통계개발원 연구보고서

개인정보보호와 데이터 활용을 위한 통계적 방법론(Ⅰ)

– 마이크로데이터에 적용한 비밀보호 기법: 이론과 활용 –

[1부] 마이크로데이터의 활용도를 높인 매스킹 기법

- 마이크로데이터 제공 현황에 관한 해외사례 연구 (2007) 김경미 · 이의규 · 정미옥
- 마이크로자료의 활용도제고를 위한 비밀보호방법 (2006) 정동명 · 강동환

[2부] 매스킹 기법의 적용과 평가

- 2005 인구주택총조사자료의 개인정보 노출제한방법 (2007) 정동명 · 정미옥
- 가계조사 마이크로데이터의 비밀보호 (2007) 정동명 · 정남수 · 한승훈
- 승법잡음모형을 이용한 가계조사 자료의 비밀보호 (2008) 정동명 · 김경미
- 가계금융 · 복지조사 마이크로데이터 제공을 위한 매스킹 방안 (2013) 박민정 · 권순필 · 심규호
- 가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가 (2014) 박민정

[3부] 매스킹 기법의 실무활용

- 마이크로데이터 매스킹 기법 실무 활용 안내서 (2015) 박민정
- 통계작성기관 마이크로데이터 제공범위 설정 가이드라인 작성 (2018) 박민정 · 성병찬 · 최유성

개인정보보호와 데이터 활용을 촉진시키는 연구서의 발간에 붙이며~

데이터는 영원한 생명체와 같다. 사람은 죽지만 데이터는 남는다. 기업은 소멸해도 데이터는 존재한다. 데이터는 4차 산업혁명에 생명을 불어넣어 데이터 세계경제를 탄생시켰으며, 21세기 미래성장의 원천이 되었다. 그러나 실제로 이의 실현적 측면에서는 「데이터 활용」의 전제 조건인 데이터의 공유에서부터 현실의 벽에 부딪히고 있다. 「개인정보보호」라는 양날의 검이 앞을 가로막고 있기 때문이다.

최근 데이터 3법이 법제화되면서 길이 열렸다. 개인정보보호법·정보통신망법·신용정보법이 모두 국회를 통과하면서 데이터 이용 규제가 대폭 완화되고, 빅데이터, 공공데이터, 국가통계데이터를 적절히 연계하여 활용할 수 있는 계기가 마련되었기 때문이다. 따라서 바이오, 금융, 인공지능 산업을 망라하는 새로운 데이터경제 성장의 지평이 열린 것이다.

그러나 다른 한편으로는 인권과 연관된 개인정보에 대해 법안이 허술하게 만들어졌다는 시민사회의 비판에 직면하고 있다. 이는 데이터의 연계가 정확성을 가질수록 개인정보 노출위험이 증폭된다는 것을 의미한다. 이렇듯 데이터의 활용과 개인정보 보호라는 두 측면의 상호충돌 관계가 현재 우리가 당면한 딜레마이다.

각 나라의 통계기관은 오랫동안 이 문제를 다루어왔다. 개별 정보의 노출위험을 제어하면서도 유용한 데이터를 제공하는 것은 국가통계기관이 수행해야 할 가장 중요한 업무 중 하나이기 때문이다. 한국 통계청은 통계개발원(Statistics Research Institute)을 중심으로 안전하고도 유용하며, 사용하기 편리한 데이터를 제공하기 위한 통계적 비밀보호 방안을 꾸준히 연구해오고 있다.

이러한 연구는 어디에서 출발하여, 어떠한 발전들이 있었을까?

통계적 비밀보호 기법의 진화

통계적 비밀보호 분야의 발전 초창기에는 데이터를 적절히 변형하여 제공하는 다양한 방법이 연구되었으며, 매스킹(masking) 기법이 그 시작이었다. 다양한 매스킹 방법들의 안정성과 유용성 정도를 판단하기 위해서 개인정보 노출위험과 자료의

정보손실을 측정하고 비교할 필요성이 대두되었으며, 이에 대한 연구가 발전하기 시작했다.

매스킹 기법을 사용할 때 노출위험과 정보손실 관점에서 합리적인 결론을 도출하기 어렵다는 것이 알려졌고, 동시에 매스킹 기법을 대체하기 위한 새로운 시도들이 이루어졌다. 그 결과 베이지안 재현자료(Bayesian synthetic data)와 차등정보보호(differential privacy)가 출현하였다. 최근에는 미국 통계청을 중심으로 차등정보보호가 국가통계에서 실용화되고 있으며, 애플, 구글 등의 데이터 관련 기업에서도 이미 이러한 방법들을 활용하고 있다.

방법론의 빠른 발전과 함께 용어 사용에도 변화가 일어났다. 초창기에는 매스킹(masking)이라는 용어가 주로 사용되지만, 노출위험과 정보손실 측도 연구를 포괄하여 통계적 노출제어(statistical disclosure control/limitation) 혹은 통계적 비밀보호(statistical data confidentiality)라는 용어가 널리 사용되었다. 차등정보보호(differential privacy)의 출현 이후 최근에는 데이터 정보보호(data privacy)라는 용어가 보편적으로 사용되고 있다.

통계개발원은 국내에서 아직 생소한 통계적 비밀보호, 즉 「데이터 정보보호」의 이해와 활용도를 높이기 위하여 그간의 주요 연구를 정리하여 두 권의 책자로 제작하게 되었다. 제1권은 「마이크로데이터에 적용한 비밀보호기법: 이론과 활용」을, 제2권은 「데이터 정보보호를 위한 최신 기법의 활용」을 다루고 있다.

제1권은 매스킹에 관한 국내·외 연구를 종합하여 정리하였다. 매스킹 기법의 소개부터 실무 활용을 위한 안내 그리고 가이드라인 프로그램의 작성을 제시하였다. 하지만 2006년 이래 초창기에 연구되었던 매스킹 기법으로 데이터 정보보호를 수행하기에는 한계가 많았다. 데이터 제공에 있어 특히 안전성과 유용성을 정량적으로 보장할 수 없었으며, 이는 데이터 제공을 위한 과학적인 비밀보호 방안으로서 매스킹 기법에 한계가 있다는 것을 의미한다.

이에 따라 제2권은 매스킹 기법의 대안으로 제시된 베이지안 재현자료 및 차등정보보호에 대한 연구결과를 중심으로 구성되었다. 즉 통계개발원이 2016년부터 발전시켜 오고 있는 데이터 정보보호를 위한 최신 기법을 활용한 대용량의 빈도표 제공, 재현자료, 차등정보보호의 이론과 적용을 다루고 있다.

연구서 1권과 2권을 각각 더 자세히 살펴보면 다음과 같다.

마이크로데이터에 적용한 비밀보호 기법: 이론과 활용 (제1권)

제1권 제1부는 데이터 제공 시 개인정보보호의 문제를 인식하고, 이를 해결하기 위한 첫 발걸음을 담고 있다. 이는 2006년 통계개발원의 개원와 함께 도전한 혁신적인 연구과제이다. 제1장에서는 통계 작성을 위하여 사용하는 상세 정보인 마이크로데이터를 통계와 함께 제공할 때 국제적으로 적용하는 제도들과 통계 기법들을 정리했다. 제2장에서는 당시 국내에서 활용도가 낮았던 마이크로데이터를 국민들이 이용할 수 있도록 제공하기 위한 노력을 담고 있다. 이를 위해 전통적인 비밀보호 방법인 매스킹 기법을 중심으로 그 개념과 상세 알고리즘 등 전반적인 내용을 국내에 처음으로 소개하고 있다.

