제5장

2005 인구주택총조사자료의 개인정보 노출제한방법

정동명 · 정미옥

제1절 서 론

통계이용자들은 심층적인 자료 분석을 위해 통계작성기관에서 마이크로자료를 제공해 줄 것을 요구하고 있으며, 이에 대응하여 통계작성기관은 작성하고 있는 통계에 대한 마이크로자료를 이용자들에게 제공하는 것을 법적으로 제도화하고 있는 실정이다. 한편, 정보통신의 발달과 대용량 DB의 구축 등으로 인해서 기업 및 여러 기관에서는 신용평가나 타깃 마케팅 등 다양한 목적을 위해 개인의 여러 가지 신상정보들을 보유하고 있으며, 이러한 정보는 통계작성기관에서 제공하는 마이크로자료의 일부 항목과 상당히 유사한 부분이 많기 때문에 마이크로자료를 그대로 제공했을 시에는 개인정보의 노출 위험성이 매우 커지게 된다. 따라서 통계작성기관에서는 마이크로자료의 비밀보호를 위해 응답자의 개인 식별이 가능한 이름, 주소 등을 모두 제거한 후에 자료를 제공하고 있다. 그렇지만 이렇게 제거된 자료도 노출의 위험성이 여전히 남아 있는데, 이는 어떤 사람이나 기업이 마이크로자료 파일로부터 특정한 개인을 확인할 수 있는 특성 정보들을 가지고 있을 수 있기 때문이다.

통계작성기관이 마이크로자료 파일을 제공할 때는 외부 이용자가 그 파일로부터 특정 개인을 연결시킬 수 없다는 것을 확실하게 보장해야만

한다. 마이크로자료에서 개인정보의 노출위험을 제거하는 가장 안전한 방법은 아예 마이크로자료를 제공하지 않는 것이지만, 이는 자료수집의 근본취지에 위배되는 행위이므로 통계작성기관은 이러한 노출위험을 최소화하면서 최대한 유용한 정보가 들어 있는 자료를 제공하기 위해 노력하고 있다. 미국・호주 등 외국의 통계작성기관에서는 이미 오래 전부터 노출의 위험성을 깊이 인식하고 개인정보의 비밀보호를 위해 다양한 방법들을 적용하여 마이크로자료를 제공하고 있다. 미국 센서스국에서는 2000년 인구센서스 결과에 다양한 기법들을 적용시킨 후 마이크로자료를 제공하고 있으며, 호주의 경우에도 인구주택총조사 마이크로자료의 비밀보호를 위해 이름이나 주소와 같이 직접적으로 개인을 식별가능하게 하는 변수를 삭제하고, 상세한 수준의 범주를 제거하는 방법을 사용하였다. 이 밖에 네덜란드와 스웨덴, 영국 등에서도 노출위험의 심각성과 비밀보호방법의 필요성 등에 대해 관심을 가지고 많은 연구가 진행되었으며, Dalenius & Reiss(1982), Kim(1986), Bethelehem et al.(1990), Marsh et al.(1991), Fuller(1993) 등 여러 학자들도 노출의 위험성과 비밀보호에 대한 여러 방법을 제시하였다.

우리나라의 경우 비밀보호방법에 대한 인식부족 등으로 인해 연구가 미흡하였으나, 최근 들어 통계청과 학계 등에서 연구가 활발히 진행되고 있다. 특히 통계청에서는 개인의 비밀을 보호하면서 마이크로자료를 제공할 수 있는 방법을 개발코자 수년 전부터 통계적 비밀보호기법에 대한 연구에 노력해 왔으며, 박원환・황조연(2004)과 정동명 외(2007) 등이 통계자료의 비밀보호에 대한 개념과 실제 자료에 적용한 사례를 연구하였다.

본 연구에서는 통계청에서 2005년도에 실시한 인구주택총조사의 표본조사 결과를 대상으로 응답자의 비밀이 보호된 2% 마이크로자료를 작성하는 일련의 과정을 소개하고자 한다. 제2절에서는 비밀보호의 기본개념과 방법을 간단히 소개하고, 제3절에서는 2005 인구주택총조사의 표본조사 결과자료에 대해 다양한 비밀보호방법을 적용하여 2% 마이크로 자료파일을 작성하는 과정을 설명한다. 그리고 본 연구의 최종적인 결론과 향후 고려할 사항 등을 제4절에서 간략히 언급하고자 한다.

제2절 노출과 노출위험

1. 노출과 유일성

통계작성기관이 자료를 수집·정리하여 다양한 형태의 통계정보로 제공할 경우 이를 통해 응답자의 특성이 파악되는 것을 노출(disclosure)이라 하는데, 노출은 어떤 경우에도 발생되지 않도록 하는 것이 바람직하다. 그리고 노출은 개인의 식별(identification)과 관련이 높은데 개인의 민감한 정보가 노출된다면 개인에 대한 식별이 가능하게 된다. 노출은 자료이용자들이 사전에 가지고 있는 유용한 정보의 성격과 양에 따라 좌우되므로, 제공되는 마이크로자료의 정보와 자료이용자들이 가지고 있는 사전정보가 서로 일치하지 않도록 자료파일을 구성하면 노출을 제한할 수 있을 것이다.

노출을 제한하는 방법은 자료의 종류와 형태에 따라 구분되는데, 자료의 종류별로 살펴보면 매크로자료인 경우 셀감추기(cell suppression), 반올림(rounding), 임의변조(random perturbation) 등이 있으며, 마이크로자료인 경우에는 익명화(anonymisation), 표본추출(sampling), 그룹화, 자료교환(data swapping) 등의 다양한 방법들이 있다. 또한 자료의 형태별로는 이산형 자료인 경우 자료교환, 코딩접근법(coding approach), 그룹화 등의 방법이 널리 활용되고 있고, 연속형 자료인 경우에는 자료교환, 반올림, 그룹화, 가법잡음(additive noise), 승법잡음(multiplicative noise) 등의 방법이 활용되고 있다. 이러한 비밀보호방법에 대한 자세한 내용은 Dalenius & Reiss(1982), Kim(1986), Kim & Winkler(2001), Bethelehem et al.(1990) 등을 참고하기 바란다.

