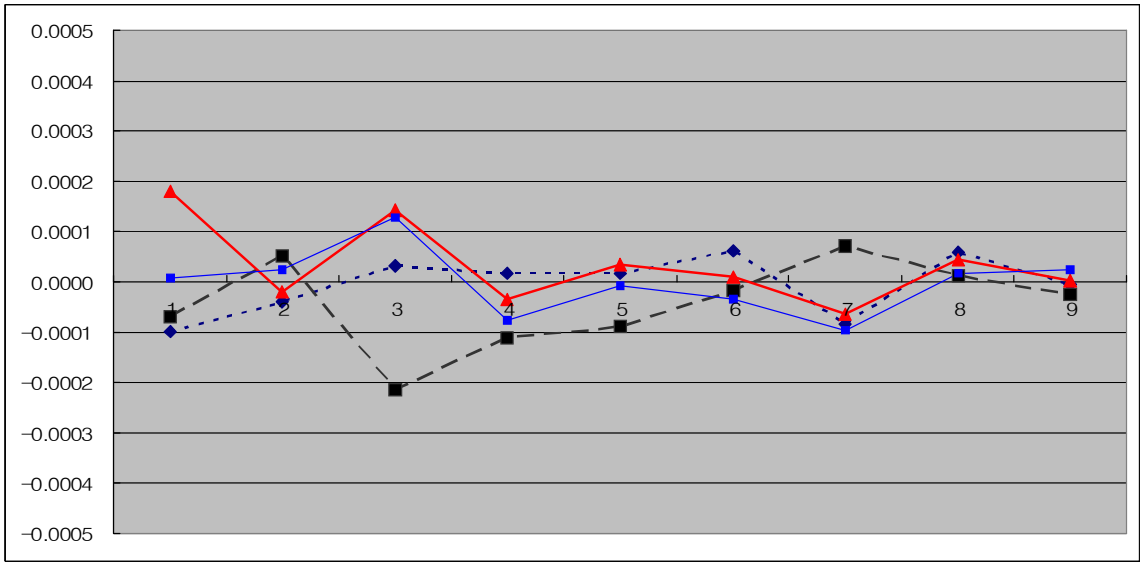
12. 집락추출과 계통추출의 비교

○ MD구성을 위한 추출방법의 선택이 정보노출에 어느 정도 영향을 미치는 가를 보기 위해 추출방법을 달리하여 결과를 비교하였다. 사실, 현재 제공되는 MD의 경우 행정구역, 거처종류(=주택유형), 농가구분 등의 변수를 이용하여 층화한 후 다시 가구를 계통 혹은 임의추출하는 방식을 취하고 있으나, 본 연구에서는 일반조사구 만을 대상으로 가구를 계통 혹은 임의추출하는 방식을 택하였다. 가구를 임의추출한 후 가구 내의 구성 인구를 모두 조사하는 방식이라면 결국 가구를 집락 추출하는 것이므로 이하에서는 '집락'으로 표현하도록 한다. 가구를 계통추출한 후, 구성원 전체를 조사하는 방식은 일종의 복합추출 기법이나 표현 편의상 이하에서는 '계통'추출이라고 표현하고자 한다.

○ <그림 2.4.1>과 <그림 2.4.2>에서 인구 및 가구주택 각각에 대한 추출방법의 비교 결과가 제시되어 있다. 두 표에서 X축은 비교를 시도하는 %구간(1~9%)을 나타낸다. 또한 Y축은 두 추출방법간의 차이(단위 %)로서, 계통추출의 결과에서 집락추출의 유일성 비율을 뺀 값을 나타낸다. 각 꺾은선 그래프는 범례에서 나타나 있듯이 현 제공 방법, 노출제한 기법 전/후 및 변수추가 후의 결과를 나타낸다.

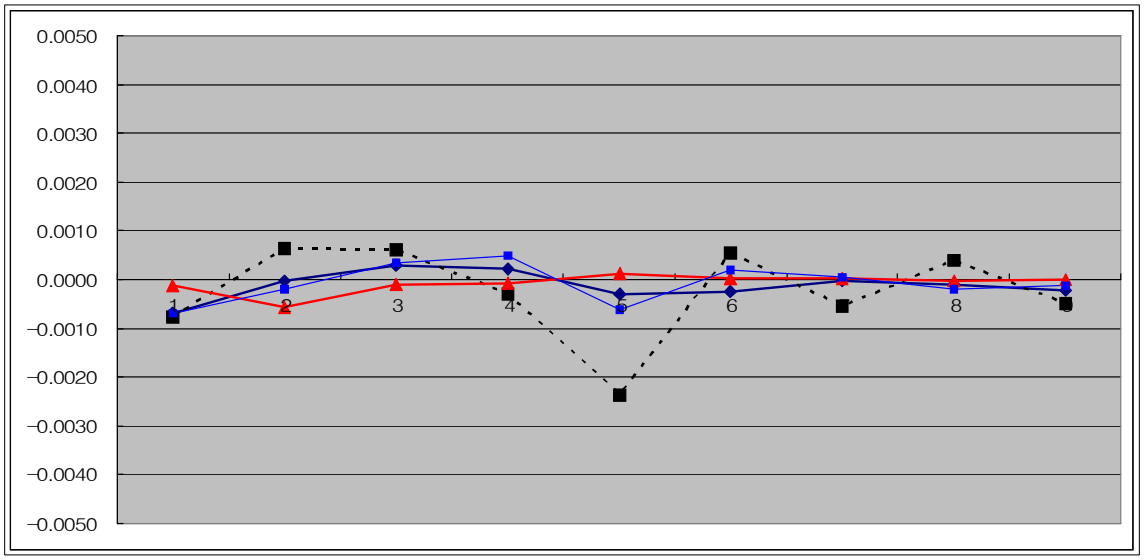
○ 인구 및 가구주택 부문의 결과가 공통적으로 추출방법에 따라 유일성 비율이 큰 차이가 없음을 나타낸다. 인구의 경우 두 방법 상의 차이가 ±0.0002%이내 이며, 가구의 경우 이보다 큰 -0.003~0.001%사이에 존재한다. 사실상 두 추출기법 사이에는 큰 차이가 없다고 보아도 무방할 것이다.

<그림 2.4.1> 표본추출 방법의 결과 비교: 인구, 단위 %



현 제공 MD 방법 적용
노출제한 기법 적용 전
노출제한 기법 적용 후
변수 추가 후

<그림 2.4.2> 표본추출 방법의 결과 비교: 가구주택, 단위 %



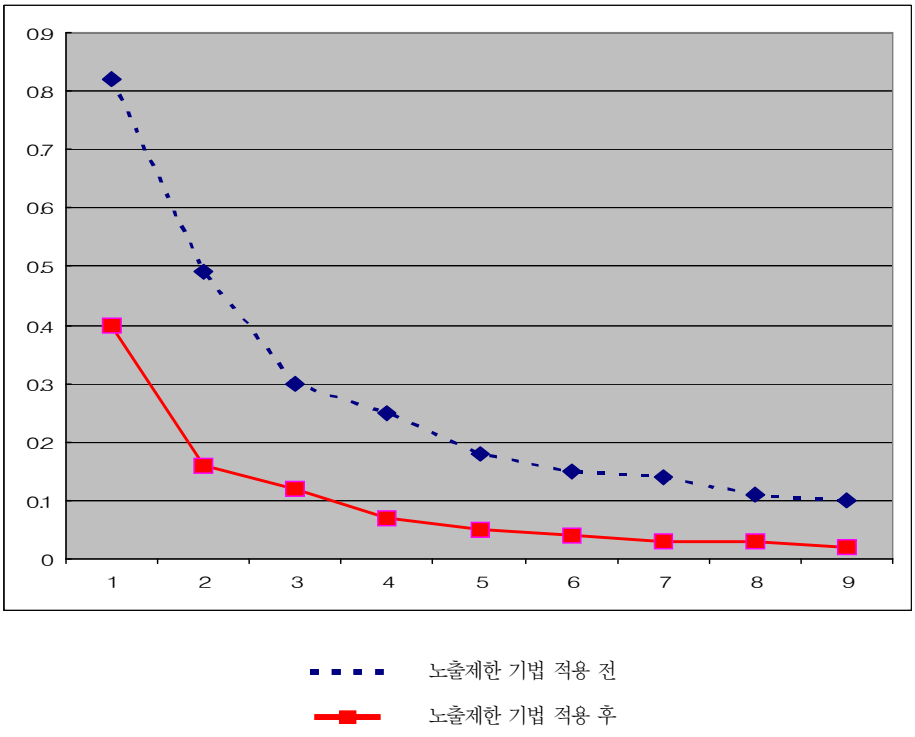
13. 표본 크기의 결정

○ 현재 인총 표본 10%데이터는 1, 2, 5%로 제공되고 있다, 이러한 제공 비율이 적정한 지에 대해서는 다른 여러 각도에서도 연구되어야 하겠으나, 본 연구에서는 정보노출과의 관련성을 살펴보고자 한다.

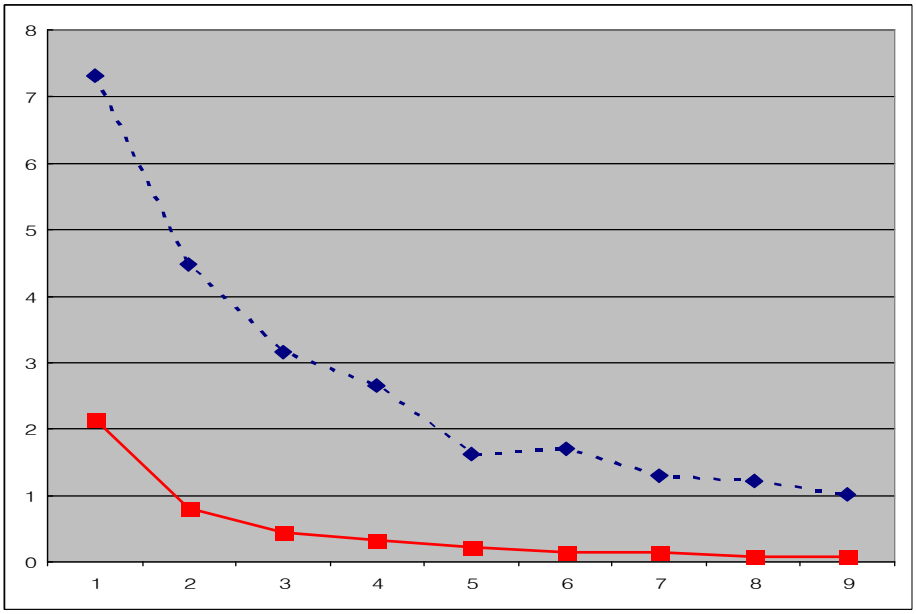
○ <그림 2.4.3>과 <그림 2.4.4>는 각각 인구와 가구주택 부문 데이터에 대한 %구간별 유일비율의 차이이다. 점선은 노출제한 기법 적용 전을, 직선을 기법 적용 후의 결과를 나타낸다. 즉, 노출제한 기법 적용 전/후의 유일성을 2%와 1% 데이터의 차이를 X축 '1'에, 3%와 2%의 차이를 '2'에 표현하는 식이다.

○ <그림 2.4.3>의 인구 부문의 각 데이터들은 실제 인구 부문의 유일비율, 건수 자체가 적은 관계로 높은 %차이를 보이고 있지는 않지만, 노출제한 기법을 적용하는 것과 관계없이 추출표본의 크기가 작은 경우에서 큰 경우로 갈수록 급격하게 줄고 있다. 이는 유일건수가 표본의 크기에 민감하게 영향을 받으며, 작은 표본일수록 정보노출이 일어나기 쉽다는 것을 시사한다. 대체로 4%로 이후 추세가 안정인 관계로 현재 제공하는 1, 2% MD에 대해서는 제공의 제고를 고려해야할 것으로 사료된다.

<그림 2.4.3> 표본크기 결정 위한 %구간별 유일비율차, 인구, 단위 %



<그림 2.4.4> 표본크기 결정 위한 %구간별 유일비율차, 가구주택, 단위 %



○ 가구주택 부문의 표본들은 인구에 비해 그 차이가 더 큰 것으로 나타났으며, 특히 1%의 노출제한 기법을 적용하기 전에는 7%이상의 유일비율의 차이를 보이고 있다. 결과상으로 유사하게 노출제한 기법 적용한 경우를 살펴보더라도 대략 3%이후에 노출비율이 안정이 되는 관계로 인구와 유사한 결론을 내릴 수 있을 것이다.

14. 읍면동 지역에 대한 마이크로 데이터 제공시 고려사항

○ 읍면동 단위로 마이크로 데이터를 제공하는 것의 가장 큰 문제는 셀 개수의 증가를 통해 정보노출을 막기가 매우 난해하다는 것이다. 즉, 인구부문의 마이크로 데이터에 적용하더라도 기존 고려 대상 키변수 11개에서 발생하는 셀 개수인 8,436만개의 셀에 다시 3,600개의 셀을 곱하여 3,036억개의 셀로서 정보노출을 막아야 되기 때문에 관측치의 개수가 아무리 많다고 하더라도 직접적인 정보제공은 거의 불가능하다고 하겠다.

○ 그러나, 정보요청자의 정보제공의 요구수준은 현재 제공되고 있는 시군구 단위 보다 더 세부적인 자료를 요청하고 있는 것이 현실이며, 이러한 요청에 부응하고 아울러 정보보호가 가능한 범위가 어느 정도인지에 대한 연구가 절실하다고 하겠다.

○ 이를 위해 본 연구에서는 다음과 같은 방향으로 읍면동 지역에 대한 MD데이터 제공을 모색하고자 한다. 읍면동 단위의 MD제공에 있어 정보노출에 대해 결정적으로 영향을 미치는 요인은 키변수의 개수라고 할 수 있다. 앞서 언급한 데로 키변수의 개수가 많아지면 정보노출이 되는 셀이 증가하며 이를 막기 위해서는 불가피하게 MD에서 제공하는 키변수의 개수를 줄여야 한다. 키변수 외의 다른 변수는 잠재적 혹은 이론적으로 정보추적이 가능하나, 현실적으로 데이터를 병합한다거나 개인적으로 정보추적을 시도하더라도 극히 일부에 해당하기 때문에 본 연구에서 제외하였다.

○ 적정한 키 변수의 개수가 과연 몇 개 정도가 타당한지를 살펴보기 위해 먼저, 인구와 가구 부문에 대해 각각 11개, 12개로 설정하였다. 이는 앞서 키변수 설정에 대한 부분과 동일하다. 여기서 도출된 키변수 군(群)에 대해 아래와 같이 위계적(hierarchical)적으로 변수를 구성하였다. 즉, 인구부문에 대한 읍면동 단위의 MD 자료를 제공하기 위해 우선 4개의 키변수를, 다음으로는 앞서 4개 변수에 2개를 추가하여 6개, 이후 8개와 마지막으로 11개의 전체 키변수를 이용하여 읍면동 단위의 정보노출을 살펴보고자 한다. 이런 식으로 고려한 변수의 위계는 아래와 같다.

<인구부문 키 변수의 구성>

- 4개의 경우 – '성별', '나이', '가구주와의 관계', '교육정도 2'
- 6개의 경우 – '종교 1', '5년전 거주지' 변수 추가 후 분석
- 8개의 경우 – '종사상 지위', '5년전 거주지 시도코드' 변수 추가 후 분석
- 11개의 경우 – '자동직업코드(대분류)', '종교 2', '혼인상태' 변수 추가 후 분석

<가구부문 키 변수의 구성>

- 4개의 경우 – '가구 구분', '거주 기간', '방수', '차량 보유'
- 6개의 경우 – '가구 소유형태', '주인가구' 변수 추가 후 분석
- 8개의 경우 – '주택 소유여부', '거처종류 2' 변수 추가 후 분석
- 12개의 경우 – '연건평', '건축년도 2', '가구내 총가구원수', '가구주의 교육정도 2' 변수 추가 후 분석

○ 앞에서 언급한 키변수의 위계를 통해 도출된 결과가 인구부문에 대해 <표 2.4.21>~<표 2.4.24>와 가구부문에 대해 <표 2.4.25>~<표 2.4.28>에 제시되었다. 우선 이 결과는 가구에 대한 계통추출에 대한 수치임을 밝혀둔다. 앞서 살펴본 바와 같이 집락추출과 결과치가 크게 다르지 않기 때문에 집락추출에 대한 결과는 제시하지 않았다. 또한 이 결과는 표본 10%에 대한 50% 표본 즉, 전국 대상 5% 표본을 대상으로 산출된 결과이다. 이 결과의 표본크기에 대한 문제 또한 앞서 살펴본바와 크게 다르지 않기 때문에 표본별로 결과를 제시하지는 않았다. 이들 표에서 3개의 기준으로 정보노출 건수 및 비율이 제시되었는데, 이중 통계청 기준이라고 표시된 항목은 현재 통계청에서 마이크로 데이터 제공 기준자료에 수록된 정보보호 기준에 의거한 것이다. 이를 정보보호 기법 적용전후의 결과와 함께 제시하였다.

○ 도출된 결과를 살펴보면 우선, 인구부문에 대해 4개의 키변수를 적용한 경우에는 대상 키변수 조합에 의한 셀의 개수가 작기 때문에 우려할 만한 정보노출이 발생하지 않았다. 인구에서 4개의 키변수 조합 적용시 정보노출 기법을 적용한 경우 전국 대비 약 3.41%이며 시도별로 1.67~4.88%의 분포를 보이고 있다. 그러나, 변수를 6개에서 8개, 11개로 늘려갈수록 정보노출 비율이 증가하여 키변수를 11개로 설정한 경우 노출제한 기법을 적용하더라도 16.45%가 정보노출이 발생하여 우려할 만한 결과를 보이고 있다. 만약 노출제한 기법을 적용하지 않은 경우는 61.66%가 노출되었으며, 현재의 통계청 기준으로도 27.79%가 노출되었다. 따라서, 현재로서 설정된 11개의 전체 키변수를 통해 MD자료를 제공하는 것은 무리라고 사료된다. 따라서, 노출 비율을 대략 10%미만으로 상정하는 경우 <표 2.4.23>에서 살펴볼 수 있는 8개 키변수 범위에서 제공하는 바람직하다고 할 수 있다. 키변수가 8개인 경우 노출제한 기법 적용 후 노출비율이 4.71%정도로 비교적 양호하기 때문이다. 11개와 8개의 노출비율이 다른 변수 개수 구간보다 급격히 차이 나는 이유는 각 변수 개수에 대한 위계를 키변수 개수가 증가할 때 추가되는 변수를 범주수가 많은 변수로 설정하였기 때문이다.

○ 가구부문에 대한 결과도 인구부문과 크게 다르지 않으나, 키변수의 개수가 증가할수록 보다 가파르게 노출비율이 증가함을 알 수 있다. 즉, 4개에서 12개의 키변수로 증가함에 따라, 5.89%, 9.18%, 14.71%, 53.54%로 노출비율이 증가하고 있다. 이는 가구에 키변수로 고려된 변수의 범주수가 인구에 비해 급격히 많아짐이 원인이라고 할 수 있다. 만약 읍면동 단위로 가구부문 MD자료를 제공하기 위해서는 10%미만의 노출비율을 상정할 때 대략 6~7개의 키변수와 기타변수로 데이터를 구성하는 것이 바람직할 것이라 사료된다.

