제4장

가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가

박민정

제1절 서 론

마이크로데이터를 공표할 때 개인정보의 노출을 제한하기 위한 통계기관의 노력은 다음 <표 4-1>에서와 같이 물리적 제한과 통계적 제한으로 나누어 생각해 볼 수 있다. 마이크로데이터의 모든 정보를 이용하고자 하는 심층 이용자를 대상으로는 보통 비밀유지를 위한 법적 계약이나 자료 입출에 대한 물리적 통제를 통해 노출제한이 이루어진다. 반면 보다 범용적인 자료 배포를 위해서는 다수 국가에서 통계적으로 노출이 제한된 공공이용파일을 작성하여 활용하고 있다.

〈표 4-1〉 마이크로데이터를 위한 노출제한의 종류와 제공되는 자료의 형식(박민정 등, 2013)

방향	물리적 제한		통계적 제한(SDL)	
이용자	심층 이용자		불특정 다수	
기법	결과통제	접근제한	매스킹처리	인위자료
자료형식	가공 통계표 원격접속	인가파일 데이터 실험실	공공이용파일	

SDL: Statistical Disclosure Limitation[1)]

1) 본 연구에서는 각종 문헌의 비밀보호(protecting confidentiality), 노출제어(disclosure control), 노출제한(disclosure limitation)을 동일한 의미로 취급하도록 한다.

지금까지 통계개발원에서는 다양한 노출제한 방안들 중 전통적인 매스킹 기법들을 활용하여 공공이용파일을 작성하는 방안을 연구해왔다. 매스킹 기법의 활용과 관련된 통계개발원 마이크로데이터 비밀보호 주요 연구 현황을 살펴보면 다음 <표 4-2>와 같다.

〈표 4-2〉 통계개발원의 마이크로데이터 비밀보호 연구 현황(박민정 등, 2013)

연구 시기		대상 자료	개념 소개	주요 방법론	측정 기준
I	2006년	인총 2%표본(충남)	식별, 유용성	그룹화	노출위험
	2007년	인총 2%표본(전국)	노출위험	재코딩	노출위험
	2007년	가계조사	민감정보	그룹화, 반올림	자료유용성
	2008년	가계조사	민감정보	승법잡음	자료유용성
II	2012년	가계금융・복지조사	R-U map	재코딩	노출위험
	2013년	해외연구동향	인위자료 활용	베이지안 기법	-
	2013년	가계금융・복지조사	-	국소통합, 잡음	위험-유용성

먼저 2006~8년에 인총 및 가계조사 자료를 대상으로 그룹화, 재코딩, 잡음추가 등의 몇몇 매스킹(masking) 기법들을 연구하였다. 이 연구들은 비밀보호의 중요성, 개념, 세부 기법을 소개한 의의를 가지기도 한다. 그 외에 표에 기록되지는 않았지만 각국 통계작성 기관의 비밀보호 현황에 관한 보고서들도 여러 편 작성되었고, 주요 물리적 노출 제한에 관한 각국의 현황도 정리되어 있다.

이어 2012년에 가계금융·복지조사 자료를 대상으로 통계적 비밀보호에 관한 연구가 다시 시작되었다(김경미와 임경은, 2012). 이 연구에서는 키변수들에 재코딩 등의 매스킹 기법들을 적용한 후에 유일성에 근거한 노출위험을 측정하였다. 이 노출위험을 시도변수의 유무에 따라 비교하여 시도변수를 공표범위에 추가했을 때 노출위험 증가가 현저함을 나타내었다. 또한 여러 노출위험에 대한 측도들과 위험-유용성 지도(R-U map) 개념을 소개하였다.

한편 가계금융·복지조사는 패널 형식으로 시작된 조사이므로, 특정 시점의 자료뿐만 아니라 종단 자료에 대한 노출제한 연구가 필요하다. 이를 위해 2013년 관련 해외연구동향 검토가 이루어졌고(박민정과 김경미, 2013), 주로 인위자료 활용에 관한 논문들과 해외 통계기관의 종단자료 노출제한 사례들을 검토하고 향후 연구 방향을 모색하였다. 그 결과 종단 자료의 노출제한 문제는 주로 인위자료를 활용하여 학계에서 연구가 진행 중이며, 이를 통계기관에서 활용한 사례는 있으나 아직 많은 어려움들이 있다고 판단된다. 따라서 현 시점에서 통계개발원에서 실제 적용을 위해 검토하기에는 적절치 않아 보이며, 인위자료 활용 및 종단 자료 비밀보호 연구는 향후 과제로 두기로 하였다.

이제 가계금융·복지조사를 포함한 마이크로데이터의 통계적 노출제한은 횡단 자료에

대해 현실적으로 전통적인 매스킹 기법들을 활용하여 공공이용파일을 작성하는 일련의 과정을 통해 접근할 수 있다는 결론을 얻게 된다. 이 과정은 각 자료의 제약 조건들을 감안한 매스킹 기법들 활용 방안 구축, 그리고 노출위험과 자료유용성 측정을 통한 방안별 비교의 두 단계로 나누어 생각할 수 있다. 아래 그림은 이러한 과정을 보여주고 있다. 다양한 매스킹 처리를 통해 M개의 비밀보호 처리된 자료들을 만드는 것이 첫 번째 단계이고, 이들 중 공표자료를 선택하기 위한 비교 작업들이 두 번째 단계에 해당된다.

[그림 4-1] 마이크로데이터 매스킹 절차

이 중 첫 번째 단계인 매스킹 기법들의 활용 방안 구축에 대한 연구가 2013년에 이루어졌다(박민정 등, 2013). 매스킹 기법들은 그 종류가 많으며 각 기법들을 적용하기 위해서는 별도의 프로그램 작성이 필요하다. 2006~8년에 개별 매스킹 기법을 중심으로 연구가 진행된 것은 개별 프로그램 작성에 많은 시간이 필요했기 때문이기도 하다고 볼 수 있다. 과거에도 해외의 마이크로데이터 매스킹에 관한 프로그램으로 네델란드 통계청의 μ-Argus가 있었으나, 새로운 프로그램을 입수하여 설치하는 불편함이나 구현된 매스킹 기법의 범위가 넓지 않은 문제점이 있었다. 이후 μ-Argus보다 많은 매스킹 기법들이 구현된 R패키지인 sdcMicro가 오스트리아 통계청을 중심으로 개발되었고, 최근 매뉴얼도 배포되었다[2]. 이를 기반으로 박민정 등(2013)에서는 실제 자료에 다양한 매스킹 기법들을

2) 다만 통계소프트웨어 R을 사용할 때는 R이 오픈소스라는 특성상 모든 명령어가 철저히 검증되지 않았을 수 있다는 단점을 감안하여야 한다. 특히 매스킹 기법 중 하나인 자료교환(swapping)에서는 sdcMicro에서 오류가 발견되어 자료교환 기법을 적용하기 위해서는 별도의 프로그래밍 작업이 필요하다.

대부분 적용하여 볼 수 있었다. 이 sdcMicro 패키지에는 노출위험과 정보손실(자료유용성) 측도도 구현되어 있어, 이를 통해 여러 매스킹 결과들을 위험-유용성 지도(R-U map) 위에서 비교하여 볼 수 있다. 이상은 특정 매스킹 기법을 이해하고 적용하는데 국한되었던 기존 연구와 달리, 다양한 매스킹 기법들을 종합적으로 적용하여 검토해 본 의의를 가진다.

본 연구에서는 매스킹 처리를 통한 마이크로데이터 노출제한 과정의 두 번째 단계를 다루고자 한다. 박민정 등(2013)에서 논의되었듯이 노출위험과 정보손실은 상충관계(trade-off)가 있어서 각 방안들을 R-U map 위에 표현하였을 때 서로 효용이 동등할 가능성이 있는 방안들이 존재할 수 있다. 때문에 R-U map만으로 매스킹 방안들 사이의 우월성을 논하기 어렵다. 뿐만 아니라 노출위험과 정보손실 측도들도 그 종류가 많아 이들을 추가로 검토할 필요도 있다. 그 외에 비밀보호 처리된 공표할 마이크로데이터를 최종적으로 결정하기 위해서는 각종 통계적 특성이 매스킹 처리 전후 어떻게 달라지는지에 관한 심층적인 검토도 필요하다. 일단 공표된 자료는 되돌리기 어려우므로 통계적 비밀보호로 특정 매스킹 방안을 선택할 때는 다양한 측면의 검증이 필요하기 때문이다.

노출위험이나 자료유용성(정보손실) 측도는 매스킹 기법을 적용한 결과를 평가하는데 사용된다. 역으로 생각하면 이러한 측도를 만족시키는 방향으로 매스킹 작업이 이루어지게 된다고 할 수 있다. 따라서 노출위험 및 정보손실 측도에 대한 검토는 매우 중요한 문제가 된다. 어떠한 최적화 알고리즘이 존재하여 하나의 매스킹 적용 방안이 도출되면 좋겠으나, 현실적으로는 가능한 모든 매스킹 방안들을 적용하여 노출위험 및 정보손실을 각각 얻고 이를 비교해야 하는 실정이다. 어느 쪽이든 매스킹 방안 선택을 위해 노출위험 및 정보손실 측도에 대한 이해는 상당히 중요하다고 할 수 있다.

본 연구에서는 2절과 3절에서 노출위험 및 자료유용성을 측정하는 각 방법들을 심층적으로 살펴본 후, 4절에서는 이들 중 실제 자료에 적용 가능한 측도들을 이용하여 이루어진 노출제한 결과를 검토하도록 한다.

제2절 노출위험 측정 방법론

노출위험은 보통 통계작성기관의 경험에 의거한 점검 목록을 체크하거나, 단순한 자료의 요약 통계량들을 이용하거나, 자료에 어떤 식별 테스트를 하는 방법들에 경험적으로(ad hoc) 의존해왔다(Skinner와 Shlomo, 2008). 하지만 노출위험은 분명한 통계적 원칙을 따라 평가되어야 하며, 그러한 노력들이 학계와 각국 통계기관들에서 꾸준히 이루어져 왔다. 대표적으로 미국의 경제활동인구조사(Current Population Survey) 자료를 위해서 Duncan과 Lambert(1989)의 방법을 적용한 연구가 있으며(Reiter, 2005), 영국에서는

로그 선형모형을(Skinner와 Shlomo, 2008), 네델란드 통계청에서는 포아송-감마 모형을 마이크로데이터의 비밀보호를 위해 적용했고(Bethlehem 등, 1990), 독일에서는 두 개의 마이크로데이터 파일에 판별분석을 적용하였다(Paass, 1988)고 알려져 있다. 이러한 노출위험 측정은 크게 ①키변수 유일성에 근거한 노출위험 측정과 ②민감변수까지 고려한 침입자 의사결정론을 이용한 노출위험 측정이라는 두 부류로 나누어 생각해 볼 수 있는데, 이 절에서는 각각에 대해 주요 노출위험 측정 방법론들을 살펴보도록 하겠다.

1. 유일성 기반 노출위험

기존의 마이크로데이터 비밀보호 관련 국내 연구에서는 유일성 기반 노출위험을 다음과 같은 정의에 따라 다루고 있다(이용희와 김용대, 2011; 김경미와 임경은, 2012). 먼저 모집단에서 유일한 경우라는 좁은 의미의 노출위험은 다음의 3가지 조건을 모두 만족하는 경우에 발생한다고 본다.

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

이때 노출위험은 확률적 모형으로 표현될 수 있고, 어떤 사람 A가 모집단에서 유일하고[3] 표본에도 있을 확률은 $\Pr[(A \in U_s) \cap (A \in U_p)]$ 이며 이를 위해 정의되는 기호들은 아래와 같다.

A : 관심의 대상인 사람
S_1 : 통계작성기관의 마이크로자료로 구성된 파일 1
S_2 : 외부인(intruder)에 의해 구성된 파일 2
U_p : 모집단에서 유일한 개체들의 모임
U_s : 표본에서 유일한 개체들의 모임

이 확률은 외부인이 자신이 가지고 있는 파일 S_2에 A가 포함된 것을 모를 경우와 알 경우로 나누어 계산할 수 있는데, 조건부 확률공식과 각 사건들의 독립성을 고려하면 노출위험은 다음과 같이 정리할 수 있다(이용희와 김용대, 2011).

3) 따라서 A는 표본에서도 유일하다.

가정		A에 대한 노출위험
$A \in S_2$ 인지 모를 경우	$A \in S_1$와 $A \in S_2$이 서로 독립	$\Pr(A \in S_1)\Pr(A \in S_2)\Pr(A \in U_p)$
	$A \in S_1$와 $A \in S_2$이 종속	$\Pr(A \in S_2)\Pr(A \in U_p)$
$A \in S_2$ 인 경우		$\Pr(A \in S_1)\Pr(A \in U_p)$

보수적으로 외부인이 A에 대하여 안다고($A \in S_2$) 가정하면 노출위험은 A가 표본에 포함되는 확률 $\Pr(A \in S_1)$과 모집단에서 유일한 확률 $\Pr(A \in U_p)$에 의해 결정된다. 표본으로 추출되는 확률 $\Pr(A \in S_1)$은 조사 단계에서 정해져 있으므로, 노출위험을 측정하는 문제는 $\Pr(A \in U_p)$를 어떻게 결정할 것인가 하는 문제와 직결된다고 할 수 있다. 만약 모집단의 정보를 가지고 있거나 이를 대체할 적절한 큰 표본이 있을 경우 $\Pr(A \in U_p)$를 비교적 쉽게 추정할 수도 있다. 예를 들어 김경미와 임경은(2012)에서와 같이 인구주택총조사 10% 표본 자료를 모집단으로 삼아 주어진 변수들의 조합이 유일한 값을 취하는 경우를 세어 $\Pr(A \in U_p)$를 추정할 수 있다. 그러나 보통의 경우 모집단에 대한 정보가 부족하여 $\Pr(A \in U_p)$를 알아내기는 어렵기 때문에 이를 추정하는 문제가 학계에서 연구되어 왔다. 즉 주어진 표본을 가지고 모집단에서 유일한 개체수를 추정하는 모형들에 관한 여러 연구들이 있다. 이제부터 모집단에서 특정 조건을 가지는 개체수의 추정과 노출위험 측정에 관한 주요 모형들을 정리하도록 한다.

가. 포아송-감마 모형을 이용한 노출위험 추정

유일성을 기반으로 노출위험을 측정할 때는 먼저 자료의 형태를 바꾸어서 생각할 필요가 있다. 예를 들어 Bethlehm 등(1990)에 나오는 마이크로데이터는 4개의 변수, H(가구구성), A(나이), M(혼인), S(성별)로 이루어진 n=8,399개의 레코드로 이루어져 있고, 각 변수 H, A, M, S는 각각 24, 14, 2, 2개의 범주를 가지고 있다. 이를 일반적인 마이크로데이터 형식으로 표현하면 아래와 같은 행렬 형식으로 표현할 수 있다.

ID	H	A	M	S
1	20	8	1	2
2	3	5	2	1
...	...	...	...	...
8399	17	2	1	2

그러나 유일성에 관한 문제를 다루기 위해서는 마이크로데이터를 키변수들에 의한 분할표로 바꾸어 이해하는 것이 필요하다. 예를 들어, M, S 2개의 키변수에 대해서는 n개의

자료를 2차원 표로 나타낼 수 있고, 이 때 어떤 n_{ij}의 값이 1이면 마이크로데이터는 해당 조건에 대해 유일한 개체를 가지는 것이 된다. M과 S는 각각 두 개의 범주를 가지므로 이 표에서 셀의 개수는 4개가 된다.

	M=1	M=2
S=1	n_{11}	n_{12}
S=2	n_{12}	n_{22}

(M,S)	(1,1)	(1,2)	(2,1)	(2,2)
n	n_{11}	n_{12}	n_{12}	n_{22}

셀의 개수는 키변수 조합의 개수를 나타내므로 위의 두 번째 표와 같은 형식으로 다르게 표현해 볼 수 있고, 변수의 수를 늘려 키변수로 H, A, M, S라는 4개 변수 모두를 사용한다면 다음 표와 같이 표현된다. 4개 변수에 대해서는 셀이 모두 K=1,108개 생기며, 이 중 어느 한 셀의 값이 1이면 해당 키변수 조합을 가지는 개체는 유일한 개체가 된다. 즉, 키변수의 개수가 늘어나면 셀의 수도 기하학적으로 증가하여 성긴 분할표(sparse table)가 되며, 어느 셀에 해당하는 레코드가 하나만 존재하는 일이 더욱 쉽게 일어날 수 있다.

(H,A,M,S)	1 = (1,1,1,1)	2 = (1,1,1,2)	...	K = (24,10,2,2)
n	$n_{1,1,1,1}$	$n_{1,1,1,2}$	...	$n_{24,10,2,2}$

이렇게 키변수 조합에 의한 분할표로 자료를 표현하면, 이제 노출위험 측정을 위해 포아송-감마 모형을 적용해 볼 수 있다. 모집단의 레코드 수는 N, 주어진 키변수들로 생성되는 셀의 개수는 K, i번째 셀에 대한 모집단의 모수(superpopulation parameter) 및 빈도수를 각각 π_i 및 F_i로 나타내자. 이제 F_i에 대해 기댓값 $\mu_i = N\pi_i$ 을 가지는 포아송 모형[4]을 가정한다. 그러면 모집단에서 유일한 레코드들의 수 U_p는 $U_p = \sum_{i=1}^{K} \mu_i e^{-\mu_i}$ 가 되고, 이를 추정하기 위해서는 이제 각 모수를 추정하면 된다.

각 모수를 추정하기 위해 모수 π_i을 감마 분포 $\Gamma(\alpha,\beta)$을 따르는 확률변수 Π_i의 실현값(realization)이라 하면,

$$F_i \sim Poisson(N\pi_i) | \pi_i = \Pi_i, \quad \Pi_i \sim \Gamma(\alpha,\beta)$$

4) 포아송 모형의 확률밀도함수는 $f(k) = \frac{\mu^k}{k!} e^{-\mu}$이다.

으로 정리되는 포아송-감마 모형을 얻는다. 이때 수리 통계학적으로 F_i의 분포(marginal distribution)는 음이항 분포가 되므로, 모집단에서 유일한 레코드들의 수는

$$U_p = K\Pr(F_i = 1) = N(1+\beta)^{-(1+N\alpha)}$$

로 나타낼 수 있다. 이제 적률 추정량(moment estimator)을 초기값으로 하여 최대우도 추정량(maximum likelihood estimator)을 계산하면 각 모수의 추정값을 얻고 따라서 U_p의 추정량을 계산할 수 있게 된다.

한편, 크기 N_j인 부분모집단의 i번째 셀의 빈도수 F_{ij}에 대하여 음이항 분포 $NB(N_j, \alpha_j, \beta_j)$를 가정할 수 있다.[5] 그러면 전체 모집단의 유일한 레코드의 수에 대한 추정량은 $\hat{U}_p = \sum_j \hat{U}_{pj}$로 계산할 수 있다. 보통 전체 모집단보다는 시도 등 지역단위 자료에 대한 수요가 많으므로 이러한 부분모집단에서 유일한 레코드 개수의 추정도 중요한 의의를 가진다.

이러한 U_p의 추정을 노출위험을 일정 수준 이하로 제어하는데 이용할 수 있다. 첫째 표본의 유일한 레코드의 수를 일정한 값 이하로($\hat{U}_{ps} < C_a$) 유지하거나, 둘째 표본과 모집단에서 유일한 레코드의 비율($U_{ps}/n = U_p/N$)을 일정한 값 이하로($\hat{U}_p/N < C_r$) 유지하는지를 그 기준으로 삼을 수 있다. 이를 통해 노출제한을 위한 최소한의 모집단 개수(critical population size)를 얻어 지역별 공표 기준 설정에 참고할 수 있다.

포아송-감마 모형의 효과를 살펴보기 위해 위에 언급하였던 Bethlehm 등(1990)에 나오는 예에 대한 결과를 살펴보면 다음과 같다. 추정 결과를 확인하기 위해 정보가 없는 모집단 대신에, 실제 유일성 개수를 알 수 있는 표본에 대해 모형을 적용하였다. 다음 표를 살펴보면 포아송-감마 모형에 의한 추정은 실제 결과를 과소추정(underestimate)하는 경향이 있음을 볼 수 있다. 또한 유일성 비율을 0.1%를 기준으로 한 모집단 최소 크기는 4개 변수에 대해 46,228임을 알 수 있다. 즉 이 기준에서 4개 변수를 안전하게 공표하기 위해서는 46,228보다 작은 인구수를 가지는 지역단위를 포함해서는 안 된다는 결론을 얻을 수 있다. 더 자세한 내용은 Bethlehem 등(1990)을 참고하라.

키변수	키변수 조합 수 (셀의 수)	표본의 유일성 개수		모집단 최소 크기
		관측 결과	추정 결과	
H	24	0	.08	1,411
H×A	288	23	21.9	14,116
H×A×M	554	50	37.7	29,252
H×A×M×S	1,108	108	80.1	46,228

5) 일반적으로 부분모집단별로 모형을 가정하는 것이 $NB(N, \alpha, \beta)$를 사용할 때보다 적합성 측면에서 더 낫다고 할 수 있기 때문이다. 또한 모형을 좀 더 단순화하기 위해 $NB(N_j, \alpha, \beta)$을 이용하기도 한다.

나. 로그 선형 모형을 이용한 노출위험 추정

주어진 키변수 조합이 만들어내는 교차 분할표의 K개 셀에 대해서 F_k를 k번째 셀의 모집단 개체수, f_k를 표본 개체수라고 하자. 포아송-감마 모형에서는 모집단에서 유일한 개체수를 추정하여 노출위험 측도로 삼았었다. 반면 로그 선형 모형에서는 표본에서 유일한 개체가 모집단에서 유일할 확률이나, 모집단 개체수 역수의 기댓값을 이용해 노출위험 측도를 다음과 같이 정의한다(Skinner와 Shlomo, 2008).

$$r_{1k} = P(F_k = 1 | f_k = 1)$$

$$r_{2k} = E[1/F_k | f_k = 1]$$

두 노출위험 측도는 레코드 수준의 측도가 되고, 파일 수준의 노출위험은 이들의 합인 $\tau_1^* = \sum_{SU} r_{1k}$ 및 $\tau_2^* = \sum_{SU} r_{2k}$ $(SU = \{k : f_k = 1\})$을 이용해 얻도록 한다.