한편, 유일성(uniqueness)이란 전체 자료파일에서 조사단위의 특성이 유일하게 존재하는 것을 말하며, 어떤 조사단위가 식별될 가능성을 나타내는 측도로 사용된다. 가령, 100명으로 구성된 모집단에서 나이가 100세인 사람이 단 한 명 있다면 그 사람은 모집단에서 유일하다고 하는데, 이렇게 유일한 사람은 다른 사람들에 비해 자료에서 식별될 가능

성이 상당히 높게 된다. 주어진 자료에서 유일성은 하나의 변수만으로도 파악할 수 있고, 여러 개의 변수들을 조합하여 파악할 수도 있다. 예를 들어, 100명 중 나이가 60세인 사람이 3명 있다고 하더라도 직업이라는 변수를 포함하여 회사원이면서 나이가 60세인 사람은 단 한 명일 수도 있다. 따라서 고려할 변수가 많아질수록 자료의 유일성은 점점 더 커지게 된다.

2. 노출위험의 확률모형

노출은 다음의 3가지 조건을 만족하는 경우에 발생하며, 이 조건에서 언급한 자료파일들 중 어느 하나에도 나타나지 않는다면 노출은 일어나지 않는다고 한다.

- 어떤 사람이 특정 변수에 대해 모집단에서 유일하다.
- 그 사람은 어떤 조사에서 마이크로자료 파일에 포함되어 있다.
- 그 사람은 외부인이 작성한 또 다른 자료파일에도 포함되어 있다.

이러한 노출의 발생조건 하에서 노출위험(disclosure risk)은 확률적 모형으로 표현할 수 있는데, 이를 위해 다음과 같은 기호를 정의한다.

- A : 관심의 대상인 사람
- S_1 : 통계작성기관의 마이크로자료로 구성된 파일 1
- S_2 : 외부인(intruder)에 의해 구성된 파일 2
- U_p : 모집단의 유일성집단
- U_s : 표본의 유일성집단

만약 금융기관이나 이웃주민 등과 같은 외부인이 자신들이 직접 작성한 파일(S_2)에 관심 있는 어떤 사람(A)이 포함되어 있다는 것을 모르고 있다면, 특정인의 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 자신들의 파일에 관심 있는 특정인(A)이 포함되어 있다는 것을 이미 알고 있다고 한다면 $\Pr(A \in S_2) = 1$ 이 될 것이며, 이

때 노출위험은 다음과 같이 된다.

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

제3절 인구주택총조사 자료의 비밀보호

1. 인구주택총조사 자료

인구주택총조사(census)는 0과 5로 끝나는 연도를 기준으로 매 5년마다 전국을 대상으로 통계청에서 실시하고 있다. 본 연구에서는 가장 최근에 실시한 "2005 인구주택총조사"의 표본조사 결과자료를 이용하여 분석하였다. 인구주택총조사의 조사표는 전수조사표와 표본조사표로 구분되어 있는데 전수조사표는 기본적인 특성을 파악하기 위해 21개 항목으로 구성되어 있으며, 표본조사표는 전수조사항목 이외에 보다 세부적인 특성을 파악하기 위한 20개 항목을 추가하여 총 41개 항목으로 구성되어 있다. 이외에 추가로 16개 시·도별로 각각 서로 다른 조사항목 3개가 포함되어 전체적으로는 44개 조사항목으로 구성되어 있다. 이에 대한 자세한 내용은 「2005 인구주택총조사 조사지침서(2005)」를 참고하기 바란다.

한편, 현행 인구주택총조사에서는 60~80가구를 하나의 조사구로 설정한 후, 이 중 10%를 표본조사구로 추출하고 표본조사구내 모든 가구는 표본조사표를 작성하도록 하고 있다. 여기서 가구란 1인 또는 2인 이상이 모여서 취사, 취침 등 생계를 같이하는 생활단위를 말하는데, 크게 일반가구와 집단가구, 외국인가구로 구분이 된다. 또한 조사구란 전국의 모든 지역에 대하여 식별이 명확한 지형지물을 기준으로 지도상에서 일정한 가구수가 포함되도록 분할한 조사담당 구역을 말한다. 조사구는 아파트조사구, 보통조사구, 섬조사구, 기숙시설조사구, 특수사회시설조사구, 관광호텔 및 외국인 거주지역 조사구 등 6개로 구분된다. 본 연구에서는 6개의 조사구 중 아파트조사구와 보통조사구, 섬조사구만을 대상으로 하였고 가구도 일반가구만을 분석하기로 하였는데, 이는 자료

활용 측면에서 집단가구보다는 개별가구의 특성을 파악하는 데 초점을 두고 있기 때문이다.

2005 인구주택총조사 결과에 의하면, <표 5-1>에 나타난 바와 같이 총 조사구 규모는 265,298개이고 인구는 45,772,054명, 가구는 15,895,481가구, 주택은 12,693,578호가 있는 것으로 나타났다. 10% 표본결과의 경우 조사구는 26,713개이고 인구는 4,455,527명, 가구는 1,582,681가구, 주택은 1,295,389호인 것으로 각각 나타났다. 이에 대한 자세한 결과는 <부록 1>을 참조하기 바란다.

