The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.6.1041
29(6), 1041–1059

Statistical disclosure control for public microdata: present and future

Min-Jeong Park[a] · Hang J. Kim[b,1]

[a]Statistical Research Institute, Statistics Korea;
[b]Department of Mathematical Sciences, University of Cincinnati

(Received August 31, 2016; Revised October 9, 2016; Accepted October 9, 2016)

Abstract

The increasing demand from researchers and policy makers for microdata has also increased related privacy and security concerns. During the past two decades, a large volume of literature on statistical disclosure control (SDC) has been published in international journals. This review paper introduces relatively recent SDC approaches to the communities of Korean statisticians and statistical agencies. In addition to the traditional masking techniques (such as microaggregation and noise addition), we introduce an online analytic system, differential privacy, and synthetic data. For each approach, the application example (with pros and cons, as well as methodology) is highlighted, so that the paper can assist statical agencies that seek a practical SDC approach.

Keywords: data privacy, masking, analytic system, differential privacy, synthetic data

1. 서론

국가 통계 기관과 기타 공공 기관들은, 수집한 정보를 정리하여 공공 자료의 형태로 제공함으로써, 정부와 민간 부문에서 요구하는 정보 수요를 충족시킨다. 정보 기술의 급격한 발전과 그에 따른 사회 환경의 변화로 인하여, 이러한 공공 자료의 수요량과 활용 범위는 최근 대폭 증가되었다. 하지만 동시에 그에 따른 개인 혹은 개별 기업의 정보 유출에 대한 염려 역시 커지고 있다. 이에 따라 각국의 통계 기관들은 개별 정보보호 혹은 노출제어(disclosure control/limitation)를 위한 연구 인력 및 시스템 개발에 대한 투자를 늘리고 있다.

국가 통계 기관에서는 1980년대까지만 하더라도 합산표로 대표되는 매크로데이터(macrodata)를 주로 제공하였다. 매크로데이터는 개별 정보가 합산되어 있고 구조가 단순한만큼, 프라이버시 침해(privacy disclosure)에 대한 위험은 크지 않다. 그러나 모집단에 대한 심층적 분석이 불가능하여 자료의 활용도가 제한적이다. 이에 따라 최근에는 개별 정보 제공 주체에 대한 원(原)자료로 구성된 마이크로데이터(microdata)를 제공하는 국가 통계 기관이 늘어나고 있다. 이러한 자료 제공 범위의 확대에 따라, 일정 수준의 자료 유용성을 확보함과 동시에 개별 정보의 노출을 제어하는 방법에 대한 연구가 활발해지고 있다.

[1]Corresponding author: Department of Mathematical Sciences, University of Cincinnati, PO Box 210025, Cincinnati, OH 45221, USA. E-mail: hang.kim@uc.edu

Table 2.1. Traditional disclosure control strategies

	User access control			Data disclosure control	
Users	Specialists			Public	
Methods	Output control	Access control	Contract	Masking techniques	Synthetic data
File type	Tables	Research data center Remote access	Licensed file	Public use microdata file	

본 논문은 마이크로데이터 공표시 개별 정보보호를 위한 기법들을 소개한다. 기존 국문 연구 문헌들에서 이미 다룬 내용들은 실제 사례와 함께 재정리하고, 최근 연구 동향을 추가하였다. 기존 연구는 노출위험의 개념을 소개하거나, 그룹화, 잡음추가 등의 매스킹 기법에 대해 소개하였고 (Park, 2004; Jeong과 Kang, 2006; Kim, 2009; Kim 등, 2011; Kim 등, 2011), 이러한 방법들을 인구주택총조사 (Jeong과 Jeong, 2008), 가계조사 (Jeong 등, 2009), 교원 정보 (Lee와 Kim, 2011), 가계금융·복지조사 (Park 등, 2013)에 대해 적용하기 위한 노력을 다루고 있다. 또한 Lee (2013)는 마이크로데이터의 정보보호 방법을 개괄적으로 요약하면서 중요 개념인 노출위험(disclosure risk)의 측정에 대하여 자세히 설명하였고, Park (2014)은 마이크로데이터의 노출위험 측도들에 대해 상세히 소개하고 이를 가계금융·복지조사 자료에 적용한 결과를 제시하였다. 이러한 연구 결과에 더하여, 본 논문은 기존 매스킹 기법들의 한계를 극복하기 위해 논의되고 있는 시스템적 대안 개발, 차등정보보호(differential privacy) 및 재현자료(synthetic data) (재현(再現)자료; Lee (2013)에서는 인위자료로, Park 등 (2013)에서는 합성데이터로 번역되었다) 방법을 중점적으로 소개한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 마이크로데이터 공표시 개별 정보보호를 위해 널리 사용되고 있는 매스킹 기법들에 대한 전반적인 내용을 소개하였다. 특히, 이러한 전통적 기법들의 한계에 대해 논의함으로써, 다음 절들에서 다루어지는 최신 방법들에 대한 필요성을 강조하였다. 3절에서는 전통적 매스킹 기법들과 시스템적 접근의 결합을 통해 통계적 노출제어를 이루려는 시도들을 정리하고, 4절에서는 차등정보보호 방법의 개념, 장·단점 및 실례를 소개하였으며, 5절에서는 재현자료 방법에 대해 자세히 소개하였다. 6절에서 향후 연구 방향을 논의하였다.

2. 전통적 개별 정보보호 전략

통계 기관에서 마이크로데이터를 공표할 때 추구하는 목표는 크게 다음 세 가지가 있다 (Reiter, 2004).

- 응답자의 식별 정보나 민감 정보를 알고자 하는 외부 공격으로부터의 안전성
- 광범위한 통계적 분석에 이용될 수 있는 정보의 충분성
- 보편적인 통계적 방법론들을 사용하는 이용자들의 편의성

그러나 정보손실없이 안전하면서도 접근하기 편리하도록 마이크로데이터를 제공하는 것은 현실적으로 쉽지 않다. 원자료를 그대로 제공할 경우 정보손실은 없지만 개별 정보가 식별될 위험이 존재하고, 노출위험을 줄이기 위해 자료를 변형하면 정보손실이 발생할 수 밖에 없기 때문이다. 또한 정보손실과 노출위험없이 자료를 제공하려면 이용자를 엄격히 통제하는 절차를 만들 수 밖에 없어, 위의 세 목표를 동시에 충족시키기는 어렵다. 때문에 대부분의 통계 기관들은 위의 세 목표 중 일부를 달성하도록 이용자의 유형을 나누어 자료를 제공하여 왔다. 지금까지 실제 자료 공표에 사용된 적이 있는 노출제어 방법들을 분류해 보면 Table 2.1과 같다.

먼저 이용자의 접근을 물리적으로 규제하는 전략(user access control)을 통해 정보손실없이 자료를 제공하고 노출위험을 제어할 수 있다. 이용자의 요청에 대해 공공정보 제공 기관이 직접 분석 결

과(tables)를 만들어서 제공하는 방법, 입출입 보안이 철저한 데이터 센터(research data center; RDC) 내에서만 자료 접근을 허용하는 방법, 특정 시간 동안 특정 IP로부터의 자료 접속(remote access)을 허가하는 방법, 기관과의 법적 계약을 체결한 자료(licensed file)의 수령, 분석, 파기 전과정을 감독하는 전략 등이 실제로 기관에 따라 시행되고 있다 (김경미 등, 2007). 이런 접근은 모두 마이크로데이터 생산 부서의 사용 허가 또는 분석 결과 검열 아래 이루어진다. 따라서 자료 이용 요청이 많은 경우 자료 제공 기관에게는 결과물에 대한 검열 업무 부담이 크다. 또한, 정보 이용자에게는 물리적 이동에 대한 비용, 시간적 제약, 행정 처리 비용 등의 부담이 발생한다.

반면 공공재로 활용되도록 공표하는 공공이용파일(public use microdata file)은 정보 이용자의 접근이 용이하며, 공공정보 제공 기관에게는 제공 후 업무 부담이 적다는 장점을 가진다. 수요가 많은 비교적 간단한 분석들에 대해 원자료 분석결과와 거의 차이가 없게 할 수 있다면, 공공이용파일 활성화는 자료 제공자와 이용자 모두에게 유익할 것이다. 이에 통계 기관들은 자료 이용 수요의 많은 부분을 만족시키도록 공공이용파일을 만들고, 심층적인 분석을 원하는 전문가들에게는 데이터 센터 등을 이용하도록 하는 이원화 전략을 취하고 있다.

원자료 분석 결과와 차이가 적으면서도 노출위험이 적은 공공이용파일을 만들기 위해서 통계 기관들은 매스킹 기법들과 재현자료 방법을 주로 사용해왔다. 재현자료 관련 실제 사례로는 2001년 프랑스의 종단 연계 자료 (Abowd와 Woodcock, 2001), 독일 IAB 기관 패널 조사 (Drechsler와 Reiter, 2009) 및 미국 사업체 종단 자료 (Kinney 등, 2011) 등이 있다. 재현자료 방법의 구체적인 내용과 최근의 발전에 대해서는 5절에서 자세히 논한다. 한편, 매스킹 처리는 전통적으로 널리 사용되어져 왔으며 변수 유형별로 다양한 기법들이 존재한다. 매스킹 처리를 많이 할수록 노출위험이 줄어들지만 자료의 변형으로 인한 정보손실도 많이 발생한다. 때문에 노출위험과 정보손실을 동시에 고려하여 적절한 절충안을 모색해야 하는 어려움이 있다. 2절에서는 매스킹 처리에 대하여 차례로 살펴본다.