<표 2.4.21> 인구 4개 키변수 적용 결과, 계통추출 5% 대상

시도	인구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**2,229,053**	**139,656**	**6.27**	**91,158**	**4.09**	**76,069**	**3.41**
서울특별시	439,727	33,845	7.70	22,608	5.14	19,276	4.38
부산광역시	160,134	6,904	4.31	4,702	2.94	3,967	2.48
대구광역시	111,136	5,410	4.87	3,538	3.18	2,982	2.68
인천광역시	117,703	7,645	6.50	4,293	3.65	3,748	3.18
광주광역시	64,815	3,264	5.04	1,838	2.84	1,538	2.37
대전광역시	64,544	4,012	6.22	2,404	3.72	2,085	3.23
울산광역시	45,707	2,414	5.28	1,403	3.07	1,192	2.61
경기도	466,264	39,904	8.56	29,383	6.30	22,766	4.88
강원도	76,261	4,198	5.50	2,295	3.01	2,197	2.88
충청북도	73,989	3,941	5.33	1,997	2.70	1,881	2.54
충청남도	94,548	5,839	6.18	3,206	3.39	2,899	3.07
전라북도	98,719	3,265	3.31	1,780	1.80	1,644	1.67
전라남도	104,421	3,690	3.53	1,925	1.84	1,755	1.68
경상북도	136,878	6,249	4.57	3,662	2.68	3,216	2.35
경상남도	149,259	7,817	5.24	5,415	3.63	4,252	2.85
제주도	24,948	1,259	5.05	709	2.84	671	2.69

<표 2.4.22> 인구 6개 키변수 적용 결과, 계통추출 5% 대상

시도	인구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**2,229,053**	**212,107**	**9.52**	**787,031**	**35.31**	**92,958**	**4.17**
서울특별시	439,727	33,128	7.53	143,679	32.67	14,044	3.19
부산광역시	160,134	13,860	8.66	55,262	34.51	5,993	3.74
대구광역시	111,136	8,725	7.85	35,869	32.27	3,748	3.37
인천광역시	117,703	8,955	7.61	34,754	29.53	3,771	3.20
광주광역시	64,815	5,495	8.48	21,867	33.74	2,346	3.62
대전광역시	64,544	5,083	7.88	20,459	31.70	2,184	3.38
울산광역시	45,707	3,111	6.81	11,851	25.93	1,444	3.16
경기도	466,264	34,067	7.31	133,171	28.56	14,378	3.08
강원도	76,261	11,169	14.65	35,955	47.15	4,941	6.48
충청북도	73,989	9,121	12.33	30,687	41.48	4,071	5.50
충청남도	94,548	12,084	12.78	40,802	43.15	5,347	5.66
전라북도	98,719	13,747	13.93	45,381	45.97	6,235	6.32
전라남도	104,421	15,916	15.24	50,173	48.05	7,264	6.96
경상북도	136,878	18,029	13.17	59,990	43.83	8,202	5.99
경상남도	149,259	17,110	11.46	57,644	38.62	7,834	5.25
제주도	24,948	2,507	10.05	9,487	38.03	1,156	4.63

<표 2.4.23> 인구 8개 키변수 적용 결과, 계통추출 5% 대상

시도	인구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**2,229,053**	**133,151**	**5.97**	**172,123**	**7.72**	**105,059**	**4.71**
서울특별시	439,727	32,154	7.31	42,319	9.62	25,813	5.87
부산광역시	160,134	6,642	4.15	8,255	5.16	5,114	3.19
대구광역시	111,136	5,211	4.69	6,603	5.94	3,920	3.53
인천광역시	117,703	7,272	6.18	8,898	7.56	5,543	4.71
광주광역시	64,815	3,154	4.87	3,964	6.12	2,285	3.53
대전광역시	64,544	3,870	6.00	4,732	7.33	2,947	4.57
울산광역시	45,707	2,320	5.08	2,763	6.05	1,757	3.84
경기도	466,264	37,456	8.03	53,230	11.42	28,544	6.12
강원도	76,261	4,102	5.38	4,542	5.96	3,735	4.90
충청북도	73,989	3,850	5.20	4,269	5.77	3,241	4.38
충청남도	94,548	5,668	5.99	6,582	6.96	4,646	4.91
전라북도	98,719	3,211	3.25	3,593	3.64	2,854	2.89
전라남도	104,421	3,604	3.45	4,059	3.89	3,059	2.93
경상북도	136,878	6,017	4.40	7,315	5.34	4,913	3.59
경상남도	149,259	7,394	4.95	9,589	6.42	5,594	3.75
제주도	24,948	1,226	4.91	1,410	5.65	1,094	4.39

<표 2.4.24> 인구 11개 키변수 적용 결과, 계통추출 5% 대상

시도	인구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**2,229,053**	**619,357**	**27.79**	**1,374,422**	**61.66**	**366,732**	**16.45**
서울특별시	439,727	115,946	26.37	281,416	64.00	67,414	15.33
부산광역시	160,134	44,499	27.79	100,776	62.93	26,051	16.27
대구광역시	111,136	29,252	26.32	68,092	61.27	17,109	15.39
인천광역시	117,703	29,293	24.89	66,950	56.88	17,062	14.50
광주광역시	64,815	17,098	26.38	38,596	59.55	10,219	15.77
대전광역시	64,544	16,360	25.35	38,938	60.33	9,640	14.94
울산광역시	45,707	9,691	21.20	24,092	52.71	5,849	12.80
경기도	466,264	110,524	23.70	258,622	55.47	65,105	13.96
강원도	76,261	29,158	38.23	53,771	70.51	17,954	23.54
충청북도	73,989	23,757	32.11	48,160	65.09	14,285	19.31
충청남도	94,548	30,566	32.33	62,132	65.71	18,331	19.39
전라북도	98,719	33,184	33.61	66,193	67.05	19,979	20.24
전라남도	104,421	35,502	34.00	70,730	67.74	21,059	20.17
경상북도	136,878	43,680	31.91	89,690	65.53	25,849	18.88
경상남도	149,259	43,419	29.09	90,499	60.63	26,321	17.63
제주도	24,948	7,428	29.77	15,765	63.19	4,505	18.06

<표 2.4.25> 가구 4개 키변수 적용 결과, 계통추출 5% 대상

시도	가구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**791,340**	**71,996**	**9.10**	**200,174**	**25.30**	**46,638**	**5.89**
서울특별시	152,748	9,959	6.52	43,329	28.37	6,430	4.21
부산광역시	55,654	4,666	8.38	15,442	27.75	2,758	4.96
대구광역시	37,904	3,039	8.02	11,081	29.23	1,776	4.69
인천광역시	39,799	2,853	7.17	8,386	21.07	1,872	4.70
광주광역시	21,998	1,774	8.06	5,142	23.37	1,083	4.92
대전광역시	22,185	1,686	7.60	5,730	25.83	1,050	4.73
울산광역시	15,341	1,289	8.40	3,963	25.83	829	5.40
경기도	154,593	10,770	6.97	34,093	22.05	7,179	4.64
강원도	29,159	4,230	14.51	8,664	29.71	2,820	9.67
충청북도	27,687	3,171	11.45	7,031	25.39	2,175	7.86
충청남도	35,515	4,159	11.71	8,347	23.50	2,789	7.85
전라북도	37,351	4,602	12.32	8,876	23.76	3,073	8.23
전라남도	41,915	5,317	12.69	9,466	22.58	3,554	8.48
경상북도	54,467	6,884	12.64	13,661	25.08	4,492	8.25
경상남도	56,292	6,561	11.66	14,170	25.17	4,136	7.35
제주도	8,732	1,036	11.86	2,793	31.99	622	7.12

<표 2.4.26> 가구 6개 키변수 적용 결과, 계통추출 5% 대상

시도	가구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**791,340**	**118,334**	**14.95**	**270,165**	**34.14**	**72,619**	**9.18**
서울특별시	152,748	20,139	13.18	62,581	40.97	11,667	7.64
부산광역시	55,654	9,190	16.51	21,786	39.15	4,931	8.86
대구광역시	37,904	5,566	14.68	15,618	41.20	3,146	8.30
인천광역시	39,799	5,368	13.49	12,104	30.41	3,249	8.16
광주광역시	21,998	3,136	14.26	7,323	33.29	1,859	8.45
대전광역시	22,185	3,240	14.60	8,234	37.12	1,870	8.43
울산광역시	15,341	2,208	14.39	5,398	35.19	1,359	8.86
경기도	154,593	19,464	12.59	48,294	31.24	12,008	7.77
강원도	29,159	6,016	20.63	10,588	36.31	4,041	13.86
충청북도	27,687	4,658	16.82	9,004	32.52	3,086	11.15
충청남도	35,515	5,782	16.28	10,131	28.53	3,832	10.79
전라북도	37,351	6,380	17.08	10,757	28.80	4,140	11.08
전라남도	41,915	6,620	15.79	10,523	25.11	4,348	10.37
경상북도	54,467	9,359	17.18	16,446	30.19	6,048	11.10
경상남도	56,292	9,600	17.05	17,903	31.80	5,993	10.65
제주도	8,732	1,608	18.42	3,475	39.80	1,042	11.93

<표 2.4.27> 가구 8개 키변수 적용 결과, 계통추출 5% 대상

시도	가구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**791,340**	**181,986**	**23.00**	**316,708**	**40.02**	**116,416**	**14.71**
서울특별시	152,748	33,347	21.83	72,892	47.72	20,607	13.49
부산광역시	55,654	14,496	26.05	25,506	45.83	8,515	15.30
대구광역시	37,904	8,617	22.73	17,724	46.76	5,339	14.09
인천광역시	39,799	9,146	22.98	15,273	38.38	5,505	13.83
광주광역시	21,998	4,480	20.37	8,171	37.14	2,883	13.11
대전광역시	22,185	5,047	22.75	9,729	43.85	3,202	14.43
울산광역시	15,341	3,496	22.79	6,210	40.48	2,219	14.46
경기도	154,593	31,562	20.42	58,738	38.00	20,114	13.01
강원도	29,159	8,647	29.65	12,152	41.67	5,955	20.42
충청북도	27,687	6,907	24.95	10,558	38.13	4,672	16.87
충청남도	35,515	8,575	24.14	12,147	34.20	5,842	16.45
전라북도	37,351	8,798	23.55	12,238	32.76	5,954	15.94
전라남도	41,915	9,057	21.61	12,021	28.68	6,168	14.72
경상북도	54,467	13,517	24.82	18,847	34.60	8,822	16.20
경상남도	56,292	13,768	24.46	20,470	36.36	8,996	15.98
제주도	8,732	2,526	28.93	4,032	46.17	1,623	18.59

<표 2.4.28> 가구 11개 키변수 적용 결과, 계통추출 5% 대상

시도	가구수	통계청기준		기법 적용 전		기법 적용 후	
		건수	비율	건수	비율	건수	비율
전국	**791,340**	**320,208**	**40.46**	**658,206**	**83.18**	**423,687**	**53.54**
서울특별시	152,748	59,537	38.98	132,109	86.49	84,235	55.15
부산광역시	55,654	25,690	46.16	46,661	83.84	33,800	60.73
대구광역시	37,904	15,209	40.13	31,564	83.27	21,790	57.49
인천광역시	39,799	15,189	38.16	32,878	82.61	21,762	54.68
광주광역시	21,998	7,834	35.61	17,331	78.78	11,169	50.77
대전광역시	22,185	8,318	37.49	18,624	83.95	11,927	53.76
울산광역시	15,341	5,785	37.71	11,896	77.54	7,966	51.93
경기도	154,593	52,145	33.73	126,413	81.77	77,376	50.05
강원도	29,159	14,530	49.83	24,055	82.50	17,167	58.87
충청북도	27,687	12,116	43.76	22,959	82.92	15,046	54.34
충청남도	35,515	15,959	44.94	29,227	82.29	19,105	53.79
전라북도	37,351	16,017	42.88	31,279	83.74	19,522	52.27
전라남도	41,915	17,624	42.05	34,738	82.88	19,694	46.99
경상북도	54,467	24,653	45.26	44,997	82.61	28,315	51.99
경상남도	56,292	25,014	44.44	46,444	82.51	29,555	52.50
제주도	8,732	4,588	52.54	7,031	80.52	5,258	60.22

Ⅱ-5. 소지역 자료 제공 방안

1. 소지역 통계 정확도 제고 방안 – 소지역 추정 방법 연구 추세

○ 정보통신기술의 획기적인 발달로 사회는 글로벌 정보화로 발전하면서 다양한 분야에서 통계정보의 활용이 증대되고 있을 뿐 아니라 통계에 대한 신뢰성, 정확성과 시의성 등 다양한 특성들이 요구되고 있으며 더욱이 1995년에 시작된 지방자치제도의 활성화로 시군구 단위 나아가서는 읍면동 단위의 지역통계에 대한 필요성이 대두되고 있다. 그러나 현재와 같은 통계청의 인력과 조직으로 시군구 단위까지 관련 통계를 생산하는 것은 어렵기 때문에 현재 주어진 여건을 감안하여 경제적으로 큰 부담 없이 일정 수준의 정확성을 갖춘 시군구 및 읍면동 단위의 지역관련 통계를 생산할 수 있는 방안의 연구가 요구되고 있다.

○ 이에 따라 각국에서는 기존의 소지역 추정 방법에 대한 연구 결과를 현장에 반영하여 이를 통해 소지역 통계의 정확도 제고에 용이한 방법을 이전부터 적용하고 있다. 본 연구에서 제시하는 여러 추정법은 실제 각국 통계현장에서 실제 적용하는 방법을 정리한 것이다.

○ 그러한 기법들로 대표적인 것이 합성추정법, 복합추정법과 최근에 실무에 적용되고 있는 EBLUP(Best Linear Unbiased Prediction)와 Empirical Bayes(EB)추정법이다. 이러한 추정법의 특징은 표본에 대한 추가적인 분포가정을 통해 기존의 직접추정(direct estimation)방법이나 비추정(ratio estimation)에서 발생하는 분산을 낮추어 상대효율을 최소화하는 방법이라고 할 수 있다. 또한 이러한 방법은 직접추정량이나 비추정량이 활용하지 못한 보조변수(auxiliary variables)를 활용하기 때문에 보조변수의 성격과 개수에 따라 얼마든지 추정의 성능을 높일 수 있다는 장점이 있다. 즉, 관심변수와 상관도가 높은 보조변수가 많으면 많을수록 소지역 추정의 정확도는 높아진다는 것이다. 이를 자세히 살펴보면 아래와 같다.

(1) 합성추정법(Synthetic Estimation)

○ 추정하기 위한 소지역의 특성과 유사한 다른 지역의 정보를 이용하여 추정값의 정도를 높이고자 하는 추정방식이다. 주변 지역이나 유사지역의 정보를 이용하기 때문에 "Borrow Strength" 라고도 한다. 합성 추정법은 인접 유사지역에 대해 연령대별, 성별 등의 범주 등으로 구분할 때 각 소지역이 동일한 특성을 갖는다는 가정 하에 관심 변수에 대하여 추정값을 산출하는 기법이다. Gonzalez(1973)는 “어느 한 대지역에 대해 표본조사를 통해 불편추정값이 얻어지고, 소지역들의

특징이 대지역과 같다는 가정 아래 대지역의 추정값이 소지역에 대한 추정값을 유도하는데 사용될 때, 소지역에 대한 추정값을 합성추정량(synthetic estimator)이라고 한다." 라고 정의하였다. 따라서 합성추정량을 구할 때 소지역 a는 a지역을 포함하는 다른 소지역 a'지역과 유사하다는 가정이 존재하므로 이 가정이 위배되면 편의(bias)가 커지게 된다.

○ 한편 합성추정량을 사용할 때 요구되는 사항은 첫째, 관심변수와 관련이 있는 보조정보가 존재해야 한다는 것이다. 보조정보와 관심변수의 관련이 높을수록 더 좋은 추정량이 만들어진다. 둘째, 모형가정이 만족되어야 한다는 것인데, 모형가정이 만족되지 않을 경우에는 편의가 발생하게 된다. 즉, 대지역에서 관찰될 수 있는 관계가 소지역에 대해서도 만족해야 한다.

○ 대지역을 $I(i=1,2,\cdots,I)$개의 소지역으로 분할하고 대지역의 특성 기준에 따라 $J(j=1,2,\cdots,J)$개의 범주로 분류한다면 j영역에 대한 총계는 $Y_{.j}=\sum_i Y_{ij}$9999999가 되고, i지역에 대한 총계는 $Y_{i.}=\sum_j Y_{ij}$ 가 된다. 표본조사로부터 영역 j에 대한 총계 $Y_{.j}$의 직접추정량 $\widehat{Y_{.j}}'$을 얻을 수 있다 (김은석, 1995).

<표 2.5.1> 합성추정량을 산출하기 위한 자료의 구조

i \ j	1	…	j	…	합계
1	Y_{11}	…	Y_{1j}	…	$Y_{1.}$
⋮	⋮		⋮		⋮
i	Y_{i1}	…	Y_{ij}	…	$Y_{i.}$
⋮	⋮		⋮		⋮
합계	$Y_{.1}$	…	$Y_{.j}$	…	Y

○ Purcell과 Linacre(1976)는 소지역 i에 대한 총계 $Y_{i.}$의 추정량으로 다음과 같은 합성추정량을 제안하였다.

$$\hat{Y}_{i.}^{s}=\sum_j\left(\frac{X_{ij}}{X_{.j}}\right)\hat{Y}_{.j}'. \qquad (2.5.1)$$

여기서 X_{ij}는 i지역 j영역에서의 보조변수이고, $X_{.j}=\sum_i X_{ij}$ 즉, j영역의 보조변수의 총계이다. 위의 추정량은, 지역별 추정량인 식 (2.5.1)을 모두 합하면 모집단 총계 Y의 직접 추정량,

$\hat{Y}'(=\sum_j \hat{Y}_{.j}')$로 일치한다.

$$\sum_i \hat{Y}_{i.}^s = \hat{Y}' = \sum_j \hat{Y}_{.j}'. \tag{2.5.2}$$

식 (2.5.1)에서 사용된 $\hat{Y}_{.j}'$은 일반적으로 비추정량의 형태를 띤다. 즉, $\hat{Y}_{.j}' = \left(\frac{\hat{Y}_{.j}}{\hat{X}_{.j}}\right) X_{.j}$이다. 이 형태의 $\hat{Y}_{.j}'$을 사용할 경우 식 (2.5.1)의 합성추정량은 $\hat{Y}_{i.}^s = \sum_j X_{ij}\left(\frac{\hat{Y}_{.j}}{\hat{X}_{.j}}\right)$가 된다. $\hat{Y}_{.j}'$이 불편추정량이 되려면 $\frac{Y_{.j}}{X_{.j}} = \frac{Y_{ij}}{X_{ij}}$를 만족해야 한다. $\hat{Y}_{.i}^s$의 설계 기반 편의(design-based bias)는 다음과 같다.

$$E(\hat{Y}_{i.}^s) - Y_{i.} = \sum_j X_{ij}\left(\frac{Y_{.j}}{X_{.j}} - \frac{Y_{ij}}{X_{ij}}\right). \tag{2.5.3}$$

○ 위의 식에서 $Y_{.j}/X_{.j} = Y_{ij}/X_{ij}$일 때 편의가 0이 됨을 알 수 있다. 특수한 경우로 X_{ij}를 i 지역에서 j영역의 모집단 크기 N_{ij}로 놓으면 $\overline{Y}_{ij} = \overline{Y}_{.j}$일 때 편의가 0이 되는 것이다. 이러한 가정은 매우 제약적이어서 합성추정량의 편의의 발생은 불가피하다고 할 수 있다. 편의가 발생하지만 식 (2.5.1)을 참고로 하면 이 추정량의 분산은 $\hat{Y}_{.j}'$의 분산에 따라서만 달라짐을 알 수 있다. 여기서 Ghosh와 Rao(1994)가 제시한 것과 같이 i 지역의 총계 $Y_{i.}$의 직접불편추정량을 $\hat{Y}_{i.}$라 하고 $cov(\hat{Y}_{i.}, \hat{Y}_{i.}^s) = 0$이라고 가정했을 때, $\hat{Y}_{i.}^s$의 평균제곱오차(Mean Squared Error; MSE)의 근사적 불편추정량은 다음과 같다.

$$mse(\hat{Y}_{i.}^s) = (\hat{Y}_{i.}^s - \hat{Y}_{i.})^2 - var(\hat{Y}_{i.}). \tag{2.5.4}$$

여기서 $var(\hat{Y}_{i.})$은 $VAR(\hat{Y}_{i.})$의 불편추정량이다.

(2) 복합추정법 (Composite Estimation)

○ 직접추정량의 불안정성과 합성추정량의 편의를 서로 보완할 수 있는 방법으로 두 추정량의 가중평균을 사용하는 방법을 생각할 수 있는데, 이렇게 두 추정량의 가중평균을 사용하는 것이 복합추정

량(composite estimator)이다. 복합추정량의 일반적인 형태는 다음과 같다.

$$\hat{Y}_{i.}^{c} = w_i \hat{Y}_{i.} + (1 - w_i)\hat{Y}_{i.}^{s} \tag{2.5.5}$$

여기서 $\hat{Y}_{i.}$는 직접추정량이며, $\hat{Y}_{i.}^{s}$은 합성추정량을 나타낸다. w_i는 가중치로 0과 1사이의 값을 갖는다. w_i는 다음과 같은 방법을 통해 결정할 수 있다. 먼저 i 지역 총계에 대한 복합추정량 $\hat{Y}_{i.}^{c}$의 MSE를 최소화하는 최적가중값을 찾는 방법이 있다. $cov(\hat{Y}_{i.}, \hat{Y}_{i.}^{s}) = 0$을 가정하며 이 때의 최적가중값 $w_i(opt)$는 다음과 같다.

$$w_i(opt) = \frac{MSE(\hat{Y}_{i.}^{s})}{[MSE(\hat{Y}_{i.}^{s}) + Var(\hat{Y}_{i.})]}. \tag{2.5.6}$$

위 식의 추정값은 분자에 식 (2.5.2)의 MSE 추정값을, 분모에 $(\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2$을 대입하여 얻는다. 즉,

$$\hat{w}_i(opt) = \frac{mse(\hat{Y}_{i.}^{s})}{(\widehat{Y_{i.}^{s}} - \hat{Y}_{i.})^2} \tag{2.5.7}$$

와 같다. 가중치 $\hat{w}_i$은 표본조사자료로부터 추정되므로 해당 소지역에 할당된 표본조사구 수에 영향을 받는다. 표본조사구 수가 충분할 경우 직접추정량의 추정분산이 감소함에 따라 $\hat{w}_i$의 값이 증가하게 되어 복합추정량은 합성추정량에 비해 상대적으로 직접추정량의 영향을 받게 되며 반대로 소지역에 할당된 표본조사구 수가 충분하지 않을 경우에는 상당 부분 합성추정량의 영향을 받게 된다.

○ Purcell과 Kish(1979)는 소지역별 최적 가중치 $\hat{w}_i(opt)$를 사용하는 대신 MSE의 평균 즉, $1/m\sum_i MSE(\hat{Y}_{i.}^{c})$을 최소화하는 w를 각 소지역의 복합 추정량에 공통적으로 적용하는 방안을 제시하였다.

$$\hat{w}(opt) = 1 - \frac{\sum_i var(\hat{Y}_{i.})}{\sum_i (\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2}. \tag{2.5.8}$$

위의 추정가중값에서 $\hat{Y}_{i.}$들의 분산이 거의 같다면 위의 식은 아래와 같이 쓸 수 있다.

$$\hat{w}\,(opt\,) = 1 - \frac{m\bar{v}}{\sum_i (\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2}. \tag{2.5.9}$$

여기서 $\bar{v} = \frac{1}{m}\sum_i var\,(\hat{Y}_{i.})$이다.