〈표 5-1〉 2005 인구주택총조사 결과

(단위: 개, 명, 가구, 호)

구분	조사구	인구	가구	주택
전수결과	265,298	45,772,054	15,895,481	12,693,578
10% 표본결과	26,713	4,455,527	1,582,681	1,295,389

2. 마이크로자료 파일의 작성

가. Key 변수의 선정

노출을 제한하여 마이크로자료 파일을 작성하고자 할 때, 우선적으로 고려할 것이 외부에서 식별가능성이 높다고 판단되는 항목인 key 변수의 선정이다. key 변수란 데이터와 특정인 간에 일대일 대응을 통해서 특정 레코드(record)가 어떤 사람에 대한 정보인지를 식별 가능하게 하는 변수를 일컫는다. 잘 알려진 key 변수로는 이름, 주소뿐만 아니라 가구구성, 연령, 인종, 성별, 거주 지역, 직업 등이 있다. 통계적 비밀보호 방법론에서 마이크로자료 파일을 작성하기 위해서는 이러한 key 변수를 선정하고, 선정된 key 변수들을 바탕으로 외부인(intruder)이 특정 개인을 식별할 수 있을지에 대해 검토하게 된다. 그러므로 key 변수를 선정할 때, 외부인이 이미 알고 있을 가능성이 높은 정보나 외부기관의 데이터에 이미 포함되어 있을 가능성이 있는 정보들을 우선적으로 선별하게

된다. 이 방법은 경험에 근거하여 key 변수를 선정하기 때문에 현실적인 가능성들을 고려하고 있다는 큰 장점을 가지고 있다.

인구주택총조사의 표본조사항목에는 개인의 특성에 관한 민감한 정보들이 많이 있기 때문에 특히 key 변수의 선정에 신중을 기하였다. key 변수의 선정을 위해 외부기관에서 보유하고 있는 자료들 중 인구주택총조사의 조사항목과 중복되어 식별될 가능성이 높다고 판단되는 항목을 우선 선정한 후, 빈도분석(frequency analysis) 등을 실시하여 각 항목의 빈도수와 분포형태 등 보다 자세한 자료의 특성을 시도별로 파악하였다. 이는 지역간 분포를 비교해 보고, 항목의 각 범주별로 최소 응답수를 확인함으로써 특정 지역이나 항목에 있어서 특징적으로 나타날 수 있는 특성들을 살펴보기 위함이었다. 이러한 면밀한 검토 작업을 거친 후, 최종적으로 12개의 key 변수를 선정하였으며, 이에 대한 결과는 다음과 같다.

1) 인구에 관한 사항

가구원에 관한 사항에는 개인에 대한 민감한 정보들이 많이 들어 있어 다른 분야에 비해 key 변수를 더 많이 고려하였으며, 외부에서 식별 가능성이 높다고 판단되는 8개 항목(성별, 나이, 가구주와의 관계, 교육정도, 혼인상태, 활동제약, 산업, 직업)을 key 변수로 선정하였다. 선정된 key 변수의 빈도분석결과를 살펴보면, 성별의 경우 남녀가 비슷한 비율을 차지하면서 지역별로도 거의 차이가 없었다. 나이는 만 100세 이상의 고령자가 총 24명으로 각 시도에 적게는 1명에서 많게는 6명 정도가 분포하고 있으며, 특·광역시보다 도지역에서 평균연령이 다소 높게 나타났다. 가구주와의 관계 항목에서는 가구주, 자녀, 가구주의 배우자의 비중은 매우 높은 반면, 그 이외의 범주에 대해서는 상대적으로 낮게 나타났다. 교육정도는 고등학교 졸업이 약 19~31.9% 정도로 지역별로 다소 차이를 보이고 있지만 가장 큰 비중을 차지하고 있었다. 다음으로 활동제약 항목은 육체적·정신적 제약과 활동제약 부분의 세부항목이 있는데, 대부분의 지역에서 제약 없음이 압도적으로 나타났다. 하지만 활동제약이 있는 인구가 매우 적더라도 활동제약이 있는 경우 외부에서 쉽게 식별 가능하기 때문에 key 변수로 선정하였다. 취업인구에 대해 직장

의 산업과 직업을 묻는 항목에서는 표준산업분류 및 표준직업분류를 기준으로 대분류에 대하여 분석하였으며, 산업항목의 경우 지역별 산업의 구성비에는 다소 차이가 있었으나 대부분의 지역에서 도매 및 소매업과 제조업, 농업이 큰 비중을 차지하고 있었다. 직업항목의 경우 군인이 강원도 1.89%를 제외하고 대부분의 지역에서 약 0.5% 이하로 매우 적게 분포하였다. 마지막으로 혼인상태는 전국적으로 약 26.6%, 사별과 이혼이 각각 8.5%, 3%인 것으로 나타났다.

2) 가구에 관한 사항

가구에 관한 사항에서는 전수항목 중 3개 항목(가구구분, 점유형태, 주택소유여부)을 외부에서 식별 가능한 정보라 판단되어 key 변수로 선정하였다. 이들 변수의 분석결과를 살펴보면, 가구구분의 경우 전국적으로 1인가구의 비율이 약 20%로 나타났고, 가족과 가족 이외의 가구는 0.45%로 그 비율이 상당히 낮게 나타났다. 점유형태의 경우, 자기 집에 살고 있는 가구의 비율은 서울지역이 약 45%로 전국평균 58.6%보다 낮고, 전세로 살고 있는 가구는 전국 평균 19.5%보다 높은 31.85%의 비율을 보이고 있었다. 또한 점유형태 항목의 세부항목인 거주하고 있는 집이 주거전용인지 영업겸용인지를 파악하는 항목은 주택에 관한 사항과 중복되는 개념이 있고, 개인을 식별해 내는 데 큰 요인이 될 수 없다는 판단으로 분석에서 제외를 하였다. 주인가구 및 주택소유여부 항목의 경우 주인가구이면서 다른 곳에 주택미소유인 가구의 비율이 전국적으로 51.8%로 나타났다.

3) 주택에 관한 사항

주택에 대한 정보는 개인을 식별해 내는 데 크게 영향을 미칠 만한 변수는 많지 않기 때문에 거처종류 항목만을 key 변수로 선정하였다. 거처종류 항목의 단독주택 비율이 가장 낮은 지역은 울산으로 약 29.8%였고, 전남은 73.2%로 가장 높게 나타났다. 또한 아파트는 전국적으로 약 40%의 비율을 차지하고 있었다.