로그 선형 모형은 우선 K개 각 셀에 대해서 모집단 개체수 F_k가 평균 λ_k인 포아송 모형을 따른다고 가정한다. 또 각 개체가 표본에 포함되는 것은 알려진 확률 π_k의 베르누이 분포에 의한다고 가정한다. 그러면 k번째 셀의 표본의 개체수 역시 포아송 분포를 따라 $f_k \sim P(\pi_k \lambda_k)$ 이게 된다. 여기서 k번째 셀의 모집단 개체수는 표본의 개체수와 표본에 포함되지 않은 모집단 개체수의 합이므로 $F_k | f_k \sim P[\lambda_k(1-\pi_k)] + f_k$ 라고 생각할 수 있다. 이러한 포아송 분포 가정을 이용하여 두 노출위험 측도는 다음과 같이 정리된다.

$$r_{1k} = exp[-(1-\pi_k)\lambda_k]$$

$$r_{2k} = \{1 - exp[-(1-\pi_k)\lambda_k]\} / [(1-\pi_k)\lambda_k]$$

이제 모수 λ_k을 추정하기 위해 각 셀의 키변수 값들로 이루어진 $q \times 1$벡터 x_k에 대해 아래와 같이 로그 선형 모형을 적용한다.

$$log \lambda_k = x_k' \beta$$

각 k번째 셀에서의 표본의 개체수 f_k가 포아송 분포 $P(\pi_k \lambda_k)$의 실현값이므로 표본 자료를 이용해 최대우도추정량 $\hat{\beta}$을 구하고 모수 추정량 $\widehat{\lambda_k} = exp(x_k' \hat{\beta})$을 얻는다.

로그 선형 모형을 이용해 노출위험 측도를 추정하는 것은 여러 연구에서 이루어져 왔으나, Skinner와 Shlomo(2008)에서는 노출위험 측도에 대한 추론이 가능하도록 검정통계량들을 제시하였다. 또한 포아송-감마 모형이나 독립 포아송 모형을 가정하는 로그 선형 모형을 사용한 추정에서 공통적으로 관찰되는 과소추정 문제를 해결하기 위해, 검정

통계량들을 이용해 로그 선형 모형에서 변수 선택을 할 수 있다는 것을 보이기도 하였다.

지금까지 모집단 유일성에 근거한 노출위험 측도에 사용될 수 있는 포아송-감마 모형과 로그 선형 모형을 살펴보았다. 이러한 측도들을 이해하기 위해서는 먼저 행렬 형태의 마이크로데이터를 키변수 조합의 교차 분할표로 바꾸어 생각할 필요가 있었다. 키변수 조합에 의한 셀들로 자료를 바꾸어 바라보면 통계학에서 범주형 자료를 다룰 때 흔히 사용하는 포아송 분포를 가정하고 자료를 분석할 수 있으며, 각 셀에서 나타나는 빈도수에 적절한 모형을 적용하여 노출위험 측도들을 정의하고 계산할 수 있기 때문이다.

각 셀에 대한 기댓값을 추정할 때 모수에 대하여 감마분포를 가정하는 것이 포아송-감마 모형이며, 포아송-감마모형을 통해서는 전체 파일에 대해서 모집단 유일성 개수를 추정한다. 한편, 노출위험 측도를 특정 키변수 조합이 모집단에서 유일할 확률(혹은 모집단 개체수 역수의 기댓값)로 정의하고, 개별 레코드에 대해 노출위험을 계산할 수 있는 것은 로그 선형 모형이다. 이 경우 파일 수준의 노출위험은 개별 노출위험의 합으로 정의한다. 포아송-감마 모형은 전반적으로 모집단 유일성 개체수를 과소 추정하는 경향이 있으며, 로그 선형 모형 역시 독립 포아송 분포를 가정할 경우 노출위험을 과소 추정하는 경향이 있다.

이 절에서 다루기에는 학술적인 측면이 강하여 생략하였으나 검정 통계량 개발 및 복잡한 로그 선형 모형을 이용해 과소 추정 문제를 어느 정도 해결할 수 있으며(Skinner와 Shlomo, 2008), 이러한 복잡한 로그 선형 모형과 비교하여도 과소 추정 문제를 더욱 완화시킨 베이지안 GoM(Bayesian version of grade of membership) 모형도 연구되어 있다(Manrique-Vallier와 Reiter, 2012). 이들은 모두 미국 센서스 자료를 이용하여 제안한 노출위험 측도를 적용하고 평가하였다.

다. 유일성 기반 노출위험 측도의 구현

지금까지 여러 유일성 기반 노출위험 측도들을 살펴보았다. 이러한 측도들을 실제 자료에 적용하기 위해서는 프로그램으로 구현하는 과정이 필요하다.

유일성 기반 노출위험들 중 첫 번째로 논의되었던 $\Pr(A\in S_1)\Pr(A\in U_p)$에서 $\Pr(A\in U_p)$을 추정하기 위해, k번째 키변수 조합에서 나타나는 모집단의 빈도수 F_k에 관한 정보를 얻을 필요가 있다. 이를 위해, 먼저 기존 통계개발원의 연구처럼 전수조사인 인총자료 등을 이용해 이를 직접 계산해 볼 수 있다. 그러나 모든 조사 자료가 동일한 키변수를 포함하고 있는 전수조사를 가지고 있는 것은 아니어서, 이를 일반적으로 효율적이라고 하기는 어렵다. 다음으로 쉽게 생각할 수 있는 방법으로는 표본 조사에서 사용했던 가중값의 합을 이용해 F_k을 계산하는 것이 있으나, 이는 적절하지 않고 모형을 통해 추정하는 것이 바람직하다고 알려져 있다(Templ과 Mendl, 2010).

모형을 통해 F_k을 추정하는 것은 지금까지는 초모집단(superpopulation) 모형[6]들을 통해 이루어져왔다고 할 수 있다. 여기에는 포아송-감마 모형(Bethlehem 등, 1990), 디리슈렛-다항정규(Dirichlet-multinomial) 모형,[7] 음이항 모형,[8] 로그-선형 모형,[9] 다항정규(mulitnomial) 모형[10] 및 포아송-역 가우시안(Poisson-inverse Gaussian) 모형[11] 등이 있다고 알려져 있다. 본 연구에서는 초기의 포아송-감마 모형(Bethlehem 등, 1990)과 최신의 로그-선형 모형(Skinner와 Shlomo, 2008)을 설명하고, 베이지안 GoM 모형(Manrique-Vallier와 Reiter, 2012)을 소개하였다.

이 중에서 본 연구에서 사용하고 있는 sdcMicro 패키지는 음이항 모형을 이용하여 개별 레코드의 유일성 기반 노출위험을 계산하여 제공하고 있다. 이는 네델란드 통계청에서 만든 마이크로데이터의 매스킹 프로그램인 μ-Argus에 구현된 노출위험을 따라 동일하게 만들어진 것이다. 전체 파일 수준의 노출위험은 개별 노출위험의 합을 이용해 계산한다. 이 개별 노출위험의 합은 노출이 기대되는 레코드 수로 해석하여 제공되고 있다. 더불어 개별 노출위험의 평균값을 전체 파일 수준의 노출위험으로 활용할 수도 있다. 한편, sdcMicro에서는 로그 선형 모형을 이용하여 전체 파일의 노출위험을 계산하는 명령어를 제공하고 있기도 하지만, 이를 위해서는 선형 모형에 포함시킬 변수 선택에 대하여 각 마이크로데이터별로 비밀보호 담당자의 검토가 필요하다.

그외에 sdcMicro에서는 k익명성(k-anonymity), l다양성(l-diversity), SUDA(Special Uniques Detection Algorithm) 측도 등을 제공하여 키변수 조합에 따라 발생하는 유일성에 근거한 노출위험 측도로 참고할 수 있다. 이러한 노출위험 측도들은 유일한 레코드를 만들지 않기 위해 국소 감추기(local suppression) 매스킹 기법을 적용하고자 할 때 활용하기에 적당하다고 할 수 있다.

---

6) 초모집단(Superpopulation) 모형은 모집단이 어떤 초모집단에서 추출되었음을(drawn) 가정하며, 이는 F_k에 대한 특정한 분포를 가정하는 것을 의미한다.

7) Takemura, A. (1999) Statistical data protection Eurostat, Luxembourg, 45-58.

8) Benedetti, R. and Franconi, L. (1998) Statistical and technological solutions for controlled data dissemination. In: Pre-Proceedings of New Techniques and Technologies for Statistics, Sorrento, Italy, 225-570.
Franconi, L. and Polettini, S. (2004) Individual risk estimation in μ-Argus: A review. In: Privacy in Statistical Databases. Lecture Notes in Computer Science, vol. 3050, Springer, New York, 262-272.

9) 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. J. and Sholomo, N. (2006) Assessing identification risk in survey microdata using log linear models. In: S3RI Methodology Working Papers, M06/14, University of Southampton, Southampton Statistical Sciences Research Institute.

10) Forster, J. J. and Webb, E. L. (2007) Bayesian disclosure risk assessment: Predicting small frequencies in contingency tables. Journal of the Royal Statistical Society C, 56, 551-570.

11) Carlson, M. (2002) Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. Statistics in Transition, 5, 901-925.

2. 침입자 의사결정론 기반 노출위험

보통은 공표되는 마이크로데이터의 키변수들과 외부인이 가지고 있는 자료가 연결될 때 개인정보의 노출이 일어난다고 간주한다. 이러한 관점에서 본다면 공표되는 표본 자료 중, 모집단에서 유일한 개체가 있을 확률이나 그 수를 추정하여 노출위험의 측도로 삼는 것은 자연스럽다. 이는 키변수 조합에 따라 다르게 계산되는 노출위험이며, 앞 절에서 이를 계산하기 위한 모형들을 살펴보았다. 현실적으로 모집단 유일성에 근거한 노출위험을 제어하는 구체적인 방법은 공표되는 표본 자료에서 키변수 범위나 각 키변수 범주의 개수를 줄이거나 혹은 감추기 기법 등을 활용하여 유일한 개체의 수를 감소시키는 것이 된다.

이러한 모집단 유일성 기반 노출위험 측도의 한계는 다음과 같다(Reiter, 2005). 첫째, 외부인이 가지고 있는 자료의 특징을 반영하지 못한다. 예를 들어 외부인이 키변수 값들을 알고 있는 표적이 표본에 있고 표본에서 유일하다면, 이 레코드가 모집단에서 유일하지 않더라도 노출이 일어나게 된다. 둘째, 연속형 키변수 등 너무 많은 표본 유일성이나 모집단 유일성이 발생하는 자료들이 존재할 때 모집단 유일성 기반의 노출위험 측정은 별다른 의미를 가지지 못한다. 셋째, 유일성 기반 노출위험은 통계기관이 시행한 통계적 노출제한 처리의 효과를 적절히 판단하지 못한다. 예를 들어 연속형 키변수들에 잡음추가 처리를 하였다고 해도 그 변수들은 여전히 유일한 레코드가 되며, 노출제한 효과를 말하기 어려운 상황이 된다. 마지막으로 표본 추출 비율이 매우 낮은 경우, 유일성 기반 노출위험 추정은 정확성을 담보하기 더욱 어렵게 된다.

유일성 기반 노출위험 측도와는 개념적으로 완전히 다르게 Duncan과 Lambert(1989)에서는 노출위험의 측정을 의사결정이론(decision theory)을 이용하여 연구하였다. 원래 마이크로데이터를 Y라고 하고, 노출제한 처리 후 공표되는 마이크로데이터를 Z, 또 Z는 Y의 부분 집합이며 크기 n이라고 하자. 그러면 표적 t에 대한 정보 일부를 가지고 있는 외부인이 알고자 하는 것은 Z에서 t 위치를 파악해 t에 대한 나머지 정보를 알아내는 것이라 할 수 있다. 이를 위해 외부인은 표적이 공표자료에 있는지 없는지에 대한 판단을 하고 자료들을 서로 연결해야 한다. 이를 통계적 의사결정이론을 이용하여 설명하면 다음과 같다.

공표자료 Z에 있는 임의의 개체 s가 표적 t라고 예측하는 확률을 $p_Z(s)$라고 하고, s을 표적 t라고 판단할 때의 손실함수를 $L(t,s)$라고 하자. 그러면 외부인은 평균손실함수 $\int L(t,s)p_Z(s)ds$을 최소화하도록 Z 내의 어떤 s를 표적 t라고 선택할 것이다. 외부인의 손실이 0이 되는 경우는 표적이 공표자료에 없을 때 없다고 판단하거나 혹은

공표자료에 있는 표적을 바르게 연결했을 때이다. 반면에 (A) 표적이 공표자료에 있는데 표적이 없다고 판단할 때나 (B) 표적을 틀리게 연결했을 때는 손실이 발생한다. 이 손실들을 각각 l_1 및 l_2라고 하자. 그러면 외부인의 표적 선택에 대한 최소평균손실 즉 불확실성은 다음과 같다.

$$U(Y) = \min\left\{ l_1 \sum_{i=1}^{n} p(z_i),\ l_2[1 - \max_{1 \le i \le n} p(z_i)] \right\}$$

즉, 상황 (A)에 대해서는 표적이 공표자료에 포함되는 확률이 $\sum_{i=1}^{n} p(z_i)$이므로 최소평균손실은 $L_A = l_1 \sum_{i=1}^{n} p(z_i)$이 되고, 상황 (B)에 대해서는 공표자료에서 표적일 확률이 가장 큰 개체가 표적이 아닐 확률을 $[1 - \max_{1 \le i \le n} p(z_i)]$으로 표현할 수 있으므로 최소평균손실은 $L_B = l_2[1 - \max_{1 \le i \le n} p(z_i)]$이게 된다. 두 상황에 대하여 $L_A < L_B$[12]이면 외부인은 손실을 적게 감수하기 위해 공표자료에 표적이 없다고 판단할 것이고, 반대의 경우 공개된 자료 중에서 표적일 확률이 가장 높은 개체를 표적이라고 판단하게 될 것이다. 이상을 표로 정리하면 아래와 같다.

	외부인에게 손실이 발생하는 상황	
	(A) 표적이 공표자료에 있는데 표적이 없다고 판단	(B) 표적을 틀리게 연결
최소평균손실	$L_A = l_1 \sum_{i=1}^{n} p(z_i)$	$L_B = l_2[1 - \max_{1 \le i \le n} p(z_i)]$
의미	표적이 공표자료에 포함되는 확률과 손실 l_1의 곱	표적으로 추정한 개체가 표적이 아닐 확률과 손실 l_2의 곱

이러한 의사결정이론을 통해 노출위험을 측정하게 되면, 통계기관은 표적을 공표자료의 어떤 개체와 연결하는 것보다는 공표자료에 표적이 없다고 판단하는 것이 외부인에게 유리하게 하는 전략을 통해 노출위험을 줄이고자 하게 된다. 즉, $L_A < L_B$가 되도록 $\sum_{i=1}^{n} p(z_i)$ 및 $\max_{1 \le i \le n} p(z_i)$이 작아지는 방법을 적용하여 외부인의 정보 노출 의지를 약화시켜 노출위험을 낮출 수 있는 것이다.

12) 표적이 공표자료에 있는데 없다고 판단할 때 발생하는 손실이 표적이 공표자료에 있다고 판단하고 잘못 연결할 경우의 손실보다 작을 경우

Reiter(2005)에서는 미국의 경제활동인구조사(Current Population Survey) 마이크로데이터에 대해서 이러한 의사결정이론을 이용해 노출위험을 측정하였다. 마이크로데이터 각 변수들에 대해 재코딩, 자료교환 혹은 잡음추가 등의 매스킹 과정을 수행하고 각 개체에 대하여, 또 파일 전체에 대하여 노출위험을 측정한 결과를 보여주어 매스킹 기법 중 무엇을 선택하여 비밀보호를 할 것인가에 대한 의미 있는 측도를 제공하고 있다. 이에 관한 구체적인 내용을 정리하면 다음과 같다.

총 p개 변수를 가지는 행렬 형태의 원래 마이크로데이터를 Y라고 하면 j번째 레코드는 $\vec{y_j} = (y_{j1}, \ldots, y_{jp})$라고 표현할 수 있다. 또한 외부인이 정보를 가지고 있는 변수들로 이루어진 부분을 A(available), 외부인이 공표자료를 통해서만 정보를 얻을 수 있는 변수들로 이루어진 부분을 U(unavailable)라고 하면, $\vec{y}_j = (\vec{y}_j^A, \vec{y}_j^U)$라고 표현할 수 있다.

A와 U를 구분하는 외부인의 정보는 모든 개체에 대해서 동일하다고 가정한다.

통계기관은 Y를 그대로 공표하지 않고 매스킹 처리를 하여 공표하며 이를 Z라고 하자. A에서 매스킹 기법 중 잡음추가나 자료교환 등의 변조적 기법(stochastic perturbation methods)을 적용한 변수들로 이루어진 부분을 Ap, 이런 기법을 적용하지 않은 부분을 Ad라고 하면, $\vec{z}_j = (\vec{z}_j^A, \vec{z}_j^U) = (\vec{z}_j^{Ad}, \vec{z}_j^{Ap}, \vec{z}_j^U)$로 공표된 마이크로데이터의 j번째 레코드를 표현할 수 있다. 매스킹 처리 후 외부인의 정보가 부족한 부분들은 Ap와 U가 되며, 이들의 합집합을 앞으로 C라고 표현한다. 정리하면 공표자료 Z은 변조적 매스킹 기법 적용 여부 및 외부인의 정보력에 따라 $Z = A \sqcup U = Ad \sqcup Ap \sqcup U = Ad \sqcup C$의 여러 조합들로 생각할 수 있다.

<table>
<tr><td colspan="3">Z</td></tr>
<tr><td colspan="2">A</td><td>U</td></tr>
<tr><td>Ad</td><td>Ap</td><td>U</td></tr>
<tr><td>Ad</td><td colspan="2">C</td></tr>
</table>

외부인이 가지고 있는 표적은 $\vec{t} = (\vec{t}^{Ad}, \vec{t}^{Ap})$이며 이는 원래 마이크로데이터 Y의 어떤 j번째 개체 $\vec{y}_j^A$와 같은 것이라 가정할 수 있다. 외부인은 이제 매스킹 처리된 n개의 개체를 가지는 공표자료 Z에서 어떤 $\vec{z}_j^A$을 찾아 $\vec{t}$라고 판단하고 $\vec{z}_j^U$을 $\vec{t}$와 연결하려는 목적을 가지게 된다. 이제 J를 $\vec{t} = \vec{z}_j \in Z$일 때 j, $\vec{t} = \vec{z}_j \notin Z$일 때 $n+1$의 값을 취하는 확률 변수라고 하면, 외부인의 목표를 $\Pr(J = j | \vec{t}, Z), \quad j = 1, \ldots, n+1$을 계산하는 것

으로 표현할 수 있다. 외부인의 의사결정을 위해 필요한 확률 $\Pr(J=j|\vec{t},Z)$은 외부인 정보력에 대한 집합별로 베이즈 법칙에 의해 다음과 같이 정리된다.

$$\Pr(J=j)=\frac{\Pr(Z^C|J=j,\vec{t},Z^{Ad})\Pr(J=j|\vec{t},Z^{Ad})}{\sum_{j=1}^{n+1}\Pr(Z^C|J=j,\vec{t},Z^{Ad})\Pr(J=j|\vec{t},Z^{Ad})}$$

같은 방식으로 계산을 위해 위 식 분자의 첫 항을 j번째 개체 및 정보 집합별로 더 나누어 정리하면 다음과 같다.

$$\Pr(Z^C|J=j,\vec{t},Z^{Ad})=\Pr(Z^C_{-j}|\overrightarrow{z_j^C},J=j,\vec{t},Z^{Ad})$$
$$\times\Pr(\overrightarrow{z_j^U}|\overrightarrow{z_j^{Ap}},J=j,\vec{t},Z^{Ad})\times\Pr(\overrightarrow{z_j^{Ap}}|J=j,\vec{t},Z^{Ad})$$

이 때 $Z^C_{-j}=(\vec{z}^C_1,...,\vec{z}^C_{j-1},\vec{z}^C_{j+1},...,\vec{z}^C_n)$을 일컫는다.

결국 주어진 정보에 대해서 j번째 개체가 표적일 조건부 확률, $\Pr(J=j|\vec{t},Z)$은 외부인의 정보력에 따라 나누어 생각할 수 있는 다음 네 개의 확률을 각각 계산하여 얻을 수 있다. 이를 이제 각각 노출위험 성분 1, 2, 3, 4라고 부르도록 하겠다.

1. $\Pr(J=j|\vec{t},Z^{Ad})$
2. $\Pr(\vec{z}_j^{Ap}|J=j,\vec{t},Z^{Ad})$
3. $\Pr(\vec{z}_j^{U}|\vec{z}_j^{Ap},J=j,\vec{t},Z^{Ad})$
4. $\Pr(Z^C_{-j}|\vec{z}_j^{C},J=j,\vec{t},Z^{Ad})$

이제 노출위험 성분을 계산하는 세부 내용에 관하여 살펴보면 다음과 같다.