제1권 제2부는 매스킹 기법의 활용을 다루는 네 개의 장으로 구성되어 있다. 제3장은 인구주택 총조사의 마이크로데이터를 안전하게 제공하기 위하여 매스킹 기법을 어떻게 적용할지 연구한 결과이다. 제4장은 이 연구결과를 바탕으로 민감한 개인정보를 포함하고 있는 가구조사 마이크로데이터의 비밀보호를 위한 매스킹 기법의 적용 방안을 제시하였다. 제5장에는 당시 최신 기법인 승법잡음모형을 가계조사 마이크로데이터에 적용한 연구 결과가 수록되어 있다. 한편, 2012년부터 본격적으로 실시되기 시작한 가계금융복지조사의 마이크로데이터 제공을 위한 매스킹 방안이 2013년에 심층연구 되었고, 그 연구 결과는 제6장에 제시되어 있다.

특정 마이크로데이터의 비밀보호를 위하여 수십 가지의 매스킹 방안이 만들어질 수 있다. 다양한 매스킹 기법이 적용된 결과들 중에서 원래 마이크로데이터를 대신하여 공표할 결과를 어떻게 선택하고 결정할 것인가? 안전성과 유용성은 그 결정을 위한 적절한 기준이다. 통계개발원은 이를 측정하기 위하여 개발된 다양한 노출위험 측도와 정보손실 측도를 정리하였다. 제1권 제3부의 제7장은 이러한 측도에 관한 정리와 더불어 제6장의 매스킹 방안들을 평가한 결과를 보여준다. 즉 제6장을 통해서는 매스킹에 관한 세부 기법이 전체적으로 정리되었다면, 제7장에서는 각종 측도를 활용한 평가 방안이 제시되었다고 할 수 있다.

제1권 제4부에서는 기존의 연구 결과를 실무에 도입하여 활용할 수 있도록 돕는 작업을 다루고 있다. 먼저 제8장에는 통계개발원의 매스킹에 관한 연구 결과를 정리하고 R패키지를 이용하여 실제 매스킹 작업을 수행할 수 있도록 작성한 안내서를 결과로 수록하였다. 이 내용은 문서로 만들어진 매스킹 실무 활용 가이드라인에

해당한다. 그러나 자료제공 업무수행 시 통계 관련 전공자가 아닌 경우에는 이 문서를 숙지해서 R프로그램을 코딩하는 것이 어렵기 때문에 보다 손쉽게 활용할 수 있는 가이드라인이 필요했다. 이에 가이드라인으로서 문서를 대신하여 엑셀로 구동되는 프로그램을 작성하였고, 그러한 가이드라인 프로그램을 작성하는 방법을 제9장에서 제시하였다. 이를 통해 업무자들이 손쉽게 제9장의 가이드라인 프로그램을 활용하여 기초적인 수준의 노출위험 측정을 수행할 수 있도록 하였다.

주지하는 바와 같이 제1권에 소개된 매스킹 기법은 데이터 정보보호를 수행하기에는 여러 한계가 노출되었다. 그러나 이러한 한계를 오히려 기회로 발전시킬 수 있지 않을까라는 사고의 전환을 통해 이 연구는 다음 단계로 나아가게 되었다. 바로 그 응답이 제2권에 제시되어 있다.

데이터 정보보호를 위한 최신 기법의 활용 (제2권)

제2권 제1부는 마이크로데이터의 집계 결과물인 매크로데이터에 대한 비밀보호 방안을 다루고 있다. 다양한 유형의 매크로데이터에 관한 매스킹 기법은 2008년에 이미 충실히 다루어졌으며, 제1장에 그 결과가 수록되어 있다. 다만 위계구조를 가지는 대용량 빈도표의 제공을 위한 비밀보호 문제의 해법은 제시되어 있지 않았다. 통계청은 통계지리정보서비스(SGIS)를 통하여 위계구조를 가지는 빈도표를 제공하고 있고, 이를 해결하기 위한 연구가 수행되었다. 제2장은 이에 대한 새로운 알고리즘을 제시하면서 노출위험과 정보손실에 관한 논의도 함께 다루었다.

제2권 제2부는 재현자료(synthetic data)를 다루고 있다. 재현자료는 베이지안 다중대체 기법을 활용하여 원래의 마이크로데이터와 유사한 분포를 가지도록 생성한 자료를 말한다. 제3장은 재현자료 생성과 관련된 이론을 정리하고, 최근 해외에서 제안된 R패키지를 활용하여 재현자료를 생성할 수 있음을 보여주었다. 재현자료에 관한 연구들에서 노출위험을 제어한 상태로 정보손실을 줄일 수 있는 다중대체 모형의 구축에 대해 다루고 있다. 그러나 재현자료의 노출위험에 대해 논란이 있으며, 모형 구축 자체가 쉽지 않은 문제가 있다. 그럼에도 자료제공의 문제를 단계적으로 해결해 나가는데 재현자료의 활용 가치는 높다고 할 수 있으며, 지속적인 연구가 필요하다.

제2권 제3부에서는 최근 가장 주목을 받고 있는 차등정보보호를 다루고 있다. 제4장은 차등정보보호를 다른 방법들과 비교하여 소개하고 있으며, 제5장에서는

차등정보보호에 관한 개념, 이론 및 활용사례를 종합하여 정리하였다. 제6장에는 매스킹 기법, 재현자료, 차등정보보호에 관하여 노출위험과 정보손실을 중심으로 이론적 체계를 제시하였다. 더불어 실제 통계청 보도자료에 차등정보보호를 적용하면 어떠한 결과를 얻는지에 대한 실험 결과와 차등정보보호 의의 및 향후 연구방향을 제시하였다.

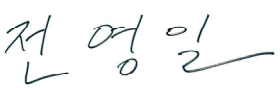
개인정보보호와 데이터 활용을 위한 핸드북

본 책자는 통계적 비밀보호, 데이터 정보보호 분야에 있어 지난 15년 가까운 기간 동안 통계개발원의 주요 연구를 집대성한 것이다. 통계개발원은 매스킹 기법, 노출위험 측도, 정보손실 측도, 대용량 빈도표의 비밀보호 등 정보보호와 관련된 실용적인 주제를 상세히 다루어 왔다. 또한 재현자료, 차등정보보호 등 혁신적인 통계적 대안을 탐구하였다. 각 연구보고서는 국가통계기관의 자료제공 관점에서 데이터 정보보호의 이론과 활용법을 제시하였다.

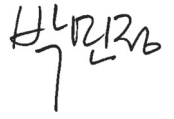
안타깝게도 국내에서 데이터 정보보호 분야의 기술적인 연구 사례를 찾기가 어려운 상황을 고려할 때 본 책자는 현 시점에서 이에 관한 거의 유일한 국문 자료일 것이다. 21세기 사회발전의 원천이 되는 「데이터 활용」과 「개인정보보호」 문제의 실타래를 풀기 위하여 '데이터 정보보호' 분야에 진입하거나 혁신적인 길을 모색하고자 하는 연구자들에게 본 책자가 도움이 되기를 바란다.

가까이 놓고 언제든지 활용할 수 있는 살아있는 핸드북이 되기를 희망한다면 지나친 기대일까? 생명체인 데이터를 다루는 모든 사람에게 유익하기를!