나. 표본의 추출

현재 통계청에서 제공하고 있는 인구주택총조사의 마이크로자료는 전체 규모의 2%에 해당하는 표본조사 결과자료이다. 이는 전체 규모로는 2%이지만 표본조사 결과만으로 보면 약 20%에 해당된다. 따라서 전체 10% 표본조사구 내에 있는 모든 가구의 명부를 추출틀(sample frame)로 사용하여 표본가구를 추출하였다. 이때 추출방법은 개인의 식별 가능성의 최소화에 적합한 계통추출법(systematic sampling)을 적용하였으며, 이렇게 추출된 결과가 <표 5-2>에 주어져 있다.

〈표 5-2〉 2% 표본현황

(단위: 개, 명, 가구, 호)

지역	조사구	인구	가구	주택
전국	**26,707**	**892,024**	**316,536**	**294,062**
서울	5,099	175,727	61,101	53,835
부산	1,881	64,597	22,269	20,905
대구	1,254	44,121	15,157	13,795
인천	1,368	47,063	15,916	14,879
광주	716	26,253	8,803	8,209
대전	749	25,579	8,874	7,918
울산	507	18,294	6,131	5,709
경기	5,146	186,697	61,840	56,658
강원	1,001	30,522	11,668	11,082
충북	930	29,476	11,072	10,488
충남	1,215	37,708	14,198	13,656
전북	1,301	39,442	14,950	14,526
전남	1,450	41,684	16,762	16,544
경북	1,872	55,182	21,790	20,993
경남	1,920	59,568	22,514	21,551
제주	298	10,111	3,491	3,314

다. 노출위험

앞 절에서 언급한 노출위험의 확률모형을 적용하여 2005 인구주택총조사 자료의 노출위험 정도를 알아보기 위해 다음과 같이 기호를 정의하였다.

- A : 관심의 대상인 사람
- S_1 : 새로 추출한 2% 표본파일
- S_2 : 외부인에 의해 구성된 파일
- U_p : 모집단(10% 표본조사결과)의 유일성집단
- U_s : 2% 표본의 유일성집단

10% 표본조사결과를 모집단으로 하고 새로 추출한 2% 표본조사결과를 표본으로 한 후, 12개 key 변수의 각 조합별로 빈도수가 1이 되는 경우를 파악한 결과가 <표 5-3>에 주어져 있다. 여러 조합들 중 12개 변수를 모두 고려한 조합이 유일성의 발생이 최대가 되는 경우인데, 모집단에서는 총 1,814,800개가 있으며 표본에서는 458,160개가 있는 것으로 각각 나타났다. 따라서 어떤 사람(A)이 모집단에서 유일하게 될 확률은 약 40.7%가 되고, 표본에서 유일하게 될 확률은 약 51.4%가 된다.

$$\Pr(A \in U_P) = \frac{1,814,800}{4,455,527} = 0.40731$$

$$\Pr(A \in U_S) = \frac{458,160}{892,024} = 0.51362$$

이 결과에 따르면 표본에서의 유일성이 모집단에서의 유일성보다 더 높게 나타났는데, 이는 모집단보다 표본에서 유일한 사람이 더 많이 발생한 것을 보여준다.

이러한 유일성에 대한 결과를 이용하여 노출위험의 발생 가능성이 최대가 되는 가장 극단적인 경우를 고려하여, 외부인이 직접 작성한 파일(S_2)에 어떤 특정인이 포함되어 있다는 것을 이미 알고 있다고 가정하고 노출위험을 계산하였다. 이 가정 하에서는 $\Pr(A \in S_2) = 1$ 이 되므로 어떤 특정한 사람이 모집단에서 유일하면서 인구주택총조사의 2% 표본에 포함될 확률은 약 8.1%가 된다.

$$DR(A) = \Pr(A \in S_1)\ \Pr(A \in U_P) = 0.2 \times \frac{1,814,800}{4,455,527} = 0.08146$$

여기서 $\Pr(A \in S_1) = 0.2$ 인데, 이는 제공하는 마이크로자료의 규모가 2%이므로 10% 표본조사결과에서 20%를 표본으로 추출하였기 때문이다. 위의 결과는, 외부인이 가지고 있는 파일에 어떤 사람이 포함되어 있다는 것을 알고 있다고 가정하고 2%를 마이크로자료 파일로 제공할 경우, 100명 중 약 8명 정도가 노출될 가능성이 있다는 것을 의미한다.

〈표 5-3〉 노출제한방법 사용 전 유일성

(단위: 명, %)

지역	모집단 (A)	U_P (B)	구성비 (B/A)	2%표본 (C)	U_S (D)	구성비 (D/C)
전국	4,455,527	1,814,800	40.7	892,024	458,160	51.4
서울	879,032	354,148	40.3	175,727	88,988	21.3
부산	320,174	135,746	42.4	64,597	34,333	25.9
대구	221,787	95,953	43.3	44,121	23,820	26.8
인천	235,621	107,132	45.5	47,063	26,281	29.4
광주	129,836	57,801	44.5	26,253	14,401	30.2
대전	128,849	61,245	47.5	25,579	14,942	33.4
울산	91,756	41,183	44.9	18,294	10,176	29.9
경기	931,851	351,056	37.7	186,697	88,621	19.3
강원	152,350	68,609	45.0	30,522	17,415	32.7
충북	148,015	65,355	44.2	29,476	16,452	30.5
충남	188,950	77,778	41.2	37,708	19,650	26.8
전북	196,976	76,003	38.6	39,442	19,733	25.7
전남	208,847	73,428	35.2	41,684	19,530	23.3
경북	273,566	104,404	38.2	55,182	27,118	23.5
경남	297,986	116,207	39.0	59,568	29,663	23.4
제주	49,931	28,752	57.6	10,111	7,037	44.4

라. 노출위험의 제한방법

위에서 2% 표본의 유일성에 대해 살펴본 결과, 총 892,024명 중 약 51.4%인 458,160명이 유일한 것으로 나타났다. 만약 추출된 2%의 표본에 노출제한방법을 적용하면 유일성은 더 감소하게 될 것이며, 이는 결국 개인의 식별과 노출에 대한 위험이 훨씬 더 줄어들게 됨을 의미한다. 노출을 제한하는 방법은 자료의 종류와 형태에 따라 다양한 방법이 있는데, 본 연구에서 자료의 제공범위를 제한하는 방법과 자료의 정보를 축소하는 방법을 이용하였다.