○ 최적가중값 외에 표본크기에 의해 결정되는 단순 가중값이 있다. 단순가중값은 모집단 크기와 표본크기 또는 보조변수의 총계에 따라서만 가중값이 달라진다. Drew 등(1982)은 다음과 같은 가중값을 사용하는 표본크기 의존 추정량(sample-size dependent estimator)을 제안하였다. i지역 모집단 크기 N_i의 직접 불편추정량을 $\hat{N}_i$이라고 할 때,

$$w_i\,(D) = \begin{cases} 1 & ,\ \hat{N}_i \geqq \delta N_i \\ \dfrac{\hat{N}_i}{\delta N_i} & ,\ otherwise \end{cases}. \tag{2.5.10}$$

여기서 N_i는 알고 있는 값이고 δ는 합성추정량의 기여도를 결정해 주는 임의로 선정되는 상수값이다. 캐나다 노동력 통계조사(Canadian Labour Force Survey)에서는 $\delta = 2/3$를 사용하고 있다.

○ 또 다른 방법으로 Sarndal과 Hidiroglou(1989)는 다음과 같은 가중치를 제안했다.

$$w_i\,(S) = \begin{cases} 1 & ,\ \hat{N}_i \geqq N_i \\ \left(\dfrac{\hat{N}_i}{N_i}\right)^{h-1} & ,\ otherwise \end{cases}. \tag{2.5.11}$$

여기서 h는 δ와 같이 주관적인 방법으로 결정되는 상수값이며 Sarndal과 Hidiroglou(1989)는 $h = 2$의 사용을 제안하였다.

(3) Empirical Best Linear Unbiased Prediction(EBLUP) 추정법

○ $BLUP$(Best Linear Unbiased Prediction)추정량은 Henderson(1950)이 제안하였으며 선형불편(linear unbiased)추정량 중 MSE를 최소로 한다. 즉, 최량선형불편추정량(Best Linear Unbiased Prediction; $BLUP$)과 유사한 개념의 추정량이다. Ghosh와 Rao(1994)는 식 (2.5.12)과 같은 일반화혼합선형모형(Generalized Linear Mixed Model; GLMM)을 사용하여 $EBLUP$(Empirical Best Linear Unbiased Prediction)추정량을 구했다.

$$\hat{\theta_i} = x_i^T\beta + v_i + e_i \tag{2.5.12}$$

위의 결합모형은 고정효과 β와 소지역 랜덤효과 v_i를 갖는 선형혼합효과모형의 일종이며, 특히 설계기반 확률변수(design-based random variable) e_i와 모형 기반 확률변수(model-based random variable) v_i를 동시에 포함하고 있는 모형이다. 여기서 모수 σ_v^2는 소지역들의 동질성을 나타내는 측도이다. $EBLUP$추정량은 랜덤오차 e_i와 v_i의 분포에 대한 가정을 필요로 하지 않으나 MSE 추정을 위해 정규분포를 가정하기도 한다. 또한, $EBLUP$추정량과 EB추정량은 e_i와 v_i를 정규분포로 가정했을 경우에는 동일하며, HB추정량과는 근사적으로 같게 된다. 그러나 추정량들의 변동을 나타내는 측도들은 동일하지 않다. 고정계수 l_t를 갖는 θ_i의 선형추정량 $\sum_t l_t\hat{\theta_i}$가 식 (2.5.12)에 대해서 $\sum_t l_t\hat{\theta_i} - \theta_i$의 기댓값이 0을 만족할 때, $\sum_t l_t\hat{\theta_i}$을 θ_i의 선형불편예측(LUP) 추정량이라 한다. θ_i의 최량선형불편예측($BLUP$)추정량은 선형불편예측(LUP)추정량들 중 최소평균제곱오차를 갖는 추정량을 말한다. 식 (2.3.1) 하에서 θ_i의 $BLUP$추정량은 다음과 같이 주어진다(Prasad and Rao, 1990).

$$\tilde{\theta_i}(\sigma_v^2) = \gamma_i\hat{\theta_i} + (1-\gamma_i)x_i^T\tilde{\beta}(\sigma_v^2). \tag{2.5.13}$$

여기에서 $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$이고, $\tilde{\beta}(\sigma_v^2)$은 가중치 $(\sigma_v^2 + \psi_i)^{-1}$을 갖는 가중최소제곱 추정량으로 아래와 같이 주어진다.

$$\tilde{\beta}(\sigma_v^2) = (\sum_i \gamma_i x_i x_i^T)^{-1}(\sum_i \gamma_i x_i y_i). \tag{2.5.14}$$

식 (2.5.13)의 $BLUP$추정량은 가중치 γ_i를 갖는 직접추정량 $\hat{\theta_i}$과 가중치 $1-\gamma_i$를 갖는 회귀합성추정량 $x_i^T\tilde{\beta}(\sigma_v^2)$의 가중결합으로 볼 수 있다. 또한, 표본분산 ψ_i가 작을 때(σ_v^2이 클 경우) $BLUP$추정량은 직접추정량 $\hat{\theta_i}$에 큰 가중치가 부여되고, 반대의 경우에는 회귀합성추정량 $x_i^T\tilde{\beta}(\sigma_v^2)$에 큰 가중치가 부여된다. 표본이 추출되지 않은 지역들에 대해서는 $BLUP$추정량은 회귀합성추정량만으로 주어질 수 있다. $BLUP$추정량의 변동의 측도는 추정량의 $MSE = E(\hat{\theta_i} - \theta_i)^2$에 의해 주어지며 다음과 같다.

$$MSE(\tilde{\theta}_i(\sigma_v^2)) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) \tag{2.5.15}$$

여기서

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i,$$

$$g_{2i}(\sigma_v^2) = \sigma_v^2 (1-\gamma_i)^2 x_i^T (\sum_i \gamma_i x_i x_i^T)^{-1} x_i \qquad (2.5.16)$$

로 주어진다.

○ 식 (2.5.13)와 식 (2.5.15)는 랜덤오차 v_i와 e_i에 관한 분포의 가정을 필요로 하지 않는다. 주요 항 $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$는 $O(1)$, $g_{2i}(\sigma_v^2)$은 $O(m^{-1})$에 유계인 항이며, 이로부터 $BLUP$ 추정량의 MSE 값은 γ_i나 모형분산 σ_v^2이 표본분산 ψ_i에 비해 작을 경우 직접추정량의 MSE 값보다 훨씬 작아질 수 있다는 사실을 알 수 있다. 따라서 소지역 추정의 정확도는 표본분산에 비해 모형 분산을 작게 할 수 있는 보조변수에 크게 의존한다고 볼 수 있다. 대부분의 문제에서는 모형분산 σ_v^2은 미지이므로 적절한 σ_v^2을 추정하여 $EBLUP$ 추정량 $\tilde{\theta} = \widetilde{\theta_i}(\hat{\sigma}_v^2)$을 산출한다. 이 때 소지역의 평균 $\overline{Y}_i$의 추정량은 $g^{-1}(\tilde{\theta}_i)$로, σ_v^2의 추정량은 $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$로 주어진다. 여기에서 $\tilde{\sigma}_v^2$은 다음 식을 만족한다.

$$(m-p)\tilde{\sigma}_v^2 = \sum_i (\tilde{\theta}_i - x_i^T \beta^*)^2 - \sum_i \psi_i h_{ii}. \qquad (2.5.17)$$

식 (2.5.17)에서 $h_{ii} = x_i^T (\sum_i x_i x_i^T)^{-1} x_i$ 이고, β^*는 β 의 Ordinary Least Squares(OLS) 추정량이다. 한편, $\tilde{\sigma}_v^2$ 은 다음과 같은 비선형 방정식의 반복적인 해로써 구할 수도 있다.

$$a(\sigma_v^2) = \sum_i \hat{\theta}_i - x_i^T \tilde{\beta}(\sigma_v^2)^2 / (\sigma_v^2 + \psi_i) = m - p. \qquad (2.5.18)$$

여기서 $\tilde{\beta}(\sigma_v^2)$은 식 (2.5.14)에 의해 주어졌고, 식 (2.5.17)의 가운데 항은 가중잔차제곱합, $m-p$는 가중잔차제곱합과 관계가 되는 자유도이다. 만약 $\hat{\sigma}_v^2 = 0$ 이면, $EBLUP$ 추정량 $\widetilde{\theta_i}$는 회귀합성추정량 $\widetilde{\theta_i}$로 축약된다. 단, $\tilde{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$이며, 식 (2.5.14)에서 σ_v^2 대신에 $\hat{\sigma}_v^2$으로 대체하여 산출한다. 물론 위의 식 (2.5.16)와 식 (2.5.17)으로부터 얻게 되는 추정량들도 v_i와 e_i의 분포에 대한 가정을 필요로 하지는 않는다. 만약 랜덤오차 v_i와 e_i가 정규분포를 따른다고 가정한다

면, $\widehat{\theta_i}$는 평균이 $x_i^T\beta$이고 분산이 $\sigma_v^2+\psi_i$인 서로 독립인 정규분포를 따르게 된다. 이러한 분포 가정 하에서 계산된 β와 σ_v^2의 최대우도추정량을 제한최대우도추정량(Restricted Maximum Likelihood Estimator; $REMLE$)이라 하며, 선형혼합모형 하에서도 근사적으로 유효하다. 따라서 $\widetilde{\theta_i}$의 $BLUP$ 추정량 산출 시 σ_v^2의 $REML$ 추정량을 이용해도 근사적으로 타당하다.

(4) Empirical Bayes(EB) 추정법

○ 경험적 베이즈(Empirical Bayes; EB) 추정법은 랜덤오차 v_i와 e_i가 정규분포를 따른다는 가정 하에서 출발한다. $(\widehat{\theta_i},\theta_i)$의 결합분포가 평균이 $(x_i^T\beta, x_i\beta)$, 분산이 $(\sigma_v^2+\psi_i, \sigma_v^2)$, 상관계수가 γ_i인 이변량 정규분포를 따른다고 가정하자. 이 때, θ_i의 평균제곱오차를 최소로 하는 베이즈 추정량은 다음과 같다.

$$\begin{aligned}\tilde{\theta}_i^B(\beta,\sigma_v^2) &= E(\theta_i|\hat{\theta}_i,\beta,\sigma_v^2)\\ &= \gamma_i\hat{\theta}_i + (1+\gamma_i)x_i^T\beta.\end{aligned} \tag{2.5.19}$$

○ 식 (2.5.18)의 베이즈 추정량은 선형성 또는 불편성을 만족하지는 않는다. 여기에서 모수 β와 σ_v^2을 제한최대우도($REML$)추정량으로 대체하여 다음과 같은 θ_i의 경험적 베이즈(EB) 추정량을 얻는다.

$$\tilde{\beta}_i^{EB} = \tilde{\beta}_i^B(\hat{\beta},\hat{\sigma}_v^2). \tag{2.5.20}$$

경험적 베이즈(EB) 추정량 $\tilde{\beta}_i^{EB}$는 정규분포 가정 하에서는 $EBLUP$ 추정량 $\widetilde{\theta_i}$와 같다. 그러나 경험적 베이즈 방법은 $\widehat{\theta_i}$과 θ_i의 임의의 결합분포에 대해서도 일반적으로 응용할 수 있다는 장점이 있다. $EBLUP$ 추정량 $\widetilde{\theta_i}=\widetilde{\theta_i}(\hat{\sigma}_v^2)$의 MSE 추정량은 식 (2.5.15)에서 σ_v^2 대신 $\hat{\sigma}_v^2$을 대체하여 얻어질 수 있으나, 이 경우에는 σ_v^2에 대한 추정효과가 무시되기 때문에 MSE의 추정값은 과소추정되는 경향을 보인다. 이러한 문제 때문에 Prasad와 Rao(1990)는 v_i와 e_i에 대해 정규성을 가정하여 근사적으로 불편인 $EBLUP$ 추정량 $\widetilde{\theta_i}$의 MSE 추정량을 제안하였다. Prasad와 Rao(1990)가 제안한 MSE 추정량은 식 (2.5.16)의 σ_v^2의 적률추정량을 사용하였을 경우 다음과 같이 주어진다.

$$mse(\tilde{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \tag{2.5.21}$$

여기서

$$\begin{aligned} g_{1i}(\hat{\sigma}_v^2) &= \gamma_i \psi_i \ , \\ g_{2i}(\hat{\sigma}_v^2) &= \sigma_v^2 (1-\gamma_i)^2 x_i^T (\sum_i \gamma_i x_i x_i^T)^{-1} x_i \ , \\ g_{3i}(\hat{\sigma}_v^2) &= [\psi_i^2 / (\sigma_v^2 + \psi_i)^3] h(\sigma_v^2) \ , \\ h(\sigma_v^2) &= 2m^{-2} \sum_i (\sigma_v^2 + \psi_i)^2 \end{aligned} \tag{2.5.22}$$

○ 최근 들어, Jiang, Lahiri와 Wan(2002)은 근사적으로 불편인 잭나이프(Jackknife) MSE 추정량을 제안하였다. 잭나이프 방법은 랜덤인 지역효과들을 갖는 로지스틱 회귀와 같은 보다 더 복잡한 모형들에 대해서도 쉽게 적용할 수 있다는 장점을 갖고 있다. θ_i의 EB추정량을 $\tilde{\beta}_i^{EB} = k(\hat{\theta_i}, \hat{\phi})$로 표현할 때, 잭나이프 절차는 다음과 같다. 여기에서 $\phi = (\beta, \sigma_v^2)$은 모형에서의 모수 β와 σ_v^2을 나타낸다.

(1) l번째 지역의 자료 $(\hat{\theta}_l, z_l)$을 제외한 ϕ의 추정량 $\hat{\phi}(l)$을 계산한다. 이때의 EB추정량을 $\tilde{\beta}_i^{EB}(l) = k(\hat{\theta}_i, \hat{\phi}(l))$로 나타내자.

(2) $\hat{M}_{1i} = g_{1i}(\hat{\sigma}_v^2) - \frac{m-1}{m} \sum_{l=1}^{m} [g_{1i}(\hat{\sigma}_v^2(l)) - g_{1i}(\hat{\sigma}_v^2)]^2$을 계산한다.

(3) $\hat{M}_{2i} = \frac{m-1}{m} \sum_{l=1}^{m} [\tilde{\theta}_i^{EB}(l) - \tilde{\theta}_i^{EB}]^2$를 계산한다.

(4) 잭나이프 MSE 추정량 $mse_J(\tilde{\theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i}$를 계산한다.

$\hat{M}_{1i}$은 ϕ가 기지일 때 MSE에 대한 추정량이며, $\hat{M}_{2i}$는 모형 모수 ϕ를 추정할 때 추가적으로 발생하는 MSE에 대한 변화량을 추정한다.(조란, 2003)

2. 소지역 통계 적용 해외 사례 연구

○ 사회가 글로벌 정보화로 발전하면서 좀더 정확하고 시의적절한 통계의 필요성은 더욱 강조되고 있으며 또한 다양한 종류의 통계 생산이 요구되고는 있지만 조사환경은 점점 어려워지고 있는 실정이다. 조사에만 의존하던 통계작성 기법은 개선되어야 하고 국가행정 전산망이 활성화되면서 행정업무나 보고를 통해서 얻어진 정보를 통계화하여 조사에 의해서 생산되는 통계와 상호 보완하여 통계의 정확성을 높일 수 있는 방법이 강구되어야 할 것이다.

○ 이에 따라 각국에서는 경제에 직접적으로 연관되는 경제활동 관련된 노동력과 소득 등에 대한 소지역 통계를 앞에서 상술한 여러 통계 기법에 의해 생산하고 있다. 앞에서 살펴보았듯이 소지역 통계를 직접추정하는 것보다는 보조정보를 이용하여 합성 혹은 복합추정 등의 방법을 통해 분산을 최소화하여 상대효율을 낮추어 가는 방법이 그러한 통계들에 적용되고 있다. 본 연구에서는 소지역 관련 각국의 통계 적용 사례를 살펴보고 이를 통해 우리에게 적용가능한 방향을 연구하고자 한다.

(1) 미 국

○ 1972년 통계국에서 주 단위에 대해서 이용 가능한 노동력, 취업자 및 실업자 추정에 대한 방법과 개념을 연구하기 시작하였고, 1973년에는 통계국이 주관이 되어, 경상인구조사(CPS:Current Population Survey)의 개념, 정의, 추정과 이전의 소책자방법을 결합하여 주 단위와 주의 세부단위까지 관련 통계를 추정할 수 있는 기법을 개발하였다. 1976년 이후에는 모든 주 단위 실업 관련 통계의 추정값에 대한 신뢰도를 높이기 위해서 각 주별로 표본 가구 수를 몇 배씩 증가시켰으며, 이후부터 CPS 자료를 이용한 관련 추정값을 공식적으로 발표하기 위해 실업 관련 추정값에 대한 변동계수의 최대 허용기준을 설정하였다. 1978년부터 규모가 큰 10개주(캘리포니아, 플로리다, 일리노이스, 메사츄셑, 미시간, 뉴저지, 뉴욕, 오하이오, 펜실베니아, 텍사스)와 2개 지역(로스엔젤레스, 뉴욕시)의 노동력 관련 통계는 CPS자료만으로 추정된 결과를 공식 통계로 사용토록 하였다.

○ UI 신청자료의 데이터베이스를 지속적으로 유지하고 개선하였으며, 1976년과 1978년 사이에 걸쳐 UI 신청자료를 모든 주에 대해서 표준화하고, CPS조사의 조사 주 간(매월 12일을 포함한 주)에 실업자로 인정받는 UI 신청자는 자동으로 데이터베이스에 등록되도록 하는 데이터베이스 관리체계를 개발하여 CPS자료와 연계한 추정법을 개발하였다. 1985년에는 1980년 센서스 자료를 근거로 하여 주 단위들에 대한 CPS 표본설계를 완성하였으며, 이 때 추가적으로 노스캐롤라이나 주를 CPS 자료에서 관련 통계를 직접 추정하여 공표 하는 주로 포함시켰고, 표본크기가 충분한 총 11개 대규모 주의 관련 통계에 대한 목표변동계수를 8%로 낮추었다.

❍ CPS는 매월 미국 내의 세부항목들(연령대별, 성별, 인종별 등)에 대한 특성을 파악하기 위해 실시되는 경상인구조사로써 층화 다단확률추출에 의해 추출된 가구단위들에 대해 조사가 이루어진다. CPS의 최종추출단위는 가구조사 단위들로써 약 50,000 조사가구들에 대해 조사가 실시되며, 조사가구들은 총 8개의 패널로 구성되어 표본으로 관리된다. CPS의 표본교체방식은 4-8-4체계를 따른다. 표본으로 추출된 최종 추출단위(가구단위)들은 4개월 간 연속조사가 진행되며, 이후 8개월 간 조사에서 제외되었다가 다시 4개월 간 연속조사가 진행된 후 표본목록에서 영구히 삭제된다. 매달 8개 패널의 표본가구에 거주하는 만 15세 이상의 사람들을 대상으로 조사가 이루어지며, 최종 추정값들은 여러 단계의 추정과정을 통해 작성된다. CPS 자료로부터 전국 단위와 주 단위에 대한 관련 추정값들을 작성하기 위한 개략적인 추정과정은 다음과 같다.

(a) CPS 가중치를 이용하여 관련 추정값들의 불편성조정
(b) 무응답에 대한 보정
(c) 1차 추출단위들(PSUs)의 분산을 줄이기 위한 일단계 비보정
(First-stage ratio adjustment)
(d) CPS 추정값들의 분산을 줄이기 위한 이단계 비보정(Second-stage ratio adjustment)
(e) 추정값들의 분산을 줄이기 위해 전 달의 조사자료를 이용하는 복합추정
(f) 주요한 노동력 통계에 대한 계절요인 조정(Seasonal adjustment)

(2) 캐나다

❍ 캐나다 노동력조사(LFS:Labour Force Survey)는 대규모 노동시장의 변화 양상 및 시의성 있는 노동시장의 정보를 파악하기 위해 2차 세계대전 이후 도입되었고, 주로 주(Province) 지역 및 국가 단위의 고용 및 실업통계를 생산할 목적으로 설계되었다. LFS는 1945년 분기별 조사로 시작하여 1952년 월별 조사로 변경되었고, 1960년부터 캐나다 실업통계를 생산하기 위한 공식조사로 승인되었다. 그 후 LFS를 통해 노동시장의 다양한 통계를 작성할 수 있도록 표본개편 및 조사방법에 관한 연구가 지속적으로 진행되었고, 현재는 캐나다 노동시장의 세부적인 변화에 관한 정보를 제공할 수 있을 정도로 발전을 거듭하였다. 매월 고용인구와 실업인구 총계 및 실업률에 관한 추정치, 노동인구의 특성(연령, 결혼여부, 교육정도, 가족현황) 등에 관한 공식통계는 LFS를 통해 작성된다.