SRI 통계개발원장 전영일



SRI 데이터정보보호센터장 박민정



목 차

[1부] 마이크로데이터의 활용도를 높인 매스킹 기법

제1장 마이크로데이터 제공 현황에 관한 해외사례 연구 ····· 3

제1절 서 론 ····· 3

1. 마이크로데이터 제공의 개요 ····· 3
2. 연구의 목적 및 필요성 ····· 4

제2절 통계청 현황 ····· 7

1. 제공방법 ····· 7
2. 제공자료 ····· 7
3. 가격책정 ····· 9
4. 자료신청절차 ····· 9
5. 제공 후 관리 ····· 10

제3절 국외 현황 ····· 11

1. 미국 센서스국(Census Bureau) ····· 11
2. 미국 노동통계국(Bureau of Labor Statistics:BLS) ····· 16
3. 캐나다 ····· 19
4. 호주 ····· 22
5. 영 국 ····· 26
6. 일 본 ····· 29

제4절 시사점 ····· 30

참고문헌 ····· 32

〈부 록〉 ····· 34

제2장 마이크로자료의 활용도제고를 위한 비밀보호방법 ····· 44

제1절 서 론 ····· 44

제2절 비밀보호방법 ····· 46

1. 노출과 식별 ····· 46
2. 비밀보호방법 ····· 47
3. Uniqueness ····· 49

제3절 마이크로데이터의 Masking ····· 52
1. 분석대상 자료 ····· 52
2. 마이크로자료 파일의 작성 ····· 53
제4절 결 론 ····· 62

참고문헌 ····· 64
〈부 록〉 ····· 65

[2부] 매스킹 기법의 적용과 평가

제3장 2005 인구주택총조사자료의 개인정보 노출제한방법 ····· 103
제1절 서 론 ····· 103
제2절 노출과 노출위험 ····· 104
1. 노출과 유일성 ····· 104
2. 노출위험의 확률모형 ····· 105
제3절 인구주택총조사 자료의 비밀보호 ····· 106
1. 인구주택총조사 자료 ····· 106
2. 마이크로자료 파일의 작성 ····· 107
제4절 결 론 ····· 114

참고문헌 ····· 115
〈부 록〉 ····· 116

제4장 가계조사 마이크로데이터의 비밀보호 ····· 122
제1절 서 론 ····· 122
제2절 노출과 유일성 ····· 123
1. 노출과 노출위험 ····· 123
2. 비밀보호방법 ····· 125
제3절 비밀보호된 마이크로자료 작성 ····· 127
1. 분석대상 ····· 127
2. 식별정보의 비밀보호 ····· 129
3. 민감정보의 비밀보호 ····· 133
제4절 결 론 ····· 151

참고문헌 ····· 153

제5장 승법잡음모형을 이용한 가계조사 자료의 비밀보호 ····· 154
제1절 서 론 ····· 154
제2절 잡음을 이용한 비밀보호 ····· 155
1. 변환모형 ····· 155
2. 잡음의 분포와 그 특성 ····· 156
제3절 적용사례 ····· 162
1. 분석대상 ····· 162
2. 승법잡음모형에 의한 비밀보호 ····· 164
제4절 결 론 ····· 170

참고문헌 ····· 172
〈부 록〉 ····· 173

제6장 가계금융·복지조사 마이크로데이터 제공을 위한 매스킹 방안 ····· 179
제1절 서론 ····· 179
1. 연구배경 및 목적 ····· 179
2. 국내 연구 현황 ····· 181
제2절 통계적 노출제어와 매스킹 ····· 182
1. 노출제어의 종류 ····· 182
2. 매스킹 기법의 종류 ····· 185
제3절 민감변수의 매스킹 방안 ····· 189
1. 매스킹 기법별 노출위험과 정보손실 측정 ····· 189
2. 국소통합 기법 적용 결과 ····· 195
3. 잡음 추가 기법 적용 결과 ····· 197
4. 결합안 적용 결과 및 비교 ····· 200
제4절 키변수의 공표 범위와 노출위험 ····· 204
제5절 결론 ····· 207

참고문헌 ····· 208
〈부 록〉 ····· 210

제7장 가계금융·복지조사 마이크로데이터 매스킹 방안 평가 ····· 224
제1절 서 론 ····· 224
제2절 노출위험 측정 방법론 ····· 227
1. 유일성 기반 노출위험 ····· 228
2. 침입자 의사결정론 기반 노출위험 ····· 235
제3절 자료유용성 측정 방법론 ····· 242
1. 거리 기반 정보손실 측도 ····· 243
2. 특정 통계량 기반 자료유용성 측도 ····· 245
3. 기타 자료유용성 측도 ····· 247
제4절 가계금융·복지조사 마이크로데이터 매스킹 방안 평가 ····· 248
1. 자료의 특징 ····· 250
2. 민감변수의 매스킹 방안 ····· 253
3. 노출위험 및 자료유용성 비교 ····· 260
제5절 요약 및 향후 연구방향 ····· 264

참고문헌 ····· 267
〈부 록〉 ····· 269

[3부] 매스킹 기법의 실무활용

제8장 마이크로데이터 매스킹 기법 실무 활용 안내서 ····· 285
제1절 서론 ····· 285
제2절 통계적 노출제어와 매스킹 ····· 287
1. 노출제어의 종류 ····· 287
2. 매스킹 기법의 종류 ····· 290
제3절 노출위험 및 정보손실 측정 ····· 301
1. 유일성 기반 노출위험 측정 ····· 302
2. 거리 기반 정보손실 측정 ····· 311
3. 특정 통계량 기반 정보손실 측정 ····· 315
제4절 요약 및 제언 ····· 320

참고문헌 ····· 321
〈부 록〉 ····· 323

제9장 통계작성기관 마이크로데이터 제공범위 설정 가이드라인 작성 ····· 331
제1절 서 론 ····· 332
제2절 마이크로데이터의 노출위험 측정 및 가이드라인 ····· 335
1. 키변수와 민감변수 ····· 335
2. 노출의 종류 ····· 336
3. 프라이버시 모형 ····· 338
4. 노출위험을 추정하는 통계적 모형 ····· 340
5. 노출위험 측정 및 해석 ····· 342
6. 가이드라인 프로그램 이용법 ····· 345
제3절 엑셀과 R을 연동한 프로그램(가이드라인) 작성 방안 ····· 349
1. 가이드라인용 프로그램 개발을 위한 벤치마킹 ····· 349
2. 엑셀과 R을 연동하는 프로그램의 구현 ····· 352
제4절 결론 ····· 365

참고문헌 ····· 366
〈부 록〉 ····· 367

표 목차

〈표 1-1〉 데이터 제공 형식 ····· 4
〈표 1-2〉 통계청 제공통계 목록 ····· 8
〈표 1-3〉 Data Ferrett 제공자료 목록 ····· 12
〈표 1-4〉 주요 지역 RDC 개요(2006년 10월 기준) ····· 14
〈표 1-5〉 호주의 제공방법별 이점 및 제약사항 ····· 23
〈표 1-6〉 가격책정 기준 ····· 24

〈표 2-1〉 2005 인구주택총조사 결과(충청남도) ····· 53
〈표 2-2〉 표본추출 현황 ····· 57
〈표 2-3〉 Key 변수의 grouping 결과 ····· 60

〈부표 A-1〉 전수조사 결과 ····· 65
〈부표 A-2〉 10% 표본조사 결과 ····· 66
〈부표 A-3〉 2% 표본조사 결과 ····· 67
〈부표 B-1〉 성별의 Entropy ····· 68
〈부표 B-2〉 나이의 Entropy ····· 68
〈부표 B-3〉 가구주와의 관계의 Entropy ····· 70
〈부표 B-4〉 혼인상태의 Entropy ····· 70
〈부표 B-5〉 교육정도의 Entropy ····· 71
〈부표 B-6〉 가구구분의 Entropy ····· 72
〈부표 B-7〉 거주층의 Entropy ····· 72
〈부표 B-8〉 주인가구여부의 Entropy ····· 72
〈부표 B-9〉 점유형태의 Entropy ····· 73
〈부표 B-10〉 거처의 종류의 Entropy ····· 74
〈부표 B-11〉 연건평의 Entropy ····· 75
〈부표 B-12〉 대지면적의 Entropy ····· 76
〈부표 C-1〉 모집단(전수조사)의 Uniqueness ····· 77
〈부표 C-2〉 2% 표본자료의 Uniqueness ····· 85
〈부표 C-3〉 Grouping한 2% 표본의 Uniqueness ····· 93

〈표 3-1〉 2005 인구주택총조사 결과 ····· 107
〈표 3-2〉 2% 표본현황 ····· 109
〈표 3-3〉 노출제한방법 사용 전 유일성 ····· 111
〈표 3-4〉 노출제한방법 ····· 113
〈표 3-5〉 노출제한방법 사용 후 유일성 ····· 113

〈부표 3-1〉 전수조사 결과 ····· 116
〈부표 3-2〉 10% 표본조사 결과 ····· 117
〈부표 3-3〉 인구에 관한 사항 ····· 118
〈부표 3-4〉 가구에 관한 사항 ····· 120
〈부표 3-5〉 주택에 관한 사항 ····· 121