2.1. 매스킹 기법

마이크로데이터에서 이름, 주소 등의 직접적인 식별 변수를 제거하더라도 지역, 성별, 연령, 직업 등 몇몇 변수를 결합하면 특정 개인을 식별할 수 있는 경우가 종종 발생한다. 매스킹(masking)이란 원자료에 적절한 변환을 하여 이러한 간접적인 식별 정보를 감추는 것을 말한다. 예를 들어서 부분적인 샘플 자료만 제공하거나, 특정 자료값을 감추거나, 그룹을 구성해 그룹별로 동일한 값을 제공하는 방법 등이 있고, 연속형·범주형 등 변수 유형에 따라 적용할 수 있는 기법이 다르다. 또한 각 기법별로 다양한 세부 알고리즘이 존재하고, 각 알고리즘마다 관련 논문들이 다양하다. 자세한 내용은 Duncan 등 (2011)의 5장 및 관련 참고 문헌들이나 Kim 등 (2011)을 참고하고, 대표적인 몇몇 기법을 소개하기로 한다.

2.1.1. 국소 감추기(local suppression) 마이크로데이터에서 보통 행은 개체 혹은 레코드를, 열은 변수를 나타낸다. 지역정보 변수처럼 특정 변수가 노출위험을 지나치게 높이거나, 고소득자와 같이 특정 개체가 지나치게 민감한 값을 가질 경우 변수나 개체를 통째로 감출 수 있다. 이는 정보손실을 많이 야기하므로, 마이크로데이터에서 몇몇 셀만 감추어 노출위험을 낮추고 정보손실을 줄이는 방법이 국소 감추기이다. 국소 감추기를 사용할 때 보통은 익명성을 2이상 확보하도록 유일한 개체를 찾아 적절한 변수에 대한 값을 감춘다. 이때 익명성이 2 이상이라는 것은, 간접적인 식별 변수의 조합이 동일한 개체수가 2개 이상인 것을 말한다 (Sweeney, 2002).

2.1.2. 전반적 재코딩(global recoding) 특정 변수의 범주를 더 상위 범주로 묶는 것을 전반적 재코딩이라고 한다. 예를 들어 연령을 각 나이별로 제공할 경우 노출위험이 커지므로 5세 단위로 묶어 제

공하는 경우를 생각할 수 있다. 또한 특정 값 이상/이하를 묶는 것을 top/bottom 코딩이라고 한다. 예를 들어서, 소득을 100만원 단위로 분류하다가, 1천만원 이상의 소득자는 "1 천만원 이상"이라는 하나의 그룹으로 분류하는 것이 top 코딩이다. 재코딩을 실행할 시, 빈도수가 너무 적은 그룹이 생기지 않도록 범주를 묶는 것이 바람직하다.

2.1.3. 국소통합(microaggregation) 보통 3개 이상의 개체를 한 그룹으로 묶고, 각 그룹의 개체 값들을 그룹의 평균값이나 중앙값 등 동일한 한 값으로 대체하는 것을 국소통합이라 한다. 최대한 유사한 개체들을 한 그룹으로 묶는 것이 좋으며, 단일 변수를 기준으로 그룹을 만들거나 주성분 등을 이용해 여러 변수를 한번에 고려하여 그룹을 편성할 수도 있다. 제안된 여러 알고리즘들 중 네델란드 통계청에서 제안한, 자료간 거리가 평균에서 먼 것부터 그룹을 형성하여 나가는 multivariate microaggregation based on maximum distance to average vector(MDAV) (Statistics Netherlands, 2007) 기법이 비교적 효율적이라 판단된다.

2.1.4. 잡음추가(noise addition) 주어진 변수들의 결합분포와 동일한 잡음을 생성하여 더하고 적절한 상수로 나누어 원래 분포를 보존하는 것이 이상적인 잡음추가 기법이다. 현실적으로는 매스킹 처리를 하고자 하는 변수들의 공분산 행렬에 비례하는 공분산을 가지는 분포에서 잡음을 생성할 수 있으나, 원자료가 정규성을 따르지 않으면 자료의 구조가 심각하게 왜곡될 수도 있다. 이에, Templ과 Meindle (2008)은 보다 로버스트한 알고리즘을 제시하였다. 이 외에도 자료의 크기를 고려하거나 자료 중에서 이상값들을 골라내어 잡음을 더하는 알고리즘 (Templ, 2008), 승법 잡음(multiplicative noise) 추가 기법 (Jeong 등, 2009; Kim 등, 2011) 등이 있다. 잡음추가 기법은 종종 다른 매스킹 기법들과 함께 사용된다. 예를 들어서, 잡음추가와 국소통합의 결합방법의 경우, 국소통합에 의하여 각 그룹의 개체 값들을 동일한 한 값(평균 등)으로 대체한 후, 각각의 값에 대하여 잡음을 추가한다.

2.1.5. 기타 기타 매스킹 기법들로는 자료 교환(data swapping), 자료순위 교환(rank swapping) 및 자료섞기(shuffling) 등이 있다. 자료 교환은 보통 범주형 변수들에 대해서 사용되며, 예를 들어 서로 다른 지역에서 비슷한 조건을 가지는 두 개체를 맞교환함으로써 노출위험을 줄인다. 자료순위 교환은 연속형 변수에서 자료를 정렬하고 제한된 범위(%) 내에서 자료를 서로 교환하는 것을 말한다. 자료섞기는 조건부 분포를 이용하여 새로운 자료셋를 생성하여 제공하는 것으로 재현자료 활용과 유사한 측면이 있다. 위에서 소개된 마이크로데이터를 위한 매스킹 기법들은 네덜란드 통계청의 μ-Argus (Statistics Netherlands, 2007) 혹은 R패키지 sdcMicro (Templ, 2008)등을 통해서 사용할 수 있다.

2.2. 노출위험과 정보손실 측도

공공정보 제공 기관은 하나의 원자료에 대하여 다양한 매스킹 기법들을 적용한 후, 정보손실 측도와 노출위험 측도를 이용하여 매스킹된 자료들(masked datasets) 중 어떤 것을 실제 공공이용파일로 사용할 것인가를 결정한다 (Figure 2.1). 정보손실 측도는 매스킹 전후 자료의 변화를 거리를 이용하여 측정하는 경우가 많으며 관련 이론은 비교적 간단하다고 할 수 있다. 노출위험 측도는 유일성에 근거하거나, 공격자에 대해 의사결정이론을 적용하여 측정한다. 노출위험 측정은 무척 중요한데, 개별 개체의 노출위험이 적절히 측정될 경우 노출위험이 높은 개체만 매스킹 처리하여 정보손실을 좀 더 낮출 수도 있기 때문이다.

2.2.1. 정보손실 정보손실은 매스킹된 자료가 원자료와 얼마나 달라졌는가를 평가하여 측정한다.

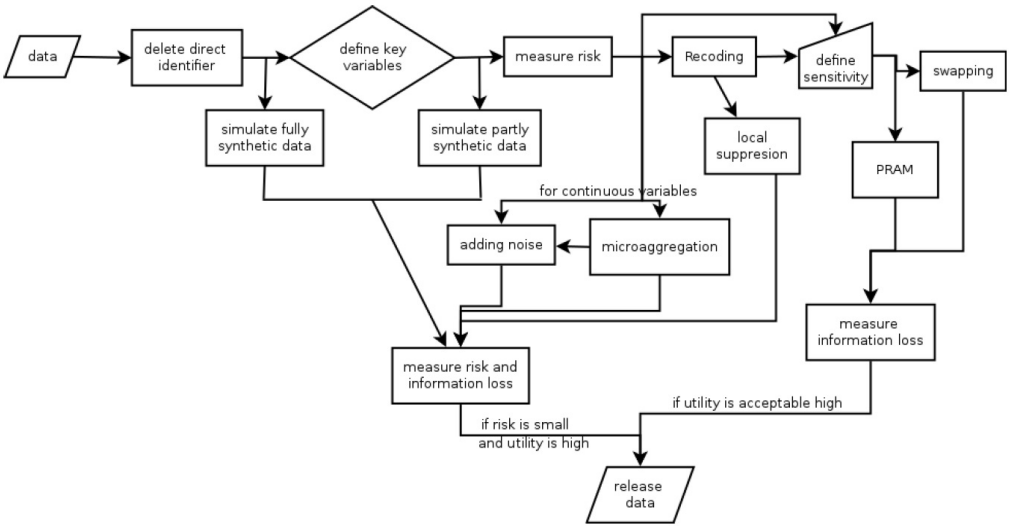

Figure 2.1. Traditional masking procedure (Meindl *et al.*, 2013).

료값 자체, 상관계수, 주성분 행렬 등에 대하여 평균제곱오차 등을 계산하여 정보손실 측도로 삼을 수 있다. 또한 매스킹 전후 확률밀도함수를 추정하여 쿨백-라이블러 거리(Kullback-Liebler divergence)를 측정하거나, 두 경험적 분포 사이의 절대 거리를 활용하기도 한다. 한편, 주요 통계량에 대한 신뢰구간 중복(confidence interval overlap)을 이용하거나 (Karr 등, 2006), 매스킹 전후 자료셋에 대한 성향 점수(propensity score)를 활용하는 연구 (Woo 등, 2009)도 있다.

2.2.2. 유일성에 근거한 노출위험 보통 모집단 내 유일한 개체의 수를 추정하여 마이크로데이터 파일의 노출위험 측도로 이용한다. 이를 위해 마이크로데이터를 변수 조합에 대한 빈도표로 바꾸어 포아송 모형을 적용하고, 각 셀에 대한 기대값을 추정하기 위해 모수에 대해 감마 분포를 가정한다. 이러한 포아송-감마 모형 (Bethlehem 등, 1990)을 통해 모집단 내 유일한 개체수를 추정할 수 있고, 그 외에도 음이항 모형 (Franconi와 Polettini, 2004), 로그-선형 모형 (Skinner과 Holmes, 1998) 등을 사용할 수 있다. 이 모형들은 과소추정이라는 단점을 가진다.

