제2장

가계조사 마이크로데이터의 비밀보호

정동명 · 정남수 · 한승훈

제1절 서 론

통계작성기관에서 수집하여 작성하는 통계자료는 크게 원시데이터(raw data)와 마이크로데이터(microdata), 그리고 매크로데이터(macrodata)로 구분할 수 있다. 원시데이터란 통계조사 자료에서 최초 입력한 전산파일 자료로서 입력오류, 조사오류 등이 걸러지기 이전 단계의 자료를 말한다. 마이크로데이터란 원시데이터에서 입력오류 등을 제거하여 공표 통계표 작성 등 자료가공의 기초자료로 사용되는 자료를 말하며 통계원시자료라고도 한다. 그리고 매크로데이터란 마이크로데이터를 임의의 기준에 따라 집계한 자료를 말하는데, 집계의 정도에 따라 세분화된 자료에서 통합된 자료까지 다양하게 제공할 수 있기 때문에 통계작성기관에서 자료를 제공할 경우 주로 매크로자료의 형태로 제공하고 있으며, 일부 제한적인 경우에만 마이크로데이터를 제공하고 있다.

그러나 갈수록 급변하는 경제·사회현상에 대한 심층적 자료분석에는 매크로데이터의 한계가 있기 때문에 여러 분야의 통계이용자들은 마이크로데이터의 제공에 대한 요구가 커지고 있다. 따라서 최근 들어 통계청을 비롯한 여러 통계작성기관들은 그들이 작성하고 있는 통계에 대한 마이크로데이터를 이용자들에게 제공하는 것을 법적으로 제도화하고 있는 실정이다. 하지만, 마이크로데이터를 그대로 제공할 경우 응답

한 개인의 특성을 나타내는 정보들이 그대로 노출되어 심각한 개인 사생활의 침해가 발생할 가능성이 높다. 그러므로 마이크로데이터를 제공할 경우에는 사전에 미리 응답자의 개인정보가 노출되지 않도록 통계적으로 보호방법을 적용하는 등 각별한 노력을 기울여야 한다.

외국의 통계작성기관에서는 1970년대부터 마이크로데이터의 제공에 대한 요구와 개인정보의 비밀보호라는 서로 상반된 요인을 동시에 고려하여 적정한 수준의 자료를 제공할 수 있는 방법을 찾기 위해 노력해 왔으며, 현재까지도 이에 대한 다양한 연구논문들이 꾸준히 발표되고 있다. 가령, 미국 상무부 센서서국에서는 마이크로데이터 제공시 통계적 비밀보호방법을 적용하고 있으며, 네덜란드나 스웨덴 등 유럽의 여러 나라에서도 개인사생활의 개방과 보호에 대해 각별한 노력을 기울이고 있다. 대학교나 연구기관 등에서도 비밀보호방법에 대한 연구가 활발하게 진행되고 있는데, Bethelehem et al.(1990)은 자료파일에서 개인의 식별(identification)과 노출(disclosure)의 문제에 대해 언급하고 이를 해결하기 위한 방법으로 모집단 유일성(uniqueness)의 추정을 위한 통계적 모형을 설정하고 예제를 통해 제시하였다. 그리고 Marsh et al.(1991)은 영국의 센서스에서 개인 비밀보호를 위해 익명화된 마이크로데이터 파일작성의 필요성과 효과 등을 설명하였다. 이외에도 Dalenius(1977)나 Kim (1986), Fuller(1993) 등 여러 학자들이 노출의 위험성과 비밀보호에 대한 다양한 방법들을 연구하였다.

한편, 우리나라의 경우 개인정보 노출의 위험성에 대한 인식부족 등으로 인해 비밀보호방법에 대한 연구가 상당히 미흡하였으나, 최근 들어 통계청을 중심으로 이에 대한 연구가 활발히 진행되고 있다. 특히 통계청에서는 개인의 비밀을 보호하면서 마이크로데이터를 제공할 수 있는 방법을 개발코자 수년 전부터 통계적 비밀보호기법에 대한 연구에 노력하였으며, 정동명 등(2007)이 통계자료의 비밀보호에 대한 개념과 실제 자료에 적용한 사례를 지속적으로 연구하였다.

본 연구에서는 개인정보의 노출위험성과 비밀보호방법들을 살펴보고, 몇 가지 방법을 실제 자료에 적용하여 비밀이 보호된 마이크로데이터를 작성하고자 한다. 제2절에서는 개인정보의 노출과 노출의 위험성, 그리고 비밀을 보호하는 다양한 방법을 소개한다. 제3절에서는 다양한

비밀보호방법들 중 연속형자료의 비밀보호방법을 가계조사 결과에 적용하여 비밀보호된 마이크로데이터를 작성하는 과정을 설명하고, 이에 대한 결과를 분석한다. 마지막으로 제4절에서는 본 연구의 결론과 향후 고려할 사항들을 간략히 언급하고자 한다.

제2절 노출과 유일성

1. 노출과 노출위험

가. 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 마이크로데이터의 정보가 서로 일치하지 않도록 한다면 노출이 발생될 가능성은 아주 낮게 될 것이다.

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 1명 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 1명일 수도 있다.

나. 노출위험의 확률모형

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

< 노출의 발생조건 >

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

노출의 발생조건하에서 노출위험(disclosure risk)은 통계적 확률모형으로 표현할 수 있다. 먼저, A를 관심의 대상인 사람, S_1을 통계작성기관의 마이크로데이터로 구성된 파일 1, S_2를 외부인(intruder)에 의해 구성된 파일 2, U_P를 모집단의 유일성집단, 그리고 U_S를 표본의 유일성집단이라고 각각 정의하자. 만약 금융기관이나 이웃주민 등과 같은 외부인이 자신들이 직접 작성한 파일(S_2)에 관심있는 어떤 사람 A 가 포함되어 있다는 것을 모르고 있다면, 특정사람의 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 관심있는 특정사람 A 가 포함되어 있다는 것을 이미 알고 있다고 한다면 $\Pr(A \in S_2) = 1$ 이 될 것이며, 이때 노출위험은 다음과 같이 된다.

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

이러한 노출위험의 확률적 모형에 대한 자세한 내용은 Kim & Jeong (2007)을 참고하기 바란다.

2. 비밀보호방법

마이크로데이터에서 개인정보의 비밀을 보호하는 방법은 자료의 종류와 형태에 따라 다양한 방법들이 있다. 자료의 종류가 매크로자료인 경우 셀 감추기(cell suppression), 반올림(rounding), 임의변조(random perturbation) 등의 방법이 있고, 마이크로데이터인 경우에는 익명화(anonymisation), 표본추출(sampling), 그룹화(grouping), 자료교환(data swapping) 등의 방법이 널리 활용되고 있다. 또한 자료를 형태에 따라 구분하면 자료교환, 코딩접근법(coding approach), 그룹화 등의 방법이 이산형 자료에 적용되고, 자료교환, 반올림, 구간그룹화(grouping into intervals), 가법잡음(additive noise), 승법잡음(multiplicative noise) 등의 방법이 연속형 자료에 활용된다. 여기서는 널리 활용되는 방법을 몇 가지 소개하기로 한다.

가. 자료교환

자료교환은 Dalenius(1979)가 제안한 방법으로서, 마이크로데이터 내의 민감한 항목에 대한 자료노출을 방지하기 위하여 동일한 key변수의 조합을 갖는 레코드(record)간에 자료값을 상호 교환하는 방법이다. 가령, HIV 상태 등과 같은 민감한 항목에 대하여 인종이나 성별, 나이, 수입 등과 같은 항목을 key변수로 사용하여 값을 적절하게 교환하는 것이다. 이 방법은 노출제한을 시킨 후에도 각 key변수별로 민감한 항목의 빈도수는 교환하기 전의 빈도수와 동일하게 되어 자료의 분석이 용이하다.

나. 코딩접근법

이 방법은 원자료에 적절한 값(noise)을 추가하여 다른 자료값으로 변형하는 방법이다. 가령, X_i와 Y_i, 그리고 e_i를 0 또는 1을 갖는 이산형 변수라고 가정하자. 만약 X_i를 원자료, e_i를 X_i와 서로 독립이고 동일한 분포를 갖는 값, 그리고 Y_i를 변형된 자료라고 한다면 $Y_i = (X_i + e_i)_{\mathrm{mod}2}$이 되고, X_i와 e_i가 서로 같은 값이면 0, 아니면 1의 값을 각각 가진다.

다. 그룹화

자료파일에서 어떤 변수들은 특성상 노출되기가 쉬운 범주로 구성된 경우가 있다. 가령, 나이 변수에서 100세인 사람이 단 1명뿐이라면 누구인지 쉽게 식별할 수가 있을 것이다. 이러한 경우에 인근의 다른 범주들과 통합함으로써 쉽게 식별이 되지 않도록 할 수 있는데, 이러한 방법을 그룹화라고 한다. 즉, 나이 변수의 값이 각 세별로 되어 있어 식별이 용이할 경우, 5세 또는 10세 간격으로 통합하여 그룹화함으로써 식별이 용이하지 않도록 할 수 있다.

라. 반올림

이 방법은 원자료를 적절한 몫과 나머지의 형태로 나눈 후, 나머지를 사사오입 등으로 반올림시켜 자료를 변환시키는 것으로, 특히 연속형 자료에 적용하기가 매우 용이하다. 가령 X를 정수, B를 base, q를 몫, 그리고 r을 나머지라고 하면, $X = qB + r$의 형태로 나타낼 수 있다. 만약, 나머지 r이 균등(uniform)분포를 따른다고 하면, 즉 $r \sim U(0, B-1)$, r의 기대값은 $E(r) = (B-1)/2$ 이고 분산은 $Var(r) = (B^2-1)/12$ $Var(r)$이 된다. 반올림방법에서 나머지 r을 처리하는 방법은 크게 4가지로 구분이 되는데, 이에 대한 자세한 내용은 Kim(2003)을 참고하기 바란다.