가. $\Pr(J=j|\vec{t},Z^{Ad})$의 계산

노출위험 성분 1은 외부인이 표적과 변조적으로 처리되지 않은 키변수들의 정보를 가지고 공표변수의 특정 개체가 표적과 일치할 확률을 추정하는 것으로 해석할 수 있다. 표적 $\vec{t}^{Ad}$와 같은 값을 가지는 $\vec{z}_j^{Ad}$들의 모집단에서 개수를 F_t, 표본에서 개수를 f_t라고

하면, 표적이 표본에 있다고 알려져 있는 경우는 $\Pr(J=j|\vec{t}, Z^{Ad}) = 1/f_t$, 표적이 표본에 있는지 모르는 경우는 $\Pr(J=j|\vec{t}, Z^{Ad}) = 1/F_t$이게 된다. 표적 $\vec{t}^{Ad}$와 같은 값을 가지지 않는 $\vec{z}_j^{Ad}$들에 대해서는 노출위험 성분 1은 당연히 0의 값을 가지게 된다.

예를 들어 공표자료에 외부인 자료와 공통인 변수로 나이(G), 인종(R), 성별(X) 및 수입(I)이 있고, 외부인이 가진 표적은 각 변수의 값이 57세 남성 아시아인으로 $125,000의 수입을 가진다고 하자. 또한 변수 (G, R, X)의 값이 (57세, 아시아인, 남성)인 개체가 11,000명, (G, R, X, I)의 값이 (57세, 아시아인, 남성, $100,000 이상)인 개체는 1,200명, (G, R, X, I)의 값이 (57세, 아시아인, 남성, $125,000)인 개체는 3명이 각각 모집단에 있다고 하자. 네 개의 변수를 모두 그대로 공표하는 경우 구하는 확률은 1/3이 되며, 수입(I) 변수를 상한 코딩하는 경우는 1/1200, 수입(I) 변수를 모두 감추는 경우는 1/11000의 확률 값을 가지게 된다. 즉, 각 매스킹 방안에 따라 노출위험을 구성하는 확률 성분에 대해 값을 구할 수 있게 된다.

보통 모집단에서 특정 변수 조합의 값을 가지는 개체수를 구하는 것이 가능한 경우는 많지 않아 표본 자료의 가중값을 이용하거나 1.에서 설명한 모형을 이용하여 F_t를 추정하게 된다. 이 경우 가중값이나 모형이 가지는 단점들을 그대로 가지게 되며, 보통은 보수적으로 표적이 표본에 있다는 것을 외부인이 안다고 가정하여 해당 노출위험 성분을 계산한다.

나. $\Pr(\vec{z}_j^{Ap}|J=j, \vec{t}, Z^{Ad})$의 계산

노출위험 성분 2는 외부인이 변조적으로 처리되지 않은 변수들 및 표적의 정보를 가지고 공표자료에서 j번째 개체의 변조적으로 처리된 변수들을 예측하는 확률로 해석할 수 있다. 그러면 외부인이 가지고 있는 표적 $\vec{t}$가 가지는 Ap에 속하는 변수들의 값을 근거로 역으로 노출위험 성분 2를 계산하게 된다. 예를 들어 Ad에 (G, R, X)가 포함되고 Ap에 잡음추가 처리된 I가 있다고 하자. 그러면 $\Pr(z_j^I|J=j, \vec{t}, Z^{Ad}) = N(z_j^I|t_j^I, \sigma_j^2)$이게 된다. 노출위험 성분 2의 계산은 변조적 매스킹 처리 방법에 따라 그 계산이 달라진다.

다. $\Pr(\vec{z}_j^{U}|\vec{z}_j^{Ap}, J=j, \vec{t}, Z^{Ad})$의 계산

노출위험 성분 3은 공표자료의 j번째 개체의 외부인과 공통으로 가지는 변수들 및 모든 개체의 변조적으로 처리되지 않은 변수들에 대한 정보를 기반으로 j번째 개체에 대해 외부인이 가지지 않은 변수들의 분포를 예측하는 것이라고 해석할 수 있다. 성분 3은 U에 속하는 변수들에 매스킹 처리를 하여 공표한다고 하면 다음과 같이 생각할 수 있다.

$$\Pr(\overrightarrow{z_j^U} \mid \overrightarrow{z_j^{Ap}}, J=j, \vec{t}, Z^{Ad})$$
$$= \int \Pr(\overrightarrow{z_j^U} \mid \overrightarrow{y_j^U}, \overrightarrow{z_j^{Ap}}, J=j, \vec{t}, Z^{Ad}) \Pr(\overrightarrow{y_j^U} \mid \overrightarrow{z_j^{Ap}}, J=j, \vec{t}, Z^{Ad}) d\overrightarrow{y_j^U}$$

주어진 정보를 가지고 $\vec{y}_j^U$의 분포를 예측하는 것은 회귀분석이나 시뮬레이션을 통해 이루어지며, $\vec{y}_j^U$까지 주어져 있을 때 $\vec{z}_j^U$의 분포는 적용하는 매스킹 기법에 따라 계산하게 된다. 이러한 과정은 복잡하며, 보수적으로 노출위험을 계산하기 위해 성분 3은 보통 1의 값을 가진다고 가정할 수 있다.

라. $\Pr(Z_{-j}^C \mid \vec{z}_j^C, J=j, \vec{t}, Z^{Ad})$의 계산

노출위험 성분 4는 변조적으로 처리되지 않은 변수들, 표적 및 j번째 개체의 공표된 값들에 대한 정보를 바탕으로 외부인이 가지지 않은 변수들에 대해 나머지 개체들의 분포를 예측하는 것으로 해석할 수 있다. 성분 3에서와 마찬가지로 성분 4는 다음과 같이 생각할 수 있다.

$$\int \Pr(Z_{-j}^C \mid \vec{y}^C, \vec{z}_j^C, J=j, \vec{t}, Z^{Ad}) \Pr(\vec{y}^C \mid \vec{z}_j^C, J=j, \vec{t}, Z^{Ad}) d\vec{y}^C$$

이 성분은 $Z_{-j}^C = (\vec{z}_1^C, ..., \vec{z}_{j-1}^C, \vec{z}_{j+1}^C, ..., \vec{z}_n^C)$에 대해 각 개체를 독립으로 가정하고, 성분 3에서와 마찬가지로 시뮬레이션과 매스킹 기법을 감안하여 계산한다.

지금까지 침입자의 의사결정 관점에서 통계적 의사결정이론을 이용해 제안된 노출위험 측도를 살펴보았다. 이는 유일성 기반 노출위험의 한계를 뛰어넘고 연속형 변수에 대해서도 매스킹 기법 적용의 효과를 평가할 수 있는 장점이 있다. 의사결정이론에 근거한 노출위험 측정의 의미를 보다 구체적으로 알 수 있도록 <부록 1>에 Reiter (2005)에 제시된 CPS 자료에 대한 노출위험 계산 결과를 정리하였으니 참고하기 바란다. 한편, 침입자 의사결정론 기반 노출위험은 매스킹 모형에 따라 노출위험의 계산이 달라지며 일반적인 프로그램을 작성하기는 현재로서는 어렵다고 판단된다. 따라서 본 연구에서는 침입자 의사결정론 기반 노출위험을 소개하는데 머무르고, 실제로 연속형 민감변수의 노출위험은 sdcMicro에 구현되어 있는 거리기반 노출위험을 활용하도록 한다. 이는 매스킹된 레코드마다 자료 순위에 근거한 구간을 만들고 해당 원자료의 값이 그 구간 내에 있는지 없는지를 보는 것으로, 전체 레코드 중 각 구간 내에 원자료가 있을 비율을 이용해 파일 단위의 노출위험을 계산하는 방식으로 이루어져 있다(Templ과 Meindl, 2008b). 참고로, 다음 <표 4-3>은 지금까지 언급된 노출위험 측도들의 특징을 정리한 것이다.

〈표 4-3〉 노출위험 측도 정리

1. 유일성 기반 노출위험 측도의 종류

기존 연구의 노출위험 측도	$\Pr(A \in S_1)\Pr(A \in U_p)$
모집단의 유일한 개체수 추정 (포아송-감마 모형)	$U_p = K\Pr(F_i = 1) = N(1+\beta)^{-(1+N\alpha)}$
모집단 유일성에 의한 각 개체의 노출위험 추정 (로그 선형 모형 → 베이지안 GoM으로 확장)	$r_{1k} = exp[-(1-\pi_k)\lambda_k]$ $r_{2k} = \{1-exp[-(1-\pi_k)\lambda_k]\}/[(1-\pi_k)\lambda_k]$ $log\lambda_k = x_k'\beta$, $r_{1k} = P(F_k = 1 \mid f_k = 1)$, $r_{2k} = E[1/F_k \mid f_k = 1]$
기타	k-anonymity l-diversity SUDA measure

2. 침입자 의사결정론 기반 노출위험 측도

자료 구분	Z A | U Ad | Ap | U Ad | C
특정 표적에 대하여 각 개체의 노출위험 추정 (어떤 개체가 표적일 확률)	$\Pr(J=j) = \dfrac{\Pr(Z^C \mid J=j, \vec{t}, Z^{Ad})\Pr(J=j \mid \vec{t}, Z^{Ad})}{\sum_{j=1}^{n+1}\Pr(Z^C \mid J=j, \vec{t}, Z^{Ad})\Pr(J=j \mid \vec{t}, Z^{Ad})}$
계산 과정	1. $\Pr(J=j \mid \vec{t}, Z^{Ad})$ 2. $\Pr(\vec{z}_j^{Ap} \mid J=j, \vec{t}, Z^{Ad})$ 3. $\Pr(\vec{z}_j^{U} \mid \vec{z}_j^{Ap}, J=j, \vec{t}, Z^{Ad})$ 4. $\Pr(Z_{-j}^{C} \mid \vec{z}_j^{C}, J=j, \vec{t}, Z^{Ad})$

주: 키변수 조합에 대해서는 개별 노출위험이 구해지나, 민감변수에 대해서는 구현되어 있는 개별 노출위험 측도가 없다(침입자 의사결정론 기반 노출위험 측도는 자료 별로 구현해야 한다).

제3절 자료유용성 측정 방법론

공공이용파일을 작성하여 배포할 때는 노출위험뿐만 아니라 자료의 유용성(정보손실)을 동시에 고려해야 한다. 그러나 노출위험을 낮추기 위해 여러 매스킹 기법들을 적용할수록 자료의 정보손실 역시 계속해서 증가하며, 이러한 노출위험과 정보손실 측도 사이의 상충관계(trade-off)가 언제나 존재한다. 그럼에도 불구하고 적절한 비밀보호 처리를 위해서는 두 측도 모두를 낮추어야 하기 때문에 적절한 해결 방안을 찾기가 쉽지 않은 측면이 있다.

Duncan 등(2001)에서는 노출위험과 유용성의 변화를 동시에 살펴볼 수 있도록 노출위험-유용성 지도를 제안하여 매스킹 방안 선택의 기준을 제시한 바 있다. 이러한 노출위험과 자료유용성을 동시에 고려하는 개념은 매우 적절히 제시된 것으로 이후 관련 연구들에서 꾸준히 활용되어 왔다. 한편, Duncan 등(2001)에서는 노출위험-유용성 지도와 함께 노출위험과 유용성 측도도 제안하였다. 제안된 노출위험은 외부인이 얻고자하는 통계량 추정값과 참값 사이의 최소제곱오차의 역수이며, 유용성은 매스킹된 자료가 가지는 모평균에 대한 최소제곱오차의 역수이다. 이는 노출위험과 유용성을 평균 중심의 단순한 통계량을 통해 정의한 것으로 자료 전체를 설명하기에는 부족한 점이 많다고 할 수 있다. 이후 많은 문헌들에 노출위험-유용성 지도에서 사용할 노출위험 및 정보손실 측도를 어떻게 정의할지에 관한 연구 결과들이 나타나 있으며, 그 중 노출위험에 관하여는 앞 절에서 관련 측도들이 어떻게 발전하였는지 주요 문헌 연구를 통해 다루었다. 참고로 지금까지 노출위험 측도에 대한 연구가 매우 활발하였던 반면 자료유용성 측도에 대한 연구는 그 양이 비교적 많지 않다고 할 수 있다. 이 절에서는 자료유용성 측도가 어떻게 정의되어 사용될 수 있는지 살펴보도록 하겠다.

자료유용성에 관해서 보통 추구되는 목표는 비밀보호 처리된 공표 자료가 원래 자료와 동일한 구조를 유지하도록 하는 것과 각종 통계 분석의 결과가 매스킹 이후에도 상당히 높은 정확도를 가지는 것이다. 먼저 매스킹 처리된 자료가 원래 자료와 얼마나 다른지를 평가하여 일반적인 정보손실의 측도로 삼으려는 노력이 꾸준히 이루어져 왔는데, 평균과 같은 일반적인 통계 추정값들을 비교하거나 매스킹 전후 고유값들 사이의 거리를 계산하는 방법 등이 여기에 속한다고 할 수 있다. 또한 각종 통계 분석을 통한 특정 추정값들의 신뢰구간을 비교하거나 전체 혹은 세부 영역에서 주요 통계량(benchmarking indicators)을 비교하는 것을 통해 정보손실을 측정하려는 노력도 있어 왔다. 이를 차례로 살펴보도록 하겠다.

1. 거리 기반 정보손실 측도

매스킹 처리된 자료가 원래 자료와 얼마나 다른지를 평가하기 위해 가장 먼저 생각할 수 있는 것은 기초 통계량들의 변화를 살펴보는 것이다. 변수 p개를 가지는 연속형 자료에 대해서는 자료값 행렬(X), 공분산 행렬(V), 상관계수 행렬(R), 각 변수와 주성분 사이의 상관계수 행렬(RF), 각 변수별 첫 번째 주성분의 설명 비율 벡터(C), 주성분 행렬(F) 등을 생각할 수 있다. 이러한 각 기초 통계량들을 원래 자료(X)와 매스킹 처리된 자료(X')에서 계산하고 두 결과 사이의 거리를 평균제곱오차(mean square error), 평균절대오차(mean absolute error), 평균변동(mean variation) 등을 통해 계산할 수 있다. 다음 표는 각 통계량 및 거리를 정리하여 보여주고 있다(Domingo-Ferrer와 Torra, 2001).

	Mean square error	Mean abs.error	Mean variation
$X-X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$V-V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$R-R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF-RF'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C-C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F-F'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

범주형 자료에 대해서는 연속형 자료와 같은 통계량들 대신 자료값, 교차표 비교, 엔트로피 기반 측도 등을 이용한다. 먼저 범주형 변수의 자료값을 직접 이용하여 거리를 측정할 때는 명목형 변수에 대해서는 매스킹 전후 값이 같으면 거리는 0, 다르면 거리는 1인 지시 함수를 사용한다. 반면 순서형 변수에 대해서는 두 값의 차이를 변수 범위로 나누는 방법 등을 이용해 거리를 정의한다. 다음으로 변수 조합에 따라 만들 수 있는 교차표의 빈도수를 가지고 매스킹 전후 거리를 측정하기도 한다. 마지막으로 잡음추가, 감추기, 재코딩 방법들을 일반화시킨 PRAM(Post-Randomization) 기법을 적용한 경우에는 엔트로피(Shannon's entropy)를 이용하여 거리를 측정하기도 한다(Domingo-Ferrer와 Torra, 2001).

그 외에도 매스킹 전후 자료 각각에 대하여 확률밀도 함수를 추정하여 아래와 같은 쿨백-라이블러 거리(Kullback-Liebler divergence)를 측정하여 자료유용성 측도로 삼을 수 있다.

$$d_{KL}(X', X) = \int log\left[\hat{f}_{X'} / \hat{f}_X\right] \hat{f}_{X'}$$

이러한 KL 측도는 자료가 다변량 정규분포를 따르지 않고, 변수의 개수가 많으면 계산하기 어려운 단점을 가진다(Karr 등, 2006).

이와 유사하게 경험적 분포를 이용한 유용성 측도도 연구되어 있다. N_X개 레코드를 가지는 자료 X의 p개 변수들에 대한 경험적 분포가 아래와 같고,

$$S_X(x_1, \dots, x_p) = \frac{1}{N_X} \sum_{i=1}^{N_X} I(x_{i1} \leq x_1, \dots, x_{ip} \leq x_p)$$

매스킹 전후 자료를 합한 자료를 $T = (X', X)$라고 할 때 다음과 같은 측도들을 생각해 볼 수 있다(Woo 등, 2009).

$$U_m = max_{1 \leq i \leq N_T} | S_X(\vec{t}_i) - S_{X'}(\vec{t}_i) |$$

$$U_s = \frac{1}{N_T} \sum_{i=1}^{N_T} | S_X(\vec{t}_i) - S_{X'}(\vec{t}_i) |^2$$

이는 두 경험적 분포 사이의 절대 차이의 최대값 혹은 차이 제곱의 평균을 측정하는 것이나, 분포의 차이를 감지하는데 변별력이 약한 것(low power)으로 알려져 있다.

2. 특정 통계량 기반 자료유용성 측도

매스킹 전후 자료의 유용성이 얼마나 변화하는가를 측정하는 또 다른 측도로 각종 통계분석을 시행한 결과가 매스킹 전후 얼마나 충실하게 유지되는지를 살펴보는 것을 고려할 수 있다. 자료 이용자가 임의의 통계 분석을 시행했을 때, 공표된 자료 Z가 원래 조사된 자료 Y와는 완전히 다른 결과를 만들어낸다면 어느 누구도 공표된 자료를 이용하기 원하지 않을 것이기 때문이다. 매스킹 전후 거의 동일한 통계 분석 결과를 만들어내야 한다는 관점에서, 관심의 대상이 되는 주요 통계량들을 매스킹 전후의 각 자료에서 얻고 이를 비교하는 것은 의미가 있다. 이와 관련된 측도들을 하나씩 정리하면 다음과 같다.

먼저 주요 통계량에 대한 신뢰구간 중복(confidence interval overlap, IO)을 살펴보자. 예를 들어 회귀분석을 시행하고 주요 통계량이 k번째 회귀계수일 때, 매스킹 전후 자료에 대해 k번째 회귀계수는 각각 β_k^Y 및 β_k^Z, 그리고 95% 신뢰구간은 각각 $(L_k^Y,\ U_k^Y)$ 및 $(L_k^Z,\ U_k^Z)$라고 하자. 이제 자료가 정규분포를 따를 때 회귀계수 통계량들은 t분포를 따른다고 볼 수 있고, 확률적인 신뢰구간 중복 IO을 p개 회귀계수에 대해 다음과 같이 정의할 수 있다(Karr 등, 2006).

$$I = \frac{1}{p}\sum_{k=1}^{p} I_k, \qquad I_k = \frac{1}{2}\left[\int_{L_k^Y}^{U_k^Y} f_k^Z(t)\,dt + \int_{L_k^Z}^{L_k^Z} f_k^Y(t)dt\right]$$

여기서 I_k의 첫째 항은 매스킹 후 자료 Z로부터 얻어진 β_k^Z의 사후분포에 매스킹 전 자료 Y에서 얻은 β_k^Y에 대한 신뢰구간 $(L_k^Y,\ U_k^Y)$을 적용할 때의 누적확률을 본다는 의미이며, 둘째 항은 그 반대의 경우에 해당한다. 두 회귀계수 β_k^Y, β_k^Z 및 사후분포들이 서로 유사하다면 두 항 모두 신뢰구간의 길이인 0.95에 가까운 값을 가지게 된다. 즉, I_k는 $[0, 0.95]$의 범위를 가지게 되며 두 신뢰구간이 완전히 겹치면 0.95의 값을, 전혀 겹치지 않으면 0의 값을 가지게 되어 회귀계수 통계량을 중심으로 한 자료유용성을 측정하여 보여주게 된다.

이 측도에 대한 이해를 돕기 위해 예를 들어 매스킹 전 자료에서 얻은 신뢰구간이 $(L_k^Y,\ U_k^Y) = (8, 10)$일 때, 매스킹 처리된 두 자료 Z_1 및 Z_2에 대해서 $(L_k^{Z_1},\ U_k^{Z_1}) = (-12, 30)$ 및 $(L_k^{Z_2},\ U_k^{Z_2}) = (3, 15)$라고 하자. 상식적으로 매스킹 방안 Z_2가 Z_1보다 자료유용성 측면에서 더 낫다고 할 수 있으며, I_k값 역시 동일한 결과를

제시한다. 만약 I_k의 두 번째 항만을 사용한다면 모두 1에 가까운 값을 가지게 되어 Z_1 및 Z_2 중에서 더 나은 매스킹 방안을 판단하기 어려운 것과 비교하여 I_k를 유용성 측도로 사용하기에 적절하다는 것을 알 수 있다.

한편, 측도 I을 사용할 때의 단점은 다음과 같다. 만약 매스킹 전후 신뢰구간들이 전혀 겹치지 않는다면 I_k의 첫째 항과 둘째 항 모두 0에 가까운 값을 가지게 되는데, 이때 각 매스킹 방안별로 얻어진 신뢰구간들의 적절성을 서로 구별하지 못하여 변별력이 떨어지게 된다. 즉, 매스킹 전후 자료에서 얻어진 통계량의 사후분포가 얼마나 멀리 떨어져 있든지 간에 I_k의 값은 모두 0에 가깝게 되어 각 매스킹 방안에 대한 비교가 어렵게 된다.

이를 보완하여 $U_k^{over} = min(U_k^Y, U_k^Z)$ 및 $L_k^{over} = max(L_k^Y, L_k^Z)$라고 하면 다음과 같은 측도를 생각할 수 있다(Karr 등, 2006).

$$J = \frac{1}{p}\sum_{k=1}^{p} J_k, \qquad J_k = \frac{1}{2}\left[\frac{U_k^{over} - L_k^{over}}{U_k^Y - L_k^Y} + \frac{U_k^{over} - L_k^{over}}{U_k^Z - L_k^Z}\right]$$

측도 J는 측도 I의 단점을 보완하여 만들어진 것이나, Fryzlewicz와 Oh(2011)에 제안된 두 시계열의 중복에 대한 측도인 Thick-Pen Measure of Association(TPMA)이 좀 더 적합한 측면이 있어 소개하고자 한다. TPMA를 신뢰구간에 맞게 정리하여 IO라 하면 아래와 같다.

$$IO = \frac{1}{p}\sum_{k=1}^{p} T_k, \qquad T_k = \frac{min(U^Y, U^Z) - max(L^Y, L^Z)}{max(U^Y, U^Z) - min(L^Y, L^Z)}$$

신뢰구간 중복에 대한 측도 T_k는 $(-1, 1]$의 범위를 가지며, 두 사후분포가 겹칠수록 1에, 겹치지 않고 멀어질수록 -1에 가까운 값을 가지게 된다.