한편 개체별 노출위험은, 표본 내 유일한 개체가 모집단에서도 유일할 확률(r_{1k})이나, 모집단 개체수 역수의 기대값(r_{2k})으로 정의한다 (Skinner과 Shlomo, 2008). 변수 조합에 대한 빈도표가 총 K개 셀을 가질 때, F_k를 k번째 셀의 모집단 개체수, f_k를 표본 개체수라 하면, k번째 셀에 속하는 개체들의 노출위험은 다음과 같이 표현된다.

$$r_{1k} = \Pr(F_k = 1 | f_k = 1),$$
$$r_{2k} = E\left(\frac{1}{F_k}\middle| f_k = 1\right).$$

추정을 위해서, F_k는 평균 λ_k인 포아송 모형을, 각 개체가 표본에 포함될 확률 π_k은 베르누이 분포를 따른다고 가정하면, $f_k \sim \text{Poisson}(\pi_k\lambda_k)$가 된다. 또한 k번째 셀의 모집단 개체수는 표본의 개체수와 표본에 포함되지 않은 개체수의 합이므로 $F_k = f_k + u_k$, $u_k \sim \text{Poisson}(\lambda_k(1-\pi_k))$가 되며, 이를 종합

하면 두 노출위험의 측도는 다음과 같이 정리된다.

$$r_{1k} = \exp\{-(1-\pi_k)\lambda_k\},$$
$$r_{2k} = \frac{1-\exp\{-(1-\pi_k)\lambda_k\}}{(1-\pi_k)\lambda_k}.$$

이제 각 셀별 변수 범주값들의 벡터 x_k를 이용해 로그선형모형을 적용하면 $\log\lambda_k = x_k'\beta$가 되고, 표본의 개체수 f_k가 Poisson$(\pi_k\lambda_k)$의 실현값이므로 표본 자료를 이용해 최대우도추정량 $\hat{\beta}$을 구하면, 모수 추정량 $\hat{\lambda}_k = \exp(x_k'\hat{\beta})$을 통해 노출위험 추정량을 얻을 수 있다. 얻어진 개체별 노출위험의 합이나 평균을 파일 단위의 노출위험 값으로 활용할 수도 있다.

이러한 로그선형모형을 활용한 노출위험 측도 연구에서는 추론이 가능하도록 검정 통계량들이 제시되어 있고, 기존 연구들의 과소추정 문제를 해결하기 위해 검정 통계량들을 이용해 변수 선택을 할 수 있기도 하다. 이 외에도 과소추정 문제를 더욱 완화시킨 Bayesian version of grade of membership(베이지안 GoM) 모형도 연구되어 있다 (Manrique-Vallier과 Reiter, 2012). 이러한 연구들은 주로 미국 센서스 자료에 적용, 평가되었다.

2.2.3. 의사결정이론에 근거한 노출위험 위에서 소개한 유일성에 근거한 개체별 노출위험 측정은 연속형 변수들을 고려할 경우 유일성이 너무 많이 발생하여 의미를 가지기 어렵다. 따라서 잡음추가 등의 매스킹 기법들에 대하여는 그 효과를 적절히 설명하지 못하고, 공격자가 가지고 있는 자료의 특징을 반영하지 못하는 단점도 가지고 있다. 이를 극복하기 위해 Duncan과 Lambert (1989)는 의사결정이론(decision theory)를 이용해 노출위험을 측정하고자 하였다. 원래 마이크로데이터를 $\boldsymbol{Y}$, 매스킹된 자료를 $\boldsymbol{Z}$, 표적 개체를 t라고 하면, 공격자는 $\boldsymbol{Z}$에서 t의 위치를 파악해 t에 대한 추가 정보를 얻고자 한다. 이를 위해 표적이 공표 자료에 있는지 없는지를 판단하고 자료를 서로 연결해야 하는데, 이런 과정을 통계적 의사결정이론을 이용하여 설명하면 다음과 같다.

매스킹된 자료 $\boldsymbol{Z}$ 내의 임의의 개체 z가 표적 t라고 예측하는 확률을 $p_Z(z)$, 손실 함수를 $L(t,z)$라고 하면 평균손실함수는 $\int L(t,z)p_Z(z)dz$이다. 공격자는 표적이 공표 자료에 있을 때 없다고 판단하거나(상황1, 손실 ℓ_1), 표적을 틀리게 연결하는(상황2, 손실 ℓ_2) 의사결정을 원하지 않는다. 상황1에서 최소 평균손실(불확실성)은 $L_1 = \ell_1\sum_{i=1}^n p(z_i)$, 상황2에서는 표적일 확률이 가장 큰 개체가 표적이 아닐 확률을 고려하여 $L_2 = \ell_2[1-\max_{1\le i\le n} p(z_i)]$가 된다. 통계 기관에서는 $L_1 < L_2$가 되도록 매스킹 처리를 하여 공격자의 정보 노출 의지를 약화시켜 노출위험을 낮추고자 하게 된다. 즉, $\sum_{i=1}^n p(z_i)$ 및 $\max_{1\le i\le n} p(z_i)$를 낮추도록 노력하게 되며, Reiter (2005)는 미국의 경제 활동 인구 조사(Current Population Survey) 마이크로데이터를 이용해 이러한 내용을 설명하였다.

2.3. 평가

매스킹 처리를 통한 노출제어의 어려움 중 하나는, 어떠한 노출위험 측도와 정보손실 측도를 사용하는지에 대한 합의가 이루어지지 않아서, 어떠한 측도에 근거하여 매스킹된 자료를 선택해야 할지 모른다는 것이다. Figure 2.2에서는 마이크로데이터에 대해 잡음추가, 국소통합 및 두 방법의 결합기법들(nsX, ma, combiX로 각각 표시됨)의 세부 알고리즘별 노출위험과 정보손실 측도를 나타내었다 (Park, 2014). 노출위험은 개별 노출위험의 평균을 사용했으며, 정보손실은 자료간의 거리를 이용한 경우(왼쪽 그래프) 및 고유값 사이의 거리를 이용한 경우(오른쪽 그래프)로 각각 나타내었다. 원점에 가까운 ns2는 ma보다 노출위험과 정보손실 양 측면에서 좋은 방안임을 알 수 있다. 하지만, ns1, ns2, ns3 사이에서는 두 측도 모두에서 더 좋은 방안을 특정하기 어렵다. 또 고유값 사이의 거리를 이용한 경우 정보손실

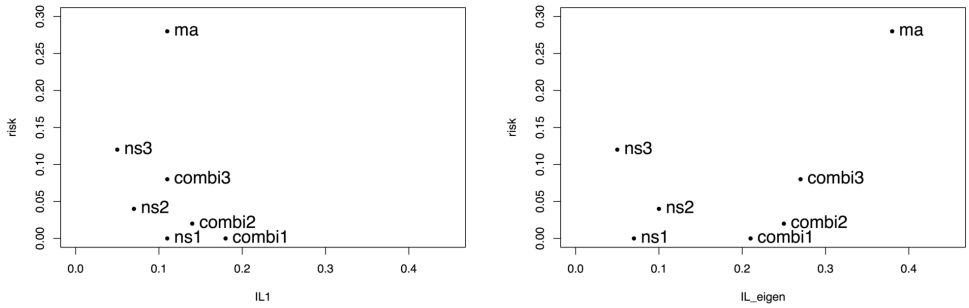

Figure 2.2. Example of risk-utility map. The masked dataset with microaggregation is denoted as `ma`. Randomly generated replicates from noise addition and combined approaches are denoted as `nsX` and `combiX`, respectively. The left panel uses the original data values while the right panel uses the eigenvalues, respectively, in order to measure information loss, IL (Park *et al.*, 2013).

은 자료의 구조 보존 측면을 나타내는데 이것이 자료간 거리 기반 정보손실 결과와 일치하지 않아 어느 방안이 더 좋은 것인지 판단하기도 쉽지 않다. 이렇듯 매스킹 처리에는, 노출위험과 정보손실 모두를 낮춘 최적의 공공이용파일을 객관적으로 선택하기 어렵다는 문제가 있다.

3. 시스템적 대안 모색

앞 절에서 전통적 매스킹 기법만으로는 노출위험과 정보손실을 동시에 낮추기 어렵다는 점을 설명하였다. 이제 매스킹의 한계를 벗어나기 위한 방법론적 대안들을 분야별로 3, 4, 5절에서 차례로 살펴보고자 한다. 이번 3절에서는 먼저, 분석 시스템을 통해 마이크로데이터를 제공하는 노력을 시스템적 대안 모색이라 명명하고 정리하도록 한다. 이 시스템은 기존 매스킹 기법들을 활용하여 노출을 제어하고, 공표용 마이크로데이터보다 제공 범위를 확대하여 자료 유용성을 높이고, 이용자의 접근 편의성을 증진시키는 것을 목적으로 한다. 이는 자유로운 접속에 대해서 노출제어 가능한 요청만을 수용하여 그 처리된 결과물을 제공하는 것이기 때문에 사용자의 접근을 규제하는 원격 접속과는 다르다. 또한 주로 분석 결과물인 매크로데이터에 노출제어 기법들을 활용한다는 점에서 4, 5절에 언급될 마이크로데이터 자체를 안전하게 제공하고자 하는 대안들과도 다른 접근이다.

시스템적 대안은 대부분 통계 기관을 중심으로 연구되었다. 분석 결과를 제공할 때 매스킹 처리에 관하여는 호주 통계청 (Chipperfield와 Yu, 2011)이나 미센서스국 (Lucero 등, 2011)의 연구 등이 있고, 붓스트랩을 응용하는 방안에 대한 연구 (Muralidhar 등, 2013)도 있다. 또한 재현자료를 활용하는 아이디어의 제안 (Retier, 2003a) 이래, 재현자료를 활용해 잔차를 제공하는 방안 (Lucero 등, 2011)도 모색되었다. 본 논문에서는 실제 사례로 가장 최근 연구인 미국 National Center for Health Statistics(NCHS)에서 현재 검토 중인 실시간 온라인 분석 시스템에 관한 연구 (Krenzke 등, 2013)를 통해 시스템적 대안을 이해해 보도록 한다.