마. 구간그룹화

구간그룹화는 연속형 자료를 적당한 구간(interval)으로 그룹화하여 각 구간을 대표하는 계급값으로 원자료를 대체하는 방법이다. 이 방법은 구간을 넓게 할수록 자료의 노출방지에는 용이하지만 정보가 그만큼 더 제한되어 회귀모형 또는 상관계수 등의 자료분석을 할 경우 제약이 따른다. 그리고 자료의 특성에 따라 다소 차이는 있지만 일반적으로 구간의 중앙값(median)을 계급값으로 많이 사용한다.

바. 가법 및 승법잡음

가법잡음은 원자료에 적절한 잡음(noise)을 더해 주어 다른 자료값으

로 변형하는 방법으로서 X_i를 원자료, e_i를 잡음, 그리고 Y_i를 변형된 자료라 하면 Y_i는 다음과 같이 나타낼 수 있다.

$$Y_i = X_i + e_i$$

여기서 잡음 e_i의 분산과 공분산은 원자료 X_i에 의존된다. 한편, 승법잡음은 가법잡음과 달리 원자료 X_i에 잡음 e_i를 곱해 주어 자료를 변형시키는 방법으로서 변형된 Y_i는 다음과 같다.

$$Y_i = X_i \, e_i$$

승법잡음 모형에서 잡음 e_i는 양수값으로 1을 중심으로 값을 취하도록 한다. 가법잡음과 승법잡음은 모두 자료가 연속형인 경우 용이하게 적용된다. 만약 비밀보호된 자료값이 항상 0 또는 양의 값이 되어야 하고, 더 많은 변화를 주어야 할 경우에는 가법잡음모형보다 승법잡음모형이 더 유용하다. 이에 대한 자세한 내용은 Kim(2001)을 참고하기 바란다.

제3절 비밀보호된 마이크로자료 작성

마이크로데이터를 비밀보호된 자료로 변환하기 위해서는 개인별 통계정보에 초점을 맞추는데, 이 때 통계정보는 자료의 레코드를 서로 중복되지 않은 '식별정보(identifying information)'와 '민감정보(sensitive information)'로 구분한다. 식별정보란 데이터와 특정인 간에 일대일 대응을 통해서 특정 레코드(record)가 어떤 사람에 대한 정보인지를 식별 가능하게 하는 것으로 '식별변수' 또는 'key변수'라고 한다. 잘 알려진 key변수로는 이름이나 주소, 가구구성, 연령, 인종, 성별, 거주지역, 직업 등이 있다. 이러한 key변수들은 외부인(intruder)이 이미 알고 있을 가능성이 높거나 외부기관의 데이터에 이미 포함되어 있을 가능성이 있다고 판단되는 것들이다. 민감정보란 응답자들의 개인 범주에 속하며 그들이 노출되기를 꺼려 하는 것으로 '민감변수'라고 하며, 범죄기록이나 개인 소득금액 등이 이에 해당된다.

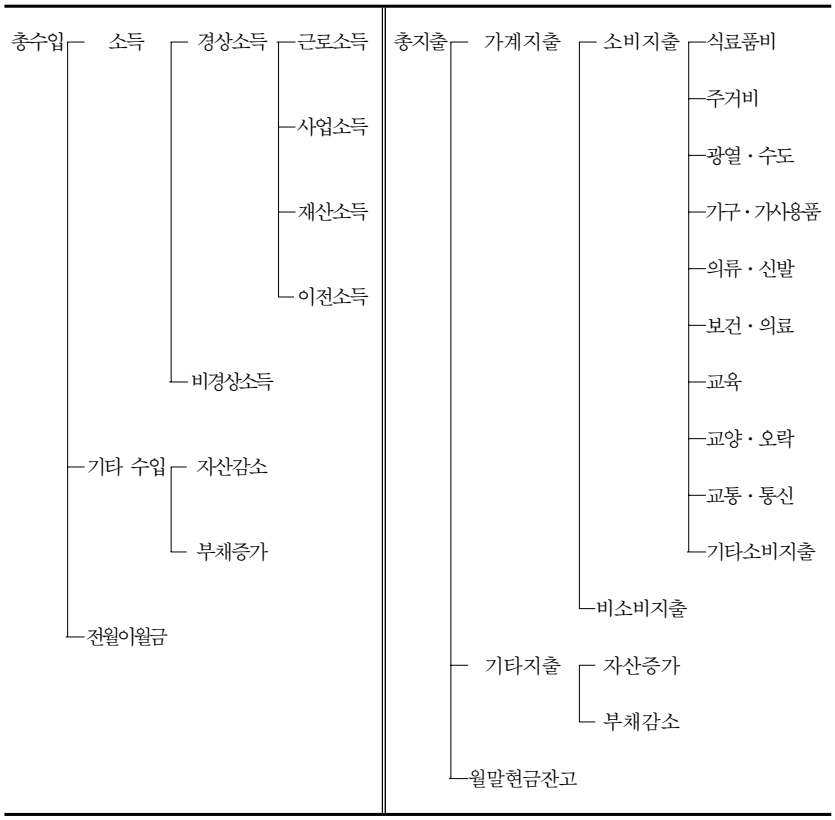
1. 분석대상

본 연구에서는 최근에 작성된 2006년 가계조사결과를 대상으로 비밀보호된 마이크로데이터를 작성코자 한다. 통계청에서 실시하는 가계조사는 우리나라 가구에 대한 가계수지의 실태를 파악하여 국민의 소득과 소비수준 변화의 측정 및 분석 등에 필요한 자료를 제공하기 위해 매월 전국의 표본가구를 대상으로 한다. 2005 인구주택총조사의 10% 표본조사구를 모집단으로 하여 전국을 25개로 층화한 후, 확률비례계통추출법으로 표본조사구를 우선 추출하고, 조사구당 10가구 추출을 기준으로 부적격 가구를 제외한 약 8,000가구를 최종 표본가구로 선정하였다. 이렇게 선정된 전국의 표본가구를 대상으로 가구의 수입과 지출관련 항목들을 응답가구에서 직접 가계부를 매일 작성하도록 하고 있으며, 가계부의 결과는 매월단위로 집계·정리한 후 분기별 및 연도별로 공표하고 있다.

가계부를 통해 조사되는 항목은 <표 2-1>에 주어져 있는 바와 같이 가구사항을 제외하면 총수입 항목과 총지출 항목으로 구분되는데, 총수입은 다시 소득과 기타수입, 그리고 전월이월금으로 나누어지고, 총지출은 가계지출과 기타지출, 그리고 월말 현금잔고로 구분된다. 조사항목들을 세분화하여 분류한 후 이를 다시 품목별로 정리하면 수입관련 품목이 47개이고 지출관련 품목이 510개로서 가계조사에서 조사되는 품목은 모두 557개가 된다. 이에 대한 자세한 내용은 가계조사 지침서(2006)를 참고하기 바란다.

한편, 2006년 가계조사의 조사대상은 <표 2-2>에 나타난 바와 같이 동일한 가구의 중복을 제외하면 총 12,458가구이며, 이 중 서울이 1,638가구이고 서울 이외 지역이 10,820가구이다. 이 결과는 분기별 및 연별로 조사결과가 집계되어 공표되는 가구 수와 다른데, 이는 1년간 매월 조사에 응답한 동일가구는 12가구가 아닌 1가구로 처리하여 중복성을 제외하였기 때문이다.

〈표 2-1〉 가계수지 항목분류



〈표 2-2〉 2006년도 가계조사의 표본규모

(단위: 가구)

지역	연간	1/4분기	2/4분기	3/4분기	4/4분기
전국	12,458	9,392	7,718	8,054	8,360
서울	1,638	1,259	947	965	987
기타(서울이외)	10,820	8,133	6,771	7,089	7,373

2. 식별정보의 비밀보호

가. key변수의 선정

가계조사의 조사항목에는 월소득이나 지출규모 등 개인 및 가구의 특성에 관한 민감한 정보들이 많이 있기 때문에 key변수의 선정에 신중을 기하였다. key변수의 선정을 위해 외부기관에서 보유하고 있는 자료들 중 가계조사의 조사항목과 중복되어 식별될 가능성이 높다고 판단되는 항목을 우선 선정한 후, 빈도분석(frequency analysis) 등을 실시하여 각 항목의 빈도수와 분포형태 등 보다 자세한 자료의 특성을 지역별 및 분기별로 파악하였다. 이는 지역별 및 분기별 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 가구유형, 거처구분, 가구원수, 배우자 유무, 가구주성별, 가구주연령, 가구주직업 등 모두 7개 항목을 key변수로 최종 선정하였다.

나. key변수의 노출위험

앞 절에서 언급한 노출위험의 확률모형을 적용하여 가계조사 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

2005 인구주택총조사의 10% 표본조사결과를 모집단으로 하고 2006년 가계조사의 마이크로데이터를 표본으로 한 후, 7개 key변수의 각 조합별로 빈도수가 1이 되는 가구를 지역별 및 분기별로 파악한 결과가 <표 2-3>에 주어져 있다. 결과해석의 편의를 위해 여기서는 연간자료인 경우에 대해서만 설명하기로 한다. 여러 조합들 중 7개 변수를 모두 고

려한 조합이 유일성의 발생이 최대가 되는 경우인데, 모집단에서는 총 21,728가구가 있으며 표본에서는 4,267가구가 있는 것으로 각각 나타났다. 따라서 어떤 가구(A)가 모집단에서 유일하게 될 확률은 약 1.52%가 되고, 표본에서 유일하게 될 확률은 약 34.25%가 된다.

$$\Pr(A \in U_P) = \frac{21,728}{1,427,563} = 0.0152$$

$$\Pr(A \in U_S) = \frac{4,267}{12,458} = 0.3425$$

이 결과에 따르면 표본에서의 유일성이 모집단보다 약 22.5배 더 많은데, 이는 표본에서 유일한 가구가 더 많이 발생한다는 것을 의미한다.