지금까지 각 회귀계수에 대한 신뢰구간 중복을 비교하는 측도로 I_k, J_k 및 T_k을 소개하였다. 그러나 여러 회귀계수를 동시에 고려하여 신뢰구간의 중복을 측정하는 것을 생각해 볼 수 있다. 이를 타원형 신뢰구간 중복(Ellipsoid Overlap, EO)라고 하며 I_k와 비슷한 개념으로 정의 되나 아래와 같은 결합 사후분포를 고려한다.

$$\frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{p\hat{\sigma}^2} \leq F(\alpha; p, n-p)$$

여기서 β는 회귀계수들의 벡터이며, 매스킹 이전 자료 Y 및 $\hat{\beta}_Y$, $\hat{\sigma}^2_Y$을 이용해 추정한

사후분포에 대해 매스킹 이후 자료 Z로부터 얻는 회귀계수에 대한 타원형 신뢰구간(ellipsoid)의 확률을 계산하고, 또 그 반대의 계산을 하여 얻는 평균으로 I_k와 유사한 측도 값을 구할 수 있다. 다만 측도 EO을 계산한 결과를 얻기 위해서는 단계별로 몬테카를로 모의실험이 필요하다.

3. 기타 자료유용성 측도

매스킹 전후 자료들에 대해서 두 자료 사이의 거리나 개별 통계량에 근거하는 대신, 전체적인 변화를 측정하여 유용성 측도로 삼을 수도 있다. 매스킹 전후 자료 Y 및 Z를 합친 자료에서 성향 점수를 이용하거나, 군집 분석을 통해 측도를 고안하는 방법이 있는데 이 절에서는 이들을 간략히 살펴보자(Woo 등, 2009).

먼저, 성향점수(propensity score)를 이용한 유용성 측도는 다음과 같은 세 단계를 거쳐 얻어진다. ①매스킹 전후 자료를 합친 $T=(Y,Z)$을 만들고, ②로지스틱 회귀분석을 통해 $\vec{t}_i \in T$가 Z에 속할 확률을 성향점수 $\hat{p}_i$로 추정하고, ③ Y와 Z에 각각에 대해 성향점수의 분포를 비교한다. 단계 ③에서 성향점수의 분포를 비교할 때는 분위수 등을 비교할 수도 있으면 다음과 같은 간단한 통계량을 만들어 활용할 수도 있다.

$$U_P = \frac{1}{N_T}\sum_{i=1}^{N_T} |\hat{p_i} - c|^2, \qquad c = \frac{N_Z}{N_T}$$

다음으로 군집분석을 통한 유용성 측도를 소개한다. 자료 $T=(Y,Z)$을 임의의 G개 군집으로 나누는 분석을 수행하고 j번째 군집에 대한 가중값을 w_j, 개체수를 n_j, 개체들 중 매스킹 후 자료인 Y에 속한 것을 n_{Yj}라고 하면 아래와 같은 측도를 생각할 수 있다.

$$U_c = \frac{1}{G}\sum_{j=1}^{G} w_j \left[\frac{n_{Yj}}{n_j} - c\right]^2, \qquad c = \frac{N_Y}{N_T}$$

이러한 측도들은 앞에 언급한 다른 측도들과 함께 자료유용성을 평가하는데 참고할 수 있다. 마지막으로, 다음 <표 4-4>는 지금까지 다루어진 자료유용성 측도들을 정리한 것이다.

〈표 4-4〉 자료유용성 측도 정리

자료유용성(정보손실) 측도의 종류	
거리 기반 정보손실 측도	1. 자료행렬, 주성분 등을 활용한 거리 $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$, $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$ 등 2. Kullback-Liebler divergence $d_{KL}(X', X)=\int log\left[\hat{f}_{X'}/\hat{f}_X\right]\hat{f}_{X'}$ 3. 경험적 분포를 이용 $U_m = max_{1\le i\le N_T}\lvert S_X(\vec{t}_i)-S_{X'}(\vec{t}_i)\rvert$
특정 통계량의 신뢰구간 중복 측정	1. $I=\frac{1}{p}\sum_{k=1}^{p}I_k,\quad I_k=\frac{1}{2}\left[\int_{L_k^Y}^{U_k^Y}f_k^Z(t)\,dt+\int_{L_k^Z}^{L_k^Z}f_k^Y(t)dt\right]$ 2. $J=\frac{1}{p}\sum_{k=1}^{p}J_k,\quad J_k=\frac{1}{2}\left[\frac{U_k^{over}-L_k^{over}}{U_k^Y-L_k^Y}+\frac{U_k^{over}-L_k^{over}}{U_k^Z-L_k^Z}\right]$ 3. $IO=\frac{1}{p}\sum_{k=1}^{p}T_k,\quad T_k=\frac{min(U^Y,U^Z)-max(L^Y,L^Z)}{max(U^Y,U^Z)-min(L^Y,L^Z)}$
기타	1. 성향점수 이용 $U_P=\frac{1}{N_T}\sum_{i=1}^{N_T}\lvert\hat{p_i}-c\rvert^2,\quad c=\frac{N_Z}{N_T}$ 2. 군집분석 활용 $U_c=\frac{1}{G}\sum_{j=1}^{G}w_j\left[\frac{n_{Yj}}{n_j}-c\right]^2,\quad c=\frac{N_Y}{N_T}$

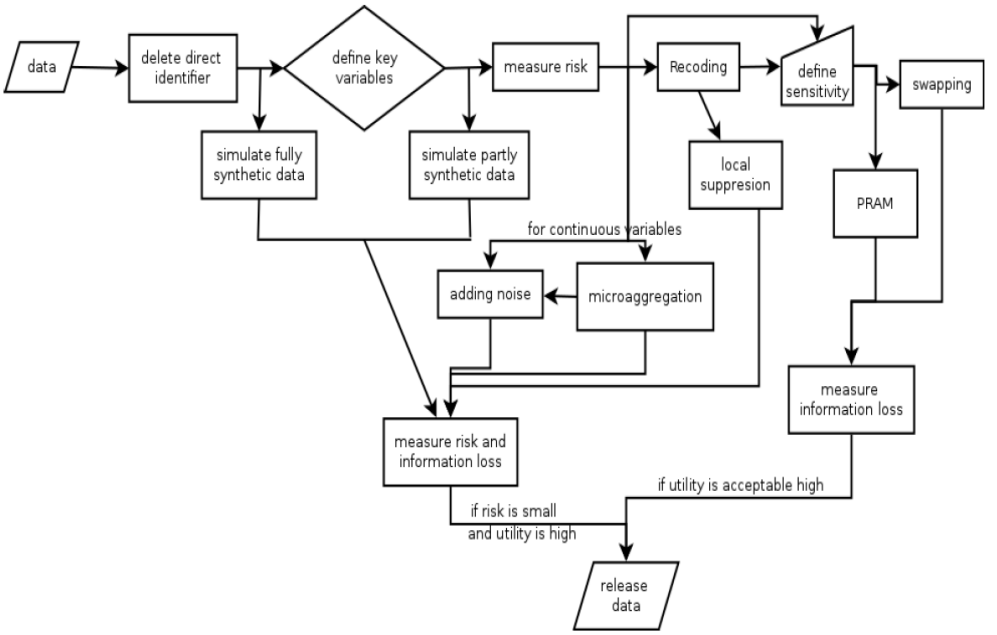
제4절 가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가

지금까지 국내 마이크로데이터 비밀보호 연구 현황, 노출위험 및 자료유용성 측도에 대해 차례로 살펴보았다. 그간 비밀보호 연구가 국내에서 방대하게 이루어져 왔다고 하기 어렵고, 노출위험 및 자료유용성 측도들은 그 수가 많고 상당히 복잡하여 실무적으로 사용하기에는 쉽지 않음을 알 수 있었다. 또한, 해외의 동향을 살펴보아도 특정한 기법

이나 측도에 대한 세부적인 방법론 위주로 연구가 이루어지고 있고, 최적의 매스킹 방안을 결정해주는 이론적 근거가 확립되어 있다고 하기는 어려운 실정이다. 수많은 가능한 매스킹 방안들에 대해서 다양한 노출위험과 정보손실 측도들을 기준으로 통계적 의사결정론을 이용해 최적안을 선택하는 알고리즘을 찾고자 하는 노력들이 계속되고 있으나 실제로 해결 방안을 제시하는 것이 그만큼 어렵기 때문이라고 할 수 있다.

최적 매스킹 방안을 제시해 주는 알고리즘이 없는 현실에서 가장 널리 사용되고 있는 방식은 여러 매스킹 방안들을 모두 적용하여 보고, 모든 방안에 대해 노출위험 및 정보손실을 측정한 후 두 측도 모두 작은 값을 가지는 방안을 자료 공표를 위해 사용하는 것이라 하겠다. 다만 노출위험과 정보손실 사이에 상충관계가 존재하여 둘 모두를 줄이는 것이 어렵고, 작은 노출위험과 큰 정보손실 혹은 그 반대의 조합이 전체적으로 같은 효용을 가질 수 있으므로, 이러한 매스킹 방안을 선택하는 방식 역시도 복잡한 정책적 의사결정 과정을 요구하게 되는 어려움이 있다.

참고로, 일반적으로 활용되고 있는 마이크로데이터 노출제한에 관한 절차는 아래 [그림 4-2]와 같이 나타낼 수 있다(Meindl 등, 2013). 이 작업 흐름도를 살펴보면 마이크로데이터



[그림 4-2] 일반적인 마이크로데이터의 매스킹 절차(Meindl 등, 2013)

에서 직접적인 식별정보를 삭제한 후, 키변수 및 민감변수에 대해 각종 매스킹 기법들을 적용하고, 노출위험과 정보손실을 측정하여 각 매스킹 방안들을 비교하며, 최종적으로

노출위험이 충분히 작고 유용성이 상당히 크면, 자료를 공표하는 절차를 따름을 알 수 있다.

이 절에서는 몇몇 노출위험 및 자료유용성 측도들을 기준으로 기존에 연구된 가계금융·복지조사의 매스킹 방안들을 평가해 보고자 한다. 더불어 추가로 연구된 매스킹 방안도 함께 적용하고 검토한다. 위에 언급한 이론적인 어려움들뿐만 아니라 연구 기간 및 인력의 한계로 인해 모든 측도들을 검토할 수는 없었으나, 현재 구현되어 있어 이용 가능한 측도들을 실제 자료에 적용하여 보기로 한다. 이를 위해 먼저 자료의 특성을 살펴보고 기존에 연구된 매스킹 방안 및 새로운 매스킹 방안들을 정리한 후, 각 매스킹 방안별로 노출위험 및 정보손실 정도를 분석하기로 한다.

1. 자료의 특징

가계금융·복지조사의 변수는 상당히 많으나, 이 중 노출제한 연구의 대상으로 다루는 변수는 가구주 특성과 관련된 시도, 성별, 연령, 교육정도, 종사상지위, 동거여부, 혼인상태, 직업, 가구원수, 주택유형, 주거면적, 입주형태 등의 키변수 12개 및 자산, 부채, 소득의 하위변수들로 구성된 12개 민감변수들이다. 이들은 모두 가계금융·복지조사의 금융부문 및 복지부문에서 공통으로 조사되는 변수들이며, 자료의 총 개수는 $n = 19{,}744$이다. 한편, 가계금융·복지조사에 대한 노출제한 연구는 2012년부터 시작되었으므로 일관성을 위해 본 연구에서는 2012년 마이크로데이터를 대상으로 분석을 수행하였다.

총 12개의 키변수에 대해서는 특정 범주에 대해서 빈도수가 상대적으로 크게 작아지는 일이 없도록 기본적인 재코딩 작업을 하였고, 변수별로 빈도수를 정리하면 다음 <표 4-5>와 같다. 괄호 안은 변수별로 범주의 개수가 표시되어 있고, *은 재코딩 되었음을 의미한다. 키변수에 대한 매스킹 기법은 재코딩만을 고려하였다.

〈표 4-5〉 키변수들의 빈도표

변수명	범주								
성별	남자	여자							
(2)	15349	4395							
연령(세)	~30	~35	~40	~45	~50	~55	~60	~65	~70
(12)*	977	1563	2212	2552	2587	2573	1859	1557	1337
	~75	~80	~85						
	1195	805	527						
교육정도	1	2	3	4	5	6	7		
(7)	1039	2478	2125	6471	1904	4579	1148		

변수명	범주								
종사상지위	1	2	3	4	기타				
(5)*	7941	2830	1167	4045	3761				
동거여부	1	2	3						
(3)*	3578	15914	252						
혼인상태	1	2	3	4					
(4)	1762	13749	2402	1831					
직업	1	2	3	4	5	6	7	8	9
(9)	573	2493	2448	1236	1675	1353	1792	2236	2085
가구원수	1	2	3	4	5~				
(5)*	3578	5027	3968	5348	1823				
주택유형	1	2	3	4					
(4)	7486	9514	2423	321					
주거면적(m^2)	~60	~85	~110	~135	~160	~185	185~		
(7)*	8623	6200	2060	1448	491	451	471		
입주형태	1	2	3	4	5				
(5)	11455	3737	3007	447	1098				

주: 교육정도: 1(안받음), 2(초등학교), 3(중학교), 4(고등학교), 5(대학, 3년제 이하), 6(대학, 4년제 이상), 7(대학원 이상)
종사상지위: 1(상용근로자), 2(임시, 일용근로자), 3(고용원이 있는 자영업자), 4(고용원이 없는 자영업자), 기타(무급가족종사자, 기타종사자, 무직자, 가사, 학생 등)
동거여부: 1(1인 가구), 2(같이 살고 있음), 3(따로 살고 있음)
혼인상태: 1(미혼), 2(배우자 있음), 3(사별), 4(이혼)
직업: 1(관리자), 2(전문가 및 관련 종사자), 3(사무 종사자), 4(서비스 종사자), 5(판매 종사자), 6(농림어업 숙련 종사자), 7(기능원 및 관련 기능 종사자), 8(장치, 기계 조작 및 조립 종사자), 9(단순 노무 종사자), 무응답 수(3853)
주택유형: 1(다가구 주택 포함 단독 주택), 2(아파트), 3(연립 및 다세대 주택), 4(기타)
입주형태: 1(자기집), 2(전세), 3(보증금이 있는 월세, 사글세), 4(보증금이 없는 월세, 사글세), 5(기타: 무상 주택, 무상 사택 등)

민감변수는 자산, 부채, 소득의 하위(중간)변수들 12개로 구성되어 있으며 변수명은 다음 <표 4-6>과 같다.[13)]

13) 가구 자료는 일반적으로 노출위험이 상대적으로 낮기 때문에 가계금융·복지조사 마이크로데이터에 대해서는 대부분의 변수들이 현재 통계청 마이크로데이터 서비스 시스템(MDSS)을 통해서 배포되고 있다. 다만 노출위험 증가에 대한 우려로 시도 변수는 제외되어 있다. 본 연구는 시도 변수를 포함하여 통계적으로 노출제어 처리된 공공이용파일 생산에 관하여 논의하고 있으며, 이는 가구 자료 이외의 다양한 마이크로데이터의 노출제한을 위해 필요한 연구이기도 하다.

〈표 4-6〉 민감변수명

자산	변수명	부채	변수명	소득	변수명
자산총액	asset	부채액	debt	가구소득	income
실물자산	asset01	금융부채	debt01	근로소득	income1
부동산평가액	asset11	담보대출	debt11	사업소득	income2
기타실물자산	asset12	신용대출	debt12	재산소득	income3
금융자산	asset02	기타	debt134	이전소득	income4
저축액	asset21	임대보증금	debt02		
전월세보증금	asset22				

각 변수별 상관계수를 살펴보면 아래 <표 4-7>과 같다. 단편적으로도 부동산과 관련된 변수끼리는 자산, 부채, 소득 범주를 뛰어넘어 상관관계가 상대적으로 높은 경향이 있음을 볼 수 있다. 따라서 큰 범주인 자산, 부채, 소득별로 매스킹 처리를 하는 것이 적절하지 않다고 판단된다. 이에 상관관계를 전체 변수들을 대상으로 고려하여 매스킹 기법을 적용하거나, 아예 상관관계는 전혀 고려하지 않고 각 변수별로 매스킹 기법을 적용을 하는 것을 비교·검토하도록 한다.

〈표 4-7〉 민감변수들의 상관계수 행렬

	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4
asset11	**1**	**0.1**	**0.45**	**0.29**	0.2	0.06	0.06	0.21	0.25	0.26	0.47	0.01
asset12	**0.1**	**1**	**0**	**0.03**	0.01	0.04	-0.01	0.23	0.24	0.01	0.02	-0.06
asset21	**0.45**	**0**	**1**	**0.3**	0.5	0.07	0.03	0.5	0.18	0.26	0.52	0.08
asset22	**0.29**	**0.03**	**0.3**	**1**	0.27	0.14	0.19	0.09	0.04	0.41	0.15	-0.03
debt11	0.2	0.01	0.5	0.27	**1**	**0.13**	**0.1**	**0.23**	0.09	0.27	0.23	-0.03
debt12	0.06	0.04	0.07	0.14	**0.13**	**1**	**0.06**	**0.06**	0.06	0.11	0.02	-0.04
debt134	0.06	-0.01	0.03	0.19	**0.1**	**0.06**	**1**	**0.02**	0	0.13	0.06	-0.04
debt02	0.21	0.23	0.5	0.09	**0.23**	**0.06**	**0.02**	**1**	0.15	0.09	0.32	0.01
income1	0.25	0.24	0.18	0.04	0.09	0.06	0	0.15	**1**	**-0.16**	**0.08**	**-0.21**
income2	0.26	0.01	0.26	0.41	0.27	0.11	0.13	0.09	**-0.16**	**1**	**0.09**	**-0.07**
income3	0.47	0.02	0.52	0.15	0.23	0.02	0.06	0.32	**0.08**	**0.09**	**1**	**0.03**
income4	0.01	-0.06	0.08	-0.03	-0.03	-0.04	-0.04	0.01	**-0.21**	**-0.07**	**0.03**	**1**

본격적으로 매스킹 방안들을 적용하기에 앞서, 각 변수들의 분포를 살펴보는 것은 필요한 과정이다. 먼저 각 변수들은 0응답(0으로 응답하거나 무응답으로 값이 0인 경우)이 많은 특징을 가진다. 변수별로 0응답인 레코드의 개수를 정리하면 아래 <표 4-8>과 같고, 0이 아닌 값들에 대한 히스토그램은 <부록 2>에 첨부한다. 히스토그램을 살펴보면 대부분 변수들은 왼쪽으로 치우친 분포를 보여 정규성을 따른다고 보기 어렵다. 이에 변수들을 로그 변환을 하면 분포의 대칭성 측면에서 좀 더 나은 결과를 보이는 변수들이 있음을 알 수 있다. 로그 변환하여 얻은 히스토그램도 함께 <부록 2>에 첨부하였다. 효율적인 표현을 위하여 상위 10%는 제외하고 히스토그램을 작성하였다.

〈표 4-8〉 변수별 0응답의 개수

자산		부채		소득	
asset11	40	debt11	13148	income1	6495
asset12	13000	debt12	15171	income2	13547
asset21	6524	debt134	13592	income3	15271
asset22	4538	debt02	16625	income4	11446

민감변수들의 특징을 종합하면, 먼저 변수들 간의 상관관계는 자산, 부채, 소득 각 분류내에서 강하지 않고, 분류와 상관없이 상관관계가 큰 하위변수들의 조합이 존재한다. 따라서 상관관계를 고려할 때는 전체 변수들을 함께 고려하거나, 상관관계를 무시하고 개별 변수별로 매스킹 처리를 하여 비교하는 것이 바람직하다. 다음으로 각 민감변수들은 0인 값들을 많이 가지고 있거나, 0이 아닐 경우 양수인 제한조건을 가지고 있어 이러한 제한조건을 고려하여 매스킹 처리를 할 필요가 있다.

2. 민감변수의 매스킹 방안

현재 적용 가능한 연속형 변수를 위한 매스킹 기법들로는 국소통합(microaggregation), 잡음추가, 자료순위 교환(rank swapping), 자료섞기(shuffling) 등이 있으며, 기존 연구에서는 이러한 기법들을 가계금융·복지조사 마이크로데이터에 적용하여 노출위험-정보손실 지도를 통해 각 방법들을 비교하였다(박민정 등, 2013). 그 결과를 살펴보면 국소통합과 잡음추가 기법을 이용할 경우 노출위험 및 정보손실 측면에서 상대적으로 효율적인 결과를 얻음을 알 수 있었다. 한편, 노출위험과 정보손실의 상충관계를 고려하여 국소통합과 잡음추가의 결합안도 함께 검토하였다.[14]

14) 참고로, 민감변수의 상한코딩(top-coding)은 정책적 고려가 필요한 측면을 가지고 있으므로 이를 배제하고

이 절에서는 먼저 기존 연구의 매스킹 방안에 대해 정리하고, 0응답 정보에 대한 처리를 좀 더 이론적으로 검토하도록 하겠다. 다음으로 양수 조건을 가지는 민감변수들에 대한 매스킹 기법을 추가로 검토하여 노출위험 및 자료유용성을 심층 검토할 매스킹 방안들을 선정하도록 한다.