〈표 4-1〉 가계수지 항목분류 ····· 128
〈표 4-2〉 2006년도 가계조사의 표본규모 ····· 129
〈표 4-3〉 key변수의 유일성(그룹화 前) ····· 130
〈표 4-4〉 key변수의 그룹화 ····· 132
〈표 4-5〉 key변수의 유일성(그룹화 後) ····· 133
〈표 4-6〉 민감변수 ····· 134
〈표 4-7〉 반올림 적용결과 ····· 136
〈표 4-8〉 구간별 분포와 계급값 ····· 139
〈표 4-9〉 구간그룹화 적용결과 ····· 140
〈표 4-10〉 승법잡음의 유형 ····· 144
〈표 4-11〉 생성난수의 결과비교 ····· 144
〈표 4-12〉 승법잡음 적용결과 ····· 146
〈표 4-13〉 3가지 방법별 결과비교 ····· 148

〈표 5-1〉 2006년 가계조사의 표본규모 ····· 162
〈표 5-2〉 가계수지 항목분류 ····· 163
〈표 5-3〉 민감변수 ····· 164
〈표 5-4〉 생성된 난수의 결과비교 ····· 165
〈표 5-5〉 품목별 변환결과 ····· 168
〈표 5-6〉 품목별 변환자료의 상대차이 ····· 168

〈표 6-1〉 국내 마이크로데이터 비밀보호 연구 현황 ····· 181
〈표 6-2〉 노출제어의 종류와 제공되는 자료의 형식 ····· 182
〈표 6-3〉 매스킹에 고려된 가계금융·복지조사 마이크로데이터의 민감변수 ····· 189
〈표 6-4〉 민감변수를 위한 매스킹 기법 및 세부 알고리즘 ····· 192
〈표 6-5〉 세부 알고리즘별 노출위험과 정보손실 측도 결과 ····· 193
〈표 6-6〉 국소통합(mdav) 기법 적용 전후 소득 변수들의 상관계수 ····· 197
〈표 6-7〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 소득 변수들의 상관계수 ····· 199
〈표 6-8〉 국소통합, 잡음 추가 및 결합안의 노출위험과 정보손실 측도 결과 ····· 200
〈표 6-9〉 결합안3 적용 전후 소득 변수들의 상관계수 ····· 203
〈표 6-10〉 키변수의 재코딩과 빈도수 ····· 204
〈표 6-11〉 공표 범위에 따른 유일성과 노출위험 ····· 206

〈부표 6-1〉 매스킹 기법 적용에 따른 표본 S 민감변수들의 평균 ····· 210
〈부표 6-2〉 매스킹 기법 적용에 따른 표본 S 민감변수들의 표준편차 ····· 211
〈부표 6-3〉 국소통합(mdav) 기법 적용 전후 자산 변수들의 상관계수 ····· 213
〈부표 6-4〉 국소통합(mdav) 기법 적용 전후 부채 변수들의 상관계수 ····· 215
〈부표 6-5〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 자산 변수들의 상관계수 ····· 217
〈부표 6-6〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 부채 변수들의 상관계수 ····· 219
〈부표 6-7〉 결합안3 적용 전후 자산 변수들의 상관계수 ····· 221
〈부표 6-8〉 결합안3 적용 전후 부채 변수들의 상관계수 ····· 223

〈표 7-1〉 마이크로데이터를 위한 노출제한의 종류와 제공되는 자료의 형식(박민정 등, 2013) · 224
〈표 7-2〉 통계개발원의 마이크로데이터 비밀보호 연구 현황(박민정 등, 2013) ····· 225
〈표 7-3〉 노출위험 측도 정리 ····· 241
〈표 7-4〉 자료유용성 측도 정리 ····· 248
〈표 7-5〉 키변수들의 빈도표 ····· 250
〈표 7-6〉 민감변수명 ····· 252
〈표 7-7〉 민감변수들의 상관계수 행렬 ····· 252
〈표 7-8〉 변수별 0응답의 개수 ····· 253
〈표 7-9〉 노출위험 및 정보손실 측도 결과값 ····· 261

〈표 8-1〉 노출제어의 종류와 배포되는 자료의 형식 ····· 287
〈표 8-2〉 sdcMicro의 예제 데이터 francdat 설명 ····· 292

〈표 8-3〉 국소 감추기 명령어 형식 및 사용 예제 ····· 293
〈표 8-4〉 전반적 재코딩 명령어 형식 및 사용 예제 ····· 294
〈표 8-5〉 국소통합 명령어 형식 및 사용 예제 ····· 295
〈표 8-6〉 잡음 추가 명령어 형식 및 사용 예제 ····· 297
〈표 8-7〉 자료순위 교환 명령어 형식 ····· 298
〈표 8-8〉 가계금융·복지조사 마이크로데이터 매스킹을 위한 프로그램 일부 ····· 299
〈표 8-9〉 행렬 형식의 마이크로데이터 예제 ····· 304
〈표 8-10〉 분할표 형식으로 정리된 마이크로데이터, 변수가 2개일 경우 ····· 304
〈표 8-11〉 변수 조합별 레코드수 형식으로 정리된 마이크로데이터, 변수가 2개일 경우 ····· 304
〈표 8-12〉 변수 조합별 레코드수 형식으로 정리된 마이크로데이터, 변수가 4개일 경우 ····· 305
〈표 8-13〉 유일한 레코드 수에 근거한 파일 단위의 노출위험 측정 ····· 308
〈표 8-14〉 유일성에 근거한 레코드별 노출위험 측도 ····· 309
〈표 8-15〉 거리 기반 노출위험 측도 ····· 310
〈표 8-16〉 거리 기반 정보손실 측도 ····· 314
〈표 8-17〉 가계금융·복지조사 마이크로데이터 매스킹 방안의 노출위험
-정보손실 지도 작성을 위한 프로그램 일부 ····· 317
〈표 8-18〉 가계금융·복지조사 마이크로데이터 매스킹 방안의 특정 통계량 기반
-정보손실 측정 프로그램 일부 ····· 318

〈표 9-1〉 다양한 키변수와 민감변수의 구분 ····· 336
〈표 9-2〉 마이크로데이터 예제 ····· 337
〈표 9-3〉 2개 변수를 제공할 때 이차원 빈도표 ····· 337
〈표 9-4〉 2개 변수 및 4개 변수를 제공할 때 빈도표 ····· 338
〈표 9-5〉 마이크로데이터에서 k-익명성과 l-다양성 ····· 339
〈표 9-6〉 이차원 분할표에서 k-익명성 파악 ····· 340
〈표 9-7〉 빈도표(flattened frequency table)에서 k-익명성 파악 ····· 340
〈표 9-8〉 마이크로데이터와 빈도표 ····· 343
〈표 9-9〉 개체별 노출위험 계산 결과 ····· 344
〈표 9-10〉 노출위험 계산 결과 예제 ····· 345
〈표 9-11〉 shell 명령어를 통한 RScript의 실행 ····· 358

그림 목차

[그림 1-1] 통계청 제공자료수 ····· 5
[그림 1-2] 통계청 제공자료 구성비 ····· 5
[그림 1-3] 요청기관별 자료제공 현황(2007) ····· 6
[그림 1-4] 마이크로데이터의 이용목적 ····· 6

[그림 4-1] 좌우대칭의 절단된 삼각분포 ····· 143

[그림 5-1] 삼각분포 ····· 156
[그림 5-2] 절단된 삼각분포 ····· 157
[그림 5-3] 사다리꼴분포 ····· 159
[그림 5-4] 이중삼각분포 ····· 161