〈표 2-3〉 key변수의 유일성(그룹화 前)

(단위: 가구, %)

구분	시기	지역	총 가구(A)	유일성(B)	구성비(B/A)	노출위험
인구총조사 (10%표본)		전국	1,427,563	21,728	1.52	
		서울	302,852	15,520	5.12	
		기타	1,124,711	20,681	1.84	
가계조사	연간	전국	12,458	4,267	34.25	0.013
		서울	1,638	1,023	62.45	0.028
		기타	10,820	3,934	36.36	0.018
	1분기	전국	9,392	3,080	32.79	0.010
		서울	1,259	617	49.01	0.021
		기타	8,133	2,842	34.94	0.013
	2분기	전국	7,718	3,738	48.43	0.008
		서울	947	842	88.91	0.016
		기타	6,771	3,392	50.10	0.011
	3분기	전국	8,054	3,824	47.48	0.009
		서울	965	857	88.81	0.016
		기타	7,089	3,485	49.16	0.012
	4분기	전국	8,360	3,917	46.85	0.009
		서울	987	831	84.19	0.017
		기타	7,373	3,566	48.37	0.012

이러한 유일성에 대한 결과를 이용하여 노출위험의 발생 가능성이 최대가 되는 가장 극단적인 경우를 고려하여 외부인이 직접 작성한 파일(S_2)에 어떤 특정한 가구가 포함되어 있다는 것을 이미 알고 있다고 가정하고 노출위험을 계산하였다. 이 가정하에서는 $\Pr(A \in S_2) = 1$ 이 되므로 특정 가구가 모집단에서 유일하면서 가계조사에 포함될 확률은 약 0.013%가 된다.

$$DR(A) = \Pr(A \in S_1)\ \Pr(A \in U_P)$$
$$= \frac{12,458}{1,427,563} \times \frac{21,728}{1,427,563} = 0.00013$$

여기서 $\Pr(A \in S_1) = 0.00873$은 인구주택총조사 10% 표본조사결과에서 가계조사 표본가구의 추출률을 말한다. 위의 결과는 외부인이 가지고 있는 파일에 어떤 가구가 포함되어 있다는 것을 알고 있다고 가정하고 가계조사 마이크로데이터 파일을 제공할 경우 10,000가구 중 1.3가구가 노출될 가능성이 있다는 것을 의미한다.

위에서 설명한 방법으로 지역별 노출위험도를 계산하면, 서울이 0.028%로 서울 이외의 기타지역(0.018%)보다 더 높게 나타났다. 또한 분기별로 살펴보면 1/4분기는 0.01%, 2/4분기는 0.008%, 3/4분기는 0.009%, 그리고 4/4분기는 0.009%가 됨을 알 수 있다.

다. key변수의 그룹화

앞에서 언급한 바와 같이 가계조사의 경우 총 12,458가구 중 약 34.25%인 4,267가구가 유일한 것으로 나타났다. 만약 key변수에 적절한 비밀보호방법을 적용하면 유일성은 더 감소하게 되어 가구의 식별과 노출에 대한 위험이 훨씬 더 줄어들게 될 것이다.

key변수로 선정된 7개 변수의 경우 나이를 제외하고는 모두 이산형 변수이고 나이항목도 구간으로 변환이 가능하다. 따라서 key변수의 비밀보호를 위해 주로 이산형 변수에 효과적인 그룹화와 상한그룹화(top-coding) 등의 방법을 이용하였는데, 이렇게 할 경우 자료의 정보량은 다소 축소된다. 7개 key변수 중 가구유형(4개), 거처구분(4개), 배우자유무(3개), 가구주성별(2개) 등 4개 변수는 원래 범주를 축소하지 않고 그대

로 사용하기로 하였으며, 나머지 3개 변수는 범주의 특성을 고려하여 적정규모의 범주로 축소하였다. 먼저 가구원수의 경우 5인 이하는 그대로 두고 6인 이상은 상한그룹으로 하나의 범주로 하여 총 6개 범주로 구분하였다. 가구주연령 변수의 경우, 각 세별로 되어 있던 것을 15~79세까지는 5세 간격(13개)으로 그룹화하고 80세 이상은 상한그룹으로 1개 범주로 처리하여 총 14개 범주로 통합하였다. 끝으로 가구주직업 변수의 경우 직업대분류에 의한 12개 범주로 구분된 것을 그대로 유지하되 범주의 빈도수를 고려하여 직업군인은 기타에 분류하여 총 11개 범주로 구분하였다. 이렇게 선정된 key변수에 대해 그룹화한 결과를 정리하여 <표 2-4>에 나타내었다.

〈표 2-4〉 key변수의 그룹화

(단위: 개)

key 변수	그룹화 前		그룹화 後	
	개수	범주	개수	범주
가구유형	4	① 노인가구 ② 모자가구 ③ 맞벌이가구 ④ 일반가구	4	좌 동
거처구분	4	① 단독주택(다가구주택 포함) ② 아파트 ③ 연립 및 다세대(빌라, 맨션 등) ④ 기타(오피스텔, 비거주용 주택)	4	좌 동
가구원수	9	각 人별(1~9인)	6	ⓐ~ⓔ (각 人별) ⓕ 6인 이상
배우자 유무	3	① 있음(동거) ② 있음(비동거) ③ 없음	3	좌 동
가구주 성별	2	① 남자 ② 여자	2	좌 동
가구주 연령		각 세별(15세 이상)	14	ⓐ~ⓜ (15~79세, 각 5세별) ⓝ 80세 이상
가구주 직업	12	① 의회의원, 고위임직원 및 관리자 ② 전문가 ③ 기술공 및 준전문가 ④ 사무종사자 ⑤ 서비스종사자 ⑥ 판매종사자 ⑦ 농·림·어업 숙련 종사자 ⑧ 기능원 및 관련 기능종사자 ⑨ 장치, 기계조작원 및 조립종사자 ⑩ 단순노무종사자 ⑪ 직업군인 ⑫ 기타(무직 및 분류불능)	11	ⓐ~ⓙ (①~⑩) ⓚ 기타 (⑪+⑫)

한편, 그룹화가 노출제한에 어느 정도 영향을 미치는지를 파악하기 위해 key변수를 그룹화한 후에 유일성을 파악한 결과, <표 2-5>에 나타난 바와 같이 모집단의 경우 총 1,427,563가구 중 약 0.28%인 3,991가구가 유일한 것으로 나타났다. 이는 노출제한방법을 적용하기 전의 21,728가구보다 약 81.6%가 감소하여 상당히 줄어들었음을 알 수 있다. 또한 가계조사의 경우 연간자료를 전국단위로 비교한 결과 그룹화하기 전(4,267가구)에 비해 약 64.7%가 줄어든 1,510가구가 유일한 것으로 나타났다.

이러한 결과를 바탕으로 외부인이 직접 작성한 파일에 어떤 특정가구(A)가 포함되어 있다는 것을 이미 알고 있다는 가정하에서 가계조사

〈표 2-5〉 key변수의 유일성(그룹화 後)

(단위: 가구, %)

구분	시기	지역	총 가구(A)	유일성(B)	구성비(B/A)	노출위험
인구총조사 (10%표본)		전국	1,427,563	3,991	0.28	
		서울	302,852	3,677	1.21	
		기타	1,124,711	4,032	0.36	
가계조사	연간	전국	12,458	1,510	12.12	0.002
		서울	1,638	606	37.00	0.007
		기타	10,820	1,455	13.45	0.003
	1분기	전국	9,392	1,259	13.41	0.002
		서울	1,259	405	32.17	0.005
		기타	8,133	1,214	14.93	0.003
	2분기	전국	7,718	1,578	20.45	0.002
		서울	947	558	58.92	0.004
		기타	6,771	1,491	22.02	0.002
	3분기	전국	8,054	1,555	19.31	0.002
		서울	965	581	60.21	0.004
		기타	7,089	1,456	20.54	0.002
	4분기	전국	8,360	1,646	19.69	0.002
		서울	987	600	60.79	0.004
		기타	7,373	1,533	20.79	0.002

의 노출위험을 계산하면 약 0.002%가 된다. 이는 노출제한방법을 적용하기 전의 노출위험(0.013%)보다 약 84.6% 정도 감소하였음을 보여주고 있다.

$$DR(A) = \Pr(A \in S_1)\ \Pr(A \in U_P)$$
$$= \frac{12,458}{1,427,563} \times \frac{3,991}{1,427,563} = 0.00002$$

3. 민감정보의 비밀보호

가. 민감변수의 선정

앞에서 언급한 바와 같이 민감변수란 주어진 자료에서 개인정보의 노출가능성이 높거나 노출을 제한해야 할 필요성이 있는 변수를 말하는데, 자료제공시 이러한 민감변수의 자료값을 적절한 비밀보호방법으로 변환시켜 줌으로써 정보의 노출을 제한시킬 수 있다. 가계조사의 경우 앞의 <표 2-1>에 나타난 바와 같이 조사항목은 크게 가구의 수입과 지출로 구분되는데, 다시 세분화하면 47개의 수입관련 항목과 510개의 지출관련 항목으로 나눌 수 있다. 그러나 이 항목들은 가계수지의 현황을 파악하기에는 매우 유용하지만 어떤 특정가구의 식별을 위한 특성을 나타내는 데 모든 항목들이 사용되지는 않는다. 가령, 수입부문의 가구주 급여소득 등과 같은 항목은 어떤 특정가구만이 해당될 수 있어 그 가구의 식별이 가능하지만, 지출부문의 식료품비를 나타내는 일부 항목들은 대부분의 가구에서 쉽게 파악될 수 있어 이 항목만으로는 가구의 식별이 용이하지 않다.