❍ LFS 추정치들은 매월 "Labour Force Information"라는 책자를 통해 공표된다. 또한 노동시장의 좀 더 다양한 정보들은 캐나다 통계국의 전자정보 데이터베이스의 일종인 “CANSIM"을 통해 획득할 수 있으며, LFS의 결과로부터 매월 9000 항목 이상의 시계열 자료들이 정기적으로 수정 보완된다. 이외에 노동시장의 중심지표가 되는 다양한 주제에 대한 세부적인 고찰을 다룬 ”Labour Force

Update"가 1997년부터 계간지로써 출간되고 있으며, 1976년 이래로 최근까지의 방대한 시계열 자료(time series data) 및 횡단면 자료(cross-sectional data)를 포함하고 있는 "Labour Force Historical Review on CD-ROM"이 매년 제작되고 있다.

○ 취업통계의 추정값들에는 인구학적 특성, 산업과 업종, 정규직과 통상적인 근로시간 등이 포함되어 있으며 설문내용에는 비자발적 부업적 취업, 복수 직업 여부와 휴직 등에 대해서 분석할 수 있는 주제들이 포함되어 있다. 특히 1997년 이후에는 근로자들의 노조가입 여부와 임금수준에 대한 정보와 작업장의 근로자 수 및 직업의 정규직 또는 임시직 여부에 대한 정보를 제공하고 있다. 실업통계의 추정값은 인구학적 범주별, 실업기간, 구직활동 전의 활동 및 바로 이전 직장에서 이직한 이유 등에 대한 정보를 제공하고 있다. 노동력 조사에 의해서 발표되는 통계는 국가 단위와 주 단위 추정값이 핵심적인 내용이지만 경제구역(ER : Economic Region)과 센서스 도시지역(CMA ; Census Metropolitan Area)과 같은 소지역 단위에 대한 노동력 상태의 추정값을 제공하고 있다.

(3) 프랑스

○ 노동력조사로부터 국가 수준의 노동 통계는 일년에 한 번씩 발표되고 있으며, 부차관심영역에서 실업 통계를 생산할 때 일정한 오차 범위 내로 기준을 충족하도록 하는 소지역 추정법에 관한 연구가 진행되고 있다. 프랑스의 표본 설계는 지역별 층화를 기준으로 하였으나 실업자 수 또는 실업률 등의 집적된 자료의 이용 시에는 무응답을 보정하여 국가 수준으로 통계를 작성한 후 성별-연령대별로 분할했으므로 지역 통계의 신뢰성에 대해서 유의해야 한다.

○ 프랑스 통계청에서 취업과 실업에 관한 지역 통계의 생산을 위해서 노동력 조사와 센서스 뿐 만 아니라 실업 보험 신청자료 등 행정 업무 자료를 이용하는 체계를 발전시켜왔으나 소지역이나 세분화된 범주의 통계 작성은 미흡하다. 특히 취업 통계는 연말에 각 회사로부터 취업자 수에 대한 자료를 수집하고 있으나 전년말 기준으로 변화율에 대한 자료로 활용하여 국가 수준에서 비율에 대한 통계를 작성하고 있으며 실업자의 수에 대한 직접 통계를 작성하지는 않는다. 그래서 취업자 통계는 실업 보험과 센서스 자료 등이 핵심적인 역할을 하고 있다. 취업 통계는 해당 익년에 이용가능하다.

○ 실업 통계는 분기별로 지역 통계와 국가 통계를 동시에 발표하며, 주로 이용되는 자료는 노동력 조사와 구직 등록자의 수이다. ILO 기준의 실업자와 실제 구직 등록자 및 노동력 조사에서 실업자와 차이를 반영하기 위해서 노동력 조사에서 추정된 ILO 실업자와 구직 등록자 수의 국가적 비율을 추정하여 활용하고 있다.

○ 경제활동인구는 실업자와 취업자를 합해서 추정하지만 취업자는 직장을 갖고 있는 사람과 가사를 돌보는 사람을 합산해야 하므로 매 해마다 연말에 외삽법으로 조정하여 분기별 통계를 수정하여 시

계열 통계로 관리한다. 취업자 통계에서 문제점은 취업자 통계를 작성 시 연말에 센서스 자료를 이용하고 있으나 센서스 간격이 너무 길어서 인구 변동과 상황 변화를 제대로 반영할 수 없다는 점이다 (센서스:1968, 1975, 1982, 1990, 1999). 특히, 지역 취업 통계 작성 시에는 더 심각해질 수 있다. 취업 통계 작성의 취약점을 보완하기 위해서 "ESTEL"이라는 프로젝트가 진행 중에 있으며 1996년 통계를 시험 중에 있다. 이 프로젝트의 특징은 매년 말을 기준으로 조사자료 및 행정자료를 기반으로 전국 및 지역 단위 취업 통계의 질과 신뢰도를 제고하는데 있다. 센서스 자료는 센서스 해에만 적용하고 다른 해에는 연말의 행정 업무와 조사 자료를 기준으로 한다는 것이 큰 특징이다.

Ⅱ-6. 토의

○ 본 보고서는 현재 제공되고 있는 인총 MD 제공 기준에 대한 타당성을 검토해보고 새로운 노출제한의 기법을 찾고자하는 목적에서 연구를 진행하였다.

○ 특히, 통계청 뿐만 아니라 대용량의 조사나 관측에 의한 데이터 수집이 증가하여 MD 제공시 MD와 외부기관의 데이터가 병합되어 정보가 노출되는 위험을 방지하기 노력이 어느 때보다 절실하다고 할 수 있다. 이런 이유로 본 연구에서는 외부기관 데이터에서 수집가능한 변수들 중 MD와 결합가능한 키변수를 중심으로 연구를 진행하였다.

○ 기존 문헌에서 살펴볼 수 있는 노출제한 연구는 정보의 '대체'를 통한 노출방지 자체에 주안점을 두고 진행되었다. 자료교환, 그룹화나 반올림, 데이터 요약 방법 등이 그러하다. 본 연구도 그러한 틀에서 크게 벗어나지는 못했지만, 이전 연구에서 이론적으로만 확장가능하던 많은 변수에 대해 실증적인 방법을 제시하고자 하였다.

○ 실제 노출에 이용가능한 키변수가 10개 넘고, 대부분이 범주형 변수인 경우 각 변수를 조합한 셀의 개수는 기하급수적으로 늘어 인총과 같이 많은 관측치를 포함한 데이터라도 노출제한의 비율을 낮추기는 대단히 어렵다고 할 수 있다. 반면, 노출을 제한하기 위해 변수의 줄인다면 분석자의 만족도가 낮아져 서비스 제공의 의미가 반감된다.

○ 이러한 문제를 해결하기 위해, 즉 키변수의 개수를 적당히 유지하며 노출건수를 줄이기 위해 본 연구에서는 카이제곱 검정통계량과 로그선형모형을 반복적으로 연산하여 키변수의 범주를 합병해 나가는 알고리즘을 제안하였다. 이 방법은 수리적으로 해결하기 힘든 정보노출의 문제를 기존의 그룹화 방법을 변형, 발전시켜 컴퓨터 연산으로 해결하는 방법이라고 할 수 있다. 비록 많은 연산을 요구하기는 하나 노출비율을 조금이라도 낮추며, 이를 통해 추가적인 변수를 제공할 수 있는 방법이라 사료된다.

○ 제공 MD의 적정 표본크기와 표본추출에 대한 연구를 부가하였다. 표본추출에 대한 연구는 데이터상의 노출비율을 계산하여 이를 비교하였다. 본 연구에서는 계통과 집락 추출을 비교하였으나, 이외에도 많은 추출기법과 그에 따른 효율 및 노출정도를 추가적으로 연구되어야할 영역이라 판단된다.

○ MD 제공의 적정 표본 비율을 정보노출과 관련하여 고찰하였으며, 본 연구를 통해 적은 크기의 MD 제공은 정보노출의 위험이 크니 가급적 표본 크기를 늘려 제공하기를 추천하였다.

부 록

인구 부문 유일성 비교: 추출기법, %구간, 노출제한기법 적용 전후

<표 A-1> 계통추출 1%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	444,797	22,113	4.97	5,256	1.18
서울특별시	87,644	4,120	4.70	542	0.62
부산광역시	31,879	1,193	3.74	291	0.91
대구광역시	21,985	992	4.51	237	1.08
인천광역시	23,519	1,455	6.19	365	1.55
광주광역시	13,023	676	5.19	185	1.42
대전광역시	12,944	892	6.89	291	2.25
울산광역시	9,088	483	5.31	190	2.09
경기도	93,013	5,038	5.42	681	0.73
강원도	15,120	808	5.34	328	2.17
충청북도	14,840	795	5.36	325	2.19
충청남도	18,892	1,187	6.28	374	1.98
전라북도	19,685	675	3.43	305	1.55
전라남도	20,827	750	3.60	293	1.41
경상북도	27,418	1,288	4.70	372	1.36
경상남도	29,896	1,485	4.97	310	1.04
제주도	5,024	276	5.49	167	3.32

<표 A-2> 계통추출 1%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	444,797	10,754	2.42	5,256	1.18	6,158	1.38
서울특별시	87,644	1,487	1.70	542	0.62	950	1.08
부산광역시	31,879	570	1.79	291	0.91	309	0.97
대구광역시	21,985	473	2.15	237	1.08	273	1.24
인천광역시	23,519	754	3.21	365	1.55	410	1.74
광주광역시	13,023	393	3.02	185	1.42	192	1.47
대전광역시	12,944	529	4.09	291	2.25	312	2.41
울산광역시	9,088	303	3.33	190	2.09	191	2.10
경기도	93,013	1,781	1.91	681	0.73	1,084	1.17
강원도	15,120	516	3.41	328	2.17	316	2.09
충청북도	14,840	551	3.71	325	2.19	290	1.95
충청남도	18,892	727	3.85	374	1.98	393	2.08
전라북도	19,685	493	2.50	305	1.55	270	1.37
전라남도	20,827	510	2.45	293	1.41	248	1.19
경상북도	27,418	712	2.60	372	1.36	406	1.48
경상남도	29,896	742	2.48	310	1.04	401	1.34
제주도	5,024	213	4.24	167	3.32	113	2.25

<표 A-3> 계통추출 2%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	891,174	36,980	4.15	6,918	0.78
서울특별시	175,561	6,503	3.70	698	0.40
부산광역시	63,874	1,987	3.11	365	0.57
대구광역시	44,029	1,700	3.86	332	0.75
인천광역시	47,210	2,480	5.25	447	0.95
광주광역시	26,102	1,229	4.71	247	0.95
대전광역시	25,738	1,588	6.17	370	1.44
울산광역시	18,298	911	4.98	274	1.50
경기도	186,520	7,714	4.14	794	0.43
강원도	30,420	1,475	4.85	473	1.55
충청북도	29,702	1,428	4.81	455	1.53
충청남도	37,903	2,100	5.54	508	1.34
전라북도	39,330	1,206	3.07	380	0.97
전라남도	41,860	1,413	3.38	402	0.96
경상북도	54,853	2,267	4.13	523	0.95
경상남도	59,746	2,476	4.14	422	0.71
제주도	10,028	503	5.02	228	2.27

<표 A-4> 계통추출 2%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	891,174	16,274	1.83	6,918	0.78	9,186	1.03
서울특별시	175,561	2,068	1.18	698	0.40	1,286	0.73
부산광역시	63,874	893	1.40	365	0.57	446	0.70
대구광역시	44,029	757	1.72	332	0.75	392	0.89
인천광역시	47,210	1,153	2.44	447	0.95	594	1.26
광주광역시	26,102	583	2.23	247	0.95	300	1.15
대전광역시	25,738	764	2.97	370	1.44	478	1.86
울산광역시	18,298	526	2.87	274	1.50	304	1.66
경기도	186,520	2,432	1.30	794	0.43	1,513	0.81
강원도	30,420	836	2.75	473	1.55	515	1.69
충청북도	29,702	917	3.09	455	1.53	462	1.56
충청남도	37,903	1,120	2.95	508	1.34	575	1.52
전라북도	39,330	744	1.89	380	0.97	418	1.06
전라남도	41,860	816	1.95	402	0.96	418	1.00
경상북도	54,853	1,157	2.11	523	0.95	675	1.23
경상남도	59,746	1,160	1.94	422	0.71	605	1.01
제주도	10,028	348	3.47	228	2.27	205	2.04

<표 A-5> 계통추출 3%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,335,380	48,916	3.66	8,239	0.62
서울특별시	263,519	8,233	3.12	767	0.29
부산광역시	95,763	2,736	2.86	501	0.52
대구광역시	66,359	2,259	3.40	408	0.61
인천광역시	70,528	3,121	4.43	535	0.76
광주광역시	38,965	1,619	4.16	302	0.78
대전광역시	38,631	2,135	5.53	415	1.07
울산광역시	27,471	1,275	4.64	315	1.15
경기도	279,179	9,654	3.46	904	0.32
강원도	45,660	2,087	4.57	593	1.30
충청북도	44,519	1,942	4.36	478	1.07
충청남도	56,696	3,025	5.34	583	1.03
전라북도	59,258	1,741	2.94	476	0.80
전라남도	62,481	1,874	3.00	495	0.79
경상북도	82,071	3,125	3.81	584	0.71
경상남도	89,294	3,311	3.71	506	0.57
제주도	14,986	779	5.20	377	2.52

<표 A-6> 계통추출 3%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,335,380	20,336	1.52	8,239	0.62	11,532	0.86
서울특별시	263,519	2,503	0.95	767	0.29	1,527	0.58
부산광역시	95,763	1,202	1.26	501	0.52	671	0.70
대구광역시	66,359	986	1.49	408	0.61	541	0.82
인천광역시	70,528	1,368	1.94	535	0.76	747	1.06
광주광역시	38,965	706	1.81	302	0.78	372	0.95
대전광역시	38,631	939	2.43	415	1.07	538	1.39
울산광역시	27,471	680	2.48	315	1.15	382	1.39
경기도	279,179	2,908	1.04	904	0.32	1,702	0.61
강원도	45,660	1,096	2.40	593	1.30	662	1.45
충청북도	44,519	1,080	2.43	478	1.07	596	1.34
충청남도	56,696	1,455	2.57	583	1.03	765	1.35
전라북도	59,258	996	1.68	476	0.80	575	0.97
전라남도	62,481	1,032	1.65	495	0.79	543	0.87
경상북도	82,071	1,453	1.77	584	0.71	819	1.00
경상남도	89,294	1,394	1.56	506	0.57	764	0.86
제주도	14,986	538	3.59	377	2.52	328	2.19

<표 A-7> 계통추출 4%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,782,287	59,828	3.36	8,932	0.50
서울특별시	351,607	9,896	2.81	864	0.25
부산광역시	127,904	3,207	2.51	493	0.39
대구광역시	88,184	2,749	3.12	426	0.48
인천광역시	94,507	3,980	4.21	574	0.61
광주광역시	52,011	2,084	4.01	329	0.63
대전광역시	51,416	2,693	5.24	430	0.84
울산광역시	36,764	1,582	4.30	355	0.97
경기도	372,552	11,509	3.09	931	0.25
강원도	60,953	2,578	4.23	610	1.00
충청북도	59,182	2,531	4.28	587	0.99
충청남도	75,854	3,633	4.79	631	0.83
전라북도	78,712	2,117	2.69	547	0.69
전라남도	83,526	2,436	2.92	541	0.65
경상북도	109,551	3,803	3.47	674	0.62
경상남도	119,480	4,061	3.40	573	0.48
제주도	20,084	969	4.82	367	1.83

<표 A-8> 계통추출 4%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,782,287	23,861	1.34	8,932	0.50	13,326	0.75
서울특별시	351,607	2,840	0.81	864	0.25	1,709	0.49
부산광역시	127,904	1,285	1.00	493	0.39	685	0.54
대구광역시	88,184	1,087	1.23	426	0.48	590	0.67
인천광역시	94,507	1,651	1.75	574	0.61	858	0.91
광주광역시	52,011	896	1.72	329	0.63	428	0.82
대전광역시	51,416	1,100	2.14	430	0.84	642	1.25
울산광역시	36,764	813	2.21	355	0.97	436	1.19
경기도	372,552	3,313	0.89	931	0.25	1,966	0.53
강원도	60,953	1,324	2.17	610	1.00	795	1.30
충청북도	59,182	1,342	2.27	587	0.99	726	1.23
충청남도	75,854	1,664	2.19	631	0.83	913	1.20
전라북도	78,712	1,207	1.53	547	0.69	683	0.87
전라남도	83,526	1,289	1.54	541	0.65	633	0.76
경상북도	109,551	1,710	1.56	674	0.62	956	0.87
경상남도	119,480	1,719	1.44	573	0.48	924	0.77
제주도	20,084	621	3.09	367	1.83	382	1.90

<표 A-9> 계통추출 5%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,229,053	69,298	3.11	9,644	0.43
서울특별시	439,727	11,260	2.56	878	0.20
부산광역시	160,134	3,829	2.39	542	0.34
대구광역시	111,136	3,246	2.92	432	0.39
인천광역시	117,703	4,458	3.79	632	0.54
광주광역시	64,815	2,364	3.65	369	0.57
대전광역시	64,544	3,136	4.86	515	0.80
울산광역시	45,707	1,946	4.26	364	0.80
경기도	466,264	12,930	2.77	942	0.20
강원도	76,261	3,120	4.09	720	0.94
충청북도	73,989	2,974	4.02	607	0.82
충청남도	94,548	4,340	4.59	687	0.73
전라북도	98,719	2,515	2.55	607	0.61
전라남도	104,421	2,802	2.68	601	0.58
경상북도	136,878	4,406	3.22	721	0.53
경상남도	149,259	4,730	3.17	586	0.39
제주도	24,948	1,242	4.98	441	1.77

<표 A-10> 계통추출 5%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,229,053	26,729	1.20	9,644	0.43	14,831	0.67
서울특별시	439,727	3,142	0.71	878	0.20	1,831	0.42
부산광역시	160,134	1,551	0.97	542	0.34	875	0.55
대구광역시	111,136	1,256	1.13	432	0.39	671	0.60
인천광역시	117,703	1,791	1.52	632	0.54	917	0.78
광주광역시	64,815	965	1.49	369	0.57	494	0.76
대전광역시	64,544	1,297	2.01	515	0.80	717	1.11
울산광역시	45,707	915	2.00	364	0.80	469	1.03
경기도	466,264	3,564	0.76	942	0.20	2,072	0.44
강원도	76,261	1,558	2.04	720	0.94	895	1.17
충청북도	73,989	1,460	1.97	607	0.82	823	1.11
충청남도	94,548	1,884	1.99	687	0.73	1,088	1.15
전라북도	98,719	1,334	1.35	607	0.61	801	0.81
전라남도	104,421	1,454	1.39	601	0.58	700	0.67
경상북도	136,878	1,939	1.42	721	0.53	1,027	0.75
경상남도	149,259	1,864	1.25	586	0.39	966	0.65
제주도	24,948	755	3.03	441	1.77	485	1.94

<표 A-11> 계통추출 6%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,672,676	78,223	2.93	10,269	0.38
서울특별시	527,170	12,357	2.34	898	0.17
부산광역시	191,906	4,270	2.23	622	0.32
대구광역시	132,928	3,594	2.70	520	0.39
인천광역시	141,273	5,112	3.62	637	0.45
광주광역시	77,913	2,704	3.47	419	0.54
대전광역시	77,318	3,646	4.72	537	0.69
울산광역시	54,919	2,206	4.02	407	0.74
경기도	559,217	14,322	2.56	998	0.18
강원도	91,327	3,644	3.99	732	0.80
충청북도	88,988	3,454	3.88	629	0.71
충청남도	113,288	4,978	4.39	727	0.64
전라북도	118,326	2,909	2.46	599	0.51
전라남도	125,475	3,330	2.65	647	0.52
경상북도	164,218	5,033	3.06	748	0.46
경상남도	178,530	5,263	2.95	661	0.37
제주도	29,880	1,401	4.69	488	1.63

<표 A-12> 계통추출 6%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,672,676	29,495	1.10	10,269	0.38	16,323	0.61
서울특별시	527,170	3,359	0.64	898	0.17	1,996	0.38
부산광역시	191,906	1,747	0.91	622	0.32	949	0.49
대구광역시	132,928	1,413	1.06	520	0.39	766	0.58
인천광역시	141,273	1,932	1.37	637	0.45	1,027	0.73
광주광역시	77,913	1,098	1.41	419	0.54	534	0.69
대전광역시	77,318	1,441	1.86	537	0.69	799	1.03
울산광역시	54,919	1,047	1.91	407	0.74	545	0.99
경기도	559,217	3,898	0.70	998	0.18	2,255	0.40
강원도	91,327	1,676	1.84	732	0.80	989	1.08
충청북도	88,988	1,669	1.88	629	0.71	888	1.00
충청남도	113,288	2,099	1.85	727	0.64	1,111	0.98
전라북도	118,326	1,440	1.22	599	0.51	846	0.71
전라남도	125,475	1,648	1.31	647	0.52	814	0.65
경상북도	164,218	2,142	1.30	748	0.46	1,158	0.71
경상남도	178,530	2,030	1.14	661	0.37	1,131	0.63
제주도	29,880	856	2.86	488	1.63	515	1.72