1) 제공범위의 제한

마이크로자료의 제공범위를 제한하기 위해 조사구의 경우 아파트조사구와 보통조사구, 섬조사구만을 대상으로 하였다. 가구의 경우 가족으로 이루어진 가구와 가족과 가족 이외의 사람이 함께 사는 가구, 1인가구, 가족이 아닌 남남끼리 함께 사는 5인 이하의 가구만 대상으로 하였으며, 지역자료도 시·도 단위까지만 제공하기로 하였다. 그리고 남북이산가족, 임차료, 대지면적 등 3개 항목은 조사의 특성상 공표대상에서 제외하기로 하였다.

2) 자료의 축소

인구주택총조사의 특성상 이산형 변수가 대부분이고 연속형 변수인 나이, 연건평 등도 구간으로 변환이 가능하기 때문에 본 연구에서는 주로 이산형 변수에 효과적인 그룹화(grouping)와 코딩(top-coding, bottom-coding), 하위 세부항목의 통합 및 제거 등의 방법을 이용하여 자료의 정보량을 다소 축소하였다. 각 분야별 key 변수에 대해 살펴보면, 인구에 관한 항목의 경우 성별은 남녀범주를 그대로 사용하였고, 나이는 각 세별로 되어 있던 코드를 0~84세까지는 그대로 두고, 85세 이상은 top-coding하여 하나의 범주로 처리를 하였다. 가구주와의 관계 항목은 14개로 구분되어 있던 범주를 범주간 관련성(부모세대 혹은 자녀세대, 혈연관계 혹은 비혈연관계)과 각 범주의 빈도수를 고려하여 8개 범주로 그룹화하였다. 교육정도는 학력에 대한 세부항목에서 4년제 미만과 4년제 이상으로 구분되어 있던 대학범주를 대학교라는 하나의 범주로, 석사과

정과 박사과정을 대학원범주로 통합하였다. 또한 졸업, 재학, 수료, 휴학, 중퇴라는 범주로 세분화되어 있던 교육상태 항목을 졸업과 졸업이 아닌 상태로 축소하였다. 활동제약 항목의 육체적·정신적 제약 부분은 민감하면서 그 빈도가 매우 낮은 치매와 중풍을 각각 정신적 제약과 육체적 제약으로 묶어 주었다. 다음으로 종사산업은 표준산업분류 대분류 기준으로 20개의 대분류 산업을 16개로 축소하였고, 직업 항목에서는 표준직업분류 대분류 기준 10개의 범주 중에 군인은 그 빈도가 매우 낮아 직업미상과 묶어 주었다. 마지막으로 혼인상태는 기존의 범주를 그대로 사용하였다. 가구에 관한 사항에서 가구구분은 가족과 가족 이외의 사람이 함께 사는 가구와 가족이 아닌 남남끼리 함께 사는 5인 이하의 가구를 기타의 범주로 묶어 기존 4개의 범주를 3개로 조정하였다. 점유형태 항목에서는 매월 주거비용을 내는지의 개념을 적용시켜 보증금 있는 월세와 보증금 없는 월세, 사글세를 월세라는 하나의 범주로 그룹화하였다. 주택소유여부 항목에서 주인가구여부 세부항목은 현재 주택을 소유하고 있는지의 개념을 적용하여 주인가구와 주인 아닌 가구로 범주를 축소하였고, 주택소유여부 세부항목은 기존 범주를 그대로 사용하였다. 마지막으로 주택에 관한 사항에서 거처종류 항목은 기존의 10개의 범주를 5개(단독주택, 아파트, 연립 및 다세대 주택, 비거주용 건물 내 주택, 기타)로 그룹화하였다. 이에 대한 결과가 <표 5-4>에 나타나 있으며, 각 항목에 대한 자세한 내용은 <부록 3>을 참고하기 바란다.

한편, 위에서 언급한 노출제한방법을 적용한 후 모집단과 2% 표본의 유일성을 파악한 결과, <표 5-5>에 나타난 바와 같이 모집단의 경우 총 4,455,527명 중 약 24.4%인 1,089,142명이 유일한 것으로 나타났으며, 2% 표본의 경우 총 892,024명 중 38.2%인 341,168명이 유일한 것으로 나타났다. 이는 노출제한방법을 적용하기 전보다 상당히 줄어들었음을 보여주고 있다. 이를 바탕으로 외부인이 직접 작성한 파일에 어떤 특정인(A)이 포함되어 있다는 것을 이미 알고 있다는 가정 하에서 노출위험을 계산하면 약 4.9%가 되어, 노출제한방법을 적용하기 전의 노출위험(8.1%)보다 약 40% 정도 감소한 것을 알 수 있다.

$$DR(A) = \Pr(A \in S_1)\Pr(A \in U_P) = 0.2 \times \frac{1,089,142}{4,455,527} = 0.04889$$