3.1. 활용 사례

3.1.1. 배경 NCHS에서는 1957년부터 National Health Interview Survey(NHIS)라는 가구 기반 조사를 시행하여 공표하고 있다. 이 조사는 전국 단위 추정을 전제로 표본이 설계되었으나, 지역(state) 단위 정책 수립 등을 위한 정보 제공 요청이 많아 지역 단위의 자료도 제공하고 있다. 보통 지역 단위 표본이 충분치 않으므로, 이용자들은 데이터 센터(RDC)에서 2년 이상의 자료를 묶는 등의 데이터 가공

Table 3.1. Illustration of real-time Online Analytic System (rOAS)

rOAS				Exterior
Subsystem P				
	Subsystem R			
O	D	S	C	L

O denotes information contained in the public files only, D demographic variables, S survey variables, C information shared by RDC and the subsystem R, and L information outside the rOAS.

작업을 한 후 자료를 이용한다. 반면에 공공이용파일 버전은 지역정보 변수나 민감 변수를 제외하고 공표된다.

최근 NCHS에서는 이용자가 RDC를 방문하지 않아도 지역 단위 자료에 접근할 수 있도록, 실시간 온라인 분석 시스템(real-time Online Analytic System; rOAS)을 구축하기 위한 연구를 시행하였다. 시스템을 구축할 때는 인터페이스나 관리 도구 등의 IT 이슈에 대한 고민도 필요하지만, 여기서는 통계적 정보보호 대안 모색의 관점에서 공표용 자료가 따로 존재하는 가구 기반 조사 자료를 위한 분석 엔진 구축에 대한 이슈만을 정리하도록 한다.

참고로 노출위험, 자료 유용성 및 이용자 편의성 사이의 균형점을 찾고 통계적 분석을 최대화 하기 위해, 분석 엔진 구축 이전에 먼저 결정해 두어야 하는 사항들은 다음과 같다. 1) 다년간 자료 분석을 위한 도구를 제공할지 여부, 2) 분석을 허용하는 개체수의 최소 규모, 3) 쿼리당 요청할 수 있는 변수의 개수, 4) 연령 구간화 폭, 변수 범주 결합 기능 제공 여부, 반올림 규칙 등, 5) 공표용 자료와 RDC 보안 자료들의 통합 범위, 6) 지역별 통계값의 결합 결과와 전국 통계값의 일치 등 일관성 확보 여부, 7) 분석 엔진 속도 및 용량 관리를 위해 내부 자료 파일 구성, 8) 사용한 비밀보호 기법들의 공개 수준 등. 이러한 사항들은 자료 제공자가 정책적으로 사전에 결정해야 하며 관련자들의 합의가 필요하다.

3.1.2. 노출위험 요소 rOAS를 통해 사용자는 요청한 쿼리에 대해서 주로 표의 형태로 분석 결과물을 받게 된다. 표의 형태로 제공되는 자료에서는, 크기가 작은 셀들에서, 즉 해당 셀을 구성하는 개체가 한 개 혹은 두 개일 때, k-익명성 (Sweeney, 2002)의 미확보로 노출이 일어난다고 정의한다. 또한, 표 연계(table linking)에 의해서도 노출이 일어날 수 있다. 어느 셀에 개체가 하나 존재할 때, 관련 교차표들을 여러 개 생성하고 그것들을 연계하여 마이크로데이터를 복원할 수도 있기 때문이다. 한편 다양한 값을 가지는 가중값도 노출위험의 한 요소가 된다. 가중값에 대하여 여러 변수 조합에 대한 교차표를 생성하면, 변수별 범주 값을 알아내어 원래 마이크로데이터가 복원될 수도 있다. 마지막으로 표 분할(table differencing)에 의해서 노출이 일어날 수 있다. 사용자가 요청한 쿼리들에 대하여 rOAS가 제공한 표(explicit tables)들을 이용하여, 결과를 제공하지 않는 한계값(threshold) 이하의 쿼리에 대한 분석 결과(implicit table)를 알아내는 정보 노출이 발생할 수 있다.

노출 발생을 보다 체계적으로 이해하기 위하여, 시스템을 두 개의 하위 시스템, 공공이용파일만을 사용하는 하위 시스템 P 및 RDC 일부 변수도 포함하는 하위 시스템 R로 이루어져 있다고 가정하자. 또한 인구학적 변수들은 D, 설문 대상 변수들은 S, 공공이용파일에만 존재하는 정보들은 O, RDC 자료와 하위 시스템 R에 공통으로 존재하는 정보는 C, 마지막으로 시스템 외부의 정보를 L이라고 표현하자 (Table 3.1). 그러면 P는 O—D—S로 구성되고, R은 D—S—C로 구성된다.

공격자들은 시스템 P를 통해 얻은 결과물을 대상으로 (a) 표 분할 작업을 통해 목표로 하는 특정 개체들에 대한 D—S자료를 얻고, (b) 이를 연계키로 활용하여 C를 연결시켜 O—D—S—C(linking implicit

tables)를 얻은 후, (c) 이를 외부 자료 L과 개체 연계(record linking)하여 O—D—S—C—L을 얻어 노출을 일으킬 수 있다. 이러한 표 분할–표 연결–개체 연계(TD-LIT-RL)의 순서로 일어나는 공격에 의해 마이크로데이터의 노출이 일어날 수 있는데, 이러한 노출 사고를 방지하기 위해 rOAS는 노출제어 기능을 가지고 있어야 한다.

3.1.3. 노출제어 처리 하위 시스템 P는 공공이용파일만을 이용하므로 자유롭게 분석 결과를 제공하는데 문제될 것이 없지만, 하위 시스템 R을 사용하는 경우 rOAS는 분석 결과에 대해 적절한 노출제어 처리를 해야 한다. 이를 위해 주로 쿼리 제한, 한계값 적용, 반올림 및 동적 부차 표본 추출(dynamic subsampling) 등을 활용한다.

먼저 rOAS에서 쿼리 제한이란 하위 시스템 R을 이용하는 쿼리에 대해서 노출위험을 근거로 분석을 거부하거나, 분석 허용 변수 개수를 제한하는 것을 말한다. 한계값 적용(threshold rules)이란 쿼리마다 사전에 정해진 기준을 넘기는지 체크하여 분석을 허용하도록 규칙을 두는 것을 말한다. 예를 들면 분석 결과물을 생성하는데 사용된 개체 개수가 정해진 기준 이하이거나, 제공하는 결과표의 부분합이 일정 기준 이하일 때 분석을 거부하는 규칙을 둘 수 있다. 반올림 규칙은 결과값 유형에 따라 반올림 단위를 정해 두는 것으로, 예를 들어 분석 변수의 마이크로데이터 1사분위수 값이 0–1이면 제공하는 평균값은 소수점 한 자리에서 반올림한다 등의 규칙을 시스템에 내장하는 것을 말한다.

한편, 사용자 및 쿼리마다 일관된 값을 제공하면서 제공된 부분합들을 사용자가 더하여도 시스템이 제공하는 전체 합과 일치하도록 시스템을 구성하고, 동시에 TD-LIT-RL 순서로 이루어지는 공격에 대해 노출이 일어나지 않도록 하는 것은 쉬운 일이 아니다. 일관성을 최대한 확보하기 위한 방편으로 rOAS는 동적 부차 표본 추출 기법을 사용하는데 이를 이차원 교차표의 예를 들어 소개하면 다음과 같다. 먼저 각 개체에 대해서 난수를 생성한다. 쿼리가 요청하는 교차표를 생성된 난수로 만들고 셀 값을 C, 행 부분합을 M1, 열 부분합을 M2, 총합을 U라고 한 다음, C, M1, M2, U를 입력값으로 가지는 seed 함수를 만든다. 동일한 seed 값에 대하여 동일한 부차 표본 추출을 각 셀마다 시행하고, 추출된 표본 자료를 사용하여 최종 교차표를 만들어 제공하도록 한다. 물론 가중값 재조정도 함께 수반되어야 한다. 이런 방식의 장점은 일관성을 지키면서, 실제 분석에는 일부 표본을 사용하여 노출제어 효과를 가지는 것이다.

3.2. 평가

실무적 차원에서 현재 사용하고 있는 시스템과 전통적인 매스킹 기법을 활용하면서 안전하고 유용하게 마이크로데이터를 공표하기 위하여 시스템적 대안 모색이 이루어지고 있다. 이러한 접근은 앞에서 언급한대로 이용자의 접근을 용이하게 하고 제공 자료의 범위를 넓히는 장점이 있으나, 구현을 위해 고려해야 하는 세부 요소들이 무척 다양하다. 또한 기존 시스템이나 공공이용파일과의 차이에 의한 노출이 없도록 통합적 자료 제공을 위해 세심한 논의가 필요하다.

4. 차등정보보호(differential privacy)

4.1. 기본 개념

차등정보보호(differential privacy) 상태란 어떤 정보보호 처리 과정에서 생산된 모든 공표 자료가, 특정하게 정의된 정보보호 수준을 보장하는 상황을 의미한다. 컴퓨터 공학 분야에서 처음 제안된 본 개념은, 특정 공표 자료만이 아닌 공표 자료를 생성하는 과정에 대한 안전도 확보를 목적으로 한다.

4.1.1. 차등정보보호의 정의 차등정보보호 상태는 '공표 자료 생성시 이용되는 데이터베이스(원자

료)안에 특정 개인의 정보가 포함되는 경우와 포함되지 않는 경우, 양 상황에서의 프라이버시 침해정도가 서로 비슷한 상태'를 의미한다 (Dwork와 Smith, 2009). 이러한 상태는 다음과 같은 수학적 표현으로 정의될 수 있다.