<표 A-13> 계통추출 7%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,119,951	86,707	2.78	10,548	0.34
서울특별시	615,590	13,568	2.20	925	0.15
부산광역시	224,191	4,733	2.11	575	0.26
대구광역시	155,140	3,954	2.55	477	0.31
인천광역시	165,026	5,745	3.48	686	0.42
광주광역시	90,837	3,070	3.38	429	0.47
대전광역시	90,255	4,012	4.45	560	0.62
울산광역시	64,078	2,439	3.81	422	0.66
경기도	652,589	15,693	2.40	1,025	0.16
강원도	106,517	3,997	3.75	746	0.70
충청북도	103,550	3,934	3.80	676	0.65
충청남도	132,362	5,510	4.16	760	0.57
전라북도	137,995	3,263	2.36	682	0.49
전라남도	146,386	3,794	2.59	667	0.46
경상북도	191,547	5,546	2.90	785	0.41
경상남도	208,886	5,842	2.80	664	0.32
제주도	35,002	1,607	4.59	469	1.34

<표 A-14> 계통추출 7%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,119,951	31,785	1.02	10,548	0.34	17,345	0.56
서울특별시	615,590	3,573	0.58	925	0.15	2,016	0.33
부산광역시	224,191	1,741	0.78	575	0.26	949	0.42
대구광역시	155,140	1,454	0.94	477	0.31	775	0.50
인천광역시	165,026	2,091	1.27	686	0.42	1,095	0.66
광주광역시	90,837	1,209	1.33	429	0.47	591	0.65
대전광역시	90,255	1,542	1.71	560	0.62	861	0.95
울산광역시	64,078	1,108	1.73	422	0.66	559	0.87
경기도	652,589	4,191	0.64	1,025	0.16	2,326	0.36
강원도	106,517	1,832	1.72	746	0.70	1,071	1.01
충청북도	103,550	1,815	1.75	676	0.65	988	0.95
충청남도	132,362	2,228	1.68	760	0.57	1,218	0.92
전라북도	137,995	1,631	1.18	682	0.49	928	0.67
전라남도	146,386	1,829	1.25	667	0.46	888	0.61
경상북도	191,547	2,336	1.22	785	0.41	1,268	0.66
경상남도	208,886	2,275	1.09	664	0.32	1,209	0.58
제주도	35,002	930	2.66	469	1.34	603	1.72

<표 A-15> 계통추출 8%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,563,789	94,209	2.64	11,191	0.31
서울특별시	703,216	14,519	2.06	964	0.14
부산광역시	255,936	5,204	2.03	637	0.25
대구광역시	177,083	4,335	2.45	550	0.31
인천광역시	188,570	6,158	3.27	708	0.38
광주광역시	103,822	3,319	3.20	436	0.42
대전광역시	102,996	4,416	4.29	588	0.57
울산광역시	73,385	2,724	3.71	448	0.61
경기도	745,249	16,683	2.24	1,056	0.14
강원도	121,860	4,463	3.66	791	0.65
충청북도	118,468	4,279	3.61	710	0.60
충청남도	151,239	6,164	4.08	815	0.54
전라북도	157,708	3,600	2.28	700	0.44
전라남도	167,141	4,129	2.47	705	0.42
경상북도	218,916	6,049	2.76	820	0.37
경상남도	238,264	6,378	2.68	720	0.30
제주도	39,936	1,789	4.48	543	1.36

<표 A-16> 계통추출 8%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,563,789	34,036	0.96	11,191	0.31	18,737	0.53
서울특별시	703,216	3,785	0.54	964	0.14	2,163	0.31
부산광역시	255,936	1,958	0.77	637	0.25	1,086	0.42
대구광역시	177,083	1,602	0.90	550	0.31	890	0.50
인천광역시	188,570	2,207	1.17	708	0.38	1,173	0.62
광주광역시	103,822	1,282	1.23	436	0.42	623	0.60
대전광역시	102,996	1,647	1.60	588	0.57	881	0.86
울산광역시	73,385	1,205	1.64	448	0.61	642	0.87
경기도	745,249	4,364	0.59	1,056	0.14	2,484	0.33
강원도	121,860	1,976	1.62	791	0.65	1,145	0.94
충청북도	118,468	1,944	1.64	710	0.60	1,031	0.87
충청남도	151,239	2,421	1.60	815	0.54	1,312	0.87
전라북도	157,708	1,762	1.12	700	0.44	1,032	0.65
전라남도	167,141	1,950	1.17	705	0.42	962	0.58
경상북도	218,916	2,496	1.14	820	0.37	1,348	0.62
경상남도	238,264	2,426	1.02	720	0.30	1,322	0.55
제주도	39,936	1,011	2.53	543	1.36	643	1.61

<표 A-17> 계통추출 9%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	4,010,534	101,353	2.53	11,419	0.28
서울특별시	791,465	15,547	1.96	971	0.12
부산광역시	288,075	5,587	1.94	630	0.22
대구광역시	199,514	4,666	2.34	542	0.27
인천광역시	212,035	6,606	3.12	728	0.34
광주광역시	116,779	3,578	3.06	460	0.39
대전광역시	115,942	4,737	4.09	617	0.53
울산광역시	82,461	2,962	3.59	456	0.55
경기도	838,755	17,752	2.12	1,058	0.13
강원도	137,057	4,790	3.49	840	0.61
충청북도	133,229	4,667	3.50	733	0.55
충청남도	170,166	6,621	3.89	826	0.49
전라북도	177,568	3,940	2.22	736	0.41
전라남도	188,040	4,521	2.40	704	0.37
경상북도	246,200	6,530	2.65	849	0.34
경상남도	268,284	6,846	2.55	716	0.27
제주도	44,964	2,003	4.45	553	1.23

<표 A-18> 계통추출 9%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	4,010,534	35,996	0.90	11,419	0.28	19,653	0.49
서울특별시	791,465	3,972	0.50	971	0.12	2,256	0.29
부산광역시	288,075	2,051	0.71	630	0.22	1,131	0.39
대구광역시	199,514	1,672	0.84	542	0.27	914	0.46
인천광역시	212,035	2,303	1.09	728	0.34	1,220	0.58
광주광역시	116,779	1,353	1.16	460	0.39	658	0.56
대전광역시	115,942	1,759	1.52	617	0.53	938	0.81
울산광역시	82,461	1,265	1.53	456	0.55	668	0.81
경기도	838,755	4,585	0.55	1,058	0.13	2,552	0.30
강원도	137,057	2,117	1.54	840	0.61	1,220	0.89
충청북도	133,229	2,069	1.55	733	0.55	1,106	0.83
충청남도	170,166	2,556	1.50	826	0.49	1,382	0.81
전라북도	177,568	1,897	1.07	736	0.41	1,100	0.62
전라남도	188,040	2,080	1.11	704	0.37	1,026	0.55
경상북도	246,200	2,664	1.08	849	0.34	1,405	0.57
경상남도	268,284	2,549	0.95	716	0.27	1,360	0.51
제주도	44,964	1,104	2.46	553	1.23	717	1.59

<표 A-19> 집락추출 1%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	445,554	22,181	4.98	5,185	1.16
서울특별시	88,561	4,134	4.67	646	0.73
부산광역시	31,890	1,200	3.76	270	0.85
대구광역시	22,032	1,029	4.67	241	1.09
인천광역시	23,839	1,379	5.78	348	1.46
광주광역시	12,953	709	5.47	197	1.52
대전광역시	12,739	894	7.02	249	1.95
울산광역시	9,339	519	5.56	198	2.12
경기도	92,592	5,086	5.49	638	0.69
강원도	15,290	785	5.13	315	2.06
충청북도	14,390	729	5.07	284	1.97
충청남도	19,006	1,221	6.42	369	1.94
전라북도	19,679	626	3.18	281	1.43
전라남도	20,916	767	3.67	257	1.23
경상북도	27,706	1,287	4.65	365	1.32
경상남도	29,436	1,516	5.15	341	1.16
제주도	5,186	300	5.78	186	3.59

<표 A-20> 집락추출 1%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	445,554	10,816	2.43	5,185	1.16	6,165	1.38
서울특별시	88,561	1,499	1.69	646	0.73	984	1.11
부산광역시	31,890	622	1.95	270	0.85	341	1.07
대구광역시	22,032	519	2.36	241	1.09	273	1.24
인천광역시	23,839	721	3.02	348	1.46	391	1.64
광주광역시	12,953	406	3.13	197	1.52	186	1.44
대전광역시	12,739	516	4.05	249	1.95	280	2.20
울산광역시	9,339	353	3.78	198	2.12	184	1.97
경기도	92,592	1,806	1.95	638	0.69	1,113	1.20
강원도	15,290	488	3.19	315	2.06	305	1.99
충청북도	14,390	490	3.41	284	1.97	251	1.74
충청남도	19,006	763	4.01	369	1.94	411	2.16
전라북도	19,679	438	2.23	281	1.43	247	1.26
전라남도	20,916	492	2.35	257	1.23	261	1.25
경상북도	27,706	713	2.57	365	1.32	417	1.51
경상남도	29,436	755	2.56	341	1.16	395	1.34
제주도	5,186	235	4.53	186	3.59	126	2.43

<표 A-21> 집락추출 2%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	890,692	36,914	4.14	6,932	0.78
서울특별시	175,218	6,486	3.70	737	0.42
부산광역시	64,167	1,975	3.08	364	0.57
대구광역시	44,749	1,741	3.89	310	0.69
인천광역시	47,225	2,371	5.02	457	0.97
광주광역시	26,007	1,215	4.67	289	1.11
대전광역시	25,486	1,583	6.21	366	1.44
울산광역시	18,397	936	5.09	241	1.31
경기도	185,149	7,851	4.24	776	0.42
강원도	30,366	1,511	4.98	485	1.60
충청북도	29,661	1,353	4.56	442	1.49
충청남도	37,430	2,146	5.73	499	1.33
전라북도	40,260	1,169	2.90	371	0.92
전라남도	41,931	1,375	3.28	399	0.95
경상북도	54,769	2,148	3.92	462	0.84
경상남도	59,843	2,530	4.23	470	0.79
제주도	10,034	524	5.22	264	2.63

<표 A-22> 집락추출 2%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	890,692	16,300	1.83	6,932	0.78	9,159	1.03
서울특별시	175,218	2,080	1.19	737	0.42	1,243	0.71
부산광역시	64,167	935	1.46	364	0.57	485	0.76
대구광역시	44,749	747	1.67	310	0.69	430	0.96
인천광역시	47,225	1,099	2.33	457	0.97	582	1.23
광주광역시	26,007	596	2.29	289	1.11	293	1.13
대전광역시	25,486	805	3.16	366	1.44	459	1.80
울산광역시	18,397	530	2.88	241	1.31	270	1.47
경기도	185,149	2,548	1.38	776	0.42	1,499	0.81
강원도	30,366	875	2.88	485	1.60	547	1.80
충청북도	29,661	816	2.75	442	1.49	459	1.55
충청남도	37,430	1,085	2.90	499	1.33	610	1.63
전라북도	40,260	756	1.88	371	0.92	398	0.99
전라남도	41,931	817	1.95	399	0.95	432	1.03
경상북도	54,769	1,075	1.96	462	0.84	594	1.08
경상남도	59,843	1,167	1.95	470	0.79	619	1.03
제주도	10,034	369	3.68	264	2.63	239	2.38

<표 A-23> 집락추출 3%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,335,799	49,217	3.68	8,050	0.60
서울특별시	263,650	8,376	3.18	754	0.29
부산광역시	95,810	2,751	2.87	458	0.48
대구광역시	67,037	2,408	3.59	366	0.55
인천광역시	70,698	3,216	4.55	510	0.72
광주광역시	39,205	1,637	4.18	303	0.77
대전광역시	38,916	2,161	5.55	428	1.10
울산광역시	27,111	1,286	4.74	335	1.24
경기도	278,728	9,860	3.54	847	0.30
강원도	45,215	2,011	4.45	547	1.21
충청북도	44,311	1,948	4.40	481	1.09
충청남도	56,572	2,918	5.16	623	1.10
전라북도	59,548	1,682	2.82	490	0.82
전라남도	62,830	1,900	3.02	479	0.76
경상북도	81,743	2,947	3.61	552	0.68
경상남도	89,427	3,320	3.71	530	0.59
제주도	14,998	796	5.31	347	2.31

<표 A-24> 집락추출 3%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,335,799	20,299	1.52	8,050	0.60	11,364	0.85
서울특별시	263,650	2,454	0.93	754	0.29	1,506	0.57
부산광역시	95,810	1,163	1.21	458	0.48	641	0.67
대구광역시	67,037	928	1.38	366	0.55	544	0.81
인천광역시	70,698	1,395	1.97	510	0.72	721	1.02
광주광역시	39,205	738	1.88	303	0.77	392	1.00
대전광역시	38,916	983	2.53	428	1.10	519	1.33
울산광역시	27,111	717	2.64	335	1.24	376	1.39
경기도	278,728	2,863	1.03	847	0.30	1,751	0.63
강원도	45,215	1,117	2.47	547	1.21	653	1.44
충청북도	44,311	1,060	2.39	481	1.09	562	1.27
충청남도	56,572	1,495	2.64	623	1.10	805	1.42
전라북도	59,548	930	1.56	490	0.82	558	0.94
전라남도	62,830	1,085	1.73	479	0.76	549	0.87
경상북도	81,743	1,367	1.67	552	0.68	714	0.87
경상남도	89,427	1,494	1.67	530	0.59	767	0.86
제주도	14,998	510	3.40	347	2.31	306	2.04

<표 A-25> 집락추출 4%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,781,084	59,984	3.37	8,987	0.50
서울특별시	350,274	9,822	2.80	830	0.24
부산광역시	128,372	3,343	2.60	541	0.42
대구광역시	88,588	2,744	3.10	420	0.47
인천광역시	94,631	3,970	4.20	588	0.62
광주광역시	51,833	2,011	3.88	343	0.66
대전광역시	51,187	2,706	5.29	441	0.86
울산광역시	36,929	1,649	4.47	352	0.95
경기도	373,092	11,513	3.09	930	0.25
강원도	61,210	2,637	4.31	636	1.04
충청북도	59,561	2,486	4.17	583	0.98
충청남도	75,890	3,821	5.03	655	0.86
전라북도	78,082	2,116	2.71	539	0.69
전라남도	83,404	2,420	2.90	545	0.65
경상북도	108,865	3,713	3.41	639	0.59
경상남도	119,671	4,051	3.39	572	0.48
제주도	19,495	982	5.04	373	1.91

<표 A-26> 집락추출 4%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,781,084	23,813	1.34	8,987	0.50	13,451	0.76
서울특별시	350,274	2,815	0.80	830	0.24	1,728	0.49
부산광역시	128,372	1,424	1.11	541	0.42	746	0.58
대구광역시	88,588	1,099	1.24	420	0.47	608	0.69
인천광역시	94,631	1,649	1.74	588	0.62	889	0.94
광주광역시	51,833	855	1.65	343	0.66	406	0.78
대전광역시	51,187	1,149	2.24	441	0.86	627	1.22
울산광역시	36,929	824	2.23	352	0.95	448	1.21
경기도	373,092	3,303	0.89	930	0.25	2,004	0.54
강원도	61,210	1,339	2.19	636	1.04	782	1.28
충청북도	59,561	1,318	2.21	583	0.98	708	1.19
충청남도	75,890	1,660	2.19	655	0.86	974	1.28
전라북도	78,082	1,162	1.49	539	0.69	665	0.85
전라남도	83,404	1,313	1.57	545	0.65	635	0.76
경상북도	108,865	1,630	1.50	639	0.59	915	0.84
경상남도	119,671	1,646	1.38	572	0.48	909	0.76
제주도	19,495	627	3.22	373	1.91	407	2.09

<표 A-27> 집락추출 5%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,228,092	69,468	3.12	9,561	0.43
서울특별시	439,283	11,311	2.57	925	0.21
부산광역시	159,892	3,764	2.35	540	0.34
대구광역시	110,473	3,148	2.85	448	0.41
인천광역시	117,975	4,582	3.88	578	0.49
광주광역시	65,098	2,389	3.67	369	0.57
대전광역시	64,134	3,189	4.97	500	0.78
울산광역시	45,559	1,931	4.24	377	0.83
경기도	467,283	12,948	2.77	929	0.20
강원도	76,241	3,153	4.14	688	0.90
충청북도	74,716	2,985	4.00	605	0.81
충청남도	94,296	4,392	4.66	681	0.72
전라북도	97,980	2,556	2.61	600	0.61
전라남도	104,763	2,974	2.84	614	0.59
경상북도	136,938	4,393	3.21	684	0.50
경상남도	149,192	4,656	3.12	611	0.41
제주도	24,269	1,097	4.52	412	1.70

<표 A-28> 집락추출 5%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,228,092	26,678	1.20	9,561	0.43	14,839	0.67
서울특별시	439,283	3,206	0.73	925	0.21	1,851	0.42
부산광역시	159,892	1,517	0.95	540	0.34	862	0.54
대구광역시	110,473	1,234	1.12	448	0.41	676	0.61
인천광역시	117,975	1,767	1.50	578	0.49	922	0.78
광주광역시	65,098	984	1.51	369	0.57	482	0.74
대전광역시	64,134	1,275	1.99	500	0.78	684	1.07
울산광역시	45,559	900	1.98	377	0.83	481	1.06
경기도	467,283	3,570	0.76	929	0.20	2,096	0.45
강원도	76,241	1,499	1.97	688	0.90	929	1.22
충청북도	74,716	1,488	1.99	605	0.81	826	1.11
충청남도	94,296	1,910	2.03	681	0.72	1,046	1.11
전라북도	97,980	1,371	1.40	600	0.61	743	0.76
전라남도	104,763	1,479	1.41	614	0.59	765	0.73
경상북도	136,938	1,934	1.41	684	0.50	1,047	0.76
경상남도	149,192	1,880	1.26	611	0.41	1,005	0.67
제주도	24,269	664	2.74	412	1.70	424	1.75

<표 A-29> 집락추출 6%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,675,073	78,341	2.93	10,252	0.38
서울특별시	528,716	12,509	2.37	923	0.17
부산광역시	191,428	4,362	2.28	582	0.30
대구광역시	133,256	3,587	2.69	513	0.38
인천광역시	141,664	5,055	3.57	619	0.44
광주광역시	78,185	2,730	3.49	415	0.53
대전광역시	77,638	3,608	4.65	531	0.68
울산광역시	54,705	2,172	3.97	415	0.76
경기도	558,650	14,453	2.59	990	0.18
강원도	91,517	3,589	3.92	729	0.80
충청북도	88,611	3,393	3.83	653	0.74
충청남도	113,410	4,995	4.40	735	0.65
전라북도	119,367	2,854	2.39	646	0.54
전라남도	125,394	3,351	2.67	634	0.51
경상북도	163,735	5,000	3.05	765	0.47
경상남도	178,679	5,302	2.97	635	0.36
제주도	30,118	1,381	4.59	467	1.55

<표 A-30> 집락추출 6%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	2,675,073	29,356	1.10	10,252	0.38	16,428	0.61
서울특별시	528,716	3,380	0.64	923	0.17	1,986	0.38
부산광역시	191,428	1,662	0.87	582	0.30	943	0.49
대구광역시	133,256	1,348	1.01	513	0.38	744	0.56
인천광역시	141,664	1,918	1.35	619	0.44	971	0.69
광주광역시	78,185	1,088	1.39	415	0.53	557	0.71
대전광역시	77,638	1,430	1.84	531	0.68	789	1.02
울산광역시	54,705	1,028	1.88	415	0.76	534	0.98
경기도	558,650	3,916	0.70	990	0.18	2,258	0.40
강원도	91,517	1,694	1.85	729	0.80	1,064	1.16
충청북도	88,611	1,651	1.86	653	0.74	904	1.02
충청남도	113,410	2,049	1.81	735	0.65	1,130	1.00
전라북도	119,367	1,501	1.26	646	0.54	846	0.71
전라남도	125,394	1,636	1.30	634	0.51	833	0.66
경상북도	163,735	2,176	1.33	765	0.47	1,191	0.73
경상남도	178,679	2,071	1.16	635	0.36	1,169	0.65
제주도	30,118	808	2.68	467	1.55	509	1.69

<표 A-31> 집락추출 7%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,119,081	86,463	2.77	10,745	0.34
서울특별시	615,513	13,459	2.19	926	0.15
부산광역시	224,304	4,713	2.10	608	0.27
대구광역시	155,137	3,899	2.51	515	0.33
인천광역시	165,185	5,664	3.43	681	0.41
광주광역시	90,645	3,077	3.39	443	0.49
대전광역시	89,790	3,989	4.44	563	0.63
울산광역시	64,735	2,450	3.78	407	0.63
경기도	652,409	15,517	2.38	1,024	0.16
강원도	107,099	4,003	3.74	778	0.73
충청북도	103,603	3,942	3.80	715	0.69
충청남도	131,963	5,540	4.20	778	0.59
전라북도	137,093	3,245	2.37	657	0.48
전라남도	146,050	3,727	2.55	656	0.45
경상북도	192,039	5,642	2.94	776	0.40
경상남도	208,558	5,942	2.85	707	0.34
제주도	34,958	1,654	4.73	511	1.46