〈표 5-4〉 노출제한방법

방 법	변 수
공표 제외	남북이산가족, 임차료, 대지면적
그룹화(grouping)	가구주와의 관계, 교육정도, 종교종류, 활동제약, 통근통학 소요시간, 경제활동상태, 산업, 직업, 고령자 생활비, 가구구분, 점유형태, 주인가구, 거처종류, 연건평
코딩(top-coding, bottom-coding)	나이, 혼인연월, 총 출생아수, 추가계획 자녀수, 사용방수, 자동차보유대수, 연건평, 편익시설수
하위 세부항목 통합 및 제외	경제활동상태, 사용방수, 거주층, 점유형태, 거처종류, 총방수

〈표 5-5〉 노출제한방법 사용 후 유일성

(단위: 명, %)

지역	모집단 (A)	U_P (B)	구성비 (B/A)	2%표본 (C)	U_S (D)	구성비 (D/C)
전국	**4,455,527**	**1,089,142**	**0.244**	**892,024**	**341,168**	**0.382**
서울	879,032	187,034	0.213	175,727	61,628	0.351
부산	320,174	83,122	0.259	64,597	26,426	0.409
대구	221,787	59,489	0.268	44,121	18,495	0.419
인천	235,621	69,152	0.294	47,063	20,849	0.443
광주	129,836	39,254	0.302	26,253	11,812	0.450
대전	128,849	43,088	0.334	25,579	12,648	0.494
울산	91,756	27,400	0.299	18,294	8,192	0.448
경기	931,851	179,385	0.193	186,697	59,123	0.317
강원	152,350	49,783	0.323	30,522	14,714	0.482
충북	148,015	45,151	0.305	29,476	13,410	0.455
충남	188,950	50,583	0.268	37,708	15,208	0.403
전북	196,976	50,658	0.257	39,442	15,403	0.391
전남	208,847	48,679	0.233	41,684	15,155	0.364
경북	273,566	64,405	0.235	55,182	20,159	0.365
경남	297,986	69,795	0.234	59,568	21,850	0.367
제주	49,931	22,164	0.444	10,111	6,096	0.603

제4절 결 론

앞에서 살펴본 바와 같이, 본 연구에서는 2005 인구주택총조사의 2% 마이크로자료 제공을 위해 다양한 비밀보호방법을 적용하여 자료파일을 작성하는 과정을 설명하였다. 먼저 외부에서 쉽게 식별이 가능할 것으로 판단되는 12개 변수를 key 변수로 선정하였는데, 인구에 관한 사항에서는 8개 항목(성별, 나이, 가구주와의 관계, 교육정도, 혼인상태, 활동제약, 산업, 직업)을, 가구에 관한 사항에서는 3개 항목(가구구분, 점유형태, 주택소유여부)을, 그리고 주택에 관한 사항에서는 1개 항목(거처종류)을 각각 선정하였다. 또한 통계이용자들에게 제공할 2% 마이크로자료 파일을 구성하기 위해 10% 표본조사결과들 중 20%를 계통추출법으로 추출하였으며, 자료파일의 유일성과 노출의 위험을 알아보기 위해 모집단(10% 표본조사결과)과 2% 표본에서 12개 key 변수의 각 조합별로 유일성을 파악하였다. 그 결과, 모집단에서는 4,455,527명 중 1,814,800명(약 40.7%)이 유일한 것으로 나타나 이에 따른 노출 위험은 약 8.1%가 됨을 알 수 있었다. 이용자들에게 제공할 자료파일인 2% 표본에서는 892,024명 중 458,160명(약 51.4%)이 유일한 것으로 나타났으나, 다양한 노출제한방법을 적용한 결과 341,168명이 유일한 것으로 나타나 노출제한방법을 적용하기 전보다 약 25.5%가 감소하였다.

한편, 본 연구에서는 마이크로자료 파일을 작성할 때 인구주택총조사의 질문항목이 대부분 이산형 형태로 구성된 것을 감안하여 이산형 변수의 노출제한방법을 주로 적용하였으나, 나이 등 일부 항목들은 연속형 변수이므로 자료를 범주화하여 제공하는 것이 더 바람직할 것이다. 또한 10% 표본조사결과를 모집단으로 하여 자료 분석 등을 하였으나, 만약 전수조사결과를 모집단으로 하고 유일성 등을 파악한다면 자료의 수가 많을수록 유일성이 감소하는 특성으로 인해 본 연구에서 계산한 것보다 훨씬 더 노출위험이 줄어들 것이다. 그러나 활동제약이나 산업 등의 key 변수는 10% 표본조사결과에만 있고 전수조사결과에는 없기 때문에 전수조사결과에서의 노출위험을 계산하지는 못하였다. 향

후 이에 대한 추가검토와 함께 연속형 자료의 노출제한방법에 대한 연구도 지속할 예정이다. 아울러 본 연구의 결과를 바탕으로 통계청에서 생산하고 있는 각종 통계들 중 마이크로자료의 제공을 많이 요구하는 통계조사를 대상으로 다양한 노출제한방법을 적용하는 연구가 활발히 진행되길 기대해 본다.

참고문헌

박원환・황조연(2004), 「통계자료의 특성과 비밀보호방법에 관한 연구」, 통계연구결과보고서, 통계청.

정동명・김종익・강동환(2007), "인구센서스자료의 비밀보호방법", 「통계연구」, 제12권 제1호.

통계청(2005), 「2005 인구주택총조사 조사지침서」.

Bethelehem, J.G., W.J. Keller, and J. Pannekoek(1990), "Disclosure Control of Microdata", *Journal of the American Statistical Association*, 85, pp.38-45.

Cox, L. H., S. McDonald, and D. Nelson(1986), "Confidentiality Issues at the United States Bureau of the Census", *Journal of Official Statistics*, 2, pp.135-160.

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

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

Kim, J.(1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp.370-374.

Kim, J.(1987), "A Further Development of the Randomized Response Technique for Masking Dichotomous Variables", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp.239-244.

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

Kim, J. and W.E. Winkler(2001), "Multiplicative Noise for Masking Continuous Data", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, CD-ROM.