정의 4.1 (Dwork, 2006) $\kappa : \boldsymbol{Y} \to \kappa(\boldsymbol{Y})$를 어떤 랜덤화 함수(randomized function)라고 정의하고, 어떤 두개의 데이터베이스 $\boldsymbol{Y}_1$ 과 $\boldsymbol{Y}_2$간에 오직 한 명의 개인 정보만 다르고, 다른 개인들의 정보는 모두 동일하다고 가정한다. 만약 모든 경우의 집합 $S \subset \text{Range}(\kappa)$에 대해서

$$\log\left(\frac{\Pr[\kappa(\boldsymbol{Y}_1) \in S]}{\Pr[\kappa(\boldsymbol{Y}_2) \in S]}\right) \leq \varepsilon \tag{4.1}$$

이 성립한다면, 랜덤화 함수 κ는 ε-차등정보보호를 보장한다. ε의 값이 0에 가까울수록, 강한 수준의 정보보호 상태가 보장된다.

위에서 정의된 차등정보보호 상태에서는, 특정 개인 한 명의 정보가 원자료(데이터베이스)에 포함되어 있는지의 여부가 공표 자료와 그 분석 결과에 유의한 영향을 미치지 못한다. A라는 특정 개인의 정보가 포함된 원자료 $\boldsymbol{Y}_1$과 포함되지 않은 원자료 $\boldsymbol{Y}_2$로부터 생성된 공표 자료들, 즉, $\kappa(\boldsymbol{Y}_1)$과 $\kappa(\boldsymbol{Y}_2)$간에 확률적으로 유의한 차이를 나타내지 않는 것이다. Dwork (2006)은 이러한 상황을 다음과 같은 예로 설명하였다. 만약 어떤 보험회사가 A라는 피보험자의 보험수여 여부를 결정하기 위해서, 공표 자료인 $\kappa(\boldsymbol{Y})$를 분석하였다. 이때, 차등정보보호를 보장하는 보호자료 생성함수 κ 하에서는, A의 정보가 포함된 원자료로부터 생성된 공표 자료 $\kappa(\boldsymbol{Y}_1)$와 A의 정보가 포함되지 않은 원자료로부터 생성된 공표 자료 $\kappa(\boldsymbol{Y}_2)$가 유의미한 차이를 가지고 있지 않기에, 공표 자료가 피보험자 A의 보험금 수령 여부에 영향을 미치지 않는다. 즉, 이러한 차등정보보호 상태에서는 A는 본인의 정보제공(의무)으로 인해 발생할 수 있는 프라이버시 침해에 대한 염려를 덜 수 있다.

이밖에도, ε-차등정보보호보다 정보보호 수준을 다소 완화한, (ε, δ)-차등정보보호 (Nissim 등, 2007), 확률적 차등정보보호 (Machanavajjhala 등, 2008) 등의 다른 정의들이 있다.

4.1.2. 차등정보보호를 보장하는 자료 생성 기법 ε-차등정보보호를 보장하는 기법들은 주로 컴퓨터 공학·암호학 분야에서 사용하는 특정 알고리즘의 이용, 잡음추가 방법의 응용 (Blum 등, 2005), 히스토그램의 응용 (Dwork 등, 2006; Machanavajjhala 등, 2008; Wasserman과 Zhou, 2012) 등이 있다. 특정 알고리즘을 이용한 정보보호는 컴퓨터 공학 분야에서 활발하게 연구되고 있지만, 통계적 추론에 이용되기에는 해당 정보보호 과정이 공표 자료의 분산 혹은 결합분포에 미치는 영향을 설명하지 못한다는 한계점을 지니고 있다. 잡음추가 방법 응용의 경우, 2절에서 소개한 잡음추가 방법과 비슷하며, 단지 식 (4.1)을 만족시킬 수 있는 충분히 큰 잡음의 크기를 찾아내는 방법을 추가적으로 제시한다. 이번 절에서는 통계학 학회지에 비교적 최근 등재된 Wasserman과 Zhou (2012)의 히스토그램을 응용한 기법을 소개한다.

Wasserman과 Zhou (2012)는 평활 히스토그램(smoothed histogram)에서 확률 표본을 생성함으로써 차등정보보호를 보장하는 방법을 제안하였다. 간략하게 그 과정을 소개하면 다음과 같다. n명의 응답자로부터 각각 p개의 항목에 대한 값을 수집하여 얻어진 원자료를 $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$로 나타내고, 각 응답자의 정보 $\boldsymbol{y}_i$는 독립적이며 f라는 동일한 분포를 따른다. 논의의 편의를 위하여 Wasserman과 Zhou (2012)는 응답값들이 0과 1사이의 값이라고 가정하였다.

1. 원자료 $\boldsymbol{Y}$에 대한 히스토그램 밀도 추정량(histogram density estimator)을 계산한다.

$$\hat{f}(\boldsymbol{z}) = \sum_{j=1}^{m} \frac{1}{h^p} \frac{\sum_{i=1}^{n} I(\boldsymbol{y}_i \in B_j)}{n} I(\boldsymbol{z} \in B_j).$$

위에서 f의 서포트(support)는 m개의 p차원 큐브 $\{B_1, \ldots, B_m\}$로 나뉘어져 있고, 각 p차원 큐브의 모든 변의 길이는 $h = 1/m$이다.

2. 적절한 수준의 δ값을 정한 후, 아래와 같이 정의된 평활 히스토그램 $\hat{f}_\delta$에서 n^*개의 공표값을 랜덤하게 생성한다.

$$\hat{f}_\delta(\boldsymbol{z}) = (1-\delta)\hat{f}(\boldsymbol{z}) + \delta.$$

즉, 원래 추정된 히스토그램에 일정 높이의 잡음 분포 δ를 더한 후, 공표값을 생성하는 것이다.

위에서 δ의 값은, 목표 ε값이 주어졌을 때, 아래의 공식에 의해서 결정된다.

$$n^* \log\left(\frac{(1-\delta)m}{n\delta} + 1\right) \leq \varepsilon.$$

Wasserman과 Zhou (2012)는 '위의 과정을 통하여 얻어진 n^*개의 공표값은 항상 ε-차등정보보호를 보장한다'는 것을 증명하였다.

4.2. 평가

차등정보보호의 장점은, (a) 생성되는 공표 자료값, (b) 정보 공격자들이 이용 가능한 외부 데이터베이스의 내용, (c) 공격 전략에 상관없이, 최소 ε수준의 정보보호가 보장된다는 점이다. 이는 공표 자료 $\boldsymbol{Z}$와 외부 데이터베이스 $\boldsymbol{D}$의 확률함수로 정의되는 통계적 노출위험과 크게 대비되는 장점이다. 왜냐하면 통계적 노출위험은, 정보 공격자들이 모의상황보다 더 많은 외부 정보를 이용할 경우, 실제 노출위험을 과소 추정하는데 반해, 차등정보보호 하에서는 정보 공격자들이 사용할 수 있는 '모든 전략과 외부 자료'에 대해 ε수준의 정보보호를 보장하기 때문이다.

이러한 장점에도 불구하고 차등정보보호는 다음과 같은 한계점을 가지고 있다. 첫 번째로, 모든 경우의 수를 고려하는 보수적 위험도 설정은 자료의 유용성을 떨어뜨리게 된다. 현재로서는 정보 공격자들이 이용하지 못하는 정보도 언젠가는 이용될 수 있다는 가정하에서 정보를 보호함으로써 지나치게 높은 수준의 보호를 하게 된다. 두 번째로, 복잡한 구조의 자료의 경우, 차등정보보호 조건을 만족하는 보호자료 생성함수 κ를 찾아내는 것이 어렵다. 세 번째로, 컴퓨터 공학 분야에서 많이 제안되고 있는 알고리즘적 접근법은 분산이나 결합분포 측면에서의 영향을 고려하지 않기에, 통계적 추론이 원할하지 않은 단점을 가지고 있다 (McClure와 Reiter, 2012).

4.3. 활용 사례

위에서 소개한 한계점들 때문에, 차등정보보호 방법이 공공정보 제공 기관들에 의해서 실무에 사용된 사례는 거의 없다. 따라서, 본 절에서는 IT 기업들에 의해서 최근 사용되기 시작한 현지 차등정보보호(local differential privacy)에 대한 개념과 사례를 대신 소개한다.

현지 차등정보보호는 Dwork (2006)이 제안한 차등정보보호와 거의 개념적으로 동일하나, 보호자료 생성함수 κ를 자료 수집 후가 아닌 자료 수집 중에 이용한다는 차이점을 가지고 있다. 즉, 차등정보보호의 개념에서는 자료 제공자로부터 수집한 원자료를 공공정보 수집/제공 기관이 보호자료 생성함수 κ로 정보보호를 한 후, 정보 이용자에게 제공한다. 하지만, 현지 차등정보보호 개념에서는 정보 수집 기관(기업)이 곧 정보 이용자로서, 자료 제공자가 정보를 제공하는 순간 보호자료 생성함수 κ가 작용하여, 기업의 데이터베이스에는 이미 잡음이 추가된 형태로 저장된다.

