응용통계연구 제 **21**권 2호, *2008*년, *pp. 313-325*

인구주택총조사 마이크로자료의 개인정보 노출제한방법

정동명[1] 정미옥[2]

요 약

통계이용자들의 마이크로자료 제공요구가 갈수록 증가하고 있으며 통계작성기관도 마이크로자료의 제공을 위해 노력을 기울이고 있는 실정이다. 그러나 마이크로자료에는 응답자의 개인정보가 많이 담겨 있으므로 자료를 그대로 제공할 경우 개인정보가 노출될 가능성이 높기 때문에 자료제공시 적절한 방법으로 노출을 제한시켜 주어야만 한다. 본 논문에서는 마이크로자료 제공시 발생하는 응답자의 정보노출에 대한 개념과 이를 제한하는 방법 등을 소개하고, 2005년에 통계청에서 실시한 인구주택총조사의 2% 마이크로자료 제공을 위해 다양한 노출제한방법을 적용하여 자료파일을 작성하는 과정을 설명하였다. 즉, 10% 표본조사결과를 모집단으로 하고 계통추출한 표본을 대상으로 외부인이 식별할 가능성이 높은 12개 항목을 key 변수로 선정한 후, 각 변수의 조합별 유일성을 파악하고 노출위험을 계산하였다. 그 결과 2% 표본을 통한 정보의 축소는 물론 그룹화, 코딩 등을 포함한 일련의 방법들을 적용함으로써 인구주택총조사 마이크로자료의 개인정보 노출을 제한하는데 상당한 효과가 있음을 알 수 있었다.

주요용어: 식별, 노출위험, 유일성, 비밀보호, 외부이용자.

1. 서론

최근 들어 통계이용자들은 다양한 경제·사회현상을 심층적으로 분석하기 위해 각종 통계조사의 마이크로자료에 대한 제공을 요구하고 있으며, 이에 대응하여 통계작성기관은 작성하고 있는 통계에 대한 마이크로자료를 이용자들에게 제공하는 것을 법적으로 제도화하고 있는 실정이다. 그러나 정보통신의 발달과 대용량 DB의 구축 등으로 인해 기업 및 여러 기관에서는 신용평가나 타깃 마케팅 등을 목적으로 개인의 여러 가지 신상정보들을 보유하고 있으며, 이러한 정보는 통계작성기관에서 제공하는 마이크로자료의 일부 항목과 유사한 부문이 상당히 많아 마이크로자료를 그대로 제공했을 시에는 개인정보의 노출 위험성이 매우 커지게 된다. 따라서 통계작성기관에서는 마이크로자료의 노출을 제한하기 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

1) (302-701) 교신저자. 대전광역시 서구 둔산동 선사로 139 정부대전청사, 통계청 통계개발원 연구기획실, 사무관.
E-mail: jedomy@nso.go.kr
2) (302-701) 대전광역시 서구 둔산동 선사로 139 정부대전청사, 통계청 통계개발원 연구기획실, 주무관.
E-mail: mojeong@nso.go.kr

기업이 마이크로자료 파일로부터 특정한 개인을 확인할 수 있는 특성 정보들을 가지고 있을 수 있기 때문이다.