따라서 본 연구에서는 가계수지 항목별로 그 특성을 파악하여 이들 중 응답가구의 식별에 영향을 줄 것으로 판단되는 82개 항목들을 민감변수로 선정하였는데 수입관련 항목이 31개, 지출관련 항목이 51개이다. 이에 대한 결과는 <표 2-6>에 주어져 있다.

〈표 2-6〉 민감변수

구분	항 목			
총수입 (31개)	가구주급여소득	가구주상여금	배우자급여소득	배우자상여금
	기타가구원급여소득	기타가구원상여금	가구주사업소득	배우자사업소득
	기타가구원사업소득	이자소득	배당소득	부동산임대소득
	기타재산소득	공적연금	기타사회보장수혜	사적이전
	경조소득	기타비경상소득	퇴직금및연금일시금	저축찾은금액
	보험탄금액	계탄금액	유가증권매각	부동산매각
	기타재산매각	기타자산감소	부동산관계빌린돈	기타빌린돈
	월부및외상	기타부채증가	전월이월금	
총지출 (51개)	식사대	월세	바닥재	기타수선재료
	설비용품및기구	설비수리서비스	가사사용인급료	조제약
	보건의료소모품	기타보건의료기구	병원입원치료비	산후조리원
	해외연수비	경승용차	소형승용차	중형승용차
	대형승용차	다목적승용차	기타차량	보험료
	일반전화요금	이동전화요금	기타통신	손해보험료
	관혼상제비	소득세	재산세	자동차세
	기타세금	일반기여금	국민연금	건강보험료
	기타사회보험료	지급이자	각종부담금및기타	교육비송금
	기타송금보조	저금	저축성보험료	계부은금액
	유가증권구입	부동산구입	기타재산구입	빌려준돈
	기타자산증가	주택부금상환	빌린돈갚은금액	월부및외상갚은금액
	기타부채감소	월말현금잔고	현물총액	

나. 민감변수의 비밀보호

위에서 선정된 민감변수의 관측값은 대부분 연속형 자료이므로 이들의 노출제한을 위해서는 연속형 자료의 비밀보호방법이 효과적이다. 이에 대해 여러 방법들이 연구되었지만 본 연구에서는 반올림과 구간그룹화, 승법잡음 등의 방법을 적용하여 분석하였다.

1) 반올림

반올림이란 주어진 자료를 적절한 몫과 나머지의 형태로 나눈 후, 나머지를 사사오입 등으로 반올림시켜 자료를 변환시키는 방법을 말한다.

<표 2-6>에 주어진 82개 민감변수는 자료특성에 따라 조사된 자료값의 단위가 매우 다양하다. 가령, 저금항목의 경우 연중 최소값은 백단위지만 최대값은 억단위가 된다. 따라서 본 연구에서는 이러한 자료의 특성을 고려하여 원자료를 백단위와 천단위, 그리고 만단위에서 반올림하는 3가지 방법을 고려하였다. 그리고 응답가구에서 가계부를 작성할 때 정확한 값이 아닌 일정단위 이상으로 반올림하여 응답한(즉, 반올림된 응답자료) 경우가 있는데, 이런 경우에는 반올림하고자 하는 일정단위 이하에서 반올림된 응답자료는 일정단위로 반올림하고, 일정단위 이상에서 반올림된 응답자료는 별도의 반올림 없이 그 값을 그대로 사용하기로 하였다. 가령, 어떤 변수의 자료값을 천단위에서 반올림한다고 할 때 응답자료가 13,000원이면 반올림하여 10,000원으로 변환하고, 응답자료가 130,000원이면 별도의 반올림 없이 130,000원을 그대로 사용하는 것이다.

가계조사의 민감변수에 대해 백단위와 천단위, 그리고 만단위로 각각 반올림하여 자료를 변환해서 시기별 및 지역별로 평균과 표준편차를 구하였으며, 이 중 수입관련 항목과 지출관련 항목 1개에 대한 결과를 <표 2-7>에 나타내었다. 여기서 원자료를 백단위에서 반올림한 것을 방법 1, 천단위로 반올림한 것을 방법 2, 그리고 만단위로 반올림한 것을 방법 3으로 나타내었다. 이 표의 결과에 의하면, 가구주급여소득 항목이나 식사대 항목에서 시기별 및 지역별로 방법 1이 원자료와 가장 유사한 결과를 보여주고 있는데, 이는 낮은 단위에서의 반올림이 높은 단위에서의 반올림보다 원자료와 더 유사하기 때문이다. 일반적으로 자료분석 측면에서는 높은 단위보다 낮은 단위에서 반올림하는 것이 더 효과적일 수 있지만, 비밀보호의 측면에서는 반올림의 단위를 낮게 할수록 정보노출의 가능성이 더 높게 된다. 따라서 반올림을 이용한 자료변환의 경우, 무조건 낮은 단위에서 반올림하는 것보다는 가급적 원자료의 특성을 고려하여 적정 수준의 단위에서 반올림하도록 하는 것이 더 바람직할 것이다.

2) 구간그룹화

구간그룹화란 연속형 자료를 적당한 구간으로 그룹화하여 각 구간을 대표하는 계급값으로 자료를 변환하는 방법을 말한다. 자료를 구간으로

〈표 2-7〉 반올림 적용결과

(단위: 가구, %)

변수	시기	지역	평 균			
			원자료	방법1	방법2	방법3
가구주 급여소득	연간	전국	1,967,254	1,967,268	1,967,467	1,970,537
		서울	2,133,225	2,133,231	2,133,380	2,137,963
		기타	1,941,942	1,941,958	1,942,164	1,945,004
	1분기	전국	1,956,740	1,956,757	1,956,953	1,960,227
		서울	2,134,227	2,134,234	2,134,355	2,138,873
		기타	1,926,376	1,926,394	1,926,603	1,929,665
	2분기	전국	1,949,293	1,949,307	1,949,481	1,952,440
		서울	2,121,116	2,121,119	2,121,261	2,125,093
		기타	1,923,359	1,923,375	1,923,554	1,926,381
	3분기	전국	1,974,921	1,974,932	1,975,128	1,978,347
		서울	2,128,457	2,128,463	2,128,663	2,133,483
		기타	1,952,518	1,952,531	1,952,726	1,955,711
	4분기	전국	1,989,891	1,989,906	1,990,133	1,992,931
		서울	2,149,759	2,149,764	2,149,907	2,155,143
		기타	1,967,586	1,967,602	1,967,841	1,970,299
식사대	연간	전국	175,213	175,298	175,551	174,942
		서울	194,296	194,373	194,595	195,359
		기타	172,460	172,543	172,801	171,993
	1분기	전국	171,462	171,545	171,758	171,321
		서울	189,925	190,007	190,225	191,209
		기타	168,529	168,612	168,824	168,161
	2분기	전국	174,400	174,483	174,758	174,038
		서울	192,698	192,776	193,059	193,261
		기타	171,813	171,896	172,170	171,320
	3분기	전국	178,967	179,050	179,286	178,903
		서울	200,560	200,637	200,807	201,111
		기타	175,977	176,061	176,306	175,828
	4분기	전국	176,723	176,803	177,100	176,172
		서울	195,503	195,574	195,791	197,300
		기타	174,154	174,235	174,543	173,283

〈표 2-7〉의 계속

(단위: 가구, %)

변수	시기	지역	표준편차			
			원자료	방법1	방법2	방법3
가구주 급여소득	연간	전국	1,238,689	1,238,689	1,238,633	1,237,661
		서울	1,371,615	1,371,613	1,371,579	1,370,280
		기타	1,215,169	1,215,169	1,215,111	1,214,155
	1분기	전국	1,261,683	1,261,686	1,261,666	1,260,466
		서울	1,377,804	1,377,801	1,377,787	1,377,020
		기타	1,238,243	1,238,248	1,238,230	1,236,908
	2분기	전국	1,203,192	1,203,192	1,203,097	1,202,252
		서울	1,334,241	1,334,238	1,334,232	1,332,743
		기타	1,180,049	1,180,050	1,179,940	1,179,179
	3분기	전국	1,228,238	1,228,238	1,228,155	1,227,213
		서울	1,366,841	1,366,839	1,366,816	1,365,610
		기타	1,205,120	1,205,120	1,205,028	1,204,093
	4분기	전국	1,256,676	1,256,673	1,256,637	1,255,782
		서울	1,408,206	1,408,209	1,408,108	1,406,168
		기타	1,232,480	1,232,477	1,232,452	1,231,717
식사대	연간	전국	149,618	149,623	149,602	153,621
		서울	169,632	169,635	169,609	172,644
		기타	146,297	146,303	146,283	150,447
	1분기	전국	143,972	143,977	143,949	148,115
		서울	141,027	141,022	141,081	144,144
		기타	144,221	144,227	144,185	148,495
	2분기	전국	153,213	153,218	153,199	156,890
		서울	134,458	134,463	134,417	137,451
		기타	155,512	155,518	155,501	159,264
	3분기	전국	149,177	149,183	149,181	153,401
		서울	205,578	205,578	205,567	208,789
		기타	139,327	139,334	139,336	143,800
	4분기	전국	152,831	152,838	152,800	156,797
		서울	195,212	195,223	195,113	198,134
		기타	145,901	145,907	145,885	150,038

그룹화하기 위해서는 계급의 수와 계급간격, 그리고 계급의 한계를 정해 주어야 하는데 이들을 정하는 과정은 다음과 같다. 먼저 주어진 자료의 수를 n, 계급의 수를 k라 하면 적정규모의 계급 수는 $2^k \geq n$을 만족시키는 최소의 정수값이 된다. 가령, $n=50$ 이면, k는 6이 되어 적정규모의 계급 수는 6개가 된다. 다음으로 계급 수가 결정되면 각 계급별 간격을 결정해야 하는데, 간격을 동일하게 할 경우 계급간격은 자료의 범위를 계급 수로 나누어 주면 된다. 즉,

$$\text{계급간격} = \frac{(\text{최대값} - \text{최소값})}{\text{계급 수}}$$

마지막으로 계급의 시작점과 끝점을 계급의 한계라고 하며 시작점을 하한계(lower limit), 끝점을 상한계(upper limit)라고 한다. 주어진 자료값이 계급의 한계에 놓이지 않게 하기 위해서는 자료의 최소값보다 조금 작은 값을 선택하게 되는데, 자료의 최소단위의 반을 자료의 최소값에서 뺀 값을 하한계로 사용한다. 가령, 자료가 10단위이고 최소값이 100이면 하한계는 95(=100-5)가 된다.