이러한 현지 차등정보보호 방법은 실제로 Google사의 Chrome browser에 적용되어서, Chrome 사용자들의 '예/아니오'에 대한 응답은 잡음이 추가된 상태로 Google사의 데이터베이스에 저장된다. 그리고, Google사 내부의 정보 이용자들은 사용자들이 '예'라고 응답한 실제 비율이 아닌, 그 비율에 대한 추정량과 분산추정량을 자신들의 분석에 이용한다. 이러한 Google사의 현지 차등정보보호 시스템은 몇 가지 한계점이 있는데, 그 중 첫째는 이산형 자료에 대해서만 적용될 수 있다는 점이고, 둘째는 실제 비율의 단순 추정량 이외에 시스템 분석에 필요한 통계 분석들, 예를 들어서 로지스틱 회귀분석 등에 적용되지 못한다는 점이다. 이를 개선하기 위해 Nguyen 등 (2016)은 Harmony 라고 명명된 현지 차등정보보호 알고리즘을 제안하였으며, 해당 알고리즘이 삼성 핸드폰 사용자의 사용 데이터 정보를 안전하게 보관하면서도, 시스템 오류 등을 위한 필수 통계분석을 비교적 정확하게 할 수 있다는 것을 실험을 통해 보였다.

5. 재현자료(synthetic data)

5.1. 기본 개념

차등정보보호 기법과 함께 최근 가장 많은 주목을 받고 있는 방법은 원자료로부터 추정된 결합밀도함수에서 재현자료(synthetic data)를 생성하는 것이다. 기존 정보보호기법과 비교하여, 재현자료 기법에 의해서 생성된 보호자료는, 원자료와는 다른 개별값을 가지고 있지만 원자료 전체의 확률분포를 비교적 정밀하게 보존한다. 이러한 특징에 기인하여, 본 저자는 synthetic data를 재현(再現)자료로 번역하여 사용하는 것을 제안한다.

방법론적 측면에서 재현자료 기법은 크게 두가지 주제에 대한 연구가 주로 진행되었다. 첫 번째는, 앞에서도 서술한 것처럼, 추정된 원자료의 결합밀도함수에서 재현값(synthetic value)을 생성하는 기법에 관한 것이고, 두 번째는 원자료 대신 재현자료를 사용함으로써 발생되는 정보의 불확실성이 사용자의 최종 분석에 자동으로 반영되게 하는 이론적 근거를 제시하는 것이다.

5.1.1. 재현자료 생성 응답자 i로부터 수집된 p변수에 대한 응답값을 $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip}) \overset{iid}{\sim} p(\boldsymbol{y}_i|\theta)$로 정의하고, 이로부터 생성될 재현값을 $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})$로 정의한다. 전체 n명의 응답자로부터 수집된 원자료는 $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, 재현자료는 $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$로 정의한다.

부분 재현자료(paritally synthetic data)의 생성은 크게 다음과 같은 단계로 요약할 수 있다.

1. 응답값들 $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$이 동일한 우도 함수를 따른다고 가정하고, 모형 모수에 대한 사전분포를 설정한 후, 베이지안 추정 기법에 의하여 사후분포 $p(\theta|\boldsymbol{Y})$를 얻는다.

   만약 자료이용자가 원하는 최종 결과물이 베이지안 분석이라면, 사전분포가 proper prior라는 것을 확인하면 된다. 만약 자료이용자가 재현자료를 통하여 (빈도론의) 신뢰구간을 얻고자 한다면, Rubin (1987)이 제안한 combining rule을 만족하는 사전분포를 사용하여야 한다. 이 때 요구되는 조건은 비교적 약해서, 관측치의 숫자가 크거나 사전분포가 weakly informative할 경우, 재현자료로부터 명목수준에 근접한 신뢰구간을 구할 수 있다 (Rubin, 1984, 1987; Rubin과 Schenker, 1987).
2. 원자료 노출에 의한 비용이 크다고 판단되는 변수 j^*를 선정한다. 경우에 따라서 특정 변수 대신 특정 응답자 i^*의 모든 응답값을 정보보호의 대상으로 결정할 수도 있다.
3. 모든 응답자의 항목 중 노출위험이 큰 변수에 대한 재현값을 예측분포에서 생성한다. 즉,

$$z_{ij^*} \sim \int p\left(z_{ij^*}|\boldsymbol{y}_{i,-j^*}, \theta\right) p(\theta|\boldsymbol{Y})\mathrm{d}\theta, \quad i = 1, \ldots, n,$$

여기서 $\boldsymbol{y}_{i,-j^*}$은 응답자 i의 응답값 중 변수 j^*를 제외한 나머지 응답값들을 가리킨다. 생성된 재현값을 이용하여, 응답자 i에 대한 재현자료 $\boldsymbol{z}_i = (z_{ij^*}, \boldsymbol{y}_{i,-j^*})$를 생성한다.

재현자료의 유용성은 1단계와 3단계에서의 통계적 추정의 정확도에 달려있다. 재현자료에 대한 초기 연구들은 주로 모수적 모형추정방법과 순차회귀모형(sequential regression modeling) (Raghunathan 등, 2001)을 이용한 방법들을 제안하였다. 하지만, 이러한 방법들은 모수적 모형에 대한 가정이나, 회귀분석에 포함되는 변수 등에 결과가 민감하게 반응하는 한계점을 지니고 있다. 이를 극복하기 위해, 최근에는 원자료에 대한 결합밀도함수를 비모수모형으로 추정한 후, 이로부터 재현값을 생성하는 비모수적 방법들이 제안되었다 (Drechsler와 Reiter, 2011).

노출위험의 최소화에 중점을 둬야 하는 상황에서는 2단계에서 전체 원자료, 즉, 모든 응답자($i = 1, \ldots, n$)의 모든 변수($j = 1, \ldots, p$)를 재현 대상으로 삼을 필요가 있다. 이처럼 전체 자료를 재현값으로 대체하는 경우를 fully synthetic data(완전 재현자료), 특정 항목/응답자만을 재현하는 경우를 partially synthetic data(부분 재현자료)라고 부른다. 완전 재현자료의 경우, 3단계 작업에서 다음과 같이 모든 변수에 대한 재현값을 생성한다.

$$\boldsymbol{z}_i \sim \int f(\boldsymbol{z}_i|\theta)p(\theta|\boldsymbol{Y})\mathrm{d}\theta, \quad i = 1, \ldots, n.$$

완전 재현자료의 개념, 자료 유용성 및 노출위험 측정은 Rubin (1993), Raghunathan 등 (2003), Reiter (2005), Drechsler 등 (2008) 등에서 자세히 소개되어져 있다.

5.1.2. 재현자료 사용에 대한 불확실성의 측정 앞 절에서 서술하였듯, 재현자료는 원자료의 결합분포와 가까운 자료를 제공한다. 이에 더하여, 재현자료의 또 다른 유용성은, 원자료 대신 재현자료를 사용함으로써 발생하는 정보의 불확실성이 사용자의 최종 분석 결과에 자동으로 반영된다는 것이다.

Reiter (2003b)는 무응답 자료 처리에 사용되는 다중대체법(multiple imputation)에서 착안하여, 다음과 같이 부분 재현자료의 불확실성을 측정하는 방법을 제안하였다.

1. 한 개의 원자료 $\boldsymbol{Y}$에 대하여 복수의 재현자료 $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(m)}$를 예측분포에서 생성해낸다.
2. 이용자가 추정하고자 하였던 모수 θ 또는 모수의 함수를 원자료 대신 각각의 재현자료를 이용하여 추정한다. 이 때, 각각의 재현자료에 대하여 점추정값 $\hat{\theta}^{(l)}$과 분산추정값 $\hat{V}(\hat{\theta}^{(l)})$을 계산한다. 예를 들어서, 이용자가 원자료를 이용하여 단순회귀분석 $y_i = \alpha + \beta x_i + \varepsilon_i$을 하고자 하였다면, m개의 재현자료에 대하여 각각 회귀분석을 실시하여, $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(m)}$와 회귀계수 추정량들의 분산 추정값을 m개 계산한다.
3. 위의 다중 계산값들을 이용하여, 최종 점추정값과 분산 추정값을 다음의 공식에 의해서 계산해낸다.

$$\hat{\theta} = \frac{1}{m}\sum_{l=1}^{m}\hat{\theta}^{(l)}, \tag{5.1}$$

$$\hat{V}(\hat{\theta}) = \frac{\sum_{l=1}^{m}\hat{V}\left(\hat{\theta}^{(l)}\right)}{m} + \frac{\sum_{l=1}^{m}\left(\hat{\theta}^{(l)} - \hat{\theta}\right)^2}{m(m-1)}. \tag{5.2}$$

Reiter (2003b)는 식 (5.1)의 $\hat{\theta}$이 근사적으로 정규분포를 따름을 증명하였다. 즉,

$$\hat{\theta} \sim N\left(\theta, \mathrm{E}\left[\hat{V}\left(\hat{\theta}\right)\right]\right).$$

식 (5.2)에서 첫 번째 항은 표본추출에 의한 불확실성을, 두 번째 항은 재현자료를 사용함으로써 추가로 발생하는 불확실성을 측정한다.

위에서 제안한 분산 추정 공식은 사용자로 하여금, 표집오차 뿐 아니라 정보보호 과정에서 발생하는 오차를 최종 분석 결과에 반영함으로써, 올바른 분산 추정량(혹은 표준오차)을 사용할 수 있도록 해준다. 특히, 제안된 분산 추정 공식은 사용자가 원하는 대부분의 일반적 통계 분석 방법(예. 회귀분석 등)에 자동으로 적용 가능하기에, 활용성 측면에서 장점이 있다. 재현자료의 불확실성 추정 공식과 제반 논의에 대해 더 관심있는 독자들에게 Reiter (2005), Reiter와 Raghunathan (2007)를 추천한다.