통계작성기관은 통계를 작성하는 과정에서 알려진 개인이나 법인 또는 단체 등의 비밀에 속하는 사항을 통계법 제33조 (비밀의 보호)에 따라 철저히 보호해야만 하고, 특히 마이크로자료 파일을 제공할 경우 외부 이용자가 그 파일로부터 특정 개인을 연결시킬 수 없도록 조치해야만 한다. 마이크로자료에서 개인정보의 노출위험을 제거하기 위해 마이크로자료의 제공 자체를 하지 않는 것이 가장 안전하겠지만, 이는 자료수집의 근본취지에 위배되는 행위이므로 통계작성기관은 이러한 노출위험을 최소화하면서 최대한 유용한 정보가 들어 있는 자료를 제공하기 위해 노력하고 있다. 외국의 통계작성기관에서는 이미 오래 전부터 정보노출의 위험성을 깊이 인식하고 개인정보의 노출제한을 위해 다양한 방법들을 적용하여 마이크로자료를 제공하고 있다. 미국 센서스국과 호주 통계청의 경우 인구센서스 결과의 비밀보호를 위해 개인의 식별이 가능한 변수는 삭제하고 자료를 제공하고 있으며, 네덜란드와 영국 등 여러 유럽 국가들도 정보노출의 심각성과 노출제한방법에 대해 많은 연구를 수행하고 있다. 또한 여러 학자들도 노출의 개념과 노출제한방법을 연구하였는데, 특히 마이크로자료의 노출제한을 위해 Dalenius와 Reiss (1982)는 자료교환(data swapping)방법을 소개하였고, Kim (1986)은 임의잡음과 변환을 이용하는 방법을 제시하였다. Bethlehem 등 (1990)과 Fuller (1993)는 마이크로자료 노출제한의 개념정의와 진행과정을 설명하였으며, Skinner 등 (1994)은 센서스 마이크로자료의 노출제한에 대해 연구하였다. 우리나라의 경우 이에 대한 인식부족 등으로 인해 연구가 미흡하였으나 최근 들어 통계청에서 응답자의 비밀을 보호하면서 마이크로자료를 제공할 수 있는 방법을 개발코자 노력해 왔으며, 정동명 (2007) 등은 2005 인구주택총조사의 일부 지역자료 (충청남도)를 대상으로 노출제한방법의 적용사례를 연구하였다.

한편, 통계청에서는 인구주택총조사가 완료되면 집계결과를 공표하면서 아울러 통계이용자들이 활용할 수 있도록 2% 마이크로자료도 함께 제공하고 있는데, 이전까지는 시·군·구단위의 표본조사 결과를 가중값 없이 제공하였으나 2007년부터는 노출제한방법이 적용된 2% 마이크로자료도 함께 제공하고자 노력하였다. 본 논문에서는 2005 인구주택총조사의 표본조사 결과를 대상으로 응답자 개인정보의 노출이 제한된 2% 마이크로자료를 작성하는 과정을 설명하고자 한다. 2장에서는 개인정보의 노출과 노출위험, 노출제한방법 등을 간단히 소개하고 3장에서는 2005 인구주택총조사를 대상으로 외부의 식별에 영향을 미치는 key 변수의 선정과 노출위험의 계산, 노출제한방법의 적용 등을 살펴보고자 한다. 그리고 4장에서는 본 연구의 결론과 향후 고려해야할 추가사항 등을 간략히 언급하였다.

2. 노출과 노출위험의 확률모형

2.1. 노출과 유일성

통계작성기관에서 응답자로부터 수집·정리된 자료를 다양한 형태의 통계정보로 제공할 경우 이를 통해서 응답자의 특성이 파악되는 것을 노출(disclosure)이라 하는데, 만약 개인의 민감한 정보가 노출된다면 이는 개인에 대한 식별(identification)이 가능하게 된다는

것을 의미한다. 이러한 노출은 외부이용자(intruder)가 사전에 가지고 있는 유용한 정보의 성격과 양에 따라 좌우되므로, 그들이 가지고 있는 사전정보와 통계작성기관에서 제공하는 마이크로자료의 정보가 서로 일치하지 않도록 한다면 노출이 발생될 가능성은 아주 낮게 될 것이다. 노출을 제한하는 방법은 자료의 종류와 형태에 따라 다양하게 적용되는데 셀 감추기(cell suppression)와 반올림(rounding) 등은 집계표나 통계표 형태인 매크로자료에 이용되고 익명화(anonymisation)나 표본추출(sampling), 그룹화(grouping), 자료교환(data swapping) 등은 마이크로자료인 경우에 주로 사용된다. 또한 자료가 이산형인 경우에는 자료교환이나 코딩접근법(coding approach), 그룹화 등이 활용되고 연속형인 경우에는 반올림과 그룹화, 가법잡음(additive noise), 승법잡음(multiplicative noise) 등이 널리 활용되고 있다.