본 연구에서는 이러한 구간그룹화 과정에 따라 앞 절에서 선정된 민감변수에 대해 그룹화하였으며, 각 구간별 계급값은 구간의 분포를 어떻게 가정하는가에 따라 6가지의 방법으로 계산하였다. 먼저 각 변수별로 일반구간은 분포를 가정하지 않거나 균등분포를 따른다고 가정하여 이들의 중앙값을 계산하였다. 또한 각 변수별 최상위구간은 균등분포와 파레토분포, 그리고 log-normal분포라고 가정하여 이들의 중앙값을 이용하였다. 이러한 분포를 가정하여 중앙값을 계급값으로 사용한 이유는 우선 자료의 가장 기본적인 분포가 균등분포이고, 파레토분포나 log-normal분포는 경제관련 자료에 널리 적용되는 분포이기 때문이다. 그리고 소득이나 지출 등과 같이 한쪽으로 치우친 특성을 나타내는 경우 자료의 대표값으로 평균보다 중앙값이 더 효과적이기 때문이다. 이렇게 가정한 3가지 분포에 대한 중앙값은 다음과 같이 구할 수 있다.

가) 균등분포인 경우

변수 X가 균등분포를 따른다고 한다면, X의 중앙값 X_m^U은 다음과 같다.

$$X \sim Uniform(a, b) \Rightarrow X_m^U = \frac{(a+b)}{2}$$

나) 파레토분포인 경우

만약 변수 X가 파레토분포를 따른다고 한다면, X의 중앙값 X_m^P은 다음과 같이 된다. 이 때 t와 k의 값은 그들의 최우추정량(MLE) $\hat{t}$와 $\hat{k}$을 각각 대입하여 구할 수 있다.

$$X \sim Pareto(t, k) \Rightarrow X_m^P = t \cdot \sqrt[k]{2}\,,$$
$$\hat{t} = \min x_i\,, \quad \hat{k} = \frac{n}{\sum_i (\ln x_i - \ln \hat{t}\,)}$$

다) log-mormal분포인 경우

변수 X가 log-mormal분포를 따른다고 한다면, X의 중앙값 X_m^L은 다음과 같이 되며, 이 때 $\ln(X)$의 평균인 $\mu(\geq 0)$의 추정량으로는 최우추정량 $\hat{\mu}$을 사용한다.

$$X \sim \log\text{-}normal(\mu, \sigma^2) \Rightarrow X_m^L = e^{\mu}\,, \quad \hat{\mu} = \frac{\sum_i \ln x_i}{n}$$

이상의 결과를 정리하여 구간별 분포와 계급값에 따른 6가지의 구간 그룹화 방법을 <표 2-8>에 나타내었다.

〈표 2-8〉 구간별 분포와 계급값

방법	일반 구간	최상위 구간
1	균등분포의 중앙값 : X_m^U	균등분포의 중앙값 : X_m^U
2	〃	파레토분포의 중앙값 : X_m^P
3	〃	log-normal 분포의 중앙값 : X_m^L
4	구간내 자료의 중앙값 : $med(x_i)$	구간내 자료의 중앙값 : $med(x_i)$
5	〃	파레토분포의 중앙값 : X_m^P
6	〃	log-normal 분포의 중앙값 : X_m^L

〈표 2-9〉 구간그룹화 적용결과

(단위: 원)

변수	시기	지역	평 균						
			원자료	방법1	방법2	방법3	방법4	방법5	방법6
가구주 급여소득	연간	전국	1,967,254	2,112,772	1,944,156	1,961,531	1,949,609	1,944,656	1,962,031
		서울	2,133,225	2,182,667	2,103,509	2,122,965	2,105,207	2,107,763	2,127,219
		기타	1,941,942	2,102,113	1,919,854	1,936,911	1,925,880	1,919,782	1,936,839
	1분기	전국	1,956,740	2,203,261	1,929,686	1,948,729	1,936,187	1,930,872	1,949,915
		서울	2,134,227	2,187,596	2,102,385	2,121,533	2,105,069	2,107,572	2,126,720
		기타	1,926,376	2,205,941	1,900,141	1,919,165	1,907,295	1,900,643	1,919,667
	2분기	전국	1,949,293	2,065,299	1,927,684	1,943,357	1,932,444	1,929,344	1,945,016
		서울	2,121,116	2,179,062	2,093,163	2,111,771	2,083,963	2,095,797	2,114,405
		기타	1,923,359	2,048,128	1,902,708	1,917,937	1,909,574	1,904,220	1,919,450
	3분기	전국	1,974,921	2,061,874	1,953,697	1,970,724	1,960,675	1,954,406	1,971,434
		서울	2,128,457	2,178,192	2,098,780	2,119,007	2,103,341	2,100,771	2,120,998
		기타	1,952,518	2,044,902	1,932,527	1,949,088	1,939,859	1,933,049	1,949,610
	4분기	전국	1,989,891	2,105,513	1,968,047	1,985,530	1,971,460	1,966,378	1,983,861
		서울	2,149,759	2,184,308	2,121,061	2,141,063	2,130,136	2,128,093	2,148,095
		기타	1,967,586	2,094,519	1,946,698	1,963,830	1,949,321	1,943,815	1,960,946
식사대	연간	전국	175,216	228,515	172,236	174,189	172,048	171,880	173,833
		서울	194,296	234,354	190,450	192,565	190,227	190,182	192,298
		기타	172,460	227,671	169,605	171,534	169,423	169,236	171,166
	1분기	전국	171,462	225,469	168,374	170,351	168,242	168,129	170,106
		서울	189,925	213,255	186,784	188,822	186,922	186,856	188,894
		기타	168,529	227,410	165,449	167,416	165,274	165,154	167,121
	2분기	전국	174,400	245,818	171,349	173,229	171,084	170,977	172,856
		서울	192,698	207,585	190,247	192,160	190,300	189,669	191,582
		기타	171,813	251,223	168,678	170,553	168,367	168,334	170,209
	3분기	전국	178,967	210,595	176,089	178,118	175,912	175,639	177,668
		서울	200,560	275,035	195,357	197,788	193,881	194,264	196,694
		기타	175,977	201,672	173,420	175,394	173,424	173,060	175,033
	4분기	전국	176,723	232,708	173,839	175,762	173,655	173,462	175,385
		서울	195,503	249,261	190,640	192,752	190,913	191,068	193,179
		기타	174,154	230,444	171,541	173,439	171,294	171,055	172,952

〈표 2-9〉의 계속

(단위: 원)

변수	시기	지역	표준편차						
			원자료	방법1	방법2	방법3	방법4	방법5	방법6
가구주 급여소득	연간	전국	1,238,689	1,729,843	1,150,619	1,194,896	1,155,391	1,144,076	1,188,589
		서울	1,371,615	1,504,904	1,270,077	1,321,745	1,251,458	1,258,945	1,310,988
		기타	1,215,169	1,761,396	1,129,335	1,172,412	1,138,176	1,123,465	1,166,759
	1분기	전국	1,261,683	2,133,913	1,147,754	1,197,199	1,153,611	1,141,093	1,190,795
		서울	1,377,804	1,533,137	1,271,995	1,324,269	1,248,495	1,255,663	1,308,514
		기타	1,238,243	2,220,508	1,122,524	1,171,591	,134,123	1,117,580	1,166,847
	2분기	전국	1,203,192	1,553,086	1,129,128	1,168,006	1,130,026	1,124,149	1,163,171
		서울	1,334,241	1,483,380	1,231,403	1,279,739	1,195,309	1,226,519	1,275,002
		기타	1,180,049	1,562,684	1,110,798	1,148,120	1,118,118	1,105,764	1,143,231
	3분기	전국	1,228,238	1,471,895	1,152,398	1,195,548	1,158,910	1,143,974	1,187,419
		서울	1,366,841	1,496,738	1,264,444	1,317,376	1,259,801	1,255,365	1,308,634
		기타	1,205,120	1,467,536	1,133,639	1,175,222	1,142,019	1,125,274	1,167,147
	4분기	전국	1,256,676	1,600,653	1,172,940	1,217,652	1,178,543	1,166,913	1,211,872
		서울	1,408,206	1,498,114	1,314,422	1,367,466	1,305,499	1,301,837	1,355,270
		기타	1,232,480	1,614,202	1,150,270	1,193,718	1,158,064	1,145,079	1,188,759
식사대	연간	전국	149,618	392,873	124,863	130,085	125,116	124,743	129,976
		서울	169,632	367,032	125,298	130,983	124,924	124,987	130,690
		기타	146,297	396,462	124,581	129,741	124,926	124,487	129,656
	1분기	전국	143,972	368,914	121,754	127,08	121,719	121,498	126,845
		서울	141,027	225,632	123,925	129,528	123,216	123,269	128,899
		기타	144,221	386,791	121,151	126,450	121,218	120,950	126,262
	2분기	전국	153,213	504,455	123,366	128,347	123,617	123,425	128,410
		서울	134,458	183,126	122,568	127,422	123,887	122,426	127,294
		기타	155,512	534,301	123,248	128,256	123,341	123,341	128,350
	3분기	전국	149,177	279,353	127,887	133,324	128,350	127,701	133,152
		서울	205,578	509,763	128,882	135,620	127,122	128,205	134,996
		기타	139,327	228,517	127,523	132,778	128,324	127,420	132,684
	4분기	전국	152,831	389,078	126,775	131,898	127,147	126,710	131,841
		서울	195,212	471,546	126,153	131,685	125,993	126,541	132,050
		기타	145,901	376,356	126,690	131,761	127,125	126,547	131,629