5.2. 평가

재현자료 방법이 처음 소개되었을 때, 국내외적으로 용어에서 오는 부정적 어감(synthetic data, 한국에서는 이전 논문들에서 인위자료 혹은 합성데이터로 번역되었음)으로 인해, 정보의 유용성 측면에 대한 우려의 목소리가 있었다. 특히, 원자료와는 전혀 상관없는 '인위'적인 자료를 '합성'하여 대중에 공표하는 것으로 인식되어지는 경우도 많았다. 하지만 이러한 일반적 인식과는 반대로 재현자료 기법은, 정보유용성과 정보보호 능력, 양측 모두에서 전통적 정보보호 기법들에 비해 우수하다는 점이 계속된 모의실험과 실증 연구를 통하여 확인되었다 (Raghunathan 등, 2003; Reiter, 2005; Drechsler, 2012; Kim 등, 2015). 또한 앞에서도 서술하였듯이, 이용자가 원자료에 적용하려던 분석 방법을 그대로 재현자료에 적용한 후, 마지막에 분산추정 공식만을 계산하며 되기에, 분석 방법의 용이성 역시 장점으로 거론된다 (Matthews와 Harel, 2011).

몇몇 통계 기관에서는 시범적으로, 이용자가 재현자료를 이용한 분석 결과와 기관이 원자료를 동일한 방법으로 분석한 결과를 비교하여 이용자에게 알려주고 있다 (Drechsler, 2012). 물론 이러한 서비스는 업무 부담 및 노출위험의 문제로 영구히 제공될 수는 없겠지만, 재현자료에 대한 이용자들의 신뢰가 확보될 때까지는 당분간 운용할 계획인 것으로 알려져 있다.

차등정보보호와 재현자료는 컴퓨터 공학 분야와 통계 분야에서 각각 독립적으로 발전되었고, 이에 따라 전체적인 목적 등에서 큰 차이를 나타낸다. 가장 뚜렷한 차이 중 하나는 차등정보보호 방법들이 주로 노출위험의 최소화에 초점을 맞추는데 비해, 재현자료 방법은 자료의 유용성 확보에 더 비중을 둔다는 점이다. 최근에는 몇몇 연구팀에 의하여 차등정보보호 조건을 만족하는 재현자료를 구현하는 방법에 대한 연구가 소개되었다. 관심있는 독자들에게 Machanavajjhala 등 (2008), Abowd와 Vilhuber (2008), McClure와 Reiter (2012)를 추천한다.

5.3. 활용 사례

미국 Census Bureau는 코넬 대학교 연구팀과 공동으로, Survey of Income and Program Participation(SIPP)에 대한 부분 재현자료를 제공하고 있다 (Abowd 등, 2006; Reeder 등, 2015).

조사 자료 이용시 흔히 겪게 되는 어려움은, 소득이나 사회 보장 혜택 여부와 같은 민감한 질문에 대해서, 응답자들이 응답하지 않거나 왜곡된 응답값을 제공한다는 것이다. 이에 반해, 국세청 등이 보유하고 있는 과세 자료에는 비교적 정확하게 개인의 소득 상황이 파악되어 있다. 이러한 동기에 기인하여, Census Bureau는 SIPP의 자료와 과세 표준 자료인 Social Security Administration(SSA)/Internal Revenue Service(IRS)의 자료를 결합한 데이터셋을 만들었다. 하지만 이 결합자료는 민감한 소득 자료의 노출에 대한 위험(속성 위험)과 특정 개인을 찾을 수 있는 식별 위험을 동시에 증가시킬 가능성이 매우 커서, 결합된 원자료 형태로는 공공에게 제공되지 않고 있다. 이러한 노출위험들을 줄이면서도, 유용한 자료를 공공이 이용하게하기 위한 목적으로, SSB라고 불리우는 재현자료 방법을 개발하게 되었다.

SSB의 기본 재현 과정은 5.1.1절에서 소개한 바와 같다. 즉, 원자료의 결합분포를 추정한 후, 추정된 모형으로부터 성별과 기혼 여부를 제외한 다른 모든 항목들에 대한 재현값을 생성해낸다. 현재 SSB에서

는 16개의 다중 재현자료셋 $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(16)}$을 제공하고 있으며, 이용자가 각각의 재현자료로부터 얻어진 점추정값과 분산 추정값을 입력하면, 자동으로 식 (5.1), (5.2)의 최종 점추정값과 분산 추정값을 계산해주는 분석 도구도 함께 제공하고 있다. 또한, 버전이 업데이트 될 때 마다 최근 자료를 추가하여, 현재 공표되어 있는 SSB Version 6.0에서는 1984년부터 2008년까지의 SIPP 자료에 대한 재현자료를 제공하고 있다.

노출위험에 대한 염려를 낮추기 위해, Census Bureau에서는 SSB 공표로 인한 특정 개인의 식별 위험이 충분히 낮음을 확인하는 작업을 공표 전에 반드시 거치고 있다. 한편 정보유용성을 담보하기 위한 방편으로, 이용자들이 SSB로 이용한 분석결과를 논문 등에 개재하기 전에, 재현자료의 분석 결과가 원자료로부터 얻은 결과와 비슷한지 Census Bureau에게 확인할 것을 강력하게 권고하고 있다.

위의 소개된 SSB 사례 이외에, 재현자료 기법은 미국의 경우, 종단면경제활동자료(LBD, Kinney와 Reiter, 2007), 종단면 고용가구 동적자료(LEHDD, Abowd 등, 2004) 등에 적용되고 있다. 타국가의 사례로는, Drechsler 등 (2008)이 독일 고용 통계에 대한 재현자료를 생성하였다.

6. 마침

최근 통계적 정보보호에 대한 연구가 활발히 이루어져 다수의 논문들이 조사방법/공공자료 학회지(Survey Methodology, Journal of Official Statistics, 등), 정보보호 학회지(Privacy in Statistical Databases, Transactions on Data Privacy, Journal of Privacy and Confidentiality, 등), 그리고 컴퓨터 공학 학회지를 중심으로 발표되고 있다. 본 논문에서는 지면 관계상 일부 내용을 개괄적으로 소개할 수 밖에 없었음을 밝힌다. 또한, 보다 자세한 내용에 관심있는 독자들에게 도움이 되었으면 하는 바람으로, 다소 많은 수의 참고문헌들을 포함하였음을 밝힌다.

본 논문에서는 마이크로데이터의 통계적 정보보호 방안에 대한 최근 방법론들을 실제 활용 사례들과 더불어 소개하였다. 정보보호 방법론은 실제로 자료를 공표하는데 사용되어야해서 기관이나 회사에서 구현할 수 있어야 하고, 실무진에게 이론적으로 쉽게 설명되며 이해하기에 명확해야 한다. 매스킹 기법들은 이론적으로는 비교적 쉬우나, 유용성과 안전성을 동시에 갖춘 자료를 만드는데 한계를 가질 뿐만 아니라 최적안을 판단하는 기준도 명확하게 세우기 어렵다. 시스템적 대안은 일단 만들어 놓으면 대용량 자료를 공표하기에도 편리하지만, 자료 종류 및 분석 기법별로 상황을 나누어 시스템을 설계하는 과정이 매우 복잡하며 매스킹의 단점을 온전히 극복하기에는 한계를 가진다. 차등정보보호 방안은 공격자의 전략이나 외부 자료들의 종류에 상관없이 특정 수준 이하의 노출위험을 보장한다는 점에서 안전하지만, 분석 활용도가 떨어진다는 단점이 있다. 재현자료 방법은 안전성과 자료 유용성을 비교적 균형있게 확보하지만, 한 개의 원자료에 대해서 다수의 재현자료를 공표하는 것에 대한 이용자의 혼란 및 정서적 거부감이 단점이다.

과거 매크로데이터 위주의 자료를 공공에게 제공하던 통계 기관들은 이제는 마이크로데이터를 폭넓게 제공하기 위해 노력하고 있으며, 나아가 전수 자료나 빅데이터와 같은 대용량 자료를 공급해야 하는 시대를 맞이하였다. 크기 문제로 대용량 자료를 기존의 마이크로데이터처럼 다운로드 방식으로 공급하는 것이 효율적이지 않으므로, 자료를 바로 분석하는 시스템의 활용에 대한 연구가 더욱 필요하다. 다만 서버에 대한 지속적인 공격을 방어하는 문제, 노출위험 제어 및 자료 유용성 확보를 이루는 세부적인 대안이 아직은 만족스럽게 제시되어 있지는 않다. 대용량 자료를 안전하고 쓸모있게 그리고 이용하기 편리하게 공급하기 위해서, 자료 제공 기관들은 통일된 공표 기준과 효율적인 방법론 확보를 위해 계속해서 노력해야 할 것이다.

끝으로 통계 기관에서 정보보호 방법을 고려할 시, 강조되어야 할 두 가지 사안에 대해 언급하며 글을

맺으려 한다. 첫째로 공공으로부터 수집된 어떠한 정보도 외부에 제공되지 않을때, 통계 기관은 완벽한 정보보호를 달성할 수 있다. 역으로 이야기하면, 연구와 정책 결정 등의 공익을 위해서 자료가 공공에게 제공될 경우에는, 프라이버시 침해의 위험은 어쩔 수 없이 존재한다. 따라서, 특정 개인에 대한 노출위험과 사회적 비용이 일정 수준을 넘지 않는 한, 통계 기관이 수집한 자료가 공공에게 적극적으로 제공되는 것이 공익를 증가시키는 길이다. 둘째로 모든 이용자의 분석 목적을 만족시키려는 공공자료는 결국 노출위험을 크게 증가시킨다. 이에 따라 많은 정보보호 관련 연구들은, 지나치게 작은 범위나 특정 대상을 목적으로 자료를 분석하려는 이용자를 자료 공격자로 분류한다. 예를 들어, 경제활동자료에서 가장 높은 소득을 가진 특정인에 대한 분석을 목적으로 할 경우, 그것이 비록 학술적 연구나 공익을 추구하는 언론활동 등이어도, 공공정보 제공 기관의 입장에서는 이러한 이용자를 자료 공격자로 분류하여야 한다. 따라서, 원자료와 보호자료의 분포가 꼬리 부분에서 동일한 모양을 가지고 있을 경우, 이 보호자료는 유용성이 높다고 평가되기보다는, 노출위험이 지나치게 높아 위험한 상태로 평가되어야 한다. 다시 말해서, 보호자료의 유용성 척도로서 회귀계수 추정량의 보존은 고려대상이 될 수 있으나, 99% 분위수 등은 유용성의 측도로 사용되어져서는 안된다.