가. 기존 연구의 매스킹 방안

기존 연구(박민정 등, 2013)의 국소통합과 잡음추가 기법 적용 과정을 요약하면 다음과 같다. 국소통합에 여러 알고리즘이 존재하는데 그 중 MDAV[15] 알고리즘이 가장 효율적인 것으로 나타났다. 잡음추가 기법 적용 시에는 상관관계를 고려한 로버스트한 알고리즘[16]이 가장 효율적인 것으로 판단된다. 한편, 앞 절에서 언급했듯이 민감변수들은 값이 0인 경우가 많고, 0이 아닌 경우 모두 양수인 조건을 가지고 있다. 여기에 잡음을 추가할 경우 음의 값이 생기거나 0인 자료들이 거의 없어지는 문제점이 있다. 이에 기존 연구에서는 0응답에 대한 다음 세 가지 시나리오를 검토하였다. 이는 0응답에 대한 정보손실을 어느 정도 허용하느냐에 따르고 있다.

① 잡음추가 후 발생하는 음의 값들을 절대값으로 대체한다.
② 잡음추가 후 원래 무응답 개수만큼 절대값이 작은 자료들을 0으로, 나머지는 절대값으로 대체한다.
③ 잡음추가 후 원래 0인 자료들을 0으로, 나머지는 절대값으로 대체한다.

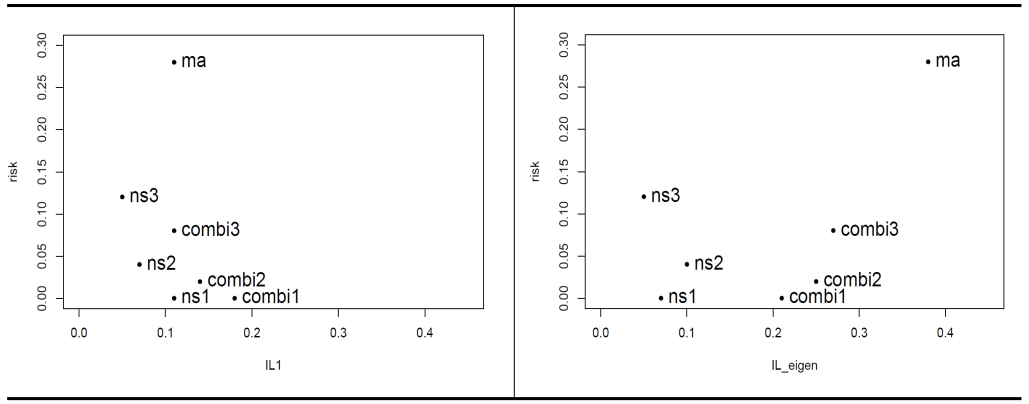
기존 연구의 결합안(combi)에서는 국소통합의 결과 값에 후 잡음추가 알고리즘에서 발생시킨 잡음을 더하였다[17]. 이때 잡음추가 후 민감변수에 대한 음의 값 처리는 위의 세 시나리오를 따라 수행하였다. 이들을 차례로 ma, ns1, ns2, ns3, combi1, combi2, combi3으로 각각 표기하면 상대적인 노출위험-정보손실은 다음 [그림 4-3]과 같이 나타난다.

나머지 매스킹 기법들을 살펴보았다.

15) 네델란드 통계청에서 제안한 자료 간 거리가 먼 것부터 묶어가는 국소통합 알고리즘

16) 마이크로데이터 매스킹 처리를 위한 R패키지인 sdcMicro의 잡음추가에서 correlated2 옵션으로 구현되어 있는 알고리즘(Templ과 Meindl, 2008a)

17) 원자료를 X_o, 국소통합 처리한 자료를 X_{ma}, 잡음추가 처리한 자료를 X_{ns}라고 할 때, 검토한 결합안은 개념적으로 표현하면 $X_{combi} = X_{ma} + X_{ns} - X_o$에 해당한다. 따라서 국소통합 결과와 마찬가지의 평균을 가지는 매스킹 결과를 만들게 된다. 만약 $X_{combi} = X_{ma} + E$이고, E는 평균 $E(X_o - X_{ma})$, 분산 $V(X_o - X_{ma})$인 정규분포에서 만들어진 잡음이라면 원자료의 평균과 분산을 보존하여 바람직하다. 그러나 이를 실제 자료에 적용했을 때 매스킹된 자료가 원자료에 매우 가깝게 되어 노출위험이 0.9 이상이었기 때문에 $X_{combi} = X_{ma} + E$ 방식의 결합안은 검토 대상에서 제외하였다.



[그림 4-3] 노출위험-정보손실 지도(기존 연구의 매스킹 방안)

그림에 표현된 risk는 자료순위 교환 매스킹 기법의 아이디어와 유사하게 만들어진 거리 기반 노출위험 측도이다.[18] 먼저 매스킹된 각 레코드에 대하여 자료순위에 근거하여 구간을 정의하고 해당 원자료의 레코드 값들이 구간 안에 있는지 없는지를 살핀다. 다음으로 전체 파일 중에서 원자료의 값이 매스킹된 자료를 가지고 만든 구간 안에 있는 비율을 통해 파일 단위의 노출위험을 정의하였다(Temple과 Meindl, 2008b). 이는 침입자 의사결정론 기반 노출위험을 구현하기가 어려운 실정에서 연속형 변수의 노출위험을 측정하는 현실적으로 사용할 수 있는 거의 유일한 방안으로 볼 수 있다.[19]

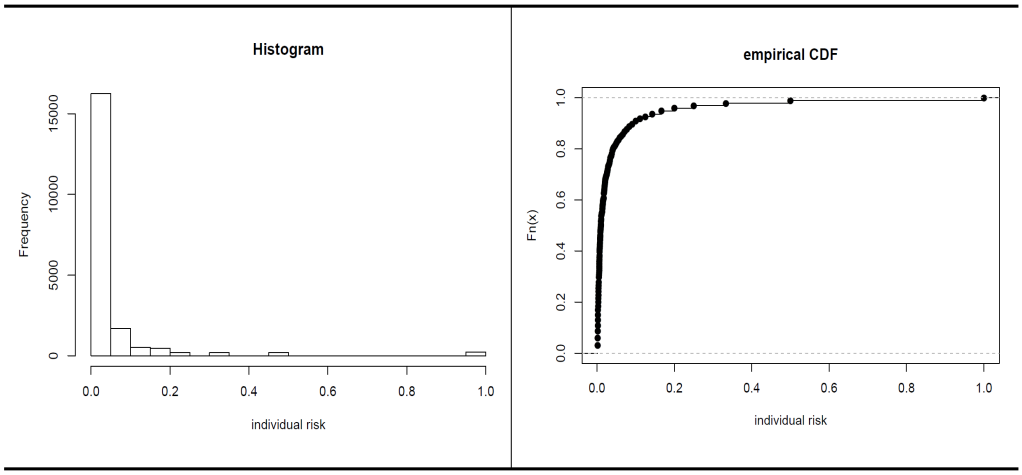
정보손실(information loss) 측도 IL1이나 IL_eigen 역시 거리 기반 측도이다. 매스킹 전후 레코드 사이의 거리를 계산하고 이를 표준 편차로 나눈 것이 IL1이며, 자료의 구조가 얼마나 변화되었는지 파악하고자 고유값들을 비교한 것이 IL_eigen에 해당한다. 매스킹의 목적은 노출위험과 정보손실 모두를 작게 만드는 것이나, 매스킹된 자료가 원자료와 가까울수록 노출위험(risk)은 커지고 정보손실(IL1)은 작아지게 되며 어느 하나를 줄이려면 다른 하나가 늘어나는 상충관계(trade-off)가 존재한다. 따라서 위 그림의 ns3과 combi3의 경우처럼 두 매스킹 방안이 하나는 노출위험이 작고 다른 하나는 정보손실이 작다면, 둘 중 어느 매스킹 방안이 우월한 것인지 논의하기 어렵다.

18) 2절 노출위험 측정 방법론 끝부분에 언급되어 있다.

19) 2절에서 설명하였듯이 연속형 변수가 아닌 키변수의 노출위험은 모집단 유일성에 근거하여 여러 노출위험 측도가 있다. 그러나 민감변수인 자산, 부채, 소득과 같은 연속형 변수에 대한 노출위험 측정은 침입자 의사결정론을 이용한 노출위험 측도를 이용해야 한다. 침입자 의사결정론에 근거한 노출위험 측도는 자료별로 또는 매스킹 기법별로 구현이 복잡한 반면에, 그 외의 연구되어 있는 연속형 변수에 대한 노출위험으로 sdcMicro에 구현되어 있는 것이 유용하다고 할 수 있다.

나. 0응답 정보 처리에 대한 노출위험 검토

기존 연구에서 0응답에 대한 시나리오는 0응답 정보를 보존할 것인지 아닌지에 따라 경험적으로(heuristic) 정리하였는데, 본 연구에서는 0응답 여부를 0과 1의 값을 가지는 변수로 변환하여 유일성에 근거한 노출위험을 측정하여 보았다. 예를 들어 변수 debt11의 경우, 0응답을 한 13,148개의 개체는 0의 값을 나머지는 1의 값을 가지게 하는 것이다. 12개 민감변수 모두를 0과 1 두 개의 범주를 가지는 변수들로 변환하여 먼저 보수적으로 가중값을 고려하지 않고 표본 유일성을 모집단 유일성으로 간주하여 노출위험을 산출하면 개별 노출위험[20]의 분포는 아래 [그림 4-4]와 같다.



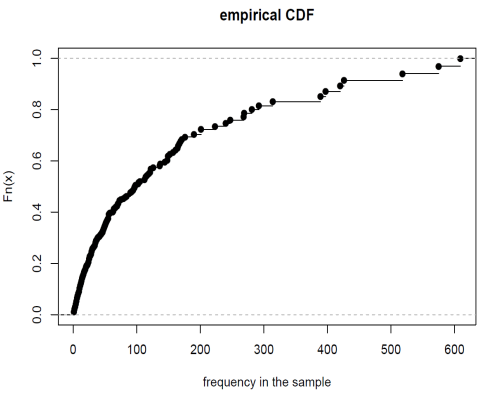
[그림 4-4] 0응답 여부에 따른 개별 노출위험 측도 값들의 분포

히스토그램이나 누적분포함수를 살펴보면 개별 노출위험(individual risk)은 비교적 작은 값을 가지는 개체가 대다수임을 알 수 있다. 이때 개별 노출위험의 평균은 0.0461(4.61%)이며, 총합은 911로 이는 전체 표본 중에서 911개가 식별될 레코드수의 기댓값이라는 의미로 해석하기도 한다.

참고로 표본 유일성 개수에 대한 표와 누적분포함수는 다음 [그림 4-5]와 같다. 표본 유일성 개수에 관한 표를 보면, 표본 유일성이 1인 개체는 총 221개로 전체의 약 1.12%가 가장 심각한 노출위험을 가진다고 할 수 있다.

20) 2.1절 마지막 부분에 언급한 대로 본 연구에서 사용하고 있는 sdcMicro 패키지는 음이항 모형을 이용하여 개별 레코드의 유일성 기반 노출위험을 계산하여 제공하고 있다. 이는 네델란드 통계청에서 만든 마이크로데이터 매스킹 프로그램인 μ-Argus에 구현된 노출위험과 동일하게 만들어진 것이다.

1	2	3	4	5
221	206	183	200	210
6	7	8	9	10
258	182	168	162	250
...				
420	426	518	575	609
420	426	518	575	609



[그림 4-5] 0응답 여부에 따른 유일성 측도 값들의 분포

한편, 가중값을 고려하여 동일하게 개별 노출위험을 계산을 하면 평균은 0.000077 (0.01% 미만), 총합은 1.523으로 식별될 레코드수의 기댓값이 2개 미만이라 해석할 수 있다. 즉, 어느 응답자가 표본에 포함되어 있다는 정보가 없다면 무응답 유무에 의한 노출위험은 상당히 낮다고 할 수 있다.

따라서 0응답 정보를 보존하는 것은 심각한 노출위험을 초래한다고 보기는 어렵다고 간주하여, 기존 연구로부터는 무응답 정보를 보존하는 매스킹 방안인 3번 시나리오를 심층 검토할 대상으로 삼도록 한다.

다. 양수 조건을 고려한 매스킹 방안

매스킹 기법들을 전반적으로 소개하고 적용하여 보는 것이 주된 목적이었던 기존 연구에서는 잡음추가 시 발생하는 음의 값들을 처리하기 위해 이들의 절대값을 취하여 양수로 변환하는 경험적인(heuristic) 방식을 취하였다. 이러한 방법은 적용하기에는 손쉬우나 분포의 구조를 왜곡할 수 있는 문제가 있다. 본 연구에서는 평균과 분산뿐만 아니라 양수 조건을 보존하는 매스킹 기법을 추가로 검토하였다. 이는 ①변수에 로그를 취하고 ②잡음을 추가한 후 ③다시 지수 변환을 하는 과정으로 설명될 수 있어 승법 잡음추가 기법을 적용하는 효과를 가지게 되는 방식이다[21](Oganian과 Karr, 2011). 즉, 승법 잡음추가처럼 양수조건을 만족하지만 평균과 분산 또한 보존하는 방법으로 제안되었다. 이를 설명하면 다음과 같다.

원래 자료를 p개 변수로 이루어진 행렬 X_o, 분산을 Σ_o라고 할 때, 일반적으로 잡음

21) 승법 잡음추가 기법은 평균과 분산을 보존하기 어려운 것으로 알려져 있다.

추가 기법을 이용해 매스킹된 자료 X_m은 다음과 같이 만들어진다.

$$X_m = E(X_o) + \frac{(X_o - E(X_0)) + E}{\sqrt{1+c}}$$

이때 c는 통계작성기관이 결정하는 공표되지 않는 값이며, E는 $N(0, c\Sigma_0)$을 따르는 잡음을 표현한다. 이렇게 잡음을 추가할 때 $E(X_m) = E(X_o)$ 및 $V(X_m) = V(X_o)$을 만족하게 된다.

이러한 평균 및 분산의 보존뿐만 아니라 $X_m > 0$ 조건 또한 만족시키는 방법으로 제안된 것은 주어진 $E(X_o)$ 및 $V(X_o)$에 대하여 원자료 X_o와 조건부 독립인 잡음 E가 다음을 만족하는 것이다. 이때 연산자 $\circ$는 성분별 곱을 의미한다.

$$E(X_o \circ \exp(E)) = E(X_o)$$
$$V(X_o \circ \exp(E)) = (1+c)\, V(X_o)$$

위 조건을 만족하는 매스킹된 자료 X_m은

$$X_m = \frac{(\sqrt{1+c} - 1)\, E(X_o) + X_o \circ \exp(E)}{\sqrt{1+c}} \qquad (4\text{-}1)$$

이 되며, 잡음 E는 $N(\mu_E, \Sigma_E)$을 따르고 Σ_E 및 μ_E은 다음과 같이 계산된다(Oganian과 Karr, 2011).

$$\Sigma_E(i,j) = \log\left(1 + \frac{c\Sigma_o(i,j)}{E(X_o(i)\, X_o(j))}\right)$$
$$\mu_E(i) = \sigma_E(i)/2$$

한편, Orgnian과 Karr(2011)에서는 자료의 범위에 따라 c의 값을 다르게 적용하여 (two-zone masking) 매스킹 전후 분산의 크기 차이를 줄이는 방안을 제시하기도 하였다.

라. 민감변수의 매스킹 방안들

앞에서 원자료의 0응답을 이진변수 취급하여 유일성에 근거한 노출위험을 측정한 결과, 노출위험이 지나치게 크다고 판단되지 않아 0응답을 보존하기로 하였다. 이는

정보손실을 감소시키기 위한 선택이라고 할 수 있다. 또한 전체적인 상관관계 고려 및 양수 조건 반영 여부에 따라 심층 검토할 매스킹 방안으로 다음과 같은 M0~M3을 고려하기로 한다.

	매스킹 방안
M0	- 국소통합(microaggregation) MDAV기법(ma)
M1	- 전체 상관관계를 고려한 로버스트한 잡음추가 기법(ns3)
M2	- 전체 상관관계를 고려하고, 양수 조건 반영
M3	- 전체 상관관계를 고려하지 않고, 변수별로 양수 조건 반영
C1	- M0의 결과에 M1에서 얻은 잡음추가
C2	- M0의 결과에 M2에서 얻은 잡음추가
C3	- M0의 결과에 M3에서 얻은 잡음추가

M0는 기존 연구의 ma에 해당하며 국소통합 기법만 적용한 경우이다. M1은 기존의 ns3에 해당하며 평균 및 상관계수 행렬을 보존하는 로버스트한 잡음추가 알고리즘을 적용한 경우이다. 이때 원래 0인 값은 0으로 환원하고 음의 값이 된 자료는 절대값으로 변환시켰다.

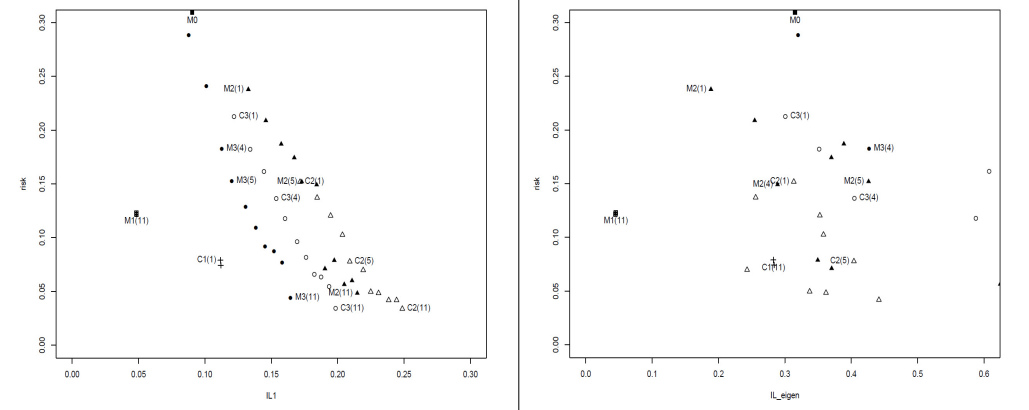
M2는 앞 절의 식 (4-1)을 전체 자료에 적용하며 매스킹 과정에서 음의 값이 발생하지 않는 잡음추가 기법이다. 이때 원래 0인 자료에 대해서는 c를 0으로 주어 2개 구역 차별처리(two-zone masking)를 통해 0응답 정보를 보존하였다. 다만 변수의 개수가 많아 원자료의 공분산 행렬이 양정치(positive definite) 행렬이 아니며, 이를 양정치 행렬로 변환시키기 위한 과정에서 많은 정보손실이 발생하는 문제가 있다. 따라서 식 (4-1)을 위해 잡음을 발생시킬 경우, 이론과 달리 매스킹 전후 상관계수 행렬에서 상대적으로 큰 차이가 발생하게 된다. 다만, 변수의 개수가 적어 공분산 행렬이 양정치일 때는 이론과 같이 공분산 행렬이 보존될 것으로 기대된다.

마지막으로 M3은 식 (4-1)을 각 변수별로 적용한 경우이다. 매스킹 과정에서 음의 값이 발생하지 않으며, 원래 0인 자료에 대해서는 c를 0으로 주어 2개 구역 차별처리를 통해 0응답 정보를 보존하였다. 그러나 각 변수별로 매스킹 처리가 되어 전체적인 상관관계 보존에는 한계가 있다. 이상 각각의 매스킹 방안들과 더불어 국소통합 M0의 결과에 M1~M3에서 발생시킨 잡음을 더하는 C1~C3의 결합안들도 함께 검토하도록 한다.

3. 노출위험 및 자료유용성 비교

가. 노출위험-정보손실 비교

앞에서 논의된 민감변수의 매스킹 방안 별로 노출위험과 정보손실을 구하여 그림으로 나타내면 다음과 같다. M0는 네모, M1 및 C1은 +, M2 및 C2는 동그라미, M3 및 C3은 세모로 표시되어 있다. M1의 경우 잡음의 강도(scale)를 많이 하거나 작게 하여도 노출위험 및 정보손실 측도 값에서 차이가 나지 않아 잡음을 가장 작게 더한 경우와 가장 크게 더한 경우를 표시하였다.



[그림 4-6] 노출위험-정보손실 지도

노출위험-정보손실 지도에서 전체적으로 M2보다는 M3이 원점에 가까워 매스킹 효과가 좋다고 판단할 수 있다. 결합안들 중 C2와 C3은 대부분 원점을 기준으로 M1 및 M3 바깥에 존재하므로 매스킹 효과가 좋다고 하기 어렵다. 상충관계를 감안하였을 때, 전체적 효용이 동등할 수 있는 매스킹 안들은 M1(1), C1(1), M3(11) 및 C3(11)이라고 할 수 있겠다. 이 중 M3과 C3은 IL_eigen 측도를 기준으로는 상당한 정보손실이 있다고 판단된다.

한편, M2 및 M3의 경우 식 (4-1)을 $E(X_o)$와 $\exp(E)$의 가중합으로 생각할 수 있고, 가중값을 각각 $w_1 = (1 - 1/\sqrt{1+c}\,)$와 $w_2 = 1/\sqrt{1+c}$로 생각할 수 있다. 잡음에 대한 가중값 w_2가 커짐에 따라 괄호안의 숫자가 증가하도록 표현하였다. 이때 가중값 w_2가 커질수록 노출위험은 작아지고 정보손실은 커지는 것을 알 수 있다. 주요 매스킹 방안들에 대해 노출위험이 0.1 이하인 것들을 중심으로 노출위험 및 정보손실 측도 결과 값을 정리하면 다음 <표 4-9>와 같다. 전체 방안에 대한 노출위험 및 정보손실 측도에 대한

표는 <부록 3>에 첨부하였다.