가계조사의 민감변수에 대해 구간그룹화의 6가지 방법에 따라 자료를 변환하고 각 구간별 계급값을 계산한 후, 시기별 및 지역별로 평균과 표준편차를 구하였으며, 이 중 수입관련 항목과 지출관련 항목 1개를 각각 선정해 <표 2-9>에 나타내었다. 이 표에 의하면, 가구주급여소득 항목에서 평균값을 비교해 보면 방법 3이 연간 및 4분기의 기타지역인 경우 원자료와 차이가 가장 작은 것을 제외하고는 거의 모든 경우에 방법 6이 원자료와 가장 유사한 것으로 나타났다. 이와 반대로 지출부문인 식사대 항목에서는 방법 3이 방법 6보다 시기별 및 지역별로 원자료와 차이가 가장 작은 것으로 나타났다.

같은 방법으로 모든 민감변수에 대해 연간자료에 대해 전국단위의 평균값을 분석해 보면, 총 82개 항목 중 방법 1이 더 우수하게 나타난 항목은 3개, 방법 2는 6개, 방법 3은 42개, 방법 4는 17개, 방법 5는 3개, 그리고 방법 6은 11개 항목이 더 우수한 것으로 각각 나타났다. 이러한 결과를 정리해 보면, 구간그룹화 방법에서 방법 3을 적용할 때 원자료와 가장 유사한 결과를 얻는 것으로 나타났는데, 이는 구간의 분포를 log-normal로 가정한 경우이다. 또한 구간의 분포를 log-normal로 가정한 방법 3과 방법 6이 모두 53개 항목에서 우수한 것으로 나타나, 가계조사의 구간그룹화에서는 구간분포를 log-normal로 가정하는 것이 매우 효과적임을 알 수 있다.

3) 승법잡음

승법잡음이란 주어진 원자료 X에 적절한 잡음 e를 곱해 주어 자료를 변환시키는 방법을 말하는데, 변환된 자료를 Y라고 하면 $Y = X \cdot e$이 된다. 이 때 잡음 e는 일반적으로 1이 아니면서 1을 중심으로 한 양의 값을 가지는 것이 바람직하다. 이것은 e가 만약 음수가 되면 변환된 변수도 음의 값이 되고, e가 1이 되면 Y는 X와 같게 되어 변환의 의미가 없기 때문이다. 또한 e가 1에서 멀리 떨어진 값을 가지면, 원자료 X와 변환된 자료 Y가 너무 차이가 나서 정보의 손실이 크게 된다. 따라서 e는 1이 아니면서 1과 가까운 값을 가지는 것이 매우 효과적이다.

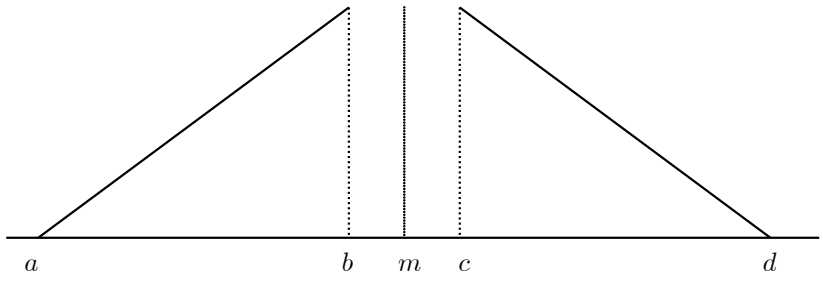
일반적으로 잡음 e의 값은 e에 대한 적절한 분포를 가정하고 이에 따라 컴퓨터로 난수를 발생시키는 방법으로 구한다. 잡음 e의 분포로

는 다양한 형태가 적용될 수 있는데, Kim & Winkler(2001)는 평균 μ를 중심으로 좌우가 절단된 정규분포(truncated normal distribution)를 적용한 결과를 소개하였다. Massell & Russell(2006)은 절단된 삼각분포(truncated triangular distribution)를 제시하였으며, Kim(2007)은 절단된 삼각분포를 변형한 다양한 형태의 분포를 제시하고 그 특성을 분석하였다. 본 연구에서는 e가 절단된 삼각분포를 따른다고 가정하고 승법잡음모형을 적용하여 선정된 민감변수의 자료를 변환하고 그 결과를 분석하였다.

가) 절단된 삼각분포

삼각분포는 자료의 최빈값을 중심으로 삼각형 모양으로 이루어진 분포를 말하고, 절단된 삼각분포는 최빈값을 중심으로 좌우로 일정부분 절단시킨 분포를 말한다. 일반적으로 삼각분포나 절단된 삼각분포는 좌우가 비대칭인 경우도 있으나, 여기서는 편의상 좌우대칭인 경우만 고려하였다. 만약 자료의 최소값을 a, 최대값을 d라 하고 최빈값 m을 중심으로 좌우대칭이면서 또한 절사점 b와 $c(>b)$도 m에서 좌우대칭이 되는 절단된 삼각분포는 [그림 2-1]에 주어진 것과 같은 형태가 된다.

[그림 2-1] 좌우대칭의 절단된 삼각분포



만약 e가 최빈값 m을 중심으로 좌우대칭이면서 절사점 b와 $c(>b)$가 m에서 각각 좌우대칭인 절단된 삼각분포를 따른다면 e의 확률밀도함수(pdf), $f(e)$는 다음과 같이 나타낼 수 있다.

$$f(e) = \begin{cases} \dfrac{(e-a)}{(d-c)^2}, & a \le e < b \\ \dfrac{(d-e)}{(d-c)^2}, & c \le e < d \end{cases}$$

또한 e의 기대값과 분산을 계산하면 다음과 같이 된다.

$$E(e) = m$$
$$Var(e) = m^2 - \frac{16mc + 8md - 2(d^2 + 2dc + 3c^2)}{12}$$

나) 난수의 발생

잡음 e가 절단된 삼각분포를 따른다는 가정하에 e에 대한 난수는 최소값(a)과 최대값(d), 최빈값(m), 그리고 두 절사점(b와 c)을 적절히 조합하여 <표 2-10>에 주어진 바와 같이 4가지 방법으로 난수를 생성하였다. 여기서 최빈값은 모두 1로 하고 난수의 범위를 다소 좁게 한 경우(방법 1과 2)와 넓게 한 경우(방법 3과 4)로 나누어 고려하였다. 아울러 1에서 가까운 지점에서 절사된 경우(방법 1과 3)와 다소 멀리 떨어진 지점에서 절사된 경우(방법 2와 4)로 구분하였다. 그리고 난수의 크기는 가계조사의 표본규모가 최대인 시점을 고려하여 약 8,615개가 되도록 하였으며, SAS프로그램을 활용하여 난수를 생성하였다.

〈표 2-10〉 승법잡음의 유형

방법	최소값(a)	절사점(b)	최빈값(m)	절사점(c)	최대값(d)
1	0.6	0.99	1.0	1.01	1.4
2	0.6	0.90	1.0	1.10	1.4
3	0.4	0.99	1.0	1.01	1.6
4	0.4	0.90	1.0	1.10	1.6

<표 2-10>에 주어진 방법에 따라 난수를 발생시킨 결과가 실제 이론에 의해 절단된 삼각분포와 어느 정도 차이가 나는지를 검토하기 위해 기초통계량을 계산해 보았다.

〈표 2-11〉 생성난수의 결과비교

방법	구분	통계량			
		평균	표준편차	최소값	최대값
1	이론적(A)	1.0000	0.1675	0.6000	1.4000
	난수결과(B)	1.0003	0.1669	0.6041	1.3916
	차이(B-A)	0.0003	- 0.0006	0.0041	- 0.0084
2	이론적(A)	1.0000	0.2121	0.6000	1.4000
	난수결과(B)	0.9994	0.2116	0.6041	1.3964
	차이(B-A)	- 0.0006	- 0.0005	0.0041	- 0.0036
3	이론적(A)	1.0000	0.2491	0.4000	1.6000
	난수결과(B)	1.0004	0.2481	0.4062	1.5874
	차이(B-A)	0.0004	- 0.0010	0.0062	- 0.0126
4	이론적(A)	1.0000	0.2915	0.4000	1.6000
	난수결과(B)	0.9987	0.2912	0.4062	1.5947
	차이(B-A)	- 0.0013	- 0.0003	0.0062	- 0.0053

그 결과 <표 2-11>에 나타난 바와 같이 4가지 방법에 의해 난수를 생성한 결과가 이론적인 값과 거의 차이가 없는데, 이는 잡음 e의 자료로 생성된 난수를 승법잡음모형에 적용해도 가능하다는 것을 의미한다.