한편, 유일성(uniqueness)이란 전체 자료파일에서 조사단위의 특성이 유일하게 존재하는 것을 말하며 어떤 조사단위가 식별될 가능성을 나타내는 측도로 사용된다. 예를 들어, 100명으로 구성된 모집단에서 나이가 100세인 사람이 단 한명 있다면 그 사람은 모집단에서 유일하다고 하는데, 이렇게 유일한 사람은 다른 사람들에 비해 자료에서 식별될 가능성이 매우 높다. 주어진 자료에서 유일성은 하나의 변수만으로도 파악할 수 있고 여러 개의 변수들을 조합하여 파악할 수도 있으며, 고려되는 변수가 많아질수록 유일성은 점점 더 커지게 된다. 가령, 100명 중 나이가 60세인 사람이 3명 있다고 하더라도 직업이라는 변수를 포함하여 회사원이면서 나이가 60세인 사람은 단 한명일 수도 있다.

2.2. 노출위험의 확률모형

제공된 마이크로자료에서 개인정보의 노출은 다음의 3가지 조건을 모두 만족하는 경우에 발생하게 되는데, 이 조건에서 언급한 자료파일들 중 어느 하나에도 나타나지 않는다면 노출은 일어나지 않는다고 한다.

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

이러한 노출의 발생조건 하에서 노출위험(disclosure risk)은 통계적 확률모형으로 표현할 수 있다. 먼저, A를 관심의 대상인 사람, S_1을 통계작성기관의 마이크로자료로 구성된 파일 1, S_2를 외부인(intruder)에 의해 구성된 파일 2, U_P를 모집단의 유일성집단 그리고 U_S를 표본의 유일성집단이라고 각각 정의하자. 여기서 모집단 (또는 표본)의 유일성집단이란 모집단 (또는 표본)에서 유일하게 존재하는 대상을 모아놓은 집합을 의미한다. 만약 금융기관이나 이웃주민 등과 같은 외부인이 자신들이 직접 작성한 파일 (S_2)에 관심있는 어떤 사람 A가 포함되어 있다는 것을 모르고 있다면, 특정사람의 노출위험 $\mathrm{DR}(A)$은 다음과 같이 정의할 수 있다.

$$\mathrm{DR}(A) = \Pr[(A \in S_1) \cap (A \in S_2) \cap (A \in U_P)]. \tag{2.1}$$

그러나 외부인이 자신들의 파일에 관심있는 특정사람 A가 포함되어 있다는 것을 이미 알

고 있다고 한다면 $\Pr(A \in S_2) = 1$이 될 것이며, 이때 노출위험은 다음과 같이 된다.

$$\mathrm{DR}(A) = \Pr(A \in S_2)\Pr(A \in U_P), \tag{2.2}$$

이러한 노출위험의 확률모형 (2.1)과 (2.2)에 대한 자세한 내용은 Kim과 Jeong (2007)을 참고하기 바란다.

3. 인구주택총조사 마이크로자료의 노출제한 방안

3.1. 인구주택총조사

본 논문에서는 통계청이 2005년에 실시한 인구주택총조사의 표본조사 결과자료를 대상으로 노출이 제한된 마이크로자료를 작성하는 과정을 설명하고자 한다. 2005 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 (통계청, 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 취사나 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 논문에서는 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 3.1에 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.

3.2. Key 변수의 선정

노출이 제한된 마이크로자료 파일을 작성할 때, 우선적으로 고려할 사항은 외부에서 식별가능성이 높다고 판단되는 항목인 key 변수를 선정하는 것이다. key 변수란 외부이용자가 그들의 자료파일을 이용하여 데이터와 특정인 간에 일대일 대응을 통해서 특정 레코드(record)가 어떤 사람에 대한 정보인지를 식별 가능하게 하는 변수를 말한다. 잘 알려진 key 변수로는 이름이나 주소뿐만 아니라 가구구성, 연령, 인종, 성별, 거주 지역, 직업 등이

표 3.1: 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

있다. 인구주택총조사의 표본조사항목에는 개인의 특성에 관한 민감한 정보들이 많이 있기 때문에 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 위함이다. 이러한 면밀한 검토 작업을 거친 후 최종적으로 인구관련 항목 8개 (성별, 나이, 가구주와의 관계, 교육정도, 혼인상태, 활동제약, 종사산업, 직업), 가구관련 항목 3개 (가구구분, 점유형태, 주택소유여부) 그리고 주택관련 항목 1개 (거처종류) 등 총 12개 항목을 key 변수로 선정하였다.