[그림 6-1] 잡음 추가 기법 적용과 무응답 정보 처리에 대한 세 가지 시나리오 ····· 191
[그림 6-2] 세부 기법별 노출위험-정보손실 지도 ····· 194
[그림 6-3] 국소통합 세부 기법별 소득 변수들의 상자그림 ····· 196
[그림 6-4] 국소통합(mdav) 기법 적용 전후 소득 변수들의 산점도 ····· 197
[그림 6-5] 잡음 추가 시나리오별 소득 변수들의 상자그림 ····· 198
[그림 6-6] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 소득 변수들의 산점도 ····· 199
[그림 6-7] 국소통합, 잡음 추가 및 결합안의 노출위험-정보손실 지도 ····· 201
[그림 6-8] 국소통합, 잡음 추가3 및 결합안3에 대한 소득 변수들의 상자그림 ····· 202
[그림 6-9] 결합안3 적용 전후 소득 변수들의 산점도 ····· 203

[부그림 6-1] 국소통합 세부 기법별 자산 변수들의 상자그림 ····· 212
[부그림 6-2] 국소통합(mdav) 기법 적용 전후 자산 변수들의 산점도 ····· 213
[부그림 6-3] 국소통합 세부 기법별 부채 변수들의 상자그림 ····· 214
[부그림 6-4] 국소통합(mdav) 기법 적용 전후 부채 변수들의 산점도 ····· 215
[부그림 6-5] 잡음 추가 시나리오별 자산 변수들의 상자그림 ····· 216
[부그림 6-6] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 자산 변수들의 산점도 ····· 217
[부그림 6-7] 잡음 추가 시나리오별 부채 변수들의 상자그림 ····· 218
[부그림 6-8] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 부채 변수들의 산점도 ····· 219

[부그림 6-9] 국소통합, 잡음 추가3 및 결합안3에 대한 자산 변수들의 상자그림 ····· 220
[부그림 6-10] 결합안3 적용 전후 자산 변수들의 산점도 ····· 221
[부그림 6-11] 국소통합, 잡음 추가3 및 결합안3에 대한 부채 변수들의 상자그림 ····· 222
[부그림 6-12] 결합안3 적용 전후 부채 변수들의 산점도 ····· 223

[그림 7-1] 마이크로데이터 매스킹 절차 ····· 226
[그림 7-2] 일반적인 마이크로데이터의 매스킹 절차(Meindl 등, 2013) ····· 249
[그림 7-3] 노출위험-정보손실 지도(기존 연구의 매스킹 방안) ····· 255
[그림 7-4] 0응답 여부에 따른 개별 노출위험 측도 값들의 분포 ····· 256
[그림 7-5] 0응답 여부에 따른 유일성 측도 값들의 분포 ····· 257
[그림 7-6] 노출위험-정보손실 지도 ····· 260
[그림 7-7] 매스킹 전후 자료의 모형적합 결과 비교 ····· 263

[그림 8-1] 마이크로데이터 매스킹 절차(박민정, 2014) ····· 301
[그림 8-2] 일반적인 마이크로데이터의 매스킹 절차(Meindl 외, 2013) ····· 302

[그림 9-1] 가이드라인 엑셀 프로그램 ····· 345
[그림 9-2] 'G_메뉴' 시트에서 노출위험 분석 수행 ····· 346
[그림 9-3] 'G_결과' 시트에 나타나는 노출위험 분석 결과 ····· 347
[그림 9-4] 'G_메뉴' 시트에서 변수 재범주화 수행 ····· 347
[그림 9-5] 'G_재범주화 설정' 시트에서 재범주화 기준(breaks) 및 새로운 범주(labels)의 입력 ····· 348
[그림 9-6] 'G_메뉴' 시트에 M행에 나타난 재범주화 처리된 변수 ····· 348
[그림 9-7] Eurostat의 PEFA tools 2017 ····· 349
[그림 9-8] KESS의 사용자 친화적인 인터페이스 ····· 350
[그림 9-9] R이 설치된 폴더 경로 복사 ····· 353
[그림 9-10] [컴퓨터]의 속성 ····· 354
[그림 9-11] 시스템 속성 ····· 354
[그림 9-12] 환경 변수 대화상자 ····· 355
[그림 9-13] 환경 변수 편집 대화상자 ····· 355
[그림 9-14] R의 path를 추가한 이후 화면 ····· 356
[그림 9-15] 명령 프롬프트의 찾기 및 실행 ····· 356

[그림 9-16] 명령 프롬프트 도스 화면에서 rscript를 실행한 화면 ····· 357
[그림 9-17] 명령단추와 매크로의 연결 ····· 359
[그림 9-18] 명령단추 만들기 ····· 359
[그림 9-19] 개발 도구 메뉴 보이기 ····· 360
[그림 9-20] 엑셀과 통신하기 위한 R 코딩 예 ····· 361
[그림 9-21] R 수행 결과로 생성된 html 파일 ····· 362

[부그림 9-1] R의 벡터, 행렬, 데이터 프레임에 대한 예제 ····· 367
[부그림 9-2] R에서 리스트의 사용 ····· 368
[부그림 9-3] R_connection_examples.xlsm의 예제 시트 ····· 370
[부그림 9-4] R_connection_examples.xlsm의 VBA 코드 ····· 370
[부그림 9-5] BERT를 이용한 엑셀과 R의 통신 ····· 371
[부그림 9-6] 엑셀 VBA에서 R 코드의 실행 ····· 372

1부

마이크로데이터의 활용도를 높인 매스킹 기법

1. 마이크로데이터 제공 현황에 관한 해외사례 연구
김경미 · 이의규 · 정미옥

2. 마이크로자료의 활용도 제고를 위한 비밀보호방법
정동명 · 강동환

제1장

마이크로데이터 제공 현황에 관한 해외사례 연구

김경미 · 이의규 · 정미옥

제1절 서 론

1. 마이크로데이터 제공의 개요

통계자료가 적시적지에 활용될 수 있도록 그 수요에 올바르게 대응하여 제공하는 것은 정확한 통계를 생산해 내는 것만큼이나 중요한 일이다. 통계자료는 일반적으로 원시자료(Raw data), 마이크로데이터(Microdata), 그리고 매크로데이터(Macrodata)로 구분된다. 통계청 MDSS(Microdata Service System)의 용어 정의에 따르면, 원시자료(Raw data)는 입력 첫 단계의 자료로서 입력오류나 조사오류 등이 걸러지기 이전 단계의 자료를 말하며, 마이크로데이터(Microdata)는 오류를 수정한 자료로서 통계표 작성 등 데이터 가공의 바탕이 되는 자료를 의미한다. 또한 매크로데이터(Macrodata)는 마이크로데이터를 임의의 기준에 따라 집계한 자료를 지칭하는데, 집계의 정도에 따라 세분된 자료부터 통합된 자료까지 다양하게 제공된다.

이용자에게 제공되는 데이터는 주로 마이크로데이터와 매크로데이터이다. 일반적으로 매크로데이터가 제공되는 형식은 통계표(Tables) 형태로 가공하여 서비스되는 형식으로 통계청에서는 간행물이나 KOSIS(Korean Statistical Information Service)를 통해 제공되고 있다. 반면 마이크로데이터가 제공되는 형식은 외부이용자를 위해 사전에 만들어진 마이크로데이터 파일로서 CD-ROM 등에 저장하여 제공하는 공공이용파일(Public Use File)의 형태, 특정한 심층이용자에게 제공되며 제공 전에 법적인 조치나 서명 등이 필요한

✍ 이 연구는 2007년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

인가파일(Licensed File)의 형태, 그리고 통신망을 통해 제공하는 원격접속파일(Remote Access Files)의 형태가 있다. 이 밖에 통계기관의 엄격한 감독하에 기관 내에서 마이크로데이터를 이용하는 제도로 데이터실험실(Data Laboratories)이 있다. 이상의 데이터 제공 형식을 정리하면 <표 1-1>과 같다.