Kim, J., M. Katzoff, J. Gonzalez Jr., and P. Williams(2003), "Techniques for Masking Microdata", National Center for Health Statistics internal memorandum.

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

Skinner, C., C. Marsh, S. Openshaw, and C. Wymer(1994), "Disclosure Control for Census Microdata", *Journal of Official Statistics*, 10, pp.31-51.

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

< 부 록 >

1. 2005 인구주택총조사 결과

〈부표 5-1〉 전수조사 결과

(단위: 개, 명, 가구, 호)

지역	조사구	인구	가구	주택
전국	265,298	45,772,054	15,895,481	12,693,578
서울	54,541	9,546,763	3,311,180	2,307,192
부산	20,060	3,428,102	1,186,651	927,413
대구	13,382	2,401,502	814,776	592,049
인천	14,185	2,470,798	823,197	700,022
광주	7,509	1,371,434	460,215	368,974
대전	8,017	1,398,799	479,098	366,547
울산	5,560	1,017,536	339,229	264,132
경기	54,950	10,100,007	3,331,307	2,676,718
강원	8,872	1,405,820	520,963	447,965
충북	8,429	1,402,094	505,718	425,564
충남	11,106	1,796,531	660,526	586,757
전북	10,599	1,718,860	620,269	555,048
전남	11,337	1,757,613	666,686	623,154
경북	15,897	2,483,816	939,591	819,045
경남	17,852	2,954,257	1,056,782	887,155
제주	3,002	518,122	179,293	145,843

〈부표 5-2〉 10% 표본조사 결과

(단위: 개, 명, 가구, 호)

지역	조사구	인구	가구	주택
전국	26,713	4,455,527	1,582,681	1,295,389
서울	5,101	879,032	305,497	214,698
부산	1,881	320,174	111,307	87,971
대구	1,254	221,787	75,808	56,055
인천	1,371	235,621	79,599	67,925
광주	716	129,836	43,996	35,818
대전	749	128,849	44,369	34,314
울산	507	91,756	30,682	24,239
경기	5,146	931,851	309,186	249,966
강원	1,001	152,350	58,319	51,119
충북	930	148,015	55,374	48,561
충남	1,215	188,950	71,029	65,352
전북	1,302	196,976	74,702	68,864
전남	1,450	208,847	83,830	80,030
경북	1,872	273,566	108,935	97,954
경남	1,920	297,986	112,584	98,228
제주	298	49,931	17,464	14,295

2. 변수별 노출제한방법

〈부표 5-3〉 인구에 관한 사항

변수	변경 전	변경 후	방법
나이	각 세	0세, 1세, 2세, … , 84세, 85세 이상	Top-Coding
가구주와의 관계	① 가구주	① 가구주	Grouping
	② 가구주의 배우자	② 가구주의 배우자	
	③ 자녀	③ 자녀	
	④ 자녀의 배우자	④ 자녀의 배우자	
	⑤ 가구주의 부모	⑤ 가구주의 부모, 배우자의 부모, 조부모	
	⑥ 배우자의 부모		
	⑨ 조부모		
	⑦ 손자녀, 그 배우자	⑥ 손자녀 및 그 배우자, 증손자녀 및 그 배우자	
	⑧ 증손자녀, 그 배우자		
	⑩ 형제자매, 그 배우자	⑦ 형제자매, 그 배우자	
	⑪ 형제자매의 자녀, 그 배우자	⑧ 기타 친인척	
	⑫ 부모의 형제자매, 그 배우자		
	⑬ 기타 친인척		
	⑭ 기타 동거인	⑨ 기타 동거인	
교육정도	① 안 받았음(미취학 포함)	① 안 받았음(미취학 포함)	Grouping
	② 초등학교	② 초등학교	
	③ 중학교	③ 중학교	
	④ 고등학교	④ 고등학교	
	⑤ 대학(4년제 미만)	⑤ 대학교	
	⑥ 대학교(4년제 이상)		
	⑦ 대학원 석사 과정	⑥ 대학원(석사과정 이상)	
	⑧ 대학원 박사 과정		
	① 졸업	① 졸업	Grouping
	② 재학	② 졸업 아님	
	③ 수료		
	④ 휴학		
	⑤ 중퇴		
종교여부	① 있다	×	Grouping
	② 없다	ⓞ 종교 없음	
종교종류	① 불교	① 불교	
	② 기독교(개신교)	② 기독교(개신교)	
	③ 기독교(천주교)	③ 기독교(천주교)	
	④ 유교	④ 유교	
	⑤ 원불교	⑤ 원불교	
	⑥ 증산교	⑥ 기타	
	⑦ 천도교		
	⑧ 대종교		
	⑨ 기타( )		

〈부표 5-3〉 인구에 관한 사항 (계속)

변수	변경 전	변경 후	방법
활동제약	① 시각·청각·언어 장애	① 시각·청각·언어 장애	Grouping
	② 치매	② 정신적 제약(치매 포함)	
	⑤ 학습의 어려움 등 정신적 제약		
	③ 중풍	③ 육체적 제약(중풍 포함)	
	④ 걷기 등 육체적 제약		
	⑥ 없음	④ 없음	
통근·통학 소요시간	□시□□분 => 분환산	~4분, 5~9분, 10~14분, …, 115~119분, 120분 이상	Grouping
경제활동 상태	① 있음	① 취업	Grouping
	② 일을 하여 왔으나, 잠시 쉬고 있음		
	③ 없음(가사, 학업 등)	② 비취업	
	① 찾아보지 않았음	×	하위 세부항목 삭제
	② 찾아보았음	×	
	① 일할 수 있었음	×	
	② 가사, 학업, 질병 등 때문에 일할 수 없었음	×	
산업	주관식	① 농업, 임업, 어업	표준산업분류의 대분류 적용 Grouping
		② 광업, 제조업, 전기,가스및수도사업	
		③ 건설업	
		④ 도매 및 소매업	
		⑤ 숙박 및 음식점업	
		⑥ 운수업	
		⑦ 통신업	
		⑧ 금융 및 보험업	
		⑨ 부동산 및 임대업	
		⑩ 사업서비스업	
		⑪ 공공행정, 국방 및 사회보장행정	
		⑫ 교육서비스업	
		⑬ 보건 및 사회복지사업	
		⑭ 오락, 문화 및 운동관련 서비스업	
		⑮ 기타 공공, 수리 및 개인 서비스업	
		⑯ 분류불능(가사서비스업, 국제및외국기관 포함)	
직업	주관식	① 의회의원, 고위임직원 및 관리자	표준직업분류의 대분류 적용 Grouping
		② 전문가	
		③ 기술공 및 준전문가	
		④ 사무 종사자	
		⑤ 서비스 종사자	
		⑥ 판매 종사자	
		⑦ 농업, 임업 및 어업 숙련 종사자	
		⑧ 기능원 및 관련 기능 종사자	
		⑨ 장치, 기계조작 및 조립 종사자	
		⑩ 단순노무 종사자	
		⑪ 분류불능(군인 포함)	