<표 A-32> 집락추출 7%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,119,081	32,041	1.03	10,745	0.34	17,639	0.57
서울특별시	615,513	3,621	0.59	926	0.15	2,109	0.34
부산광역시	224,304	1,825	0.81	608	0.27	1,007	0.45
대구광역시	155,137	1,435	0.92	515	0.33	827	0.53
인천광역시	165,185	2,105	1.27	681	0.41	1,117	0.68
광주광역시	90,645	1,257	1.39	443	0.49	592	0.65
대전광역시	89,790	1,508	1.68	563	0.63	842	0.94
울산광역시	64,735	1,079	1.67	407	0.63	568	0.88
경기도	652,409	4,223	0.65	1,024	0.16	2,303	0.35
강원도	107,099	1,869	1.75	778	0.73	1,119	1.04
충청북도	103,603	1,851	1.79	715	0.69	1,004	0.97
충청남도	131,963	2,261	1.71	778	0.59	1,208	0.92
전라북도	137,093	1,646	1.20	657	0.48	930	0.68
전라남도	146,050	1,809	1.24	656	0.45	904	0.62
경상북도	192,039	2,333	1.21	776	0.40	1,275	0.66
경상남도	208,558	2,251	1.08	707	0.34	1,229	0.59
제주도	34,958	968	2.77	511	1.46	605	1.73

<표 A-33> 집락추출 8%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,564,458	94,186	2.64	11,036	0.31
서울특별시	703,866	14,681	2.09	940	0.13
부산광역시	256,209	5,179	2.02	623	0.24
대구광역시	177,008	4,325	2.44	538	0.30
인천광역시	188,409	6,101	3.24	704	0.37
광주광역시	104,047	3,274	3.15	432	0.42
대전광역시	102,950	4,364	4.24	568	0.55
울산광역시	73,469	2,720	3.70	419	0.57
경기도	746,341	16,651	2.23	1,019	0.14
강원도	121,923	4,395	3.60	820	0.67
충청북도	118,156	4,294	3.63	706	0.60
충청남도	151,452	6,131	4.05	787	0.52
전라북도	157,101	3,569	2.27	719	0.46
전라남도	166,979	4,174	2.50	677	0.41
경상북도	218,649	6,102	2.79	840	0.38
경상남도	238,014	6,372	2.68	707	0.30
제주도	39,885	1,854	4.65	537	1.35

<표 A-34> 집락추출 8%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	3,564,458	33,832	0.95	11,036	0.31	18,683	0.52
서울특별시	703,866	3,787	0.54	940	0.13	2,149	0.31
부산광역시	256,209	1,925	0.75	623	0.24	1,095	0.43
대구광역시	177,008	1,559	0.88	538	0.30	890	0.50
인천광역시	188,409	2,200	1.17	704	0.37	1,171	0.62
광주광역시	104,047	1,264	1.21	432	0.42	620	0.60
대전광역시	102,950	1,625	1.58	568	0.55	869	0.84
울산광역시	73,469	1,188	1.62	419	0.57	612	0.83
경기도	746,341	4,378	0.59	1,019	0.14	2,450	0.33
강원도	121,923	1,913	1.57	820	0.67	1,155	0.95
충청북도	118,156	1,950	1.65	706	0.60	1,056	0.89
충청남도	151,452	2,391	1.58	787	0.52	1,301	0.86
전라북도	157,101	1,750	1.11	719	0.46	995	0.63
전라남도	166,979	1,917	1.15	677	0.41	974	0.58
경상북도	218,649	2,559	1.17	840	0.38	1,372	0.63
경상남도	238,014	2,379	1.00	707	0.30	1,297	0.54
제주도	39,885	1,047	2.63	537	1.35	677	1.70

<표 A-35> 집락추출 9%, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	4,010,450	101,450	2.53	11,409	0.28
서울특별시	790,618	15,507	1.96	958	0.12
부산광역시	288,314	5,638	1.96	607	0.21
대구광역시	199,644	4,640	2.32	542	0.27
인천광역시	211,676	6,650	3.14	717	0.34
광주광역시	116,978	3,565	3.05	457	0.39
대전광역시	116,102	4,696	4.04	598	0.52
울산광역시	82,403	2,953	3.58	461	0.56
경기도	839,208	17,808	2.12	1,049	0.12
강원도	137,075	4,815	3.51	826	0.60
충청북도	133,553	4,675	3.50	761	0.57
충청남도	169,988	6,606	3.89	855	0.50
전라북도	177,283	3,920	2.21	717	0.40
전라남도	188,214	4,562	2.42	711	0.38
경상북도	245,866	6,550	2.66	852	0.35
경상남도	268,586	6,840	2.55	732	0.27
제주도	44,942	2,025	4.51	566	1.26

<표 A-36> 집락추출 9%, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	4,010,450	36,014	0.90	11,409	0.28	19,553	0.49
서울특별시	790,618	3,957	0.50	958	0.12	2,245	0.28
부산광역시	288,314	2,056	0.71	607	0.21	1,133	0.39
대구광역시	199,644	1,637	0.82	542	0.27	927	0.46
인천광역시	211,676	2,292	1.08	717	0.34	1,239	0.59
광주광역시	116,978	1,380	1.18	457	0.39	654	0.56
대전광역시	116,102	1,737	1.50	598	0.52	906	0.78
울산광역시	82,403	1,247	1.51	461	0.56	655	0.79
경기도	839,208	4,570	0.54	1,049	0.12	2,554	0.30
강원도	137,075	2,112	1.54	826	0.60	1,195	0.87
충청북도	133,553	2,083	1.56	761	0.57	1,115	0.83
충청남도	169,988	2,587	1.52	855	0.50	1,396	0.82
전라북도	177,283	1,905	1.07	717	0.40	1,049	0.59
전라남도	188,214	2,095	1.11	711	0.38	1,036	0.55
경상북도	245,866	2,682	1.09	852	0.35	1,390	0.57
경상남도	268,586	2,536	0.94	732	0.27	1,349	0.50
제주도	44,942	1,138	2.53	566	1.26	710	1.58

<표 A-37> 10%표본, 노출제한기법 적용 전후 유일성 비교

시도	인구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	4,455,527	108,310	2.43	11,747	0.26
서울특별시	879,032	16,439	1.87	988	0.11
부산광역시	320,174	5,987	1.87	631	0.20
대구광역시	221,787	4,976	2.24	563	0.25
인천광역시	235,621	7,024	2.98	733	0.31
광주광역시	129,836	3,835	2.95	474	0.37
대전광역시	128,849	5,053	3.92	633	0.49
울산광역시	91,756	3,198	3.49	488	0.53
경기도	931,851	18,811	2.02	1,058	0.11
강원도	152,350	5,189	3.41	876	0.57
충청북도	148,015	5,026	3.40	764	0.52
충청남도	188,950	7,110	3.76	850	0.45
전라북도	196,976	4,270	2.17	766	0.39
전라남도	208,847	4,877	2.34	723	0.35
경상북도	273,566	7,015	2.56	871	0.32
경상남도	297,986	7,335	2.46	761	0.26
제주도	49,931	2,165	4.34	568	1.14

<표 A-38> 10%표본, 변수추가 후 유일성 비교

시도	인구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	4,455,527	37,826	0.85	11,747	0.26	20,607	0.46
서울특별시	879,032	4,097	0.47	988	0.11	2,336	0.27
부산광역시	320,174	2,146	0.67	631	0.20	1,182	0.37
대구광역시	221,787	1,720	0.78	563	0.25	973	0.44
인천광역시	235,621	2,417	1.03	733	0.31	1,287	0.55
광주광역시	129,836	1,445	1.11	474	0.37	694	0.53
대전광역시	128,849	1,854	1.44	633	0.49	980	0.76
울산광역시	91,756	1,335	1.45	488	0.53	713	0.78
경기도	931,851	4,799	0.51	1,058	0.11	2,637	0.28
강원도	152,350	2,248	1.48	876	0.57	1,286	0.84
충청북도	148,015	2,208	1.49	764	0.52	1,182	0.80
충청남도	188,950	2,687	1.42	850	0.45	1,459	0.77
전라북도	196,976	2,017	1.02	766	0.39	1,141	0.58
전라남도	208,847	2,176	1.04	723	0.35	1,100	0.53
경상북도	273,566	2,809	1.03	871	0.32	1,462	0.53
경상남도	297,986	2,681	0.90	761	0.26	1,425	0.48
제주도	49,931	1,187	2.38	568	1.14	750	1.50

가구주택 부문 유일성 비교: 추출기법, %구간, 노출제한기법 적용 전후

<표 A-39> 계통추출 1%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	158,268	86,741	54.81	8,396	5.30
서울특별시	30,549	18,142	59.39	757	2.48
부산광역시	11,131	6,981	62.72	553	4.97
대구광역시	7,581	4,800	63.32	490	6.46
인천광역시	7,960	4,706	59.12	521	6.55
광주광역시	4,399	2,417	54.94	362	8.23
대전광역시	4,437	2,726	61.44	430	9.69
울산광역시	3,068	1,823	59.42	351	11.44
경기도	30,919	14,959	48.38	770	2.49
강원도	5,832	3,556	60.97	532	9.12
충청북도	5,537	3,169	57.23	521	9.41
충청남도	7,103	3,570	50.26	519	7.31
전라북도	7,470	3,660	49.00	531	7.11
전라남도	8,383	3,590	42.82	485	5.79
경상북도	10,894	5,447	50.00	602	5.53
경상남도	11,258	5,912	52.51	594	5.28
제주도	1,747	1,283	73.44	378	21.64

<표 A-40> 계통추출 1%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	158,268	21,399	13.52	8,396	5.30	26,586	16.80
서울특별시	30,549	2,563	8.39	757	2.48	2,897	9.48
부산광역시	11,131	1,672	15.02	553	4.97	1,884	16.93
대구광역시	7,581	1,214	16.01	490	6.46	1,623	21.41
인천광역시	7,960	1,317	16.55	521	6.55	1,620	20.35
광주광역시	4,399	822	18.69	362	8.23	1,205	27.39
대전광역시	4,437	912	20.55	430	9.69	1,235	27.83
울산광역시	3,068	715	23.31	351	11.44	968	31.55
경기도	30,919	2,354	7.61	770	2.49	2,839	9.18
강원도	5,832	1,261	21.62	532	9.12	1,562	26.78
충청북도	5,537	1,130	20.41	521	9.41	1,412	25.50
충청남도	7,103	1,152	16.22	519	7.31	1,541	21.70
전라북도	7,470	1,215	16.27	531	7.11	1,528	20.46
전라남도	8,383	1,185	14.14	485	5.79	1,520	18.13
경상북도	10,894	1,583	14.53	602	5.53	1,918	17.61
경상남도	11,258	1,639	14.56	594	5.28	1,987	17.65
제주도	1,747	665	38.07	378	21.64	847	48.48

<표 A-41> 계통추출 2%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	316,536	150,367	47.50	9,973	3.15
서울특별시	61,099	31,717	51.91	814	1.33
부산광역시	22,261	12,278	55.15	666	2.99
대구광역시	15,162	8,787	57.95	578	3.81
인천광역시	15,920	8,012	50.33	641	4.03
광주광역시	8,799	4,329	49.20	459	5.22
대전광역시	8,874	5,089	57.35	512	5.77
울산광역시	6,136	3,319	54.09	467	7.61
경기도	61,837	25,749	41.64	878	1.42
강원도	11,664	6,137	52.61	653	5.60
충청북도	11,075	5,455	49.26	585	5.28
충청남도	14,206	6,058	42.64	612	4.31
전라북도	14,940	6,115	40.93	638	4.27
전라남도	16,766	5,800	34.59	583	3.48
경상북도	21,787	9,046	41.52	716	3.29
경상남도	22,517	10,133	45.00	703	3.12
제주도	3,493	2,343	67.08	468	13.40

<표 A-42> 계통추출 2%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	316,536	29,160	9.21	9,973	3.15	35,948	11.36
서울특별시	61,099	3,307	5.41	814	1.33	3,548	5.81
부산광역시	22,261	2,263	10.17	666	2.99	2,453	11.02
대구광역시	15,162	1,702	11.23	578	3.81	2,167	14.29
인천광역시	15,920	1,843	11.58	641	4.03	2,217	13.93
광주광역시	8,799	1,191	13.54	459	5.22	1,665	18.92
대전광역시	8,874	1,321	14.89	512	5.77	1,758	19.81
울산광역시	6,136	1,071	17.45	467	7.61	1,367	22.28
경기도	61,837	3,058	4.95	878	1.42	3,700	5.98
강원도	11,664	1,755	15.05	653	5.60	2,226	19.08
충청북도	11,075	1,519	13.72	585	5.28	1,968	17.77
충청남도	14,206	1,601	11.27	612	4.31	2,163	15.23
전라북도	14,940	1,653	11.06	638	4.27	2,159	14.45
전라남도	16,766	1,605	9.57	583	3.48	2,072	12.36
경상북도	21,787	2,133	9.79	716	3.29	2,569	11.79
경상남도	22,517	2,159	9.59	703	3.12	2,663	11.83
제주도	3,493	979	28.03	468	13.40	1,253	35.87

<표 A-43> 계통추출 3%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	474,804	204,325	43.03	11,065	2.33
서울특별시	91,649	42,960	46.87	877	0.96
부산광역시	33,392	16,828	50.40	745	2.23
대구광역시	22,742	12,227	53.76	647	2.84
인천광역시	23,880	10,790	45.18	729	3.05
광주광역시	13,199	6,002	45.47	526	3.99
대전광역시	13,310	6,933	52.09	593	4.46
울산광역시	9,205	4,667	50.70	523	5.68
경기도	92,756	35,038	37.77	940	1.01
강원도	17,495	8,296	47.42	722	4.13
충청북도	16,613	7,476	45.00	626	3.77
충청남도	21,308	8,043	37.75	729	3.42
전라북도	22,411	8,131	36.28	661	2.95
전라남도	25,149	7,715	30.68	683	2.72
경상북도	32,680	12,095	37.01	780	2.39
경상남도	33,776	13,838	40.97	774	2.29
제주도	5,239	3,286	62.72	510	9.73

<표 A-44> 계통추출 3%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	474,804	34,382	7.24	11,065	2.33	41,566	8.75
서울특별시	91,649	3,867	4.22	877	0.96	3,989	4.35
부산광역시	33,392	2,655	7.95	745	2.23	2,845	8.52
대구광역시	22,742	1,970	8.66	647	2.84	2,472	10.87
인천광역시	23,880	2,132	8.93	729	3.05	2,506	10.49
광주광역시	13,199	1,462	11.08	526	3.99	1,918	14.53
대전광역시	13,310	1,496	11.24	593	4.46	2,009	15.09
울산광역시	9,205	1,306	14.19	523	5.68	1,614	17.53
경기도	92,756	3,419	3.69	940	1.01	4,134	4.46
강원도	17,495	2,124	12.14	722	4.13	2,627	15.02
충청북도	16,613	1,824	10.98	626	3.77	2,383	14.34
충청남도	21,308	1,972	9.25	729	3.42	2,546	11.95
전라북도	22,411	1,891	8.44	661	2.95	2,434	10.86
전라남도	25,149	1,910	7.59	683	2.72	2,564	10.20
경상북도	32,680	2,557	7.82	780	2.39	2,958	9.05
경상남도	33,776	2,638	7.81	774	2.29	3,031	8.97
제주도	5,239	1,159	22.12	510	9.73	1,536	29.32

<표 A-45> 계통추출 4%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	633,072	252,491	39.88	11,874	1.88
서울특별시	122,198	52,997	43.37	935	0.77
부산광역시	44,523	20,967	47.09	751	1.69
대구광역시	30,323	15,304	50.47	685	2.26
인천광역시	31,840	13,257	41.64	749	2.35
광주광역시	17,598	7,627	43.34	544	3.09
대전광역시	17,748	8,873	49.99	658	3.71
울산광역시	12,273	5,873	47.85	547	4.46
경기도	123,674	42,880	34.67	1,020	0.82
강원도	23,328	10,255	43.96	816	3.50
충청북도	22,149	9,318	42.07	699	3.16
충청남도	28,412	9,841	34.64	736	2.59
전라북도	29,881	9,988	33.43	738	2.47
전라남도	33,532	9,299	27.73	769	2.29
경상북도	43,574	14,819	34.01	846	1.94
경상남도	45,033	17,025	37.81	803	1.78
제주도	6,986	4,168	59.66	578	8.27

<표 A-46> 계통추출 4%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	633,072	38,512	6.08	11,874	1.88	45,859	7.24
서울특별시	122,198	4,221	3.45	935	0.77	4,274	3.50
부산광역시	44,523	2,961	6.65	751	1.69	3,092	6.94
대구광역시	30,323	2,258	7.45	685	2.26	2,692	8.88
인천광역시	31,840	2,360	7.41	749	2.35	2,816	8.84
광주광역시	17,598	1,594	9.06	544	3.09	2,086	11.85
대전광역시	17,748	1,770	9.97	658	3.71	2,373	13.37
울산광역시	12,273	1,544	12.58	547	4.46	1,782	14.52
경기도	123,674	3,823	3.09	1,020	0.82	4,568	3.69
강원도	23,328	2,360	10.12	816	3.50	3,045	13.05
충청북도	22,149	1,988	8.98	699	3.16	2,630	11.87
충청남도	28,412	2,200	7.74	736	2.59	2,811	9.89
전라북도	29,881	2,197	7.35	738	2.47	2,792	9.34
전라남도	33,532	2,204	6.57	769	2.29	2,694	8.03
경상북도	43,574	2,834	6.50	846	1.94	3,207	7.36
경상남도	45,033	2,850	6.33	803	1.78	3,304	7.34
제주도	6,986	1,348	19.30	578	8.27	1,693	24.23

<표 A-47> 계통추출 5%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	791,340	294,596	37.23	12,287	1.55
서울특별시	152,748	61,358	40.17	928	0.61
부산광역시	55,654	24,575	44.16	772	1.39
대구광역시	37,904	17,807	46.98	712	1.88
인천광역시	39,799	15,784	39.66	761	1.91
광주광역시	21,998	8,789	39.95	575	2.61
대전광역시	22,185	9,925	44.74	667	3.01
울산광역시	15,341	7,002	45.64	599	3.90
경기도	154,593	49,965	32.32	1,020	0.66
강원도	29,159	11,956	41.00	839	2.88
충청북도	27,687	10,764	38.88	727	2.63
충청남도	35,515	11,438	32.21	792	2.23
전라북도	37,351	11,562	30.95	729	1.95
전라남도	41,915	11,002	26.25	841	2.01
경상북도	54,467	17,459	32.05	860	1.58
경상남도	56,292	20,122	35.75	867	1.54
제주도	8,732	5,088	58.27	598	6.85

<표 A-48> 계통추출 5%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	791,340	41,258	5.21	12,287	1.55	48,946	6.19
서울특별시	152,748	4,358	2.85	928	0.61	4,348	2.85
부산광역시	55,654	3,140	5.64	772	1.39	3,299	5.93
대구광역시	37,904	2,413	6.37	712	1.88	2,863	7.55
인천광역시	39,799	2,538	6.38	761	1.91	3,070	7.71
광주광역시	21,998	1,779	8.09	575	2.61	2,274	10.34
대전광역시	22,185	1,890	8.52	667	3.01	2,399	10.81
울산광역시	15,341	1,643	10.71	599	3.90	1,991	12.98
경기도	154,593	3,922	2.54	1,020	0.66	4,711	3.05
강원도	29,159	2,520	8.64	839	2.88	3,194	10.95
충청북도	27,687	2,156	7.79	727	2.63	2,798	10.11
충청남도	35,515	2,482	6.99	792	2.23	3,025	8.52
전라북도	37,351	2,320	6.21	729	1.95	2,950	7.90
전라남도	41,915	2,463	5.88	841	2.01	3,023	7.21
경상북도	54,467	3,011	5.53	860	1.58	3,469	6.37
경상남도	56,292	3,145	5.59	867	1.54	3,581	6.36
제주도	8,732	1,478	16.93	598	6.85	1,951	22.34

<표 A-49> 계통추출 6%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	949,608	338,161	35.61	12,634	1.33
서울특별시	183,298	70,260	38.33	982	0.54
부산광역시	66,784	28,063	42.02	814	1.22
대구광역시	45,485	21,110	46.41	748	1.64
인천광역시	47,759	17,792	37.25	847	1.77
광주광역시	26,398	10,214	38.69	596	2.26
대전광역시	26,621	12,054	45.28	704	2.64
울산광역시	18,409	8,174	44.40	629	3.42
경기도	185,512	57,532	31.01	1,047	0.56
강원도	34,991	13,710	39.18	837	2.39
충청북도	33,225	12,389	37.29	702	2.11
충청남도	42,617	12,975	30.45	835	1.96
전라북도	44,821	13,145	29.33	750	1.67
전라남도	50,298	12,283	24.42	789	1.57
경상북도	65,361	19,692	30.13	884	1.35
경상남도	67,551	22,971	34.01	853	1.26
제주도	10,478	5,797	55.33	617	5.89

<표 A-50> 계통추출 6%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	949,608	44,283	4.66	12,634	1.33	52,162	5.49
서울특별시	183,298	4,775	2.61	982	0.54	4,673	2.55
부산광역시	66,784	3,384	5.07	814	1.22	3,509	5.25
대구광역시	45,485	2,597	5.71	748	1.64	3,096	6.81
인천광역시	47,759	2,795	5.85	847	1.77	3,179	6.66
광주광역시	26,398	1,891	7.16	596	2.26	2,447	9.27
대전광역시	26,621	2,078	7.81	704	2.64	2,692	10.11
울산광역시	18,409	1,821	9.89	629	3.42	2,148	11.67
경기도	185,512	4,214	2.27	1,047	0.56	5,008	2.70
강원도	34,991	2,701	7.72	837	2.39	3,340	9.55
충청북도	33,225	2,348	7.07	702	2.11	3,038	9.14
충청남도	42,617	2,525	5.92	835	1.96	3,228	7.57
전라북도	44,821	2,499	5.58	750	1.67	3,188	7.11
전라남도	50,298	2,549	5.07	789	1.57	3,186	6.33
경상북도	65,361	3,227	4.94	884	1.35	3,656	5.59
경상남도	67,551	3,279	4.85	853	1.26	3,701	5.48
제주도	10,478	1,600	15.27	617	5.89	2,073	19.78