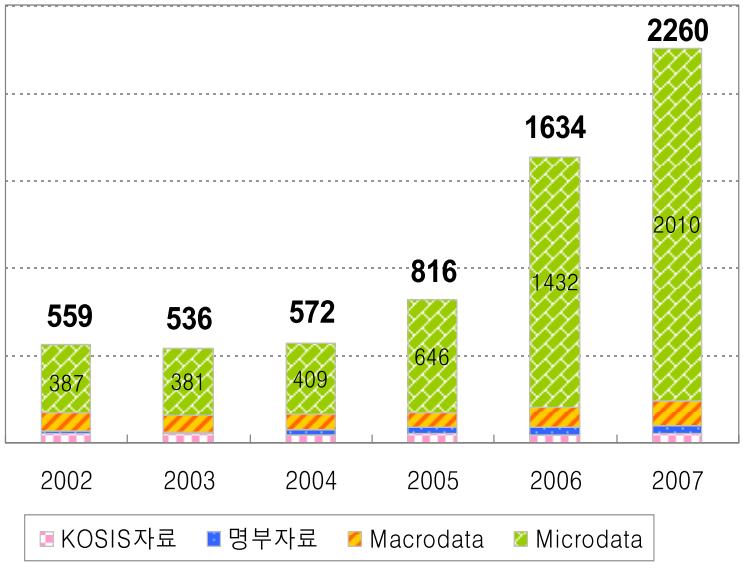
〈표 1-1〉 데이터 제공 형식

형 식		자료의 특징 및 내용	통계청 적용형태
통계표(Tables) 형태로 가공하여 서비스		· 심층 이용자 요구에 따라 통계표 작성 · 지나치게 세분화되면 비밀노출 위험	간행물 KOSIS
익명자료파일 (AMF1))	공공이용파일 (Public Use File)	· 외부이용자를 위하여 사전에 만들어진 마이크로데이터 파일로 CD-ROM 등에 저장	마이크로데이터 CD-ROM
	인가파일 (Licensed File)	· 특정한 심층이용자에게 제공되며, 제공전에 법적인 조치(서명 등)가 필요	개별자료 제공
원격접속파일 (RAF2))		· 통신망을 통하여 심층 이용자들이 마이크로데이터를 활용하는 제도	MDSS
데이터실험실 (DL3))		· 통계기관의 엄격한 감독하에 기관 내에서 마이크로데이터를 이용하는 제도	On-site

주: 1) Anonymised Microdata Files.
2) Remote Access Files.
3) Data Laboratories.
자료: 박원환·황조연, 「통계자료의 특성과 비밀보호 방법에 관한 연구」, 2004; 통계청, "통계마이크로데이터 제공현황 및 개선방안", 2006.

2. 연구의 목적 및 필요성

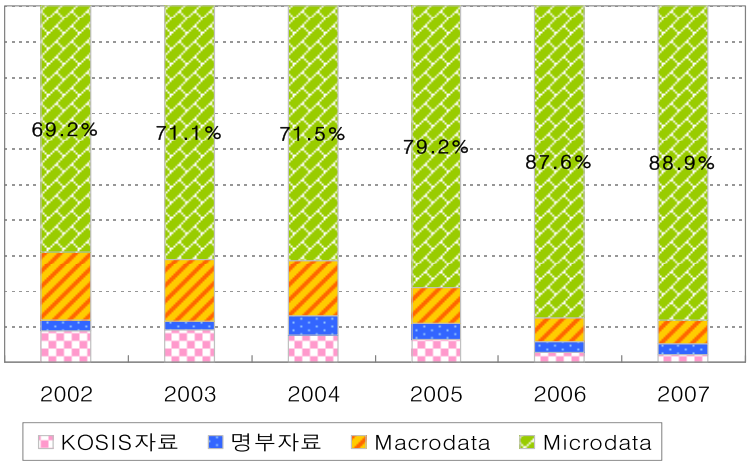
사회구조가 복잡해짐에 따라 의사결정의 기초자료로서 통계수요가 급격히 증가하고 있다. 특히, 중앙통계기관인 통계청의 자료는 각종 국가정책 수립 및 평가, 기업경영, 사회현상 탐구 등의 기초자료로 활용되는데, [그림 1-1]과 같이 2002년 559건의 자료가 이용되었음에 비해 2007년에는 2,260개의 자료가 이용되어 그 수요가 급격히 증가함을 알 수 있다.



주: 2002~04 년은 건수 , 2005~07 년은 개수 .
자료: 통계청 (2007).

[그림 1-1] 통계청 제공자료수

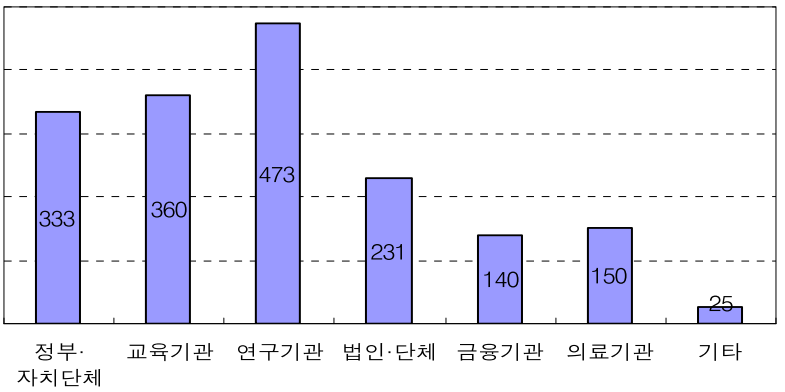
특히 초기 단순한 집계자료 요청에서 점차 다양한 분석이 가능한 마이크로데이터의 요청이 증가하고 있는데, [그림 1-2]를 보면 제공되는 전체 자료 중 마이크로데이터 비중이 점점 증가하고 있음을 알 수 있다.



주: 2002~04년은 건수, 2005~07년은 개수.
자료: 통계청(2007).

[그림 1-2] 통계청 제공자료 구성비

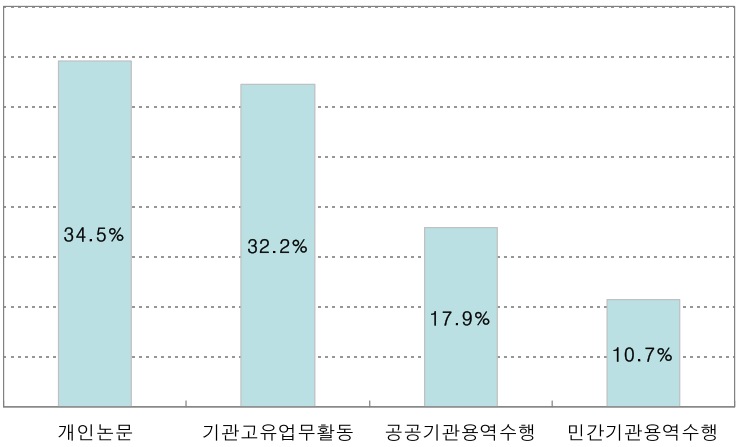
또한 제공되는 통계자료는 다양한 기관에서 이용되는데 2007년 현재 [그림 1-3]과 같이 연구기관, 교육기관, 정부・자치단체, 법인・단체의 순으로 많이 이용되고 있는 것으로 나타났다.



자료: 통계청(2007).

[그림 1-3] 요청기관별 자료제공 현황(2007)

2006년 통계청 통계정보국에서 실시한 “마이크로데이터 활용도 제고를 위한 심층 의견조사” 결과에 따르면 [그림 1-4]와 같이 그 이용목적 또한 다양하다. [그림 1-4]에서 보듯이 개인논문과 기관 고유 업무활동 부분이 약 2/3를 차지했으며, 공공기관 및 민관 기관의 용역수행을 위해 이용하는 경우도 일정 비율을 차지하였다.



자료: 통계청, “마이크로데이터 활용도 제고를 위한 심층 의견조사”, 2006.

[그림 1-4] 마이크로데이터의 이용목적

이와 같이 마이크로데이터의 이용이 증가함에 따라 마이크로데이터의 제공방법 및 비용 등에 관한 의견들도 분분하다. 또한 이용기관과 이용목적도 다양하므로 그에 적합한 제공방법에 대해 연구할 필요가 있는 바, 해외의 마이크로데이터 제공 현황에 관한 사례를 자세히 파악함으로써 제도적 또는 방법론적 측면에서 적용방안을 모색하고자 한다.