다) 자료의 변환

민감변수의 원자료를 X_i, 잡음을 e_i라 하면 승법잡음모형에 따라 변환된 자료 Y_i는 다음과 같다.

$$Y_i = X_i \cdot e_i, \quad i = 1, \cdots, n$$

그리고 e_i는 X_i와 독립이기 때문에 Y_i의 기대값과 분산은 다음과 같이 나타낼 수 있다.

$$E(Y_i) = E(X_i) \cdot E(e_i)$$
$$Var(Y_i) = Var(X_i)\ Var(e_i) + [E(e_i)]^2\ Var(X_i) + [E(X_i)]^2\ Var(e_i)$$

한편, 승법잡음모형으로 자료를 변환한 후 통계이용자들에게 변환된 자료를 제공할 경우, 이용자들은 원자료의 평균과 분산은 모르고 단지

변환된 자료 Y_i와 잡음 e_i의 평균과 분산만 알 수 있다. 따라서 원자료 X_i의 평균과 분산을 구할 수 있는 계산식이 필요한데, 이들의 평균과 분산은 위의 식을 이용하여 다음과 같이 얻을 수 있다.

$$\hat{E}(X_i) = \hat{E}(Y_i) / \hat{E}(e_i)$$
$$\widehat{Var}(X_i) = \frac{\widehat{Var}(Y_i) - [\hat{E}(X_i)]^2 \cdot \widehat{Var}(e_i)}{\widehat{Var}(e_i) + [\hat{E}(e_i)]^2}$$

이상에서 살펴본 바와 같이, 가계조사의 민감변수에 대해 <표 2-10>에 주어진 4가지의 승법잡음에 따라 자료를 변환하여 시기별 및 지역별로 평균과 표준편차를 구하였으며, 이 중 수입관련 항목인 가구주급여 소득과 지출관련 항목인 식사대에 대한 결과를 <표 2-12>에 나타내었다. 가구주급여소득 항목에서 연간자료의 평균값을 비교해 보면 전국단위에서는 방법 1이, 서울지역은 방법 4, 그리고 기타지역은 방법 2가 원자료와 가장 유사한 것으로 나타났다. 분기별 자료의 분석에서는 방법 1과 방법 2가 약간 더 좋은 것으로 나타났다. 지출부문의 식사대 항목의 경우, 어느 특정방법이 매우 우수한 것으로 나타난 것이 아니라, 시기별과 지역별에 따라 4가지 방법의 우수성이 서로 다르게 나타났다.

이러한 방법으로 연간자료에 대해 전국단위의 평균값을 비교해 보면, 총 82개 항목 중 34개 항목은 방법 1이, 14개 항목은 방법 2, 19개 항목은 방법 3, 그리고 15개 항목은 방법 4가 약간 더 우수한 것으로 나타나, 방법 1이 가장 우수함을 알 수 있다. 이러한 결과를 정리해 보면 방법 1과 방법 3은 모두 최빈값 1을 중심으로 절사점이 작은 경우인데, 이는 최빈값을 중심으로 절사점을 좌우로 작게 할수록 원자료와 더 유사하게 된다는 것을 의미한다. 따라서 승법잡음모형에서 잡음에 대한 난수를 생성할 때 최빈값에 가까운 값으로 절사점을 주는 것이 더 효과적이다.

〈표 2-12〉 승법잡음 적용결과

(단위: 원)

변수	시기	지역	평 균				
			원자료	방법1	방법2	방법3	방법4
가구주 급여소득	연간	전국	1,967,254	1,966,919	1,968,925	1,966,025	1,962,908
		서울	2,133,225	2,123,645	2,141,393	2,114,724	2,124,118
		기타	1,941,942	1,943,018	1,942,622	1,943,348	1,938,323
	1분기	전국	1,956,740	1,957,399	1,956,716	1,954,555	1,956,118
		서울	2,134,227	2,119,722	2,134,022	2,106,302	2,130,597
		기타	1,926,376	1,929,629	1,926,383	1,928,595	1,926,268
	2분기	전국	1,949,293	1,949,723	1,948,386	1,951,703	1,940,620
		서울	2,121,116	2,114,220	2,136,073	2,130,682	2,109,477
		기타	1,923,359	1,924,895	1,920,058	1,924,690	1,915,135
	3분기	전국	1,974,921	1,975,879	1,981,287	1,974,771	1,974,196
		서울	2,128,457	2,132,042	2,139,952	2,107,755	2,132,618
		기타	1,952,518	1,953,093	1,958,136	1,955,368	1,951,080
	4분기	전국	1,989,891	1,986,348	1,991,458	1,985,056	1,981,950
		서울	2,149,759	2,130,543	2,158,837	2,116,489	2,122,058
		기타	1,967,586	1,966,229	1,968,105	1,966,718	1,962,402
식사대	연간	전국	175,216	175,264	175,142	175,441	175,161
		서울	194,296	194,334	194,500	194,360	194,115
		기타	172,460	172,509	172,346	172,708	172,423
	1분기	전국	171,462	171,619	171,103	171,972	171,393
		서울	189,925	189,905	190,228	189,830	190,233
		기타	168,529	168,714	168,065	169,135	168,400
	2분기	전국	174,400	174,492	174,273	174,477	174,177
		서울	192,698	192,184	193,647	192,686	192,425
		기타	171,813	171,991	171,534	171,903	171,597
	3분기	전국	178,967	179,015	179,202	179,348	178,953
		서울	200,560	201,015	200,273	201,364	200,910
		기타	175,977	175,968	176,284	176,300	175,913
	4분기	전국	176,723	176,599	176,731	176,604	176,810
		서울	195,503	195,766	195,305	195,113	194,222
		기타	174,154	173,978	174,190	174,073	174,428

<표 2-12>의 계속

(단위: 원)

변수	시기	지역	표준편차				
			원자료	방법1	방법2	방법3	방법4
가구주 급여소득	연간	전국	1,238,689	1,297,501	1,343,543	1,357,165	1,409,158
		서울	1,371,615	1,410,215	1,494,568	1,457,358	1,542,959
		기타	1,215,169	1,277,768	1,317,028	1,339,795	1,385,992
	1분기	전국	1,261,683	1,325,573	1,363,929	1,385,849	1,441,657
		서울	1,377,804	1,418,635	1,480,452	1,483,613	1,556,807
		기타	1,238,243	1,307,033	1,340,700	1,366,802	1,418,944
	2분기	전국	1,203,192	1,257,689	1,301,839	1,328,629	1,368,854
		서울	1,334,241	1,377,420	1,468,969	1,460,723	1,527,190
		기타	1,180,049	1,236,784	1,272,385	1,305,478	1,341,570
	3분기	전국	1,228,238	1,290,521	1,344,488	1,347,158	1,405,107
		서울	1,366,841	1,421,852	1,503,083	1,419,778	1,556,333
		기타	1,205,120	1,268,691	1,318,237	1,335,195	1,380,219
	4분기	전국	1,256,676	1,310,424	1,359,443	1,361,475	1,414,505
		서울	1,408,206	1,422,484	1,533,372	1,456,533	1,527,948
		기타	1,232,480	1,292,810	1,331,781	1,346,731	1,396,905
식사대	연간	전국	149,618	154,834	154,659	164,666	163,828
		서울	169,632	178,685	170,082	191,611	181,954
		기타	146,297	150,880	152,100	160,217	160,858
	1분기	전국	143,972	149,383	146,968	158,537	158,872
		서울	141,027	146,617	151,615	155,998	151,768
		기타	144,221	149,616	145,990	158,756	159,770
	2분기	전국	153,213	159,177	157,812	172,510	165,471
		서울	134,458	138,362	146,705	144,507	151,394
		기타	155,512	161,751	159,133	175,961	167,209
	3분기	전국	149,177	155,018	155,163	165,743	161,085
		서울	205,578	219,124	196,843	245,781	206,669
		기타	139,327	143,651	148,244	151,118	153,472
	4분기	전국	152,831	156,476	159,640	162,568	170,436
		서울	195,212	207,031	186,058	213,858	217,807
		기타	145,901	148,043	155,513	154,068	162,752

다. 최적의 비밀보호방법

위에서 2006년 가계조사의 마이크로데이터 제공시 개인정보 노출에 영향을 줄 것으로 판단되는 민감변수를 선정한 후, 연속형 자료의 비밀보호방법 중 반올림과 구간그룹화, 그리고 승법잡음을 적용하여 자료를 변환시킨 후 각 방법별로 기초통계량을 계산하여 원자료와 비교·분석하였다. 여기서는 3가지 방법들 중 가장 우수한 방법을 1차로 선정한 후, 각 방법별로 서로 비교하여 어떤 방법이 가계조사의 마이크로데이터 작성에 더 효율적인지를 살펴보고자 한다. 위에서 언급한 바와 마찬가지로 3가지 방법들을 서로 비교하기 위해서는 모든 민감변수에 대해 검토해야 하지만, 여기서는 편의상 수입관련 항목인 가구주급여소득과 지출관련 항목인 식사대를 기초로 하여 시기별 및 지역별로 비교·분석해 보았다.

먼저, 반올림을 적용한 경우에는 백단위에서 반올림한 방법 1을 선정하였고, 구간그룹화를 적용한 경우에는 log-normal 분포를 가정한 방법 3을 선정하였다. 그리고 승법잡음을 적용한 방법에서는 최빈값을 중심으로 가까운 값을 절사점으로 사용한 방법 1을 선정하여 이들 3가지 방법에 의한 평균과 표준편차를 시기별 및 지역별로 구한 후, 그 결과를 <표 2-13>에 수록하였다. 이 표에 의하면 거의 모든 경우에 반올림으로 변환한 평균값이 원자료의 평균값과 차이가 가장 적다는 것을 알 수 있다. 즉, 백단위에서 반올림한 결과가 구간그룹화나 승법잡음에 의한 결과보다 거의 대부분의 경우에 더 우수한 것으로 나타났다. 또한 반올림을 제외하고 구간그룹화와 승법잡음의 두 방법만을 비교한다면 거의 모든 경우에 승법잡음에 의한 변환이 더 우수함을 알 수 있다. 이러한 결과를 정리하면, 가계조사자료의 비밀보호를 위해서는 백단위에서 반올림하는 것이 원자료와 가장 유사하다고 할 수 있으나, 자료제공 및 정보손실 등의 측면에서 보면 승법잡음을 이용한 방법도 효과적이라고 할 수 있다.