〈부표 5-3〉 인구에 관한 사항 (계속)

변수	변경 전	변경 후	방법
혼인연월	□□□□년 □□월 => 만나이 환산	18세 미만, 18세, … , 39세, 40세 이상	Bottom & Top-Coding
출생자녀 수(남)	각 명	0명, 1명, 2명, 3명, 4명, 5명 이상	Top-Coding
출생자녀 수(여)	각 명	0명, 1명, 2명, 3명, 4명, 5명 이상	Top-Coding
동거자녀 수(남)	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
동거자녀 수(여)	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
비동거 자녀 수(남)	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
비동거 자녀 수(여)	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
사망한 자녀 수(남)	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
사망한 자녀 수(여)	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
추가계획 자녀여부	① 있음	×	하위 세부항목 삭제
	② 없음	×	
추가계획 자녀 수	각 명	0명, 1명, 2명, 3명 이상	Top-Coding
고령자 생활비 원천	① 본인・배우자의 일, 직업	① 본인・배우자의 일, 직업	Grouping
	② 예금, 적금	② 예금, 적금	
	③ 국민・공무원・교직원연금	③ 국민・공무원・교직원연금	
	④ 개인연금(은행, 보험 등)	④ 개인연금(은행, 보험 등)	
	⑤ 부동산	⑤ 부동산	
	⑦ 함께 사는 자녀	⑥ 함께 사는 자녀	
	⑧ 따로 사는 자녀	⑦ 따로 사는 자녀	
	⑩ 국가・지방자치단체 보조	⑧ 국가・지방자치단체 보조	
	⑥ 주식, 채권, 증권	⑨ 기타	
	⑨ 친・인척		
	⑪ 이웃, 종교・사회단체 보조		
	⑫ 기타		

〈부표 5-4〉 가구에 관한 사항

변수	변경 전	변경 후	방법
가구구분	① 가족으로 이루어진 가구	① 가족가구	Grouping
	② 1인가구	② 1인가구	
	③ 가족과 가족 이외의 함께 사는 가구	③ 기타	
	④ 가족이 아닌 남남끼리 5인 이하의 가구		
침실 수	각 개	1개, 2개, 3개, 4개, 5개 이상	하위 세부항목 통합
침실 이외의 방 수	각 개		Top-Coding
거실(대청마루) 수	각 개	0개, 1개, 2개 이상	Top-Coding
식당 수	각 개	0개, 1개, 2개 이상	Top-Coding
거주층 수	지상 층 수	×	하위 세부항목 삭제
승용차	각 대	0대, 1대, 2대, 3대 이상	Top-Coding
승합차	각 대	0대, 1대, 2대 이상	Top-Coding
화물 및 기타자동차	각 대	0대, 1대, 2대 이상	Top-Coding
점유형태	① 주거전용	×	하위 세부항목 삭제
	② 영업겸용	×	
자기집	① 자기집	① 자가	Grouping
	② 전세(월세 없음)	② 전세	
	③ 보증금 있는 월세	③ 월세	
	④ 보증금 없는 월세		
	⑤ 사글세		
	⑥ 무상(관사,사택,친척집 등)	④ 무상	
주인가구	① 주인가구	① 주인가구	Grouping
	② 대표가구	② 주인 아닌 가구	
	③ 기타 세 들어 살고 있는 가구		

〈부표 5-5〉 주택에 관한 사항

변수	변경 전	변경 후	방법
거처종류	① 단독주택	① 단독주택	Grouping
	② 아파트	② 아파트	
	③ 연립주택	③ 연립·다세대 주택	
	④ 다세대주택		
	⑤ 비거주용 건물(상가 등)내 주택	④ 비거주용 건물(상가 등)내 주택	
	⑥ 오피스텔	⑤ 기타	
	⑦ 호텔, 여관 등 숙박업소의 객실		
	⑧ 기숙사 및 특수 사회시설		
	⑨ 판잣집, 비닐하우스, 움막		
	⑩ 기타( )		
단독주택일 경우	① 일반 단독주택	×	하위 세부항목 삭제
	② 다가구 단독주택	×	
	③ 영업겸용 단독주택	×	
연건평	각 평(㎡)	~7평, 7~9, 9~14, 14~19, 19~29, 29~39, 39~49, 49~69, 69~99, 99평 이상	Grouping
방 수	각 개	1개, 2개, … , 9개, 10개~	하위 세부항목 통합
거실 수	각 개		Top-Coding
식당 수	각 개		
부엌 편익시설 수	각 개	1개, 2개, 3개, 4개, 5개, 6개 이상	Top-Coding
화장실 편익시설 수	각 개	0개, 1개, 2개, 3개, 4개, 5개, 6개 이상	Top-Coding
출입구 편익시설 수	각 개	1개, 2개, 3개, 4개, 5개, 6개 이상	Top-Coding