다음 제2절에서는 우리나라 통계청의 마이크로데이터 제공현황에 대해 살펴보고, 이후 제3절에서 미국, 캐나다, 호주, 영국, 그리고 일본의 현황을 각각 파악하여 마지막 절에서는 이들로부터 시사점을 도출한다.

제2절 통계청 현황

1. 제공방법

해외의 현황을 살펴보기에 앞서, 우리나라 통계청에서는 어떻게 마이크로데이터를 제공하고 있는지를 알아보고자 한다. 마이크로데이터는 크게 세 가지 방법으로 이용자에게 제공된다. 먼저 개인 식별 데이터를 제거한 마이크로데이터를 CD-ROM에 저장하여 제공하는 형식이 있다. 이는 통계쇼핑몰(http://www.nso.go.kr/shop)을 통해 제공되고 있다. 둘째로 통신망을 이용한 방법으로 이용자가 인터넷을 통해 시스템에 접속하여 원하는 자료를 직접 가공하거나 다운받을 수 있도록 지원하는 원-스톱 서비스 시스템인 MDSS(MicroData Service System)를 이용하는 방법이다. MDSS는 2006년 1월 1일부터 서비스를 시작하였고 홈페이지를 통해 제공 가능한 자료의 목록과 범위, 이용방법, 비용 등을 안내하고 있으며, MDSS 사이트(http://mdss.nso.go.kr)에 접속하여 신청한 후 자료를 다운로드 받을 수 있다. 중앙정부 및 지방자치단체에서 신청할 경우에는 통계청에서 작성·제공하고, 대학 및 연구기관에서 신청할 경우에는 통계협회로 이첩하여 제공하는 방식을 취하고 있다. 마지막으로 특정장소에서 좀 더 심층적인 데이터를 제공받을 수 있는 On-site Access 방법이 있다. 현재 2개소를 운영하고 있는데 통계청(대전)과 대한통계협회(서울) 내에서 운영하고 있다.

2. 제공자료

통계청에서 작성되는 통계는 조사통계가 44종, 가공통계가 11종, 보고통계가 1종으로 총 56종의 통계가 작성되고 있다. 이 중 마이크로데이터가 제공되는 통계는 조사통계

32종과 기업체모집단명부를 포함하여 총 33종에 대해 제공하고 있다. 제공되는 조사통계 목록은 <표 1-2>와 같다.

〈표 1-2〉 통계청 제공통계 목록

No.	제공되는 조사통계	No.	제공되는 조사통계
1	가계조사	16	사회통계조사
2	가계자산조사	17	산업총조사
3	가구소비실태조사	18	생활시간조사
4	건설업통계조사	19	서비스업총조사
5	경제활동인구조사	20	양곡소비량조사
6	광업제조업통계조사	21	어가경제조사
7	국부통계조사	22	어업기본통계조사
8	기업체모집단명부	23	어업총조사
9	농가경제조사	24	운수업통계조사
10	농산물생산비조사	25	인구동태조사
11	농어업법인사업체통계조사	26	인구이동통계
12	농업기본통계조사	27	인구주택총조사
13	농업총조사	28	인력실태조사
14	도소매업 및 서비스업통계조사	29	임업총조사
15	사망원인통계	30	전국사업체기초통계조사

주: 1) 인구총조사와 주택총조사, 도소매업 통계와 서비스업 통계는 통합해 총 30종의 통계로 제공.
2) 2005년 인구주택총조사의 경우는 1%, 2%, 5% 표본자료를 제공하고 있음.

각 조사별 자료의 제공범위는 자료 제공 심의회를 거쳐서 결정되며 결정된 제공범위는 MDSS 홈페이지에 공시되어 있다. 자료 제공 심의회는 통계청 통계정보국장이 위원장이 되며 통계작성 담당과장 등 10인의 위원이 모두 통계청 내부자로 구성된다. 자료제공 심의회에서 주로 심의하는 내용은 ① 통계조사별 자료제공범위 설정에 관한 사항, ② 통계목적으로 사용여부의 판단에 관한 사항, ③ 제한적 공개가능 자료에 대한 제공여부에 관한 사항, ④ 자료제공업무의 효율성 향상과 ⑤ 이용자의 편의제고를 위한 제도개선 및 운영방법에 관한 사항이다. 자료의 제공범위를 심의함에 있어서는 응답단위의 비밀보호와 응답단위・항목별 자료의 신뢰성 등에 중점을 둔다. 심의회는 자료제공범위의 변경 필요성 등이 있을 경우 수시로 개최하여 심의한다.

3. 가격책정

마이크로데이터를 중앙정부와 지방자치단체에서 요청할 경우 무상으로 제공하며, 대학, 연구기관 및 개인이 요청할 경우 유상으로 제공한다. 마이크로데이터의 비용 산정 기준은 수익자부담원칙[1] 및 실비보상 원칙을 바탕으로 통계청통계자료규정 제5조 제3항의 규정에 의하여 정하고 있다. 자료이용료는 1MB에 기본료 20,000원이 부과되며 초과되는 1MB당 1,500원이 추가된다. 예를 들어 약 12만 개의 전국의 광업·제조업 사업체에 대해 조사되는 광업·제조업통계조사 전체자료가 130MB인 경우, 기본료 20,000원과 이를 초과하는 129MB에 대해 1MB당 각 1,500원이 추가되어 총 213,500원[2]이 부과된다.

주문형 위탁처리의 경우는 자료이용료와 별도로 소프트웨어개발비가 부과되는데 이는 정보통신부 고시 제2005-22호 「소프트웨어사업 대가의 기준[3]」을 준용하여 징수한다. 소프트웨어개발비가 추가되는 인구주택총조사의 표본추출의 경우 기본 추출틀은 조사구 100개, 조사구당 특성치 2개이며 그 산출내역은 자료이용료(기본 1MB = 20,000원)에 소프트웨어개발비 198,010원이 더해져 218,010원이 된다. 기본 특성치(2개)를 초과할 경우 특성치 1개당 37,120원이 추가 부과되며 특성치가 기본(2개)인 상태에서 기본 조사구(100개)를 초과할 경우 초과조사구 100개당기본사용료 218,010원이 추가 부과된다.

MDSS로 제공되는 경우는 마이크로데이터 비용 산정 기준이 그대로 적용되며, CD-ROM으로 제공되는 경우는 마이크로데이터 비용 산정 기준에 수수료 2%가 부과된다. On-site Access의 경우 같은 기준에 이용환경 설정, 시스템 이용 등에 대한 수수료 10%가 추가로 부과된다.[4]

4. 자료신청절차

통계청의 마이크로데이터 이용은 공표통계에서 제공할 수 없는 보다 심층적인 사회현상 분석을 원하는 다양한 계층의 이용자에게 마이크로데이터를 제한 없이 제공함을

---

1) 일반적으로 공공사업의 실시에 의하여 특별히 이익을 받는 사람에게 소요경비의 전부 또는 일부를 부담시키는 일을 의미하며, 마이크로데이터를 서비스하기 위해서는 별도의 마이크로데이터 관리, 전산처리 등의 통계청의 자원이 추가로 소요되기 때문에 추가로 소요되는 자원에 대한 비용을 이용자에게 부담시키고 있음.

2) 20,000원 + 1,500원 × 129MB = 213,500원.

3) 소프트웨어개발비 = 개발원가 + 직접경비 + 이윤.
- 개발원가 = 보정전개발원가(코드라인수×코드라인당 단가) × 보정계수.
- 직접경비 = 컴퓨터시스템 사용료, 소프트웨어 도구 사용료, 제출문서의 인쇄, 청사진비 등 특별히 소요되는 직접비용.
- 이윤 = 개발원가 × 10%.

4) 현재 총괄원가보상원칙으로 비용산정 기준의 개정을 검토 중에 있음.