3.3. 표본의 추출

현재 통계청에서 제공하고 있는 인구주택총조사의 마이크로자료는 전체 규모의 2%에 해당하는 표본조사 결과자료이며, 이는 표본조사 결과의 20%에 해당된다. 따라서 2% 마이크로자료의 제공을 위해서 전체 10% 표본조사구 내에 있는 모든 가구명부를 추출틀(sample frame)로 하고, 계통추출법(systematic sampling)을 적용하여 표본가구를 추출하였다. 이는 서로 이웃한 사람이 표본으로 추출될 경우 정보의 노출 가능성이 높기 때문에 이러한 가능성을 최소화하기 위해 계통추출법을 적용하였으며, 그 결과가 표 3.2에 주어져 있다.

3.4. 노출위험

2005 인구주택총조사의 노출위험 정도를 확률모형을 이용해 알아보기 위해 본 연구에서는 10% 표본조사결과를 모집단으로 하고 새로 추출한 2% 표본조사결과를 표본으로 가정하였다. 이러한 가정하에 관심의 대상인 사람을 A라 하고, 새로 추출한 2% 표본파일을 S_1, 외부인에 의해 구성된 파일을 S_2, 모집단의 유일성집단을 U_P 그리고 2% 표본의 유일성집단을 U_S라고 각각 정의하였다. 모집단과 표본의 유일성 파악을 위해 key 변수가 1개인 경우부터 점차 추가하여 12개를 모두 포함한 경우까지 고려하였다. 즉, key 변수가 1개일 때의 유일성을 계산하고, 또 다른 변수를 1개 추가하여 유일성을 다시 계산하였으며, 이런 방법으로 12개 변수가 모두 포함될 때까지 각 경우별로 유일성의 규모를 파악하였다. 그 결과 모든 변수가 포함되었을 때 유일성의 규모가 최대가 되었으며, 모집단에서 유일한 사람은 총 1,814,800명이고 표본에서는 458,160명인 것으로 각각 나타났다. 따라서 어떤 사람(A)이 모집단에서 유일하게 될 확률은 약 40.7%가 되고, 표본에서 유일하게 될 확률은

표 3.2: 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

약 51.4%가 된다. 이것은 모집단보다 표본에서 유일한 사람이 더 많이 발생한다는 것을 보여주고 있다.

$$\Pr(A \in U_P) = \frac{1,814,800}{4,455,527} = 0.40731,$$
$$\Pr(A \in U_S) = \frac{458,160}{892,024} = 0.51362.$$

이러한 유일성에 대한 결과를 이용하여 노출위험의 발생 가능성이 최대가 되는 가장 극단적인 경우의 노출위험을 계산해 보았다. 즉, 외부인이 직접 작성한 파일(S_2)에 어떤 특정인이 포함되어 있다는 것을 이미 알고 있다고 가정하면 $\Pr(A \in S_2) = 1$이 되므로 2.2절의 모형 (2.2)를 적용하면, 어떤 특정한 사람이 모집단에서 유일하면서 인구주택총조사의 2% 표본에 포함될 확률은 약 8.1%가 된다.

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

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 3.3: 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	50.6
부산	320,174	135,746	42.4	64,597	34,333	53.1
대구	221,787	95,953	43.3	44,121	23,820	54.0
인천	235,621	107,132	45.5	47,063	26,281	55.8
광주	129,836	57,801	44.5	26,253	14,401	54.9
대전	128,849	61,245	47.5	25,579	14,942	58.4
울산	91,756	41,183	44.9	18,294	10,176	55.6
경기	931,851	351,056	37.7	186,697	88,621	47.5
강원	152,350	68,609	45.0	30,522	17,415	57.1
충북	148,015	65,355	44.2	29,476	16,452	55.8
충남	188,950	77,778	41.2	37,708	19,650	52.1
전북	196,976	76,003	38.6	39,442	19,733	50.0
전남	208,847	73,428	35.2	41,684	19,530	46.9
경북	273,566	104,404	38.2	55,182	27,118	49.1
경남	297,986	116,207	39.0	59,568	29,663	49.8
제주	49,931	28,752	57.6	10,111	7,037	69.6

3.5. 노출제한방법

이 절에서는 자료의 제공범위를 제한하는 방법과 자료의 정보를 축소하는 방법을 이용하여 노출을 제한하는 방법을 살펴보기로 한다.