<표 A-51> 계통추출 7%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,107,876	375,515	33.90	13,131	1.19
서울특별시	213,847	77,609	36.29	953	0.45
부산광역시	77,915	31,366	40.26	836	1.07
대구광역시	53,066	23,423	44.14	768	1.45
인천광역시	55,719	19,790	35.52	856	1.54
광주광역시	30,797	11,474	37.26	646	2.10
대전광역시	31,059	13,485	43.42	723	2.33
울산광역시	21,477	9,222	42.94	647	3.01
경기도	216,430	63,552	29.36	1,094	0.51
강원도	40,824	15,258	37.38	883	2.16
충청북도	38,761	13,826	35.67	765	1.97
충청남도	49,721	14,416	28.99	844	1.70
전라북도	52,291	14,557	27.84	794	1.52
전라남도	58,681	13,652	23.26	854	1.46
경상북도	76,255	21,831	28.63	928	1.22
경상남도	78,808	25,446	32.29	860	1.09
제주도	12,225	6,608	54.05	680	5.56

<표 A-52> 계통추출 7%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,107,876	46,650	4.21	13,131	1.19	54,769	4.94
서울특별시	213,847	4,847	2.27	953	0.45	4,759	2.23
부산광역시	77,915	3,516	4.51	836	1.07	3,692	4.74
대구광역시	53,066	2,742	5.17	768	1.45	3,150	5.94
인천광역시	55,719	2,926	5.25	856	1.54	3,399	6.10
광주광역시	30,797	1,995	6.48	646	2.10	2,540	8.25
대전광역시	31,059	2,239	7.21	723	2.33	2,874	9.25
울산광역시	21,477	1,880	8.75	647	3.01	2,264	10.54
경기도	216,430	4,455	2.06	1,094	0.51	5,299	2.45
강원도	40,824	2,836	6.95	883	2.16	3,560	8.72
충청북도	38,761	2,437	6.29	765	1.97	3,200	8.26
충청남도	49,721	2,723	5.48	844	1.70	3,425	6.89
전라북도	52,291	2,648	5.06	794	1.52	3,306	6.32
전라남도	58,681	2,732	4.66	854	1.46	3,351	5.71
경상북도	76,255	3,468	4.55	928	1.22	3,856	5.06
경상남도	78,808	3,465	4.40	860	1.09	3,940	5.00
제주도	12,225	1,741	14.24	680	5.56	2,154	17.62

<표 A-53> 계통추출 8%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,266,144	412,781	32.60	13,346	1.05
서울특별시	244,397	85,034	34.79	990	0.41
부산광역시	89,046	34,385	38.61	842	0.95
대구광역시	60,646	26,184	43.18	796	1.31
인천광역시	63,679	21,638	33.98	876	1.38
광주광역시	35,197	12,770	36.28	655	1.86
대전광역시	35,495	14,985	42.22	780	2.20
울산광역시	24,546	10,250	41.76	658	2.68
경기도	247,349	69,700	28.18	1,082	0.44
강원도	46,655	16,707	35.81	879	1.88
충청북도	44,299	15,160	34.22	761	1.72
충청남도	56,823	15,681	27.60	868	1.53
전라북도	59,762	15,968	26.72	798	1.34
전라남도	67,064	14,878	22.18	859	1.28
경상북도	87,148	23,965	27.50	923	1.06
경상남도	90,067	28,181	31.29	894	0.99
제주도	13,971	7,295	52.22	685	4.90

<표 A-54> 계통추출 8%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,266,144	48,810	3.86	13,346	1.05	56,804	4.49
서울특별시	244,397	5,112	2.09	990	0.41	4,990	2.04
부산광역시	89,046	3,705	4.16	842	0.95	3,804	4.27
대구광역시	60,646	2,877	4.74	796	1.31	3,339	5.51
인천광역시	63,679	3,029	4.76	876	1.38	3,512	5.52
광주광역시	35,197	2,097	5.96	655	1.86	2,611	7.42
대전광역시	35,495	2,327	6.56	780	2.20	3,002	8.46
울산광역시	24,546	2,063	8.40	658	2.68	2,332	9.50
경기도	247,349	4,594	1.86	1,082	0.44	5,377	2.17
강원도	46,655	2,962	6.35	879	1.88	3,709	7.95
충청북도	44,299	2,564	5.79	761	1.72	3,304	7.46
충청남도	56,823	2,824	4.97	868	1.53	3,536	6.22
전라북도	59,762	2,778	4.65	798	1.34	3,423	5.73
전라남도	67,064	2,866	4.27	859	1.28	3,488	5.20
경상북도	87,148	3,600	4.13	923	1.06	4,012	4.60
경상남도	90,067	3,616	4.01	894	0.99	4,086	4.54
제주도	13,971	1,796	12.86	685	4.90	2,279	16.31

<표 A-55> 계통추출 9%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,424,412	446,997	31.38	13,708	0.96
서울특별시	274,947	91,578	33.31	1,002	0.36
부산광역시	100,176	37,332	37.27	864	0.86
대구광역시	68,227	28,503	41.78	796	1.17
인천광역시	71,639	23,555	32.88	908	1.27
광주광역시	39,597	13,842	34.96	673	1.70
대전광역시	39,932	16,222	40.62	789	1.98
울산광역시	27,614	11,224	40.65	683	2.47
경기도	278,267	75,246	27.04	1,103	0.40
강원도	52,487	18,156	34.59	897	1.71
충청북도	49,837	16,459	33.03	780	1.57
충청남도	63,926	16,964	26.54	907	1.42
전라북도	67,232	17,218	25.61	824	1.23
전라남도	75,447	16,090	21.33	912	1.21
경상북도	98,041	25,958	26.48	948	0.97
경상남도	101,326	30,600	30.20	908	0.90
제주도	15,717	8,050	51.22	714	4.54

<표 A-56> 계통추출 9%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,424,412	50,702	3.56	13,708	0.96	58,637	4.12
서울특별시	274,947	5,218	1.90	1,002	0.36	5,084	1.85
부산광역시	100,176	3,830	3.82	864	0.86	3,868	3.86
대구광역시	68,227	2,980	4.37	796	1.17	3,418	5.01
인천광역시	71,639	3,130	4.37	908	1.27	3,637	5.08
광주광역시	39,597	2,183	5.51	673	1.70	2,749	6.94
대전광역시	39,932	2,429	6.08	789	1.98	3,096	7.75
울산광역시	27,614	2,123	7.69	683	2.47	2,447	8.86
경기도	278,267	4,759	1.71	1,103	0.40	5,448	1.96
강원도	52,487	3,081	5.87	897	1.71	3,829	7.30
충청북도	49,837	2,677	5.37	780	1.57	3,440	6.90
충청남도	63,926	2,981	4.66	907	1.42	3,688	5.77
전라북도	67,232	2,879	4.28	824	1.23	3,510	5.22
전라남도	75,447	3,038	4.03	912	1.21	3,648	4.84
경상북도	98,041	3,713	3.79	948	0.97	4,168	4.25
경상남도	101,326	3,771	3.72	908	0.90	4,195	4.14
제주도	15,717	1,910	12.15	714	4.54	2,412	15.35

<표 A-57> 집락추출 1%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	158,269	86,864	54.88	8,414	5.32
서울특별시	30,665	18,144	59.17	779	2.54
부산광역시	11,076	6,864	61.97	558	5.04
대구광역시	7,560	4,840	64.02	482	6.38
인천광역시	8,089	4,734	58.52	550	6.80
광주광역시	4,432	2,508	56.59	401	9.05
대전광역시	4,415	2,874	65.10	415	9.40
울산광역시	3,126	1,909	61.07	353	11.29
경기도	30,779	15,001	48.74	761	2.47
강원도	5,901	3,605	61.09	560	9.49
충청북도	5,407	3,144	58.15	480	8.88
충청남도	7,150	3,574	49.99	523	7.31
전라북도	7,484	3,666	48.98	514	6.87
전라남도	8,380	3,579	42.71	546	6.52
경상북도	10,990	5,353	48.71	580	5.28
경상남도	11,028	5,821	52.78	562	5.10
제주도	1,787	1,248	69.84	350	19.59

<표 A-58> 집락추출 1%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	158,269	21,509	13.59	8,414	5.32	26,694	16.87
서울특별시	30,665	2,578	8.41	779	2.54	2,879	9.39
부산광역시	11,076	1,773	16.01	558	5.04	1,939	17.51
대구광역시	7,560	1,260	16.67	482	6.38	1,672	22.12
인천광역시	8,089	1,387	17.15	550	6.80	1,725	21.33
광주광역시	4,432	862	19.45	401	9.05	1,236	27.89
대전광역시	4,415	927	21.00	415	9.40	1,284	29.08
울산광역시	3,126	747	23.90	353	11.29	956	30.58
경기도	30,779	2,373	7.71	761	2.47	2,882	9.36
강원도	5,901	1,253	21.23	560	9.49	1,584	26.84
충청북도	5,407	1,042	19.27	480	8.88	1,394	25.78
충청남도	7,150	1,179	16.49	523	7.31	1,534	21.45
전라북도	7,484	1,149	15.35	514	6.87	1,499	20.03
전라남도	8,380	1,188	14.18	546	6.52	1,492	17.80
경상북도	10,990	1,551	14.11	580	5.28	1,854	16.87
경상남도	11,028	1,610	14.60	562	5.10	1,993	18.07
제주도	1,787	630	35.25	350	19.59	771	43.14

<표 A-59> 집락추출 2%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	316,537	150,164	47.44	10,150	3.21
서울특별시	60,980	31,709	52.00	811	1.33
부산광역시	22,355	12,252	54.81	677	3.03
대구광역시	15,248	8,766	57.49	561	3.68
인천광역시	15,956	7,979	50.01	612	3.84
광주광역시	8,830	4,440	50.28	511	5.79
대전광역시	8,795	5,056	57.49	536	6.09
울산광역시	6,181	3,359	54.34	460	7.44
경기도	61,618	25,735	41.77	893	1.45
강원도	11,659	6,098	52.30	671	5.76
충청북도	11,100	5,461	49.20	575	5.18
충청남도	14,121	5,978	42.33	680	4.82
전라북도	15,120	6,030	39.88	621	4.11
전라남도	16,787	5,844	34.81	666	3.97
경상북도	21,744	9,029	41.52	721	3.32
경상남도	22,509	10,106	44.90	719	3.19
제주도	3,534	2,322	65.70	436	12.34

<표 A-60> 집락추출 2%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	316,537	29,169	9.22	10,150	3.21	36,010	11.38
서울특별시	60,980	3,269	5.36	811	1.33	3,556	5.83
부산광역시	22,355	2,307	10.32	677	3.03	2,482	11.10
대구광역시	15,248	1,707	11.19	561	3.68	2,210	14.49
인천광역시	15,956	1,813	11.36	612	3.84	2,188	13.71
광주광역시	8,830	1,245	14.10	511	5.79	1,719	19.47
대전광역시	8,795	1,308	14.87	536	6.09	1,756	19.97
울산광역시	6,181	1,022	16.53	460	7.44	1,338	21.65
경기도	61,618	3,081	5.00	893	1.45	3,736	6.06
강원도	11,659	1,700	14.58	671	5.76	2,218	19.02
충청북도	11,100	1,463	13.18	575	5.18	1,980	17.84
충청남도	14,121	1,712	12.12	680	4.82	2,145	15.19
전라북도	15,120	1,630	10.78	621	4.11	2,126	14.06
전라남도	16,787	1,661	9.89	666	3.97	2,100	12.51
경상북도	21,744	2,114	9.72	721	3.32	2,560	11.77
경상남도	22,509	2,173	9.65	719	3.19	2,681	11.91
제주도	3,534	964	27.28	436	12.34	1,215	34.38

<표 A-61> 집락추출 3%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	474,805	204,036	42.97	11,106	2.34
서울특별시	91,525	43,129	47.12	905	0.99
부산광역시	33,405	16,713	50.03	694	2.08
대구광역시	22,872	12,188	53.29	655	2.86
인천광역시	23,846	10,708	44.90	665	2.79
광주광역시	13,260	6,020	45.40	520	3.92
대전광역시	13,357	7,061	52.86	578	4.33
울산광역시	9,088	4,607	50.69	515	5.67
경기도	92,475	35,019	37.87	967	1.05
강원도	17,436	8,168	46.85	801	4.59
충청북도	16,639	7,468	44.88	633	3.80
충청남도	21,268	8,113	38.15	760	3.57
전라북도	22,647	8,086	35.70	661	2.92
전라남도	25,212	7,698	30.53	705	2.80
경상북도	32,617	11,819	36.24	790	2.42
경상남도	33,882	13,893	41.00	770	2.27
제주도	5,276	3,346	63.42	487	9.23

<표 A-62> 집락추출 3%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	474,805	34,244	7.21	11,106	2.34	41,408	8.72
서울특별시	91,525	3,817	4.17	905	0.99	3,920	4.28
부산광역시	33,405	2,630	7.87	694	2.08	2,810	8.41
대구광역시	22,872	1,975	8.64	655	2.86	2,496	10.91
인천광역시	23,846	2,099	8.80	665	2.79	2,548	10.69
광주광역시	13,260	1,405	10.60	520	3.92	1,860	14.03
대전광역시	13,357	1,510	11.30	578	4.33	2,097	15.70
울산광역시	9,088	1,281	14.10	515	5.67	1,581	17.40
경기도	92,475	3,447	3.73	967	1.05	4,176	4.52
강원도	17,436	2,100	12.04	801	4.59	2,570	14.74
충청북도	16,639	1,838	11.05	633	3.80	2,287	13.74
충청남도	21,268	2,037	9.58	760	3.57	2,614	12.29
전라북도	22,647	1,889	8.34	661	2.92	2,356	10.40
전라남도	25,212	1,963	7.79	705	2.80	2,473	9.81
경상북도	32,617	2,551	7.82	790	2.42	2,990	9.17
경상남도	33,882	2,562	7.56	770	2.27	3,075	9.08
제주도	5,276	1,140	21.61	487	9.23	1,555	29.47

<표 A-63> 집락추출 4%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	633,073	252,684	39.91	11,922	1.88
서울특별시	121,977	52,787	43.28	960	0.79
부산광역시	44,779	21,144	47.22	792	1.77
대구광역시	30,360	15,325	50.48	648	2.13
인천광역시	32,060	13,307	41.51	762	2.38
광주광역시	17,560	7,516	42.80	580	3.30
대전광역시	17,673	8,818	49.90	661	3.74
울산광역시	12,323	5,922	48.06	546	4.43
경기도	123,684	43,105	34.85	1,003	0.81
강원도	23,427	10,345	44.16	827	3.53
충청북도	22,299	9,346	41.91	675	3.03
충청남도	28,543	9,824	34.42	768	2.69
전라북도	29,581	9,765	33.01	719	2.43
전라남도	33,536	9,412	28.07	755	2.25
경상북도	43,341	14,779	34.10	853	1.97
경상남도	45,106	17,123	37.96	813	1.80
제주도	6,824	4,166	61.05	560	8.21

<표 A-64> 집락추출 4%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	633,073	38,377	6.06	11,922	1.88	45,549	7.19
서울특별시	121,977	4,172	3.42	960	0.79	4,241	3.48
부산광역시	44,779	2,986	6.67	792	1.77	3,037	6.78
대구광역시	30,360	2,205	7.26	648	2.13	2,702	8.90
인천광역시	32,060	2,403	7.50	762	2.38	2,821	8.80
광주광역시	17,560	1,634	9.31	580	3.30	2,096	11.94
대전광역시	17,673	1,743	9.86	661	3.74	2,309	13.07
울산광역시	12,323	1,459	11.84	546	4.43	1,775	14.40
경기도	123,684	3,837	3.10	1,003	0.81	4,543	3.67
강원도	23,427	2,362	10.08	827	3.53	2,922	12.47
충청북도	22,299	1,979	8.87	675	3.03	2,569	11.52
충청남도	28,543	2,199	7.70	768	2.69	2,883	10.10
전라북도	29,581	2,118	7.16	719	2.43	2,637	8.91
전라남도	33,536	2,234	6.66	755	2.25	2,751	8.20
경상북도	43,341	2,863	6.61	853	1.97	3,209	7.40
경상남도	45,106	2,831	6.28	813	1.80	3,300	7.32
제주도	6,824	1,352	19.81	560	8.21	1,754	25.70

<표 A-65> 집락추출 5%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	791,341	296,472	37.46	12,198	1.54
서울특별시	152,667	61,700	40.41	989	0.65
부산광역시	55,578	24,663	44.38	772	1.39
대구광역시	37,819	18,222	48.18	685	1.81
인천광역시	39,980	15,595	39.01	791	1.98
광주광역시	21,976	8,923	40.60	554	2.52
대전광역시	22,133	10,456	47.24	687	3.10
울산광역시	15,261	7,064	46.29	568	3.72
경기도	154,820	50,465	32.60	1,033	0.67
강원도	29,125	12,104	41.56	842	2.89
충청북도	27,892	10,885	39.03	713	2.56
충청남도	35,468	11,438	32.25	786	2.22
전라북도	37,139	11,649	31.37	770	2.07
전라남도	41,923	10,924	26.06	766	1.83
경상북도	54,584	17,223	31.55	870	1.59
경상남도	56,428	20,158	35.72	791	1.40
제주도	8,548	5,003	58.53	581	6.80

<표 A-66> 집락추출 5%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	791,341	41,491	5.24	12,198	1.54	49,438	6.25
서울특별시	152,667	4,530	2.97	989	0.65	4,463	2.92
부산광역시	55,578	3,204	5.76	772	1.39	3,282	5.91
대구광역시	37,819	2,404	6.36	685	1.81	2,947	7.79
인천광역시	39,980	2,517	6.30	791	1.98	3,018	7.55
광주광역시	21,976	1,779	8.10	554	2.52	2,233	10.16
대전광역시	22,133	1,949	8.81	687	3.10	2,537	11.46
울산광역시	15,261	1,599	10.48	568	3.72	1,943	12.73
경기도	154,820	4,056	2.62	1,033	0.67	4,784	3.09
강원도	29,125	2,558	8.78	842	2.89	3,215	11.04
충청북도	27,892	2,207	7.91	713	2.56	2,883	10.34
충청남도	35,468	2,395	6.75	786	2.22	3,093	8.72
전라북도	37,139	2,323	6.25	770	2.07	2,986	8.04
전라남도	41,923	2,410	5.75	766	1.83	3,024	7.21
경상북도	54,584	3,040	5.57	870	1.59	3,480	6.38
경상남도	56,428	3,038	5.38	791	1.40	3,623	6.42
제주도	8,548	1,482	17.34	581	6.80	1,927	22.54

<표 A-67> 집락추출 6%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	949,609	337,639	35.56	12,619	1.33
서울특별시	183,528	70,104	38.20	933	0.51
부산광역시	66,462	28,045	42.20	817	1.23
대구광역시	45,479	21,091	46.38	760	1.67
인천광역시	47,702	17,790	37.29	815	1.71
광주광역시	26,460	10,187	38.50	589	2.23
대전광역시	26,713	12,203	45.68	712	2.67
울산광역시	18,357	8,127	44.27	594	3.24
경기도	185,514	57,422	30.95	1,054	0.57
강원도	34,935	13,568	38.84	842	2.41
충청북도	33,063	12,326	37.28	717	2.17
충청남도	42,578	12,982	30.49	821	1.93
전라북도	45,231	13,184	29.15	791	1.75
전라남도	50,254	12,409	24.69	817	1.63
경상북도	65,304	19,528	29.90	871	1.33
경상남도	67,540	22,871	33.86	849	1.26
제주도	10,489	5,802	55.32	637	6.07

<표 A-68> 집락추출 6%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	949,609	44,519	4.69	12,619	1.33	51,986	5.47
서울특별시	183,528	4,722	2.57	933	0.51	4,671	2.55
부산광역시	66,462	3,416	5.14	817	1.23	3,418	5.14
대구광역시	45,479	2,610	5.74	760	1.67	3,060	6.73
인천광역시	47,702	2,690	5.64	815	1.71	3,238	6.79
광주광역시	26,460	1,857	7.02	589	2.23	2,401	9.07
대전광역시	26,713	2,080	7.79	712	2.67	2,655	9.94
울산광역시	18,357	1,753	9.55	594	3.24	2,121	11.55
경기도	185,514	4,306	2.32	1,054	0.57	5,088	2.74
강원도	34,935	2,754	7.88	842	2.41	3,347	9.58
충청북도	33,063	2,314	7.00	717	2.17	2,953	8.93
충청남도	42,578	2,579	6.06	821	1.93	3,230	7.59
전라북도	45,231	2,569	5.68	791	1.75	3,145	6.95
전라남도	50,254	2,595	5.16	817	1.63	3,208	6.38
경상북도	65,304	3,297	5.05	871	1.33	3,657	5.60
경상남도	67,540	3,323	4.92	849	1.26	3,762	5.57
제주도	10,489	1,654	15.77	637	6.07	2,032	19.37