〈표 4-9〉 노출위험 및 정보손실 측도 결과값

	c	w_1	w_2	risk	IL1	IL_eigen
M0				0.3094	0.0904	0.3152
M1(1)			5	0.1217	0.0484	0.0445
M3(8)	0.225	0.0965	0.9035	0.0919	0.1453	0.876
M3(9)	0.25	0.1056	0.8944	0.0872	0.152	0.9597
M3(10)	0.275	0.1144	0.8856	0.077	0.1581	1.1294
M3(11)	0.3	0.1229	0.8771	0.044	0.1644	1.3122
C3(6)	0.175	0.0775	0.9225	0.0962	0.1694	0.8179
C3(7)	0.2	0.0871	0.9129	0.0817	0.1762	1.0497
C3(8)	0.225	0.0965	0.9035	0.066	0.1825	0.8334
C3(9)	0.25	0.1056	0.8944	0.0636	0.1876	0.9744
C3(10)	0.275	0.1144	0.8856	0.0547	0.1936	1.1368
C3(11)	0.3	0.1229	0.8771	0.0341	0.1985	1.2469

위 표를 살펴보면, 정보손실이 가장 작은 M1과 비교하여 추가적인 정보손실을 감안하고 노출위험을 줄이는 방안들을 비교 검토할 수 있겠다. 그러나 IL_eigen 측도가 보여주는 공분산 구조에 대한 정보손실은 어떠한 방안도 M1과 차이가 매우 큼을 알 수 있고, 노출위험 축소를 위해서는 상대적으로 큰 정보손실을 감수해야 한다고 판단된다.

나. 자료유용성 비교

각 통계량들의 차이를 다음과 같은 매스킹 전후 각 통계량들 사이의 절대적 차이에 대한 상대적인 비율을 이용하여 표현하고자 한다.

$$ARB = \frac{1}{h}\sum_{i=1}^{h}\left|\frac{\hat{\theta_i} - \theta_i}{\theta_i}\right|$$

주요 매스킹 방안들에 대해 각 변수별 평균의 ARB와 상관계수 차이 절대값의 합(S)을 정리하면 다음 표와 같다. 전체 방안에 대한 각 변수별 평균의 ARB와 상관계수 차이

절대값의 합에 대한 표는 <부록 4>에 첨부하였다.

	평균의 ARB												S
	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4	corr
M0	0.00	0.03	0.01	0.02	0.08	0.09	0.08	0.05	0.02	0.04	0.07	0.03	**1.06**
M1(1)	0.06	0.00	0.01	0.05	0.00	0.00	0.03	0.00	0.00	0.00	0.01	0.00	**0.10**
M3(8)	0.01	0.01	0.01	0.01	0.00	0.02	0.01	0.01	0.00	0.01	0.02	0.01	**1.34**
M3(9)	0.01	0.00	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.00	0.01	0.01	**1.37**
M3(10)	0.00	0.02	0.00	0.01	0.01	0.01	0.03	0.01	0.00	0.01	0.03	0.01	**1.86**
M3(11)	0.00	0.01	0.01	0.03	0.01	0.00	0.02	0.02	0.00	0.01	0.01	0.01	**1.83**
C3(6)	0.01	0.01	0.01	0.01	0.09	0.09	0.10	0.03	0.02	0.03	0.07	0.04	**2.05**
C3(7)	0.01	0.02	0.02	0.04	0.07	0.07	0.06	0.04	0.01	0.03	0.08	0.04	**2.40**
C3(8)	0.01	0.03	0.02	0.03	0.07	0.11	0.09	0.06	0.02	0.05	0.05	0.04	**2.20**
C3(9)	0.00	0.03	0.02	0.03	0.07	0.11	0.07	0.06	0.02	0.04	0.06	0.02	**2.30**
C3(10)	0.00	0.04	0.02	0.03	0.07	0.08	0.05	0.06	0.02	0.03	0.04	0.04	**2.70**
C3(11)	0.00	0.02	0.01	0.01	0.08	0.08	0.05	0.07	0.02	0.05	0.08	0.02	**2.73**

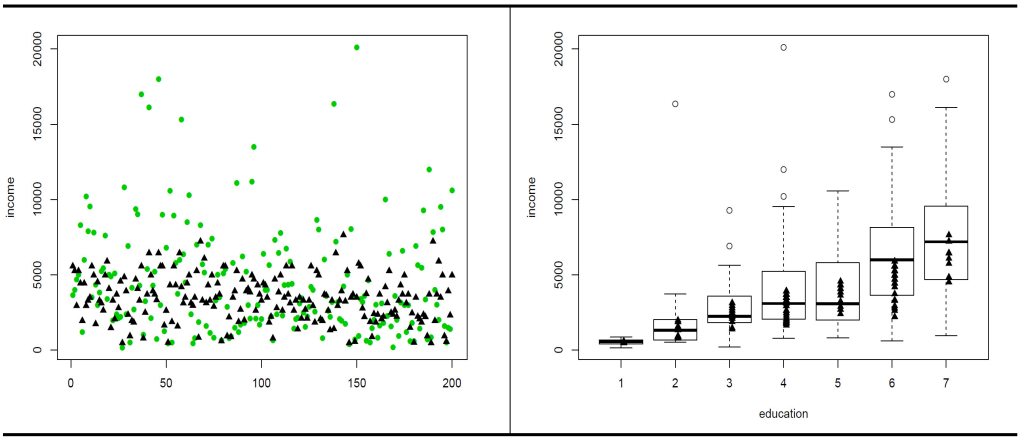
주: S는 상관계수 차이 절대값의 합

노출위험 측도에서 예상되었듯이 평균의 차이는 크지 않으나 매스킹 전후 상관계수 상이의 차이에 대해서는 M1에 비해서 다른 방안들이 큰 값을 가지는 것을 알 수 있다.

다음으로 벤치마킹 통계량에서의 차이에 대해 살펴보고자 한다. 선정한 벤치마킹 통계량은 성별에 따른 소득 차이(Gender Pay Gap, GPG) 및 지니(Gini) 계수이다. 각 통계량을 계산하는 것이 매스킹 전후 어떻게 달라지는 ARB를 기준으로 살펴보았다. 또한 층화 층이나 시도별로 ARB의 값들이 어떻게 변하는지 비교하고, 자료 자체 값뿐만 아니라 특정 모형을 자료를 통해 얻은 후, 그 모형의 예측값을 가지고 벤치마킹 통계량을 구할 때 ARB 값들도 비교하였다. 고려한 모형은 다음과 같이 세 개의 설명 변수를 가지며 상수항을 포함하는 모형이다.

$$log(income) \sim constant + age + gender + edu$$

모형을 사용한 결과를 200개 표본에 대하여 그리면 아래 [그림 4-7]과 같다.



[그림 4-7] 매스킹 전후 자료의 모형적합 결과 비교

동그라미는 자료의 값이며 세모는 모형에 의한 적합값을 나타낸다. 오른쪽 그림은 원래 소득의 교육수준별 상자줄기 그림과 적합된 값을 나타내었다. 이는 자료 이용의 한 예를 상정하여 매스킹에 따라 자료 분석 결과가 얼마나 달라지는 알아보고자 하는 것이다.

이제 층화 층은 h, 시도 변수는 d, 모형의 예측값을 이용한 경우는 m으로 표현하여 각 통계량의 ARB값들을 정리하면 다음과 같다.

GPG	ARB	ARB.h	ARB.d	ARB.m	ARB.m.h	ARB.m.d
M0	0.0031	0.0203	0.0076	0.0014	0.0012	0.0012
M1(1)	0.0005	0.0182	0.0089	0.0001	0.0016	0.0011
M3(1)	0.0360	0.0276	0.0158	0.0145	0.0160	0.0154
M3(2)	0.0005	0.0328	0.0209	0.0235	0.0248	0.0243
M3(3)	0.0380	0.0407	0.0247	0.0282	0.0301	0.0295
M3(4)	0.0474	0.0515	0.0362	0.0363	0.0399	0.0387
M3(5)	0.0174	0.0601	0.0424	0.0423	0.0455	0.0443
M3(6)	0.0381	0.0596	0.0507	0.0546	0.0603	0.0582
M3(7)	0.0751	0.0734	0.0505	0.0560	0.0617	0.0600
M3(8)	0.0917	0.0728	0.0512	0.0574	0.0624	0.0608
M3(9)	0.0430	0.0773	0.0569	0.0599	0.0646	0.0632
M3(10)	0.0057	0.0967	0.0701	0.0696	0.0737	0.0720
M3(11)	0.1845	0.1041	0.0809	0.0767	0.0825	0.0804

Gini 계수	ARB	ARB.h	ARB.d	ARB.m	ARB.m.h	ARB.m.d
M0	0.0082	0.0144	0.0086	0.0006	0.0015	0.0011
M1(1)	0.0009	0.0078	0.0034	0.0018	0.0030	0.0025
M3(1)	0.0020	0.0123	0.0050	0.0200	0.0209	0.0204
M3(2)	0.0083	0.0181	0.0105	0.0354	0.0359	0.0348
M3(3)	0.0055	0.0211	0.0118	0.0466	0.0472	0.0462
M3(4)	0.0085	0.0299	0.0124	0.0493	0.0508	0.0497
M3(5)	0.0122	0.0327	0.0164	0.0588	0.0605	0.0591
M3(6)	0.0087	0.0292	0.0191	0.0648	0.0676	0.0660
M3(7)	0.0114	0.0306	0.0143	0.0766	0.0779	0.0760
M3(8)	0.0217	0.0406	0.0220	0.0863	0.0873	0.0852
M3(9)	0.0238	0.0417	0.0255	0.0911	0.0919	0.0900
M3(10)	0.0169	0.0469	0.0215	0.0943	0.0970	0.0949
M3(11)	0.0267	0.0506	0.0320	0.1041	0.1065	0.1044

잡음에 대한 가중값 w_2가 커짐에 따라 정보손실이 커지고, 각 통계량들의 ARB값들 역시 커짐을 알 수 있다. 다만, 충분히 작은 노출위험을 위해 감수해야 하는 통계량 추정시 상대적 편차(bias)는 M3(9)의 경우 GPG에 대해서는 대략 4~8%, Gini 계수에 대해서는 대략 2~9%로 대부분 M3 매스킹 방안들에 대하여 한 자리수 %를 나타내고 있다. M1의 경우 대부분 1% 미만의 상대편차를 보여 정보손실의 측면에서 비교적 안정적으로 판단된다.

제5절 요약 및 향후 연구방향

마이크로데이터의 노출제어는 세계 각국의 통계기관에서 관심을 가지고 노력하고 있지만, 만족할 만한 방법이 정립되어 있다기보다는 적절한 대안을 계속해서 찾아가고 있는 분야라고 할 수 있다. 전통적으로는 개별 매스킹 기법들에 대한 연구가 이루어져 왔으며, 각 기법들을 실제 자료에 적용하기 위해서는 각 기법의 프로그래밍 작업이 별도로 이루어질 필요가 있었다. 다양한 매스킹 기법들을 편리하게 실제 자료에 적용하기 위해 네델란드 통계청의 μ-Argus라는 프로그램이 개발되어 사용되어 왔고, 이를 기반으로 접근성과 활용성에 장점을 가지는 무료 통계 분석 프로그램인 R을 이용한 패키지로 sdcMicro가 오스트리아 통계청을 중심으로 제작되었다. R패키지인 sdcMicro에는

μ-Argus의 매스킹 기법들뿐만 아니라 더욱 최신의 알고리즘과 몇몇 노출위험 및 자료유용성 측도들이 구현되어 있기도 하다.

매스킹 기법들이 발전하면서, 여러 매스킹 기법들을 이용한 다양한 노출제어 방안들을 비교하기 위해 노출위험과 자료유용성 측도의 중요성은 더욱 높아진다. 자료 공급자의 입장에서는 노출위험이 작다는 것이 보장될 필요가 있고, 자료 이용자 입장에서는 원래 조사된 자료와 비교하여 매스킹된 자료의 자료유용성이 충분히 보장될 필요가 있기 때문이다. 그러나 어떠한 매스킹 기법을 적용하여도 노출위험을 줄이고 자료유용성을 높이는(정보손실을 줄이는) 효과를 동시에 얻을 수는 없기 때문에 어떠한 매스킹 기법을 적용할 것인지 선택하는 문제는 쉽지 않은 것이 된다. 더구나 노출위험과 자료유용성 측도들은 그 종류가 많아 어떤 노출위험을 줄이는 매스킹 방안이 더 좋은지를 논하는 것도 쉽지 않은 문제이다.

본 연구에서는 이러한 한계 속에서 sdcMicro를 이용하여 현재 적용 가능한 마이크로데이터의 매스킹 기법들을 적용하고, 양수 조건을 가지는 자료의 특성을 반영하기 위해 최신 논문의 방법들을 검토하였다. 이에 의해 총 7개의 매스킹 방안들에 대해 거리 기반 노출위험 측도들을 비교하였다. 한편, 거리 기반 자료유용성 측도들을 함께 노출위험-자료유용성 지도에서 비교하여, 로버스트한 잡음추가 기법 및 국소통합과 잡음추가의 결합안을 전체적인 효용이 유사할 수 있는 방안들로 간주하였다. 나아가 주요 통계량으로 평균, GPG 및 Gini 계수에 대하여 매스킹 전후 자료유용성의 변화를 비교하였다. 이는 개별 매스킹 기법들을 연구하는 것을 넘어서, 매스킹 기법을 적용하는 전체 과정을 국내에서 처음으로 조명하여 각 매스킹 방안들을 평가하고 선택하는 사례를 제시한 의의를 가진다고 할 수 있다. 다만 연구 내용에 비하여 연구 기간이 짧아 본 보고서에서 전문적인 내용을 알기 쉽게 풀어쓰는데 시간적 제약이 큰 한계가 있었다. 본 보고서는 노출위험 및 자료유용성(정보손실) 측도에 관한 이론적인 내용을 대부분 담고 있으니 전문적인 내용 습득을 위해 활용하기 바라며, 매스킹 기법들의 실무 활용을 위해서는 향후 통계개발원에서 발간될 「마이크로데이터 매스킹 기법 실무 활용 안내서」를 참고하기 바란다.

마이크로데이터의 노출제어 문제를 해결하기 위해 각 통계기관이 전통적으로는 매스킹 기법들에 의존해 와서 매스킹 기법들이 가장 널리 활용되고 있지만, 매스킹 기법만으로는 자료 공급자 및 자료 이용자를 모두 만족시키기 어려운 한계를 가진다고 할 수 있다. 때문에 미국, 독일 등에서는 인위자료(synthetic data)를 활용하여 공공이용파일을 작성하는 것을 최근 꾸준히 연구해왔다. 실제 자료에 대해 인위자료를 생성하여 공공이용파일을 작성한 사례도 시작되고 있으며 학계에서 연구가 활발히 진행되고 있다고 할 수 있다. 그러나 인위자료를 생성하는 것은 베이지안 기법에 근거한 여러 이론들을 자료 공급자가

익혀야 하는 문제나 인위자료의 활용 자체가 가지는 한계들도 있어 역시 많은 어려움이 예상된다. 이외에 마이크로데이터의 노출제어 문제에 대한 연구로 가장 최근에 활발한 분야는 차등 정보보호(differential privacy)를 들 수 있다(이용희, 2013).

우리나라에서 마이크로데이터를 안전하고 편리하게 배포하기 위해 마이크로데이터의 노출제어 문제는 계속해서 연구되고 발전되어야 할 분야라고 할 수 있다. 세계 각 국에서도 이러한 어려운 문제를 풀기위해 다양한 연구가 진행되고 있다. 본 연구의 매스킹 방안을 평가하기 위해 각종 노출제어 및 자료유용성 측도를 살펴보고, 실제 자료를 분석한 결과가 이러한 발전에 디딤돌이 되기를 바란다.

<참고문헌>

이용희, 김용대 (2011). 에듀데이타의 통계적 노출관리기법 연구, 한국교육학술정보원.

김경미, 임경은 (2012). 가계금융·복지조사 자료 비밀보호방법 연구, 통계개발원.

박민정, 김경미 (2013). 종단자료 비밀보호의 국제 연구동향 및 향후 추진방향, 통계개발원.

박민정, 권순필, 심규호 (2013). 가계금융·복지조사 마이크로데이터 제공을 위한 매스킹 방안, 통계개발원.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata, Journal of the American Statistical Association, 85(409), 38-45.

Domingo-Ferrer, J. and Torra, V. (2001). Chapter5. Disclosure control methods and information loss for microdata, Confidentiality, Disclosure and Data Access, North Holland.

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

Duncan, G. T., Keller-McNulty, S. A. and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map, Chance.

Fryzlewicz, P. and Oh, H.-S. (2011). Thick-pen transform for time series, Journal of the Royal Statistical Society B, 73(4), 499-529.

Manrique-Vallier, D. and Reiter, J. (2012). Estimating identification disclosure risk using mixed memebership models, Journal of the American Statistical Association, 107(500), 1385-1394.

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

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(3), 1-9.

Oganian, A. and Karr, A. F. (2006a). Combinations of SDC methods for microdata protection, In: Privacy in Statistical Databases, Springer Lecture Notes in Computer Science Volumn 4302, 102-113.

Oganian, A. and Karr, A. F. (2006b). Multistage masking methods for microdata protection, MSRI Women in Mathematics: MAY 18-20.

Oganian, A. and Karr, A. F. (2011). Masking methods that preserve positivity constraints in microdata, Journal of Statistical Planning and Inference, 141, 31-41.

Reiter, J. (2005). Estimating risks of identification disclosure in microdata, Journal of the American Statistical Association, 100(472), 1103-1112.

Skinner, C. and Shlomo, N. (2008). Assessing identification risk in survey microdata using

log-linear models, Journal of the American Statistical Association, 103(483), 989-1001.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata, Journal of Business and Economic Statistics, 6, 487-500.

Templ, M. and Meindl, B. (2008a) Robustification of microdata masking methods and the comparison with existing methods. Lecture Notes in Computer Science, Privacy in Statistical Databases, vol 5262, 113-126, Springer.

Templ, M. and Meindl, B. (2008b) Robust statistics meets SDC: new disclosure risk measures for continuous microdata masking. Lecture Notes in Computer Science, Privacy in Statistical Databases, vol 5262, 177-189, Springer.

Templ, M. and Meindl, B. (2010) Practical applications in statistical disclosure control using R. In J. Nin and J. Herranz, editors, Privacy and Anonymity in Information Management Systems, Advanced Information and Knowledge Processing, pages 31-62. Springer London.

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(1), 111-124.

Reiter, J. P. (2004), New approaches to data dissemination: A glimpse into the future, Chance, 17:3 (Summer 2004), 12-16.

<부록 1>

침입자 의사결정론 기반 노출위험에 대한 이해를 돕기 위해 Reiter(2005)에 나오는 CPS 자료의 예를 이용하여 노출위험을 계산하는 과정을 살펴보도록 하자. 사용된 CPS 자료의 변수는 총 9개로 다음 표와 같다. 모집단의 수는 N = 104,781,947로 추정하며 조사된 자료의 크기는 n = 51,016이다.

변수	범주수 혹은 범위
X (sex)	2 (male, female)
R (race)	4 (white, black, american indian, asian)
M (marital status)	7
E (highest attained education level)	16
G (age, years)	15 ~ 90
C (child support payments, $)	0, 1 ~ 23,917
S (social security payments, $)	0, 1 ~ 50,000
P (household property taxes, $)	0, 1 ~ 99,997
I (household income, $)	-21,011 ~ 768,742

이제 다음과 같이 변수들에 대해 매스킹 기법이 적용되었다고 하자.

	A				U				
변수	X	R	M	G	E	C	S	P	I
매스킹 처리	-	자료 교환	자료 교환	재코딩	-	-	-	잡음 추가	-

외부인이 가진 표적은 (X, R, M, G)의 4개 변수에 대한 값을 가지고 있고, 나머지 5개 변수에 대해서는 정보가 없다고 하자. 또한 노출위험을 추정할 개체에 대하여 다음과 같은 4가지 유형을 생각하자.

표적의 유형		X	G	R	M	P	I
1 (Everyman)	원래 값	M	43	1	1	635	40000
	변환 값	M	40-44	1	1	596	40000
2 (Unique)	원래 값	F	39	3	3	0	12700
	변환 값	F	35-39	3	1	0	12700
3 (Big I)	원래 값	M	57	1	1	1100	768742
	변환 값	M	55-59	1	1	1210	768742
4 (Big P)	원래 값	F	79	1	4	99997	94552
	변환 값	F	75-79	1	1	100033	94552

각 유형에 대하여 매스킹 적용 방안은 다음과 같다. 아무 변수에도 매스킹 처리를 하지 않은 경우(-), 나이만 재코딩 된 경우(G), 인종과 혼인 상태만 자료 교환된 경우(R,M) 및 세 변수 모두 매스킹 처리된 경우(G,R,M)를 다루었다. 다음 표는 각 매스킹 방안 별로 노출위험 결과를 보여준다. 괄호 안에는 크기 n = 51,016인 표본에서 표적과 같은 키조합을 가지는 개체들의 개수가 나타나 있다.