3.5.1. 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, 1인 가구, 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.

3.5.2. 자료의 축소

마이크로자료에 민감한 정보들이 있을 경우 제공범위를 제한하는 것보다는 정보량이 다소 줄어들더라도 제공하는 것이 이용자들의 요구에 더 부응할 수 있다. 자료의 정보량을 축소하는 방법에는 여러 가지가 있는데, 인구주택총조사의 경우 이산형 변수가 대부분이고 연속형 변수인 나이, 연건평 등도 구간으로 변환이 가능하기 때문에 이산형 변수에 효과적인 그룹화(grouping)와 코딩(top-coding, bottom-coding), 하위 세부항목의 통합 및 제거 등의 방법을 이용하였다. 각 분야별 key 변수에 대해 살펴보면, 인구에 관한 항목의 경

표 3.4: 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	24.4	892,024	341,168	38.2
서울	879,032	187,034	21.3	175,727	61,628	35.1
부산	320,174	83,122	25.9	64,597	26,426	40.9
대구	221,787	59,489	26.8	44,121	18,495	41.9
인천	235,621	69,152	29.4	47,063	20,849	44.3
광주	129,836	39,254	30.2	26,253	11,812	45.0
대전	128,849	43,088	33.4	25,579	12,648	49.4
울산	91,756	27,400	29.9	18,294	8,192	44.8
경기	931,851	179,385	19.3	186,697	59,123	31.7
강원	152,350	49,783	32.3	30,522	14,714	48.2
충북	148,015	45,151	30.5	29,476	13,410	45.5
충남	188,950	50,583	26.8	37,708	15,208	40.3
전북	196,976	50,658	25.7	39,442	15,403	39.1
전남	208,847	48,679	23.3	41,684	15,155	36.4
경북	273,566	64,405	23.5	55,182	20,159	36.5
경남	297,986	69,795	23.4	59,568	21,850	36.7
제주	49,931	22,164	44.4	10,111	6,096	60.3

우 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, 석사과정과 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 단독주택, 아파

트, 연립 및 다세대 주택, 비거주용 건물내 주택, 기타 등 5개로 그룹화하였다. 이에 대한 결과가 부록의 표 A.1에 나타나 있다.

한편, 위에서 언급한 노출제한방법을 적용한 후 모집단과 2% 표본의 유일성을 파악한 결과 표 3.4에 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 모형 (2.2)를 적용하여 노출위험을 계산하면 약 4.9%가 되어, 노출제한방법을 적용하기 전의 노출위험보다 약 40% 정도 감소한 것을 알 수 있다.

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

4. 결론

앞에서 살펴본 바와 같이, 본 논문에서는 2005 인구주택총조사의 2% 마이크로자료 제공을 위해 다양한 방법을 적용하여 자료파일을 작성하는 과정을 설명하였다. 먼저 외부인이 식별할 가능성이 높다고 판단되는 key 변수로 12개 항목을 선정하였는데 인구관련 8개 항목, 가구관련 3개 항목 그리고 주택관련 1개 항목 등 이다. 또한 10% 표본조사결과를 모집단으로 하여 계통추출법으로 표본을 추출하였으며, 12개 key 변수의 각 조합별로 유일성을 파악하고 노출위험을 계산하였다. 그 결과 2% 마이크로자료 파일을 그대로 제공할 경우 모형 (2.2)에 의해 노출위험은 약 0.081로 나타났으나, 그룹화 등의 방법을 적용하면 약 0.049로 감소하였다. 이는 그룹화 등의 노출제한방법이 인구주택총조사 자료의 노출제한에 상당히 효과가 있음을 보여주고 있다.