〈표 2-13〉 3가지 방법별 결과비교

(단위: 원)

변수	시기	지역	평 균			
			원자료	반올림 (방법 1)	구간그룹화 (방법 3)	승법잡음 (방법 1)
가구주 급여소득	연간	전국	1,967,254	1,967,268	1,961,531	1,966,919
		서울	2,133,225	2,133,231	2,122,965	2,123,645
		기타	1,941,942	1,941,958	1,936,911	1,943,018
	1분기	전국	1,956,740	1,956,757	1,948,729	1,957,399
		서울	2,134,227	2,134,234	2,121,533	2,119,722
		기타	1,926,376	1,926,394	1,919,165	1,929,629
	2분기	전국	1,949,293	1,949,307	1,943,357	1,949,723
		서울	2,121,116	2,121,119	2,111,771	2,114,220
		기타	1,923,359	1,923,375	1,917,937	1,924,895
	3분기	전국	1,974,921	1,974,932	1,970,724	1,975,879
		서울	2,128,457	2,128,463	2,119,007	2,132,042
		기타	1,952,518	1,952,531	1,949,088	1,953,093
	4분기	전국	1,989,891	1,989,906	1,985,530	1,986,348
		서울	2,149,759	2,149,764	2,141,063	2,130,543
		기타	1,967,586	1,967,602	1,963,830	1,966,229
식사대	연간	전국	175,213	175,298	174,189	175,264
		서울	194,296	194,373	192,565	194,334
		기타	172,460	172,543	171,534	172,509
	1분기	전국	171,462	171,545	170,351	171,619
		서울	189,925	190,007	188,822	189,905
		기타	168,529	168,612	167,416	168,714
	2분기	전국	174,400	174,483	173,229	174,492
		서울	192,698	192,776	192,160	192,184
		기타	171,813	171,896	170,553	171,991
	3분기	전국	178,967	179,050	178,118	179,015
		서울	200,560	200,637	197,788	201,015
		기타	175,977	176,061	175,394	175,968
	4분기	전국	176,723	176,803	175,762	176,599
		서울	195,503	195,574	192,752	195,766
		기타	174,154	174,235	173,439	173,978

〈표 2-13〉의 계속

(단위: 원)

변수	시기	지역	표준편차			
			원자료	반올림 (방법 1)	구간그룹화 (방법 3)	승법잡음 (방법 1)
가구주 급여소득	연간	전국	1,238,689	1,238,689	1,194,896	1,297,501
		서울	1,371,615	1,371,613	1,321,745	1,410,215
		기타	1,215,169	1,215,169	1,172,412	1,277,768
	1분기	전국	1,261,683	1,261,686	1,197,199	1,325,573
		서울	1,377,804	1,377,801	1,324,269	1,418,635
		기타	1,238,243	1,238,248	1,171,591	1,307,033
	2분기	전국	1,203,192	1,203,192	1,168,006	1,257,689
		서울	1,334,241	1,334,238	1,279,739	1,377,420
		기타	1,180,049	1,180,050	1,148,120	1,236,784
	3분기	전국	1,228,238	1,228,238	1,195,548	1,290,521
		서울	1,366,841	1,366,839	1,317,376	1,421,852
		기타	1,205,120	1,205,120	1,175,222	1,268,691
	4분기	전국	1,256,676	1,256,673	1,217,652	1,310,424
		서울	1,408,206	1,408,209	1,367,466	1,422,484
		기타	1,232,480	1,232,477	1,193,718	1,292,810
식사대	연간	전국	149,618	149,623	130,085	154,834
		서울	169,632	169,635	130,983	178,685
		기타	146,297	146,303	129,741	150,880
	1분기	전국	143,972	143,977	127,08	149,383
		서울	141,027	141,022	129,528	146,617
		기타	144,221	144,227	126,450	149,616
	2분기	전국	153,213	153,218	128,347	159,177
		서울	134,458	134,463	127,422	138,362
		기타	155,512	155,518	128,256	161,751
	3분기	전국	149,177	149,183	133,324	155,018
		서울	205,578	205,578	135,620	219,124
		기타	139,327	139,334	132,778	143,651
	4분기	전국	152,831	152,838	131,898	156,476
		서울	195,212	195,223	131,685	207,031
		기타	145,901	145,907	131,761	148,043

제4절 결 론

앞에서 살펴본 바와 같이, 본 연구에서는 연속형 자료의 비밀보호방법인 반올림과 구간그룹화, 그리고 승법잡음을 적용하여 2006년 가계조사를 대상으로 비밀보호된 마이크로데이터를 작성하는 과정을 설명하고 비교·분석해 보았다. 이를 위해 먼저 조사항목들을 외부에서 쉽게 식별이 가능할 것으로 판단되는 식별정보와 자료제공시 개인정보 노출에 영향을 줄 것으로 판단되는 민감정보로 구분하였다. 식별정보의 보호를 위해 가구유형과 거처구분, 가구원수, 배우자유무, 가구주성별, 가구주연령, 그리고 가구주직업 등 모두 7개의 항목을 식별변수로 선정하여 key변수로 사용하였다. 또한 자료파일의 유일성과 노출의 위험을 알아보기 위해 모집단(2005 인구주택총조사 10% 표본조사결과)과 표본(2006년 가계조사결과)에서 key변수의 각 조합별로 유일성을 파악하였다. 그 결과 모집단에서는 약 1.52%의 가구가 유일하고, 표본인 가계조사 자료에서는 약 34.25%의 가구가 유일한 것으로 나타나 전체 자료의 노출위험이 약 0.013%가 되었다. 이에 노출위험을 줄이고자 key변수를 그룹화와 상한그룹화 등의 비밀보호방법을 적용하여 자료의 정보를 축소한 결과, 모집단은 약 0.28%의 가구가 유일하고, 표본은 12.12%의 가구가 유일한 것으로 나타나 두 집단에서 모두 유일성이 크게 줄어들었으며 노출위험도 0.002%로 감소하였다.

한편, 민감정보의 노출제한을 위해 전체 조사항목들 중 수입관련 항목 31개와 지출관련 항목 51개를 합해 총 82개 항목을 민감변수로 선정하였으며, 이들 민감변수는 모두 연속형 자료값을 가지기 때문에 반올림과 구간그룹화, 승법잡음 등 연속형 자료의 비밀보호방법을 적용하여 분석하였다. 먼저 원자료를 백단위와 천단위, 그리고 만단위에서 반올림한 후 변환된 자료의 평균을 계산한 결과, 낮은 단위에서 반올림한 결과가 더 우수한 것으로 나타났다. 구간그룹화에서는 적절한 계급수와 계급간격 및 계급의 한계 등을 정하고 각 구간별로 균등분포와 파레토분포, 그리고 log-normal분포를 가정하여 6가지의 방법에 따라 자료를 변환한 후 원자료의 평균값과 비교한 결과, log-normal로 가정한 경우가 가

장 우수한 것으로 나타났다. 마지막으로 승법잡음의 모형에서 잡음의 분포를 절단된 삼각분포로 가정하여 4가지 방법으로 난수를 생성한 후 원자료에 곱하여 자료를 변환시킨 결과, 자료의 최빈값을 중심으로 가장 가까운 값을 절사점으로 사용한 방법이 가장 우수한 것으로 나타났다.

이상에서 살펴본 바와 같이, 가계조사의 경우 연속형 자료가 대부분이므로 본 연구에서 검토한 세 가지 방법에 의한 자료변환이 비밀보호된 마이크로데이터를 작성하는 데 상당히 효과적임을 알 수 있다. 다만 가계조사의 특성상 고려해야 할 변수들이 너무 많고 변수별로도 적합한 방법이 다르기 때문에 어떤 하나의 방법을 최적이라고 하기에는 다소 무리가 있다. 그렇지만 자료제공과 정보노출 제한의 측면을 동시에 고려한다면 승법잡음을 이용한 비밀보호가 매우 유용한 방법이 될 수 있을 것으로 판단된다. 물론 잡음의 분포와 난수생성, 절사점 선정 등의 다소 복잡한 과정이 필요하지만, 모의실험 등을 통한 후속 연구가 활발히 이루어지고 이에 대한 알고리즘이 정립된다면 승법잡음모형을 이용한 방법이 연속형 자료의 비밀보호에 매우 효과적일 수 있을 것이라 사료된다. 따라서 향후에도 승법잡음모형에 대한 이론적 연구가 꾸준히 진행되어 연속형 자료의 마이크로데이터 작성에 많이 활용되길 기대해 본다.

참고문헌

정동명・강동환(2007), “마이크로자료의 활용도제고를 위한 비밀보호방법”, 통계연구결과보고서, 통계청.

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

통계청(2006), 「가계조사 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. 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, Jr., J. Gonzalez, and P. Williams(2003), "Techniques for Masking Microdata", National Center for Health Statistics internal memorandum.

Kim, J.(2007), "Application of Truncated Triangular and Trapezoidal Distributions for Developing Multiplicative Noise", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, CD-ROM.

Kim, J. and Jeong, D. M.(2007), "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.

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.

Massell, P. and N. Russell(2006), "Protecting Confidentiality of Commodity Flow Survey Tabular Data by Adding Noise to Underlying Microdata", presented at a Washington Statistical Society seminar.

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