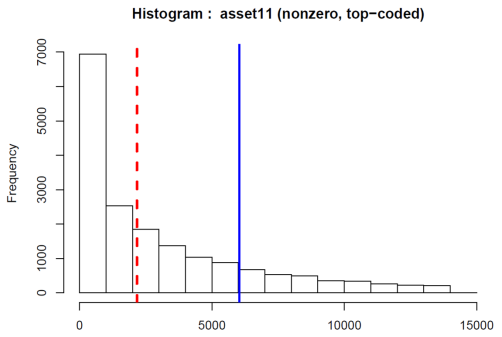
외부인이 표적이 표본에 있는 것을 아는 경우

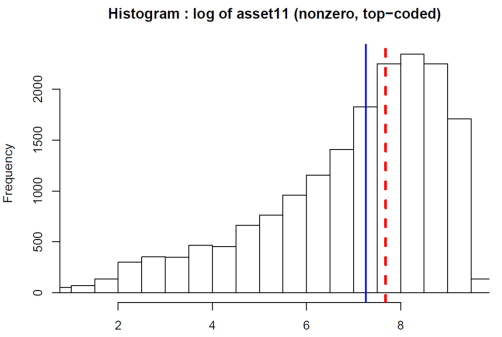
매스킹 처리	-	G	R,M	G,R,M
표적 유형 1 (Everyman)	.0022 (455)	.00045 (2230)	.0023 (410)	.00045 (2026)
표적 유형 2 (Unique)	1 (1)	1 (1)	.022 (10)	.0047 (49)
표적 유형 3 (Big I)	.0029 (345)	.00067 (1498)	.0031 (300)	.00069 (1355)
표적 유형 4 (Big P)	.0060 (166)	.0013 (776)	.0046 (169)	.00097 (836)

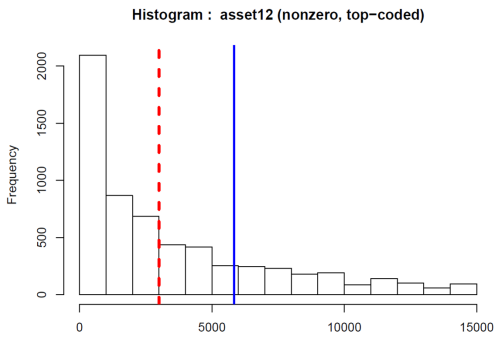
외부인이 표적이 표본에 있는지 모르는 경우

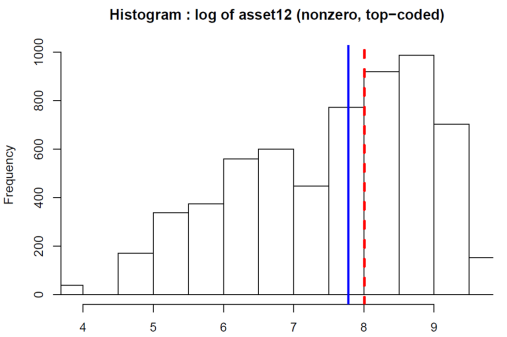
매스킹 처리	-	G	R,M	G,R,M
표적 유형 1 (Everyman)	.000001	.0000002	.000001	.0000002
표적 유형 2 (Unique)	.00032	.00032	.000001	.0000003
표적 유형 3 (Big I)	.000002	.0000003	.000002	.0000003
표적 유형 4 (Big P)	.000003	.0000006	.000002	.0000005

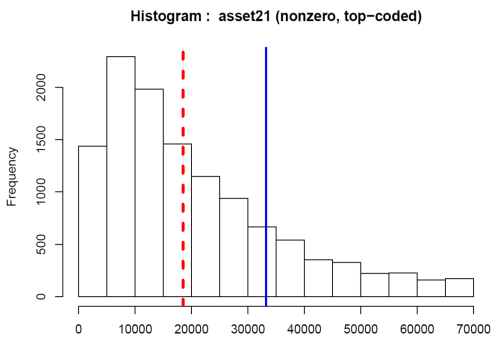
<부록 2>

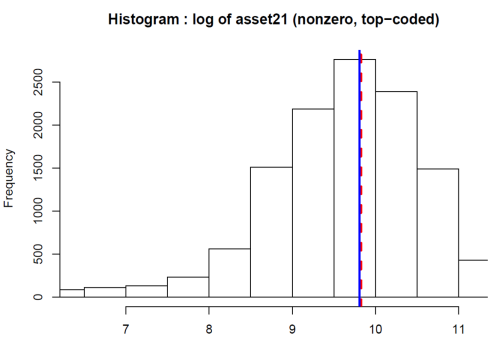

Histogram : asset11 (nonzero, top-coded)
Frequency
0
1000
3000
5000
7000
0
5000
10000
15000

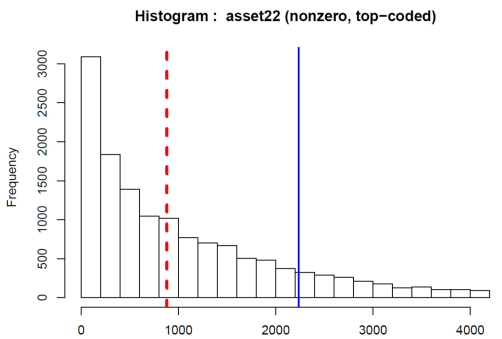

Histogram : log of asset11 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
2
4
6
8

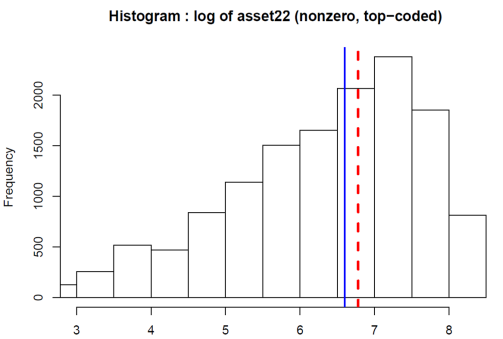

Histogram : asset12 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
0
5000
10000
15000

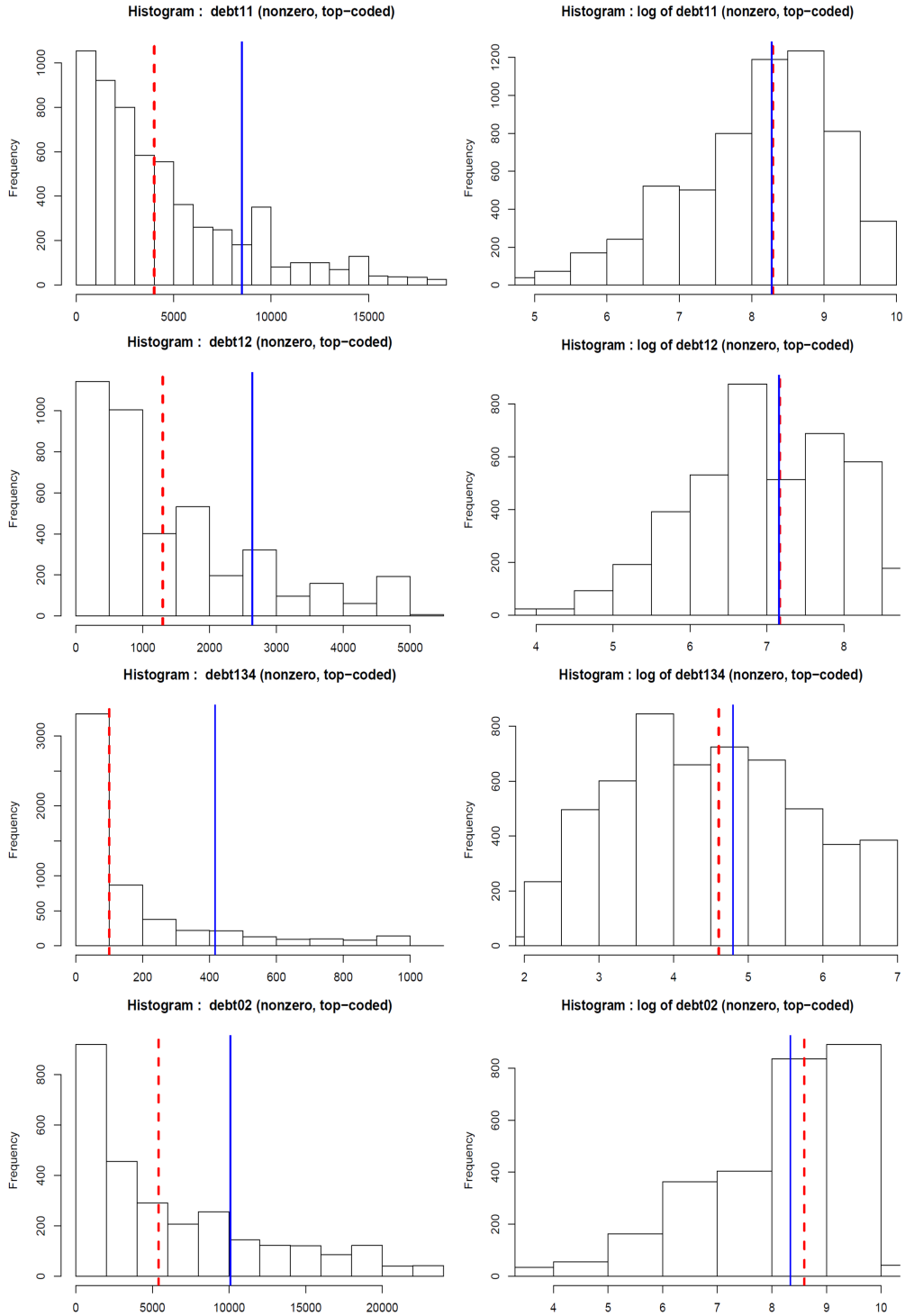

Histogram : log of asset12 (nonzero, top-coded)
Frequency
0
200
400
600
800
1000
4
5
6
7
8
9

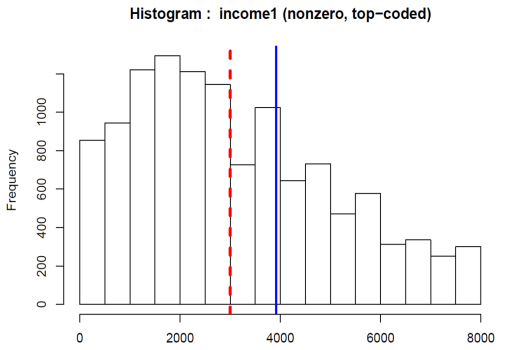

Histogram : asset21 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
0
10000
20000
30000
40000
50000
60000
70000

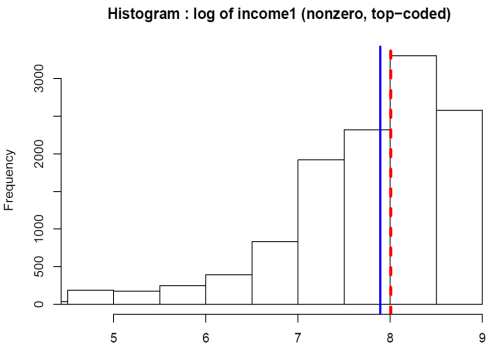

Histogram : log of asset21 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
2500
7
8
9
10
11

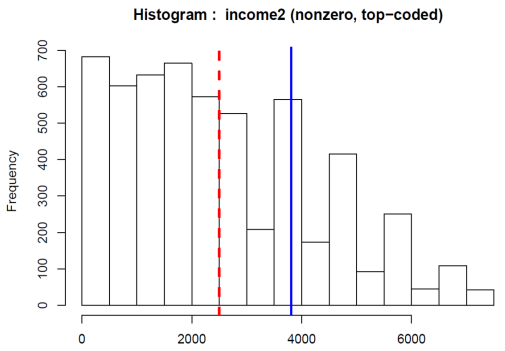

Histogram : asset22 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
2500
3000
0
1000
2000
3000
4000

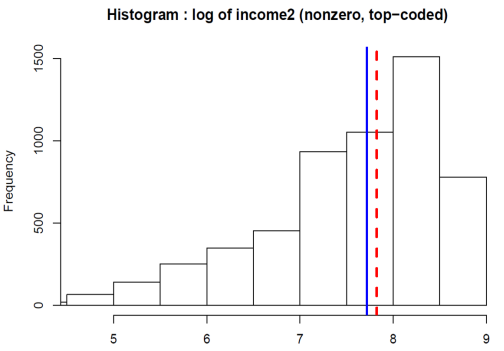

Histogram : log of asset22 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
3
4
5
6
7
8

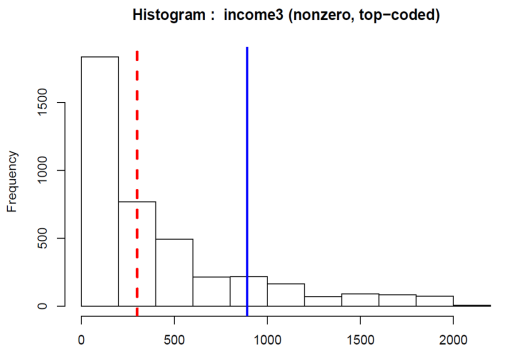

Histogram : debt11 (nonzero, top-coded)
Histogram : log of debt11 (nonzero, top-coded)
Histogram : debt12 (nonzero, top-coded)
Histogram : log of debt12 (nonzero, top-coded)
Histogram : debt134 (nonzero, top-coded)
Histogram : log of debt134 (nonzero, top-coded)
Histogram : debt02 (nonzero, top-coded)
Histogram : log of debt02 (nonzero, top-coded)
Frequency

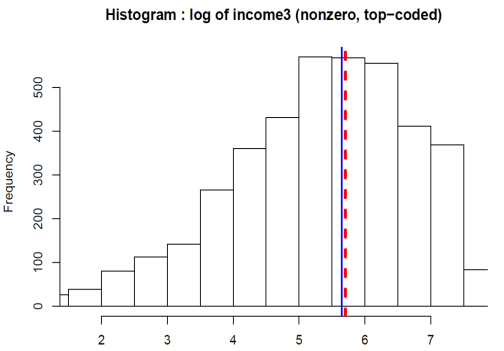

Histogram : income1 (nonzero, top-coded)
Frequency
0
200
400
600
800
1000
0
2000
4000
6000
8000

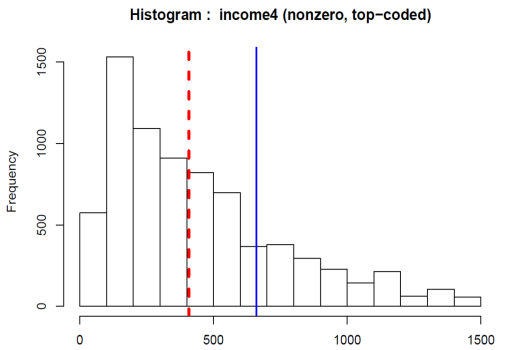

Histogram : log of income1 (nonzero, top-coded)
Frequency
0
500
1000
2000
3000
5
6
7
8
9

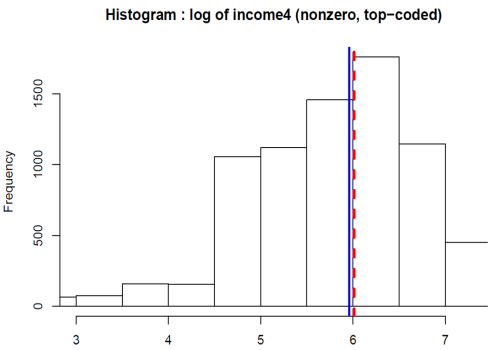

Histogram : income2 (nonzero, top-coded)
Frequency
0
100
200
300
400
500
600
700
0
2000
4000
6000


Histogram : log of income2 (nonzero, top-coded)
Frequency
0
500
1000
1500
5
6
7
8
9


Histogram : income3 (nonzero, top-coded)
Frequency
0
500
1000
1500
0
500
1000
1500
2000


Histogram : log of income3 (nonzero, top-coded)
Frequency
0
100
200
300
400
500
2
3
4
5
6
7


Histogram : income4 (nonzero, top-coded)
Frequency
0
500
1000
1500
0
500
1000
1500


Histogram : log of income4 (nonzero, top-coded)
Frequency
0
500
1000
1500
3
4
5
6
7

<부록 3>

M0 (국소통합(microaggregation) MDAV기법)						
				risk	IL1	IL_eigen
M0				0.3094	0.0904	0.3152

M1 (전체 상관관계를 고려한 로버스트한 잡음추가 기법)						
			$w(\%)$	risk	IL1	IL_eigen
M1(1)			5	0.1217	0.0484	0.0445
M1(11)			50	0.1229	0.0484	0.0452

M2 (전체 상관관계를 고려하고, 양수 조건 반영)						
	c	w_1	w_2	risk	IL1	IL_eigen
M2(1)	0.05	0.0241	0.9759	0.2379	0.1327	0.1886
M2(2)	0.075	0.0355	0.9645	0.2089	0.1459	0.2545
M2(3)	0.1	0.0465	0.9535	0.1871	0.1575	0.3888
M2(4)	0.125	0.0572	0.9428	0.1744	0.1673	0.37
M2(5)	0.15	0.0675	0.9325	0.152	0.173	0.4261
M2(6)	0.175	0.0775	0.9225	0.1492	0.1842	0.2889
M2(7)	0.2	0.0871	0.9129	0.0711	0.1904	0.3704
M2(8)	0.225	0.0965	0.9035	0.0789	0.1974	0.3495
M2(9)	0.25	0.1056	0.8944	0.0565	0.205	0.6244
M2(10)	0.275	0.1144	0.8856	0.06	0.2108	0.8243
M2(11)	0.3	0.1229	0.8771	0.0483	0.2148	0.7439

M3 (전체 상관관계를 고려하지 않고, 변수별로 양수 조건 반영)						
	c	w_1	w_2	risk	IL1	IL_eigen
M3(1)	0.05	0.0241	0.9759	0.3631	0.0717	0.2781
M3(2)	0.075	0.0355	0.9645	0.2886	0.0878	0.3198
M3(3)	0.1	0.0465	0.9535	0.2409	0.101	0.6582
M3(4)	0.125	0.0572	0.9428	0.1826	0.1127	0.4267
M3(5)	0.15	0.0675	0.9325	0.1526	0.1202	0.6479
M3(6)	0.175	0.0775	0.9225	0.1287	0.1307	0.8225
M3(7)	0.2	0.0871	0.9129	0.1091	0.1384	0.9999
M3(8)	0.225	0.0965	0.9035	0.0919	0.1453	0.876
M3(9)	0.25	0.1056	0.8944	0.0872	0.152	0.9597
M3(10)	0.275	0.1144	0.8856	0.077	0.1581	1.1294
M3(11)	0.3	0.1229	0.8771	0.044	0.1644	1.3122

C1 (M0의 결과에 M1에서 얻은 잡음추가)						
			$w(\%)$	risk	IL1	IL_eigen
C1(1)			5	0.0743	0.112	0.2837
C1(11)			50	0.0789	0.1117	0.2826

C2 (M0의 결과에 M2에서 얻은 잡음추가)						
	c	w_1	w_2	risk	IL1	IL_eigen
C2(1)	0.05	0.0241	0.9759	0.1519	0.1716	0.3134
C2(2)	0.075	0.0355	0.9645	0.1372	0.1845	0.2559
C2(3)	0.1	0.0465	0.9535	0.1204	0.1947	0.3527
C2(4)	0.125	0.0572	0.9428	0.1026	0.2036	0.3581
C2(5)	0.15	0.0675	0.9325	0.0777	0.2091	0.404
C2(6)	0.175	0.0775	0.9225	0.0697	0.2192	0.243
C2(7)	0.2	0.0871	0.9129	0.0496	0.2249	0.3373
C2(8)	0.225	0.0965	0.9035	0.0484	0.2308	0.3619
C2(9)	0.25	0.1056	0.8944	0.0418	0.2385	0.4416
C2(10)	0.275	0.1144	0.8856	0.0418	0.2444	0.8152
C2(11)	0.3	0.1229	0.8771	0.0338	0.2487	0.7889

C3 ((M0의 결과에 M3에서 얻은 잡음추가))						
	c	w_1	w_2	risk	IL1	IL_eigen
C3(1)	0.05	0.0241	0.9759	0.2126	0.122	0.3008
C3(2)	0.075	0.0355	0.9645	0.1822	0.1342	0.3517
C3(3)	0.1	0.0465	0.9535	0.1616	0.1445	0.6077
C3(4)	0.125	0.0572	0.9428	0.1363	0.1537	0.4048
C3(5)	0.15	0.0675	0.9325	0.1179	0.1604	0.5878
C3(6)	0.175	0.0775	0.9225	0.0962	0.1694	0.8179
C3(7)	0.2	0.0871	0.9129	0.0817	0.1762	1.0497
C3(8)	0.225	0.0965	0.9035	0.066	0.1825	0.8334
C3(9)	0.25	0.1056	0.8944	0.0636	0.1876	0.9744
C3(10)	0.275	0.1144	0.8856	0.0547	0.1936	1.1368
C3(11)	0.3	0.1229	0.8771	0.0341	0.1985	1.2469