한편, 본 논문에서는 인구주택총조사의 질문항목이 대부분 이산형 형태인 것을 감안하여 그룹화 등과 같은 이산형 자료의 노출제한방법을 주로 적용하였지만, 연면적 등과 같은 연속형 형태의 항목은 반올림이나 승법잡음모형 등과 같은 연속형자료의 노출제한방법을 적용하는 것이 더 효과적일 수도 있다. 또한 노출위험도 10% 표본조사결과를 모집단으로 간주하고 계산하였으나, 만약 전수조사결과를 모집단으로 한다면 자료의 수가 많을수록 유일성이 감소하는 특성으로 인해 노출위험이 훨씬 더 줄어들 것이다. 따라서 향후에는 연속형 자료의 노출제한방법에 대한 연구와 함께 인구주택총조사의 전수조사결과에 대한 노출위험의 추가연구도 지속적으로 추진하고자 한다. 아울러 이번 기회를 통해 각종 마이크로자료 제공시 개인정보의 보호에 보다 많은 관심을 갖는 계기가 되길 기대해 본다.

부록

표 A.1: 변수별 노출제한방법

변수	변경 전	변경 후	방법
나이	각 세	0세, 1세, 2세,..., 84세, 85세~	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 ⑥ 배우자의 부모 ⑨ 조부모	⑤ 가구주의 부모, 배우자의 부모, 조부모	
	⑦ 손자녀, 그 배우자 ⑧ 증손자녀, 그 배우자	⑥ 손자녀 및 그 배우자, 증손자녀 및 그 배우자	
	⑩ 형제자매, 그 배우자	⑦ 형제자매, 그 배우자	
	⑪ 형제자매의 자녀, 그 배우자 ⑫ 부모의 형제자매, 그 배우자 ⑬ 기타 친인척	⑧ 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) ⑥ 대학교(4년제 이상)	⑤ 대학교	
	⑦ 대학원 석사 과정 ⑧ 대학원 박사 과정	⑥ 대학원(석사과정 이상)	
	① 졸업	① 졸업	grouping
	② 재학 ③ 수료 ④ 휴학 ⑤ 중퇴	② 졸업 아님	
활동제약	① 시각·청각·언어 장애	① 시각·청각·언어 장애	grouping
	② 치매 ⑤ 학습의 어려움 등 정신적 제약	② 정신적 제약(치매 포함)	
	③ 중풍 ④ 걷기 등 육체적 제약	③ 육체적 제약(중풍 포함)	
	⑥ 없음	④ 없음	
종교여부	① 있다	X(삭제)	grouping
	② 없다	⓪ 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 ⑦ 천도교 ⑧ 대종교 ⑨ 기타( )	⑥ 기타	
통근·통학 소요시간	□시□□분 ⇒ 분 환산	~4분, 5~9분, 10~14분,...,115~119분, 120분~	grouping
혼인연월	□□□□년 □□월 ⇒ 만나이 환산	18세 미만, 18세,...,39세, 40세~	bottom & top-coding
변수	변경 전	변경 후	방법
경제활동상태	① 있음 ② 일을 하여 왔으나, 잠시 쉬고있음	① 취업	grouping
	③ 없음(가사, 학업 등)	② 비취업	
	① 찾아보지 않았음	X(삭제)	세부항목삭제
	② 찾아보았음	X(삭제)	
	① 일할 수 있었음	X(삭제)	
	② 가사, 학업, 질병 등 때문에 일할수 없었음	X(삭제)	

종사산업	주관식	① 농업, 임업, 어업	표준산업분류 대분류적용 및 grouping
		② 광업, 제조업, 전기, 가스 및 수도사업	
		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, 국제 및 외국기관 포함)	
직업	주관식	① 의회의원, 고위임직원 및 관리자	표준직업분류 대분류적용 및 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)	
출생자녀 수	각 명	0명, 1명, 2명, 3명, 4명, 5명~	top-coding
동거자녀 수	각 명	0명, 1명, 2명, 3명~	top-coding
비동거자녀 수	각 명	0명, 1명, 2명, 3명~	top-coding
사망한자녀 수	각 명	0명, 1명, 2명, 3명~	top-coding
추가계획자녀여부	① 있음 ② 없음	X(삭제)	세부항목삭제
추가계획자녀 수	각 명	0명, 1명, 2명, 3명~	top-coding
고령자생활비 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 친·인척 ⑪ 이웃, 종교·사회단체 보조 ⑫ 기타	⑨ 기타	
변수	변경 전	변경 후	방법
가구구분	① 가족으로 이루어진 가구	① 가족가구	grouping
	② 1인 가구	② 1인 가구	
	③ 가족과 가족 이외의 함께 사는 가구 ④ 가족이 아닌 남남끼리 5인 이하의 가구	③ 기타	
침실 수	각 개	1개,2개,3개,4개,5개~	세부항목통합
침실이외의 방 수	각 개		top-coding
거실 수	각 개	0개, 1개, 2개~	top-coding
식당 수	각 개	0개, 1개, 2개~	top-coding
거주층 수	지상 층 수	X(삭제)	세부항목삭제