References

Abowd, J. M., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project, *Technical Report*, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program.

Abowd, J. M. and Vilhuber, L. (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygın (Eds), *Privacy in Statistical Databases* (pp. 239–246), Springer-Verlag Berlin, Heidelberg.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data, In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (Eds), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (pp. 215–277), North-Holland, Amsterdam.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data, In *Privacy in Statistical Databases* (pp. 290–297), Springer Berlin, Heidelberg.

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

Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005). Practical privacy: The SuLQ framework, In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 128–138), Association for Computing Machinery, New York.

Chipperfield, J. and Yu, F. (2011). *Protecting confidentiality in a remote analysis server for tabulation and analysis of data*, Paper presented at the October 2011 UNECE Work Session on Statistical Data Confidentiality.

Drechsler, J. (2012). New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey, *Journal of Applied Statistics*, **39**, 243–265.

Drechsler, J., Bender, S., and Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel, *Transactions on Data Privacy*, **1**, 1002–1050.

Drechsler, J. and Reiter, J. P. (2009). Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB Establishment Survey, *Journal of Official Statistics*, **25**, 589–603.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, *Computational Statistics and Data Analysis*, **55**, 3232–3243.

Duncan, G. T., Elliot, M., and Gonzalez J. J. S. (2011). *Statistical confidentiality: principles and practice*, Springer.

Duncan, G. and Lambert, D. (1989). The risk of disclosure for microdata, *Journal of Business & Economic Statistics*, **7**, 207–217.

Dwork, C. (2006). Differential Privacy, In *Inference Control in Statistical Databases* (pp. 1–12), Springer, Berlin, Heidelberg.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitive in private data analysis, In *Proceedings of the 3rd Theory of Cryptography Conference* (pp. 265–284), Springer, New York.

Dwork, C. and Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn, *Journal of Privacy and Confidentiality*, **1**, 135–154.

Franconi, L. and Polettini, S. (2004). Individual risk estimation in μ-Argus: a review, In *Privacy in Statistical Databases* (pp. 262–272), Springer, New York.

Jeong, D. M. and Jeong, M. (2008). A method of masking for 2005 Korean Census microdata, *Korean Journal of Applied Statistics*, **21**, 313–325.

Jeong, D. M. and Kang, D. H. (2006). *Disclosure control methods to increase microdata usage* (the original title is written in Korean), Daejeon, Korea.

Jeong, D. M., Kim, J. J., and Kim, K. M. (2009). A method of masking based on multiplicative noise, *Korean Journal of Applied Statistics*, **22**, 141–151.

Karr, A. F., Kohnen, C. N., Oganian, A. Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician*, **60**, 1–9.

Kim, H. J., Karr, A. F., and Reiter, J. P. (2015). Statistical disclosure limitation in the presence of edit rules, *Journal of Official Statistics*, **31**, 1–18

Kim, K., Lee, E., and Jeong, M. (2007). *A case study on the overseas release system of microdata*, Statistical Research Institute.

Kim, K.-S. (2009). Release of microdata and statistical disclosure control techniques, *Communications for Statistical Applications and Methods*, **16**, 1–11.

Kim, K. Y., Kwon, D. H., Shin, J. E., and Lee. S. H. (2011). *Introduction to Statistical Disclosure Control* (the original title is written in Korean), Freeacademy, Gyeonggi-do.

Kim, Y.-W., Kim, T.-Y., and Ki, K.-N. (2011). Application of a statistical disclosure control techniques based on multiplicative noise, *Korean Journal of Applied Statistics*, **24**, 127–136.

Kinney, S. K. and Reiter, J. P. (2007). Making public use, synthetic files of the Longitudinal Business Database, In *Proceedings of the Joint Statistical Meetings*, American Statistical Association, Alexandria, VA.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: the synthetic longitudinal business database, *International Statistical Review*, **79**, 363–384.

Krenzke, T., Gentleman, J. F., Li, J. and Moriarity, C. (2013). Addressing disclosure concerns and analysis demands in a Real-Time Online Analytic System, *Journal of Official Statistics*, **29**, 99–124.

Lee, Y. (2013). Review on statistical methods for protecting privacy and measuring risk of disclosure when releasing information for public use, *Journal of the Korean Data and Information Science Society*, **24**, 1029–1041.

Lee, Y. H. and Kim, Y. D. (2011). *Statistical disclosure control for EduData* (the original title is written in Korean), Korea Eduation & Research Information Service, Daegu, Korea.

Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M., and Freiman, M. (2011). The current stage of the microdata analysis system at the U.S. Census Bureau, In *Proceedings of the 58th World Statistical Congress of the International Statistical Institute*.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: theory meets practice on the map, In *IEEE 24th International Conference on Data Engineering*, 277–286.

Manrique-Vallier, D. and Reiter, J. (2012). Estimating identification disclosure risk using mixed membership models, *Journal of the American Statistical Association*, **107**, 1385–1394.

Matthews, G. J. and Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for accessing privacy, *Statistics Surveys*, **5**, 1–29

McClure, D. and Reiter, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data, *Transactions on Data Privacy*, **5**, 535–552.

Meindl, B., Templ, M., and Kowarik, A. (2013). *Guidelines for the Anonymization of Microdata Using R-package sdcMicro*.

Muralidhar, K., O'Keefe, C. M. and Sarathy, R. (2013). A general methodology for masking output from remote analysis systems, Paper presented at the October 2013 UNECE Work Session on Statistical

Data Confidentiality.

Nguyen, T. T., Xiao, X., Yang, Y., Hui, S. C., Shin, H., and Shin, J. (2016). Collecting and analyzing data from smart device users with local differential privacy, *arXiv:1606.05052v1, cs.DB*.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis, In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, 75–84.

Park, M. J. (2014). *Evaluation of microdata masking approaches with Survey of Household Finances and Living Conditions* (the original title is written in Korean), Statistical Research Institute, Daejeon.

Park, M. J., Kwon, S. P., and Shim, K. H. (2013). *Microdata masking for Survey of Household Finances and Living Conditions* (the original title is written in Korean), Statistical Research Institute, Daejeon.

Park, W.-H. (2004). Disclosure limitation techniques for statistical tables and microdata, *Journal of The Korean Official Statistics*, **9**, 146–172.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, **27**, 85–95.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, **19**, 1–16.

Reeder, L. B., Stinson, M., Trageser, K. E., and Vilhuber, L. (2015). *Codebook for the SIPP Synthetic Beta 6.0.2.*, Cornell Institute for Social and Economic Research and Labor Dynamics Institute, Cornell University, Ithaca, NY.

Reiter, J. P. (2003a). Model diagnostics for remote-access regression servers. *Statistics and Computing*, **13**, 371–380.

Reiter, J. P. (2003b). Inference for partially synthetic, public use microdata sets, *Survey Methodology*, **29**, 181–188.

Reiter, J. P. (2004). New approaches to data dissemination: a glimpse into the future, *Chance*, **17**, 12–16.

Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study, *Journal of the Royal Statistical Society, Series A*, **168**, 185–205.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation, *Journal of the American Statistical Association*, **102**, 1462–1471.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician, *The Annals of Statistics*, **12**, 1151–1172.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, NJ.

Rubin, D. B. (1993). Statistical disclosure limitation, *Journal of Official Statistics*, **9**, 461–468.

Rubin, D. B. and Schenker, N. (1987). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association*, **81**, 366–374.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata, *Journal of Official Statistics*, **14**, 361–371.

Skinner, C. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models, *Journal of the American Statistical Association*, **103**, 989–1001.

Statistics Netherlands (2007). μ-Argus User's manual, 4.1 version.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, **10**, 571–588.

Templ, M. (2008). Statistical disclosure control for microdata using the R-package sdcMicro, *Transactions on Data Privacy*, **1**, 67–85.

Templ, M. and Meindl, B. (2008). Robustification of microdata masking methods and the comparison with existing method, *Privacy in Statistical Database*, Springer, **5262**, 177–189.

Wasserman, L. and Zhou, S. (2012). A statistical framework for differential privacy, *Journal of the American Statistical Association*, **105**, 375–389.

Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation, *The Journal of Privacy and Confidentiality*, **1**, 111–124.

마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰

박민정[a] · 김항준[b,1]

[a]통계청 통계개발원, [b]Department of Mathematical Sciences, University of Cincinnati

(2016년 8월 31일 접수, 2016년 10월 9일 수정, 2016년 10월 9일 채택)

요 약

학술 연구나 정책 입안 등을 위한 심층적 자료 활용의 확대는 동시에 개별 정보 노출에 대한 염려도 증가시킨다. 때문에 최근 이십여 년 간 통계적 노출제어(정보보호) 분야에서 많은 논문들이 발표되었다. 본 논문은 그러한 연구 내용들을 정리하여 국내 통계인들과 기관들에게 소개하고자 한다. 주요 내용으로 국소통합이나 잡음추가와 같은 전통적인 매스킹 기법 뿐만 아니라, 온라인 자료 분석 시스템에서의 정보보호 처리, 차등정보보호를 통한 노출제어 및 재현자료를 활용한 정보보호 대안 모색에 대해 다룬다. 또한 각각의 주제에 대한 방법론 소개와 함께 활용 사례 및 장단점을 논의하였다. 본 논문이 실제적인 통계적 노출제어 문제를 고민하는 통계인들에게 도움이 되기를 바란다.

주요용어: 비밀보호, 매스킹, 분석 시스템, 차등정보보호, 재현자료

[1]교신저자: Department of Mathematical Sciences, University of Cincinnati, PO Box 210025, Cincinnati, OH 45221, USA. E-mail: hang.kim@uc.edu