<표 A-69> 집락추출 7%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,107,877	376,111	33.95	13,094	1.18
서울특별시	214,001	77,568	36.25	1,011	0.47
부산광역시	77,907	31,430	40.34	808	1.04
대구광역시	52,973	23,649	44.64	737	1.39
인천광역시	55,846	19,803	35.46	878	1.57
광주광역시	30,706	11,477	37.38	639	2.08
대전광역시	31,007	13,584	43.81	753	2.43
울산광역시	21,629	9,328	43.13	659	3.05
경기도	216,499	63,837	29.49	1,072	0.50
강원도	40,927	15,312	37.41	878	2.15
충청북도	38,705	13,762	35.56	743	1.92
충청남도	49,705	14,380	28.93	820	1.65
전라북도	52,045	14,475	27.81	796	1.53
전라남도	58,600	13,623	23.25	844	1.44
경상북도	76,377	21,870	28.63	929	1.22
경상남도	78,751	25,446	32.31	884	1.12
제주도	12,199	6,567	53.83	643	5.27

<표 A-70> 집락추출 7%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,107,877	46,665	4.21	13,094	1.18	54,725	4.94
서울특별시	214,001	4,889	2.28	1,011	0.47	4,849	2.27
부산광역시	77,907	3,503	4.50	808	1.04	3,606	4.63
대구광역시	52,973	2,736	5.16	737	1.39	3,222	6.08
인천광역시	55,846	2,922	5.23	878	1.57	3,428	6.14
광주광역시	30,706	2,021	6.58	639	2.08	2,587	8.43
대전광역시	31,007	2,214	7.14	753	2.43	2,810	9.06
울산광역시	21,629	1,920	8.88	659	3.05	2,273	10.51
경기도	216,499	4,441	2.05	1,072	0.50	5,214	2.41
강원도	40,927	2,899	7.08	878	2.15	3,566	8.71
충청북도	38,705	2,408	6.22	743	1.92	3,150	8.14
충청남도	49,705	2,761	5.55	820	1.65	3,435	6.91
전라북도	52,045	2,659	5.11	796	1.53	3,266	6.28
전라남도	58,600	2,688	4.59	844	1.44	3,344	5.71
경상북도	76,377	3,423	4.48	929	1.22	3,792	4.96
경상남도	78,751	3,494	4.44	884	1.12	3,981	5.06
제주도	12,199	1,687	13.83	643	5.27	2,202	18.05

<표 A-71> 집락추출 8%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,266,145	412,274	32.56	13,376	1.06
서울특별시	244,581	84,671	34.62	1,004	0.41
부산광역시	89,003	34,325	38.57	861	0.97
대구광역시	60,561	26,100	43.10	769	1.27
인천광역시	63,626	21,695	34.10	892	1.40
광주광역시	35,236	12,729	36.12	629	1.79
대전광역시	35,501	14,992	42.23	789	2.22
울산광역시	24,527	10,299	41.99	670	2.73
경기도	247,586	69,773	28.18	1,086	0.44
강원도	46,640	16,712	35.83	908	1.95
충청북도	44,276	15,090	34.08	745	1.68
충청남도	56,916	15,735	27.65	848	1.49
전라북도	59,599	15,943	26.75	801	1.34
전라남도	66,987	14,811	22.11	843	1.26
경상북도	87,106	23,847	27.38	938	1.08
경상남도	90,025	28,203	31.33	887	0.99
제주도	13,975	7,349	52.59	706	5.05

<표 A-72> 집락추출 8%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,266,145	48,940	3.87	13,376	1.06	57,063	4.51
서울특별시	244,581	5,148	2.10	1,004	0.41	5,018	2.05
부산광역시	89,003	3,708	4.17	861	0.97	3,791	4.26
대구광역시	60,561	2,898	4.79	769	1.27	3,365	5.56
인천광역시	63,626	2,942	4.62	892	1.40	3,486	5.48
광주광역시	35,236	2,066	5.86	629	1.79	2,641	7.50
대전광역시	35,501	2,326	6.55	789	2.22	2,966	8.35
울산광역시	24,527	2,029	8.27	670	2.73	2,400	9.79
경기도	247,586	4,588	1.85	1,086	0.44	5,386	2.18
강원도	46,640	3,051	6.54	908	1.95	3,753	8.05
충청북도	44,276	2,540	5.74	745	1.68	3,310	7.48
충청남도	56,916	2,827	4.97	848	1.49	3,614	6.35
전라북도	59,599	2,794	4.69	801	1.34	3,429	5.75
전라남도	66,987	2,875	4.29	843	1.26	3,497	5.22
경상북도	87,106	3,656	4.20	938	1.08	4,033	4.63
경상남도	90,025	3,650	4.05	887	0.99	4,062	4.51
제주도	13,975	1,842	13.18	706	5.05	2,312	16.54

<표 A-73> 집락추출 9%, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,424,413	447,702	31.43	13,706	0.96
서울특별시	274,855	91,525	33.30	993	0.36
부산광역시	100,249	37,309	37.22	870	0.87
대구광역시	68,229	28,587	41.90	779	1.14
인천광역시	71,495	23,571	32.97	912	1.28
광주광역시	39,567	13,873	35.06	674	1.70
대전광역시	39,960	16,408	41.06	803	2.01
울산광역시	27,567	11,221	40.70	677	2.46
경기도	278,393	75,714	27.20	1,117	0.40
강원도	52,439	18,171	34.65	923	1.76
충청북도	49,967	16,497	33.02	760	1.52
충청남도	63,892	16,973	26.57	890	1.39
전라북도	67,236	17,326	25.77	833	1.24
전라남도	75,455	16,049	21.27	892	1.18
경상북도	97,938	25,872	26.42	950	0.97
경상남도	101,446	30,538	30.10	922	0.91
제주도	15,725	8,068	51.31	711	4.52

<표 A-74> 집락추출 9%, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,424,413	51,013	3.58	13,706	0.96	58,825	4.13
서울특별시	274,855	5,225	1.90	993	0.36	5,104	1.86
부산광역시	100,249	3,871	3.86	870	0.87	3,893	3.88
대구광역시	68,229	2,986	4.38	779	1.14	3,457	5.07
인천광역시	71,495	3,116	4.36	912	1.28	3,654	5.11
광주광역시	39,567	2,220	5.61	674	1.70	2,767	6.99
대전광역시	39,960	2,430	6.08	803	2.01	3,099	7.76
울산광역시	27,567	2,155	7.82	677	2.46	2,472	8.97
경기도	278,393	4,765	1.71	1,117	0.40	5,498	1.97
강원도	52,439	3,152	6.01	923	1.76	3,866	7.37
충청북도	49,967	2,655	5.31	760	1.52	3,404	6.81
충청남도	63,892	3,004	4.70	890	1.39	3,693	5.78
전라북도	67,236	2,871	4.27	833	1.24	3,499	5.20
전라남도	75,455	3,055	4.05	892	1.18	3,665	4.86
경상북도	97,938	3,753	3.83	950	0.97	4,104	4.19
경상남도	101,446	3,824	3.77	922	0.91	4,208	4.15
제주도	15,725	1,931	12.28	711	4.52	2,442	15.53

<표 A-75> 10%표본, 노출제한기법 적용 전후 유일성 비교

시도	가구수	노출제한기법 적용 전		노출제한기법 적용 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,582,681	480,483	30.36	13,927	0.88
서울특별시	305,497	97,919	32.05	1,002	0.33
부산광역시	111,307	40,064	35.99	883	0.79
대구광역시	75,808	30,902	40.76	801	1.06
인천광역시	79,599	25,304	31.79	928	1.17
광주광역시	43,996	14,982	34.05	685	1.56
대전광역시	44,369	17,676	39.84	817	1.84
울산광역시	30,682	12,187	39.72	702	2.29
경기도	309,186	80,829	26.14	1,121	0.36
강원도	58,319	19,598	33.60	933	1.60
충청북도	55,374	17,738	32.03	778	1.40
충청남도	71,029	18,214	25.64	916	1.29
전라북도	74,702	18,524	24.80	853	1.14
전라남도	83,830	17,223	20.55	896	1.07
경상북도	108,935	27,825	25.54	958	0.88
경상남도	112,584	32,740	29.08	915	0.81
제주도	17,464	8,758	50.15	739	4.23

<표 A-76> 10%표본, 변수추가 후 유일성 비교

시도	가구수	현 제공기준		변수 추가 전		변수 추가 후	
		유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)	유일건수 (Us)	구성비 (%)
전국	1,582,681	52,708	3.33	13,927	0.88	60,537	3.82
서울특별시	305,497	5,386	1.76	1,002	0.33	5,226	1.71
부산광역시	111,307	4,004	3.60	883	0.79	4,000	3.59
대구광역시	75,808	3,099	4.09	801	1.06	3,512	4.63
인천광역시	79,599	3,218	4.04	928	1.17	3,761	4.72
광주광역시	43,996	2,291	5.21	685	1.56	2,853	6.48
대전광역시	44,369	2,532	5.71	817	1.84	3,220	7.26
울산광역시	30,682	2,224	7.25	702	2.29	2,566	8.36
경기도	309,186	4,873	1.58	1,121	0.36	5,577	1.80
강원도	58,319	3,274	5.61	933	1.60	3,981	6.83
충청북도	55,374	2,756	4.98	778	1.40	3,525	6.37
충청남도	71,029	3,092	4.35	916	1.29	3,787	5.33
전라북도	74,702	2,998	4.01	853	1.14	3,623	4.85
전라남도	83,830	3,154	3.76	896	1.07	3,782	4.51
경상북도	108,935	3,891	3.57	958	0.88	4,239	3.89
경상남도	112,584	3,910	3.47	915	0.81	4,354	3.87
제주도	17,464	2,006	11.49	739	4.23	2,531	14.49

참 고 문 헌

김은석 (1995), “예측적 접근에 의한 소지역 통계 추정”, 고려대 석사학위논문.

조란 (2003), “시·군·구 실업률 산출을 위한 소지역 추정방법”, 숙명여대 석사학위논문.

정동명 · 강동환 (2007), “마이크로자료의 활용도 제고를 위한 비밀보호방법” 통계연구 결과보고서, 통계청.

정동명 · 김종익 · 강동환 (2007), “인구센서스자료의비밀보호방법” 「통계연구」, 제12권 제1호.

Agrawal, D. and Aggarwal, C. C. On the Design on Privacy Preserving Data Mining Algorithms, Proceedings of the ACM SIGPODS, (2001) 247-255

Agrawal, R. and Srikant, R. Privacy Preserving Data Mining, Proceedings of the ACM SIGMOD, (2000) 439-450

Bacher, J., Brand, R. and Bender, S. Re-identifying Register Data by Survey Data usingCluster Analysis: An Empirical Study, International Journal of Uncertainty, Fuzziness,

Knowledge-Based Systems, 10 (5) (2002) 589-608

Benedetti, P. and Franconi, L. Statistical and Technological Solutions to the Controlled Data Dissemation. In: Pre-proceedings of New Techniques and Technologies for Statistics, Volume 1, Sorrento (1998) 225-232

Bethlehem, J. A., Keller, W. J. and Pannekoek, J. Disclosure Control of Microdata, Journal of the American Statistical Association, 85 (1990) 38-45

Bilenko, M., Mooney, R., Cohen, W., Ravikumar P., and Fienberg, S. Adaptive Name Matching in Information Integration, IEEE Intelligent Systems, 18 (5) (2003) 16-23

Brand, R. Microdata Protection Through Noise Addition, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, Springer: New York (2002)

Dalenius, T. and Reiss, S.P. Data-swapping: A Technique for Disclosure Control, Journal of Statistical Planning and Inference, 6 (1982) 73-85

Dandekar, R. A., Domingo-Ferrer, J., and Sebe, F. LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection, in (J. Domingo-Ferrer, ed.)

Inference Control in Statistical Databases, Springer: New York (2002).

Dandekar, R., Cohen, M., and Kirkendal, N. Sensitive Microdata Protection Using Latin Hypercube Sampling Technique, in (J. Domingo-Ferrer, ed.) Inference Control in Statistical

Databases, Springer: New York (2002)

Defays, D. and Anwar, M. N. Masking Microdata Using Micro-aggregation, Journal of Official Statistics, 14, (1998) 449-461

De Waal, A. G. and Willenborg, L.C.R.J. Global Recodings and Local Suppressions in Microdata Sets, Proceedings of Statistics Canada Symposium 95, (1995) 121-132

De Waal, A. G. and Willenborg, L.C.R.J. A View of Statistical Disclosure Control for Microdata, Survey Methodology, 22, (1996) 95-103

Domingo-Ferrer, J. and Mateo-Sanz, J. M. Practical Data-Oriented Microaggregation for Statistical Disclosure Control, IEEE Transactions on Knowledge and Data Engineering, 14 (1), (2002) 189-201

Domingo-Ferrer, J. and Torra, V. A Quantitative Comparison of Disclosure Control Methods for Microdata, in (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds. Confidentiality, Disclosure Control and Data Access: Theory and Practical Applications, North Holland, (2001) 111-134

Domingo-Ferrer, J. and Torra, V. Statistical Data Protection in Statistical Microdata Protection via Advanced Record Linkage, Statistics and Computing, 13 (4), (2003) 343-354

Drew, D., Singh, M. P. and choudhry, G. H., Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. Survey Methodology, 8, (1982), 17-47.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map, Los Alamos National Laboratory Technical Report LA-UR-01-6428 (2001)

Elliott, M. A., Manning, A. M., and Ford, R. W. A Computational Algorithm for Handling the Special Uniques Problem, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 493-510

Elliott, M. A., Skinner, C. J., and Dale, A. Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk, in

Statistical Data Protection '98, Eurostat, Brussels, Belgium, (1998), 261-265, also Research in Official Statistics, 1 (2) 53-68

Fellegi, I. P., and Sunter, A. B. A Theory for Record Linkage, Journal of the American Statistical Association, 64, (1969) 1183-1210

Fienberg, S. E. Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, commissioned by Committee on National Statistics of the National Academy of Sciences (1997)

Fienberg, S. E., and Makov, U. Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, Journal of Official Statistics, 14 (1998) 385-397

Fienberg, S. E., Makov, E. U. and Sanil, A. P., A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data, Journal of Official Statistics, 14, (1997) 75-89

Fienberg, S. E., Makov, E. U. and Steel, R. J. Disclosure Limitation using Perturbation and Related Methods for Categorical Data, Journal of Official Statistics, 14, (1998), 485-502

Fuller, W. A. Masking Procedures for Microdata Disclosure Limitation, Journal of Official Statistics, 9, (1993) 383-406

Ghosh, M., and Rao, J.N.K., Small area estimation: An appraisal, Statistical Science, 9, (1994), 55-93.

Iyengar, V. Transforming Data to Satisfy Privacy Constraints, Association of Computing Machinery, Special Interest Group on Knowledge Discovery and Datamining '02 (2002)

Jiang, J., Lahiri, P., Wan, S., A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation. The Annals of Statistics, 30, (2002), 1782-1810.

Kennickell, A. B. Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances, in Record Linkage Techniques 1997, Washington, DC: National Academy Press, 248-267 (available at http://www.fcsm.gov) (1999)

Kim, J. J. A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1986) 303-308

Kim, J. J. Subdomain Estimation for the Masked Data, American Statistical Association,

Proceedings of the Section on Survey Research Methods, (1990) 456-461

Kim, J. J. and Winkler, W. E. Masking Microdata Files, American Statistical Association, Proceedings of the Section on Survey Research Methods, (1995) 114-119

Lambert, D. Measures of Disclosure Risk and Harm,Journal of Official Statistics, 9, (1993) 313-331

Little, R. J. A. Statistical Analysis of Masked Data, Journal of Official Statistics, 9, (1993) 407-426

Little, R. J. A. and Liu, F. Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata, American Statistical Association, Proceedings of the Section on Survey Research Methods, (2003)

Marsh, C., C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, and N. Walford , The Case for Samples of Anonymized Records from the 1991 Census Journal of the Royal Statistical Society A, 154, (1991) 305-340.

Muralidhar, K., Parsa, R. and Sarathy, R. A General Additive Data Perturbation Method for Database Security, Management Science, 45 (10), (1999) 1399-1415

Muralidhar, K., Sarathy, R. and R. Parsa, R. An Improved Security Requirement for Data Perturbation with Implications for E-Commerce, Decision Sciences, 32 (4), (2001) 683-698

Paas, G. Disclosure Risk and Disclosure Avoidance for Microdata, Journal of Business and Economic Statistics, 6, (1988) 487-500

Palley, M. A. and Simonoff, J. S. The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases, ACM Transactions on Database Systems, 12 (4), (1987) 593-608

Polettini, S. Maximum Entropy Simulation for Microdata Protection, Statistics and Computing, 13 (4), (2003) 307-320

Polettini, S. and Stander, J. A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation, in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, Springer: New York, (2004).

Prasad, N.G.N., and Rao, J.N.K., The Estimation of the Mean Squared Error of Small Area

Estimators, Journal of the American Statistical Association, 93, (1990), 720-729.

Purcell, N. J. and Kish, L., Estimation for Small Domain. Biometrics, 35, (1979), 365-371.

Purcell, N. J. and Linacre, S., Techniques for the Estimation for Small Area Characteristics, Unpublished Paper, (1976), Australian Bureau of Statistics, Canberra.

Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. Multiple Imputation for Statistical Disclosure Limitation, Journal of Official Statistics, 19, (2003) 1-16

Reiss, J.P. Practical Data Swapping: The First Steps, ACM Transactions on Database Systems, 9, (1984) 20-37

Reiter, J.P. Satisfying Disclosure Restrictions with Synthetic Data Sets, Journal of Official Statistics, 18, (2002) 531-543

Reiter, J.P. Inference for Partially Synthetic, Public Use Data Sets, Survey Methodology, (2003)

Reiter, J.P. Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study, Journal of the Royal Statistical Society, A, (2004)

Rinott, Y. On Models for Statistical Disclosure Risk Estimation, UNECE Work Session on Statistical Data Confidentiality, Luxembourg, April 2003, http://www.unece.org/stats/documents/ 2003/04/confidentiality/wp.16.e.pdf

Roque, G. M.. Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. Dissertation, University of California at Riverside, (2000)

Sardnal, C. E., and Hidiroglou, M. A. (1989), Small Domain Estimation: A Conditional Analysis, Jounal of the American Statistical Association, 84, 266-275.

Sarathy, R., Muralidhar, K., and Parsa, R.. Perturbing Non-Normal Attributes: The Copula Approach, Management Science, 48 (12), (2002) 1613-1627

Scheuren, F. and Winkler W. Regression Analysis of Data Files that are Computer Matched . Part II, Survey Methodology (1997) 157-165

Schlorer, J. Security of Statistical Databases: Multidimensional Transformation, ACM Transactions on Database Systems, 6, (1981) 91-112

Skinner, C. J. and Elliot, M. A. A Measure of Disclosure Risk for Microdata, Journal of the Royal

Statistical Society, B 64 (4), (2001), 855-867

Skinner, C. J. and Holmes, D. J. Estimating the Re-identification Risk per Record in Microdata, Journal of Official Statistics, 14 (1998) 361-372

Sweeney, L. Computational Disclosure Control for Medical Microdata: The Datafly System, in Record Linkage Techniques 1997, Washington, DC: National Academy Press (1999) 442-453

Sweeney, L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 571-588

Trottini, M. and Fienberg, S. E. Modelling User Uncertainty for Disclosure Risk and Data Utility, International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10 (5), (2002) 511-528

Van Den Hout, A. and Van Der Heijden, P. G. M. Randomized Response, Statistical Disclosure Control, and Misclassification: A Review, International Statistical Review, 70 (2) (2002) 269-288

Willenborg, L. and De Waal, T. Statistical Disclosure Control in Practice, Vol. 111, Lecture Notes in Statistics, Springer-Verlag, New York (1996)

Willenborg, L. and De Waal, T. Elements of Statistical Disclosure Control, Vol. 155, Lecture Notes in Statistics, Springer-Verlag, New York (2000)

Winkler, W. E. Matching and Record Linkage, in B. G. Cox (ed.) Business Survey Methods, New York: J. Wiley, (1995) 355-384

Winkler, W. E. Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Research in Official Statistics, 1, (1998) 87-104

Winkler, W. E. Issues with Linking Files and Performing Analyses on the Merged Files, Proceedings of the Sections on Government Statistics and Social Statistics, American Statistical Association, (1999) 262-265

Winkler, W. E. Single Ranking Micro-aggregation and Re-identification, Statistical Research Division report RR 2002/08 at http://www.census.gov/srd/www/byyear.html (2002)

보고서 발간 참여자

■ (사) 한국통계학회

연구진

강현철 | 호서대학교 정보통계학과 교수 (연구책임자)

한상태 | 호서대학교 정보통계학과 교수 (공동연구원)

송주원 | 고려대학교 통계학과 교수 (공동연구원)

김은석 | 호서대학교 정보통계학과 겸임교수 (공동연구원)

■ 통계청 조사관리국

김광섭 | 조사관리국 국장

강창익 | 인구총조사과 과장

이민경 | 인구총조사과 사무관

황수린 | 인구총조사과 주무관

2010 인구주택총조사 자료분석 및 활용제고방안 연구 (Ⅱ)

발 행 인 이 인 실

발 행 처 통계청

인 쇄 일 2010년 12월

발 행 일 2010년 12월