승용차	각 대	0대, 1대, 2대, 3대~	top-coding
승합차, 화물차 및 기타자동차	각 대	0대, 1대, 2대~	top-coding
점유형태	① 주거전용 ② 영업겸용	X(삭제)	세부항목삭제
자기집	① 자기집	① 자가	grouping
	② 전세(월세 없음)	② 전세	
	③ 보증금 있는 월세 ④ 보증금 없는 월세 ⑤ 사글세	③ 월세	
	⑥ 무상(관사, 사택, 친척집 등)	④ 무상	
주인가구	① 주인가구	① 주인가구	grouping
	② 대표가구 ③ 세 들어 살고있는 가구	② 주인 아닌 가구	
거처종류	① 단독주택	① 단독주택	grouping
	② 아파트	② 아파트	
	③ 연립주택 ④ 다세대주택	③ 연립·다세대 주택	
	⑤ 비거주용 건물(상가 등)내 주택	④ 비거주용 건물(상가 등)내 주택	
	⑥ 오피스텔 ⑦ 숙박업소의 객실 ⑧ 기숙사 및 특수 사회시설 ⑨ 판잣집, 비닐하우스, 움막 ⑩ 기타( )	⑤ 기타	
단독주택일 경우	① 일반 ② 다가구 ③ 영업겸용	X(삭제)	세부항목삭제
연건평	각 평(m^2)	~7평, 7-9, 9-14, 14-19, 19-29, 29-39, 39-49, 49-69, 69-99, 99평~	grouping
방 수	각 개	1개,2개,...,9개,10개~	세부항목통합 top-coding
거실 수	각 개		
식당 수	각 개		
부엌편익시설수	각 개	1개, 2개, 3개, 4개, 5개, 6개~	top-coding
화장실편익시설수	각 개	0개, 1개, 2개, 3개, 4개, 5개, 6개~	top-coding
출입구편익시설수	각 개	1개, 2개, 3개, 4개, 5개, 6개~	top-coding

참고문헌

정동명, 김종익, 강동환 (2007). 인구센서스자료의 비밀보호방법, <통계연구>, **12**, 95–121.

통계청 (2005). 2005 인구주택총조사 조사지침서, 통계청.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association*, **85**, 38–45.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control, *Journal of Statistical Planning and Inference*, **6**, 73–85.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation, *Journal of Official Statistics*, **9**, 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*, 370–374.

Kim, J. and Jeong, D. M. (2007). The Application of the Concept of Uniqueness for Creating Public Use Microdata Files, In *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, UNECE*, To appear.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata, *Journal of Official Statistics*, **10**, 31–51.

[2008년 1월 접수, 2008년 2월 채택]

A Method of Masking for 2005 Korean Census Microdata

Dong Myeong Jeong[1] Miock Jeong[2]

ABSTRACT

Large amounts of information on individuals is available to many organizations and data users and government agencies release microdata files from their survey data or administrative records data. However, if a microdata file is released without any limitation, an invasion of privacy is likely to occur. Therefore, in creating a microdata file, agencies attempt to eliminate disclosure risk of the file while maintaining maximum utility of the data. In this paper, we introduce the concept of disclosure risk, identification and uniqueness. Also, we show the method for creating a 2% microdata file using the 2005 Korean census microdata.

Keywords: Identification, disclosure risk, uniqueness, confidentiality, intruder.

1) Corresponding author. Deputy Director, Statistics Research Institute, KNSO, Government Complex Daejeon, 139 Seonsaro Seo-gu, Daejeon 302-701, Korea.
E-mail: jedomy@nso.go.kr

2) Researcher, Statistics Research Institute, KNSO, Government Complex Daejeon, 139 Seonsaro, Seo-gu, Daejeon 302-701, Korea.
E-mail: mojeong@nso.go.kr