<부록 4>

	평균의 ARB												S
	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4	corr
M0	0.0003	0.0255	0.0143	0.0172	0.0755	0.0871	0.0759	0.0463	0.0173	0.0394	0.0727	0.0301	1.0565
M1(1)	0.0608	0.0032	0.0051	0.052	0.0046	0.0005	0.0289	0.0003	0.0005	0.0006	0.0106	0.0004	0.0998
M1(2)	0.0599	0.0046	0.0042	0.0464	0.002	0.0002	0.0281	0.0018	0.0012	0.0025	0.0113	0.0021	0.0947
M1(3)	0.0596	0.0042	0.007	0.0476	0.0049	0.0024	0.0278	0.0011	0.0002	0.0012	0.0109	0.0001	0.1106
M1(4)	0.0591	0.005	0.006	0.0473	0.008	0.0014	0.0266	0.0002	0	0.001	0.0119	0.0018	0.0998
M1(5)	0.0588	0.0047	0.0057	0.0514	0.0041	0.0026	0.0265	0.0023	0.0003	0.0012	0.0094	0.0008	0.1061
M1(6)	0.0609	0.0056	0.0051	0.0483	0.0059	0.0016	0.026	0.0013	0.0001	0.0012	0.0073	0.0004	0.0943
M1(7)	0.0581	0.0031	0.0042	0.0521	0.0063	0.0022	0.0277	0.0014	0.001	0.0016	0.0113	0.001	0.0982
M1(8)	0.0583	0.0043	0.0056	0.0498	0.0072	0.0036	0.0256	0.0006	0.0004	0.0009	0.0105	0.0007	0.0987
M1(9)	0.0599	0.0052	0.0041	0.0478	0.0036	0.0015	0.0253	0.0001	0.0003	0.0011	0.0096	0.0006	0.1054
M1(10)	0.0617	0.0049	0.0058	0.0492	0.0042	0.0032	0.0262	0.0002	0.0005	0.0013	0.0101	0.0016	0.0991
M1(11)	0.0611	0.004	0.0049	0.0497	0.0066	0.001	0.029	0.0018	0.0017	0.0009	0.0113	0.0007	0.0919
M2(1)	0.0408	0.0014	0.006	0.0395	0.0031	0.0226	0.013	0.0095	0.0028	0.0044	0.0302	0.0113	0.8278
M2(2)	0.0365	0.0207	0.0072	0.0156	0.0173	0.0389	0.0122	0.027	0.0012	0.0205	0.0251	0.0024	0.8029
M2(3)	0.0395	0.0176	0.0075	0.0393	0.021	0.0037	0.0258	0.0466	0.002	0.0036	0.0354	0.0198	1.1587
M2(4)	0.038	0.0081	0.0065	0.0261	0.0131	0.0113	0.0154	0.033	0.0039	0.022	0.0327	0.0186	1.1295
M2(5)	0.0492	0.0207	0.0029	0.0384	0.0264	0.0303	0.0151	0.0381	0.0003	0.0291	0.0502	0.012	0.8297
M2(6)	0.0665	0.0207	0.0011	0.0175	0.0259	0.0253	0.0394	0.051	0.0015	0.034	0.0624	0.0272	0.8838
M2(7)	0.0466	0.0246	0.0036	0.0517	0.0662	0.0139	0.0343	0.0635	0.0149	0.0503	0.0549	0.0158	1.1127
M2(8)	0.0695	0.0457	0.0106	0.0373	0.0467	0.0229	0.0135	0.0707	0.0139	0.0475	0.0642	0.0535	1.163
M2(9)	0.0448	0.0415	0.0014	0.0134	0.0537	0.0883	0.0527	0.0735	0.0221	0.058	0.0689	0.0345	1.2381
M2(10)	0.0671	0.0498	0.0032	0.0266	0.056	0.044	0.0412	0.0788	0.0191	0.0687	0.0073	0.0523	1.6772
M2(11)	0.0493	0.0453	0.0161	0.0182	0.071	0.0707	0.0645	0.0846	0.0179	0.0711	0.0789	0.0467	1.7801

주: S는 상관계수 차이 절대값의 합

	평균의 ARB												S
	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4	corr
M3(1)	0.0005	0.0042	0.0027	0.0053	0.0027	0.0218	0.0064	0.0015	0.0005	0.001	0.0123	0.0015	0.4793
M3(2)	0.0018	0.0017	0.0045	0.0051	0.008	0.0111	0.0037	0.0037	0.0002	0.0029	0.0024	0.0051	0.5961
M3(3)	0.0003	0.0014	0.0001	0.0023	0.0031	0.0022	0.0195	0.0054	0.0007	0.0096	0.013	0.0013	0.9643
M3(4)	0.0012	0.003	0.0018	0.0142	0.002	0.013	0.0025	0.0217	0.0016	0.0032	0.0038	0.0016	0.7857
M3(5)	0.0017	0.0042	0	0.0049	0.0098	0.0111	0.0014	0.007	0.0029	0.0122	0.0188	0.0006	1.1344
M3(6)	0.0122	0.0143	0.0015	0.0115	0.0172	0.0049	0.0263	0.0147	0.0005	0.0047	0.0023	0.0086	1.1839
M3(7)	0.0085	0.0007	0.0081	0.0222	0.0081	0.0124	0.0188	0.0028	0.0024	0.0129	0.01	0.0075	1.5448
M3(8)	0.0103	0.0059	0.0074	0.01	0.0044	0.0185	0.0136	0.014	0.0035	0.0073	0.0195	0.0095	1.3397
M3(9)	0.0052	0.0032	0.009	0.0153	0.0091	0.0199	0.0083	0.0125	0.0003	0.0023	0.0096	0.0106	1.3702
M3(10)	0.0006	0.0186	0.001	0.0105	0.0072	0.0075	0.0286	0.009	0.0027	0.0114	0.0298	0.0055	1.8582
M3(11)	0.0028	0.0098	0.0055	0.025	0.0092	0.0031	0.0226	0.0229	0.0028	0.0079	0.0063	0.0079	1.8278
C1(1)	0.0604	0.0223	0.0092	0.0349	0.071	0.0865	0.047	0.046	0.0178	0.04	0.0621	0.0305	1.0852
C1(2)	0.0595	0.021	0.0102	0.0293	0.0735	0.0869	0.0477	0.0445	0.0162	0.0418	0.0614	0.0322	1.0827
C1(3)	0.0593	0.0214	0.0073	0.0305	0.0706	0.0846	0.0481	0.0474	0.0176	0.0382	0.0618	0.0302	1.0921
C1(4)	0.0587	0.0205	0.0083	0.0302	0.0675	0.0856	0.0493	0.0461	0.0174	0.0404	0.0608	0.0319	1.0856
C1(5)	0.0585	0.0208	0.0086	0.0343	0.0714	0.0844	0.0493	0.044	0.017	0.0382	0.0633	0.0309	1.0943
C1(6)	0.0606	0.02	0.0092	0.0312	0.0696	0.0854	0.0499	0.045	0.0172	0.0406	0.0654	0.0305	1.0783
C1(7)	0.0578	0.0225	0.0102	0.0349	0.0693	0.0849	0.0481	0.0477	0.0183	0.041	0.0614	0.0311	1.0888
C1(8)	0.0579	0.0212	0.0087	0.0326	0.0683	0.0834	0.0502	0.0457	0.017	0.0402	0.0622	0.0308	1.0876
C1(9)	0.0595	0.0203	0.0102	0.0306	0.0719	0.0855	0.0506	0.0462	0.0176	0.0405	0.0631	0.0307	1.0873
C1(10)	0.0613	0.0206	0.0085	0.032	0.0713	0.0838	0.0496	0.0465	0.0168	0.038	0.0626	0.0317	1.0795
C1(11)	0.0608	0.0216	0.0095	0.0326	0.0689	0.086	0.0468	0.0445	0.0156	0.0403	0.0614	0.0308	1.0836

주: S는 상관계수 차이 절대값의 합

	평균의 ARB												S
	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4	corr
C2(1)	0.0405	0.0242	0.0084	0.0223	0.0787	0.1097	0.0889	0.0558	0.0201	0.0438	0.1029	0.0414	1.3183
C2(2)	0.0361	0.0462	0.0071	0.0015	0.0928	0.1259	0.0881	0.0733	0.0161	0.0599	0.0978	0.0325	1.383
C2(3)	0.0392	0.0432	0.0068	0.0221	0.0966	0.0834	0.1016	0.0928	0.0153	0.0358	0.1081	0.0499	1.4924
C2(4)	0.0376	0.0336	0.0079	0.009	0.0886	0.0984	0.0912	0.0793	0.0212	0.0614	0.1054	0.0487	1.3729
C2(5)	0.0488	0.0463	0.0115	0.0213	0.1019	0.1173	0.091	0.0844	0.017	0.0685	0.1229	0.0421	1.4646
C2(6)	0.0662	0.0462	0.0155	0.0003	0.1014	0.1124	0.1152	0.0972	0.0188	0.0733	0.1351	0.0573	1.4178
C2(7)	0.0462	0.0501	0.0107	0.0345	0.1417	0.101	0.1101	0.1098	0.0322	0.0897	0.1276	0.0459	1.5668
C2(8)	0.0692	0.0712	0.0037	0.0201	0.1222	0.11	0.0894	0.117	0.0312	0.0869	0.1369	0.0836	1.6655
C2(9)	0.0444	0.0671	0.0157	0.0038	0.1292	0.1753	0.1286	0.1197	0.0394	0.0974	0.1416	0.0646	1.5651
C2(10)	0.0668	0.0754	0.0175	0.0095	0.1315	0.131	0.1171	0.1251	0.0364	0.1081	0.08	0.0824	2.1779
C2(11)	0.0489	0.0709	0.0304	0.001	0.1465	0.1578	0.1403	0.1309	0.0353	0.1104	0.1516	0.0768	2.2238
C3(1)	0.0001	0.0213	0.0117	0.0119	0.0782	0.0653	0.0822	0.0448	0.0168	0.0384	0.0604	0.0286	1.3686
C3(2)	0.0022	0.0272	0.0188	0.012	0.0675	0.0981	0.0721	0.0425	0.0175	0.0365	0.0703	0.0249	1.4775
C3(3)	0.0007	0.0269	0.0143	0.0149	0.0786	0.0848	0.0564	0.0516	0.0167	0.0297	0.0597	0.0288	1.7594
C3(4)	0.0015	0.0225	0.0125	0.0313	0.0776	0.0741	0.0784	0.068	0.0189	0.0426	0.0689	0.0285	1.5851
C3(5)	0.0021	0.0298	0.0144	0.0122	0.0657	0.0982	0.0773	0.0533	0.0145	0.0516	0.0915	0.0307	1.9504
C3(6)	0.0126	0.0113	0.0128	0.0056	0.0927	0.0919	0.1022	0.0316	0.0178	0.0347	0.0704	0.0387	2.0498
C3(7)	0.0088	0.0248	0.0224	0.0393	0.0674	0.0747	0.057	0.0435	0.0149	0.0265	0.0827	0.0376	2.4037
C3(8)	0.0106	0.0315	0.0218	0.0271	0.0711	0.1055	0.0895	0.0603	0.0208	0.0467	0.0532	0.0396	2.1991
C3(9)	0.0049	0.0288	0.0233	0.0324	0.0664	0.107	0.0676	0.0587	0.0176	0.0417	0.0631	0.0195	2.3005
C3(10)	0.0003	0.0442	0.0154	0.0277	0.0683	0.0795	0.0472	0.0553	0.02	0.028	0.0429	0.0356	2.6983
C3(11)	0.0032	0.0157	0.0088	0.0078	0.0847	0.0839	0.0532	0.0691	0.0201	0.0473	0.079	0.0222	2.7288

주: S는 상관계수 차이 절대값의 합

<부록 5>

GPG	ARB	ARB.h	ARB.d	ARB.m	ARB.m.h	ARB.m.d
M0	0.0031	0.0203	0.0076	0.0014	0.0012	0.0012
M1(1)	0.0005	0.0182	0.0089	0.0001	0.0016	0.0011
M1(2)	0.0073	0.0183	0.0090	0.0033	0.0038	0.0035
M1(3)	0.0268	0.0189	0.0100	0.0008	0.0015	0.0012
M1(4)	0.0116	0.0155	0.0086	0.0001	0.0019	0.0012
M1(5)	0.0251	0.0228	0.0099	0.0016	0.0033	0.0028
M1(6)	0.0009	0.0206	0.0076	0.0017	0.0015	0.0014
M1(7)	0.0153	0.0189	0.0094	0.0036	0.0035	0.0034
M1(8)	0.0153	0.0154	0.0101	0.0027	0.0027	0.0025
M1(9)	0.0154	0.0207	0.0075	0.0003	0.0008	0.0007
M1(10)	0.0089	0.0163	0.0086	0.0006	0.0010	0.0009
M1(11)	0.0067	0.0151	0.0063	0.0046	0.0059	0.0055
M2(1)	0.0139	0.0377	0.0212	0.0242	0.0268	0.0260
M2(2)	0.0831	0.0726	0.0442	0.0341	0.0383	0.0371
M2(3)	0.0470	0.0537	0.0343	0.0281	0.0306	0.0298
M2(4)	0.0605	0.0522	0.0326	0.0393	0.0430	0.0418
M2(5)	0.0273	0.0794	0.0502	0.0406	0.0459	0.0443
M2(6)	0.0661	0.0859	0.0480	0.0424	0.0477	0.0463
M2(7)	0.1663	0.0931	0.0802	0.0627	0.0725	0.0691
M2(8)	0.0615	0.0937	0.0589	0.0560	0.0607	0.0593
M2(9)	0.1097	0.1083	0.0688	0.0655	0.0726	0.0705
M2(10)	0.0039	0.0953	0.0596	0.0656	0.0709	0.0692
M2(11)	0.1260	0.1165	0.0889	0.0725	0.0810	0.0783
M3(1)	0.0360	0.0276	0.0158	0.0145	0.0160	0.0154
M3(2)	0.0005	0.0328	0.0209	0.0235	0.0248	0.0243
M3(3)	0.0380	0.0407	0.0247	0.0282	0.0301	0.0295
M3(4)	0.0474	0.0515	0.0362	0.0363	0.0399	0.0387
M3(5)	0.0174	0.0601	0.0424	0.0423	0.0455	0.0443
M3(6)	0.0381	0.0596	0.0507	0.0546	0.0603	0.0582
M3(7)	0.0751	0.0734	0.0505	0.0560	0.0617	0.0600
M3(8)	0.0917	0.0728	0.0512	0.0574	0.0624	0.0608
M3(9)	0.0430	0.0773	0.0569	0.0599	0.0646	0.0632
M3(10)	0.0057	0.0967	0.0701	0.0696	0.0737	0.0720
M3(11)	0.1845	0.1041	0.0809	0.0767	0.0825	0.0804

GPG	ARB	ARB.h	ARB.d	ARB.m	ARB.m.h	ARB.m.d
C1(1)	0.0201	0.0289	0.0118	0.0004	0.0022	0.0015
C1(2)	0.0281	0.0283	0.0124	0.0003	0.0013	0.0006
C1(3)	0.0478	0.0326	0.0133	0.0038	0.0066	0.0057
C1(4)	0.0083	0.0280	0.0109	0.0006	0.0026	0.0018
C1(5)	0.0462	0.0347	0.0128	0.0037	0.0060	0.0053
C1(6)	0.0191	0.0314	0.0104	0.0009	0.0018	0.0010
C1(7)	0.0360	0.0311	0.0130	0.0004	0.0034	0.0018
C1(8)	0.0362	0.0278	0.0130	0.0009	0.0016	0.0009
C1(9)	0.0363	0.0273	0.0135	0.0003	0.0011	0.0007
C1(10)	0.0296	0.0283	0.0123	0.0026	0.0045	0.0040
C1(11)	0.0270	0.0232	0.0094	0.0070	0.0086	0.0080
C2(1)	0.0342	0.0460	0.0243	0.0271	0.0291	0.0284
C2(2)	0.1055	0.0812	0.0502	0.0379	0.0421	0.0409
C2(3)	0.0687	0.0587	0.0365	0.0333	0.0354	0.0346
C2(4)	0.0831	0.0533	0.0358	0.0481	0.0525	0.0511
C2(5)	0.0479	0.0934	0.0547	0.0454	0.0494	0.0482
C2(6)	0.0816	0.0973	0.0523	0.0498	0.0556	0.0542
C2(7)	0.1922	0.0997	0.0848	0.0707	0.0804	0.0769
C2(8)	0.0844	0.0997	0.0629	0.0671	0.0711	0.0696
C2(9)	0.1306	0.1192	0.0730	0.0760	0.0827	0.0807
C2(10)	0.0234	0.0986	0.0641	0.0747	0.0783	0.0773
C2(11)	0.1509	0.1278	0.0937	0.0829	0.0917	0.0888
C3(1)	0.0574	0.0397	0.0189	0.0176	0.0186	0.0182
C3(2)	0.0198	0.0442	0.0242	0.0267	0.0277	0.0273
C3(3)	0.0596	0.0463	0.0282	0.0314	0.0327	0.0323
C3(4)	0.0687	0.0568	0.0396	0.0419	0.0446	0.0437
C3(5)	0.0372	0.0684	0.0458	0.0474	0.0501	0.0490
C3(6)	0.0588	0.0700	0.0544	0.0605	0.0661	0.0641
C3(7)	0.0974	0.0835	0.0541	0.0626	0.0682	0.0665
C3(8)	0.1148	0.0796	0.0549	0.0643	0.0683	0.0671
C3(9)	0.0629	0.0868	0.0606	0.0685	0.0724	0.0713
C3(10)	0.0248	0.1033	0.0741	0.0774	0.0809	0.0794
C3(11)	0.2107	0.1091	0.0853	0.0862	0.0921	0.0900

Gini 계수	ARB	ARB.h	ARB.d	ARB.m	ARB.m.h	ARB.m.d
M0	0.0082	0.0144	0.0086	0.0006	0.0015	0.0011
M1(1)	0.0009	0.0078	0.0034	0.0018	0.0030	0.0025
M1(2)	0.0014	0.0109	0.0041	0.0023	0.0033	0.0030
M1(3)	0.0008	0.0110	0.0034	0.0009	0.0030	0.0019
M1(4)	0.0006	0.0080	0.0040	0.0008	0.0035	0.0017
M1(5)	0.0005	0.0080	0.0041	0.0004	0.0009	0.0006
M1(6)	0.0018	0.0072	0.0042	0.0007	0.0030	0.0013
M1(7)	0.0003	0.0080	0.0036	0.0026	0.0036	0.0032
M1(8)	0.0002	0.0065	0.0028	0.0013	0.0040	0.0026
M1(9)	0.0014	0.0078	0.0030	0.0015	0.0020	0.0020
M1(10)	0.0001	0.0078	0.0031	0.0027	0.0033	0.0032
M1(11)	0.0000	0.0079	0.0035	0.0007	0.0017	0.0015
M2(1)	0.0255	0.0458	0.0282	0.0330	0.0336	0.0326
M2(2)	0.0267	0.0462	0.0307	0.0395	0.0397	0.0385
M2(3)	0.0366	0.0495	0.0403	0.0386	0.0394	0.0385
M2(4)	0.0322	0.0544	0.0342	0.0521	0.0526	0.0511
M2(5)	0.0428	0.0616	0.0407	0.0524	0.0538	0.0528
M2(6)	0.0498	0.0820	0.0472	0.0548	0.0548	0.0534
M2(7)	0.0463	0.0676	0.0424	0.0561	0.0586	0.0566
M2(8)	0.0468	0.0781	0.0480	0.0738	0.0736	0.0714
M2(9)	0.0506	0.0858	0.0503	0.0816	0.0826	0.0803
M2(10)	0.0547	0.0952	0.0625	0.0759	0.0761	0.0745
M2(11)	0.0557	0.0831	0.0584	0.0764	0.0776	0.0759
M3(1)	0.0020	0.0123	0.0050	0.0200	0.0209	0.0204
M3(2)	0.0083	0.0181	0.0105	0.0354	0.0359	0.0348
M3(3)	0.0055	0.0211	0.0118	0.0466	0.0472	0.0462
M3(4)	0.0085	0.0299	0.0124	0.0493	0.0508	0.0497
M3(5)	0.0122	0.0327	0.0164	0.0588	0.0605	0.0591
M3(6)	0.0087	0.0292	0.0191	0.0648	0.0676	0.0660
M3(7)	0.0114	0.0306	0.0143	0.0766	0.0779	0.0760
M3(8)	0.0217	0.0406	0.0220	0.0863	0.0873	0.0852
M3(9)	0.0238	0.0417	0.0255	0.0911	0.0919	0.0900
M3(10)	0.0169	0.0469	0.0215	0.0943	0.0970	0.0949
M3(11)	0.0267	0.0506	0.0320	0.1041	0.1065	0.1044

Gini 계수	ARB	ARB.h	ARB.d	ARB.m	ARB.m.h	ARB.m.d
C1(1)	0.0073	0.0178	0.0090	0.0024	0.0032	0.0022
C1(2)	0.0066	0.0188	0.0094	0.0015	0.0021	0.0013
C1(3)	0.0074	0.0200	0.0100	0.0017	0.0049	0.0026
C1(4)	0.0073	0.0181	0.0102	0.0045	0.0052	0.0043
C1(5)	0.0086	0.0163	0.0084	0.0058	0.0063	0.0063
C1(6)	0.0062	0.0154	0.0092	0.0047	0.0051	0.0047
C1(7)	0.0077	0.0188	0.0095	0.0013	0.0039	0.0020
C1(8)	0.0081	0.0168	0.0088	0.0001	0.0033	0.0016
C1(9)	0.0073	0.0158	0.0089	0.0030	0.0032	0.0032
C1(10)	0.0078	0.0153	0.0095	0.0001	0.0023	0.0013
C1(11)	0.0083	0.0150	0.0092	0.0046	0.0048	0.0044
C2(1)	0.0209	0.0454	0.0261	0.0362	0.0363	0.0352
C2(2)	0.0240	0.0449	0.0315	0.0411	0.0406	0.0393
C2(3)	0.0334	0.0531	0.0398	0.0408	0.0413	0.0402
C2(4)	0.0305	0.0547	0.0328	0.0592	0.0593	0.0575
C2(5)	0.0418	0.0641	0.0406	0.0602	0.0611	0.0600
C2(6)	0.0491	0.0842	0.0458	0.0582	0.0573	0.0558
C2(7)	0.0466	0.0694	0.0437	0.0594	0.0617	0.0593
C2(8)	0.0473	0.0793	0.0497	0.0789	0.0786	0.0762
C2(9)	0.0528	0.0892	0.0529	0.0902	0.0907	0.0883
C2(10)	0.0581	0.1004	0.0663	0.0894	0.0881	0.0867
C2(11)	0.0582	0.0871	0.0608	0.0808	0.0817	0.0799
C3(1)	0.0096	0.0177	0.0109	0.0218	0.0223	0.0217
C3(2)	0.0153	0.0208	0.0149	0.0376	0.0376	0.0364
C3(3)	0.0114	0.0258	0.0146	0.0494	0.0497	0.0485
C3(4)	0.0140	0.0321	0.0152	0.0567	0.0575	0.0562
C3(5)	0.0174	0.0338	0.0205	0.0631	0.0643	0.0626
C3(6)	0.0134	0.0345	0.0232	0.0702	0.0726	0.0707
C3(7)	0.0166	0.0334	0.0158	0.0825	0.0833	0.0811
C3(8)	0.0261	0.0402	0.0249	0.0949	0.0953	0.0930
C3(9)	0.0268	0.0449	0.0288	0.1003	0.1005	0.0984
C3(10)	0.0200	0.0501	0.0235	0.1045	0.1063	0.1039
C3(11)	0.0298	0.0531	0.0356	0.1120	0.1138	0.1115