원칙으로 한다. 외국기관을 포함한 외국인에 대한 제공은 국제기구, 외국정부기관 등 신뢰성 있는 기관에 한정하여 국내 기준과 동일하게 적용함을 원칙으로 하고, 사안별로 담당과에서 검토를 하고 필요시 자료 제공 심의회에서 심의한다.

MDSS를 이용해 사용자가 직접 추출하여 다운로드 받고자 하는 경우는 홈페이지에서 회원가입을 하여야 한다. KOSIS에서 무료로 제공되는 자료의 목록 확인과 이용할 자료의 자료제공 범위를 확인 후 자료 조회 및 필요한 항목을 선택하고 승인을 요청한다. 자료 제공 유무가 결정되면 비용을 납부하고 자료를 다운로드 받을 수 있다. 통계자료를 위탁처리하는 경우는 자료제공 범위 확인까지는 동일한 절차이며, 그 후 위탁신청서를 작성하여 신청하면 통계청에서 자료제공유무를 결정하여 자료를 가공하고 이용자가 비용을 납부하면 자료가 제공된다. On-site Access의 경우에는 이용을 승인받기 위해서 On-site Access 이용신청서 및 연구계획서를 같이 제출해야 한다. 이 연구계획서에서는 통계청에서 제공하는 통계조사보고서, KOSIS 등의 공표자료, 그리고 기존의 마이크로데이터 상품으로는 자료이용 목적을 충족할 수 없음을 입증하여야 한다. 이용을 승인받은 후에는 통계청 또는 대한통계협회 내의 지정장소에서 사전 설치되어 있는 전산장비를 이용하여 마이크로데이터 분석 작업을 한다. 또한 자료이용자의 분석결과에 대하여, 자료생산부서에서 응답단위의 비밀보호의무 준수 여부 및 분석 결과의 통계적 신뢰도 등을 검토 후 외부 반출 여부를 결정한다. 외부로 분석결과의 반출을 결정한 경우 자료이용자는 비용을 납부하고 자료를 제공받을 수 있다.

5. 제공 후 관리

통계청은 자료 이용 시 지켜야 할 사항을 통계자료 제공규정 제7조 제2항으로 정하고 있다. 포함되는 사항은 개별 자료에 의거 알게 된 사항에 대한 누설금지, 통계목적 이외의 사용금지, 제공 자료의 복제 및 대여금지, 자료 사용 후 즉시 폐기이다. 이러한 의무를 위반할 시에는 개정통계법[5] 제41조 3항에 의거하여 100만원 이하의 벌금이 부과된다.

CD-ROM이나 MDSS로 제공되는 경우 마이크로데이터를 이용한 분석결과에 대한 검토는 이루어지지 않고 있으며, On-site Access의 경우에 는 분석결과의 반출 전 응답단위의 비밀보호의무 준수 여부 및 분석 결과의 통계적 신뢰도 등을 검토한 후 반출 여부를 결정한다.

---

5) 개정통계법은 2007년 10월 28일자로 시행된다.

제3절 국외 현황

국외의 마이크로데이터 제공현황은 각국의 통계청과 마이크로데이터를 제공하는 각 기관의 홈페이지, 관련 문헌들을 참조하였고 각 국 통계청의 자료제공 담당자와의 이메일을 통해 자료를 수집하였다. 또한 미국과 캐나다의 On-site방법과 미국 노동통계국의 off-site방법은 많은 부분 神林 龍(2007)의 보고서를 발췌 인용하여 정리하였음을 밝혀둔다.

1. 미국 센서스국(Census Bureau)

가. 공공이용파일(Public Use File)

미국은 공공이용파일(Public Use File)[6]을 다양한 기관에서 간단한 절차를 통해 거의 무료로 제공하고 있다. 센서스국은 American Factfinder를 통해 직접 제공하고 있으며 웹사이트(http://factfinder.census.gov)에서다운로드 할 수 있다. 약간의 실비를 지급하면 CD-ROM으로 구매도 가능하다.

또한 Data Ferrett 사이트(http://dataferrett.census.gov)에서 다양한 자료를 제공받을 수 있는데 이메일주소를 기입한 후 무료로 다운로드 받을 수 있다. 이 사이트에서는 마이크로데이터의 제공뿐만 아니라 사이트 내에서 프로그램을 이용하여 집계표 작성 및 그래프 작성이 가능하다. 제공되는 자료는 총 20종이며 그 목록은 <표 1-3>과 같다.

미네소타 인구센터에서 운영하는 IPUMS(Integrated Public Use Microdata Series)에서는 주로 센서스 자료를 제공하고 있는데 간략한 연구계획서 제출을 통해 회원으로 가입한 후 무료로 다운로드 받을 수 있다. 제공 절차를 자세히 살펴보면, 먼저 사이트(http://www.ipums.org)에서 사용자 등록 후 로그인을 하고 자료 추출 요청을 한다. 이때 국가, 연도, 대상 집단, 변수를 지정하면 미네소타 인구센터에서 자료를 추출한 후(수분에서 수 시간 소요), 연구자에게 이메일로 통보하게 되는데 72시간 내에 자료를 다운로드 받아 이용할 수 있다. 자료는 128비트 암호화 장치에 의해서 전송된다. IPUMS는 미국의 센서스 자료뿐만 아니라 다른 나라의 센서스 자료도 제공하고 있다. 제공하고 있는 자료를 살펴보면, 1850년에서 2000년까지의 미국 센서스 자료, 미국의 CPS(Current Population Survey)자료, 국가별 센서스 자료(26개국, 80개의 센서스자료, 2억 2백만 개의 레코드) 등이 있다.

---

6) 외부이용자를 위하여 사전에 만들어진 마이크로데이터 파일로, 해외의 경우 비밀보호기법이 적용된 공공파일이다.

〈표 1-3〉 Data Ferrett 제공자료 목록

No.	제공되는 통계자료
1	American Community Survey (ACS)
2	American Housing Survey (AHS)
3	Behavioral Risk Factor Surveillance System (BRFSS)
4	Consumer Expenditure Survey (CES)
5	County Business Patterns (CBP)
6	Current Population Survey (CPS)
7	Decennial Census of Population and Housing (Census 2000)
8	Decennial Census of Population and Housing (Census 1990)
9	Delaware Statistics
10	Harvard MIT data Center Collection
11	Home Mortgage Disclosure Act (HMDA)
12	Maryland Statistics
13	National Ambulatory Medical Care Survey (NAMCS)
14	National Health and Nutrition Examination Survey (NHANES)
15	National Health Interview Survey (NHIS)
16	National Survey of Fishing, Hunting, and Wildlife-Associated Recreation (FHWAR)
17	Small Area Income and Program Poverty Estimates (SAIPE)
18	Social Security Administration
19	Survey of Income and Program Participation (SIPP)
20	Survey of Program Dynamics (SPD)

자료: Data Ferrett 홈페이지.

NBER(National Bureau of Economic Reseach)에서도 홈페이지(http://www.nber.org/data)를 통해 다운로드 받을 수 있도록 하고 있는데 산업별 생산성자료 등 각종 경제연구에 필요한 자료를 제공한다. 가공된 자료로는 개별 학자들이 취합하여 통합한 자료 중 검증이 된 자료를 제공하고 있다. 이들 자료는 무료로 제공되고 있으며 CPS자료를 저장매체로 요청할 경우 약 100달러의 실비가 부과된다.

CEPR(Center for Economic and Policy Research)에서는 CPS자료와 SIPP(U.S. Census Survey of Income and Program Participation)자료, 그리고 1980년에서 2000년까지의 센서스 1% 자료를 홈페이지(http://www.ceprdata.org)에서 제공하고 있다. 역시 무료로 제공되며 CPS자료를 저장매체로 요청할 경우 약 20~30달러의 실비가 부과된다.

Unicon Research라는 연구기관을 통해서도 제공되는데 제공되는 자료는 CPS자료, SIPP 자료, NLTCS(National Long Term Care Survey)자료, CPS자료에서 추출한 MDU(Medicare Data Utilities)자료가 있다. 홈페이지(http://www.unicon.com)의 회원가입 후 제공받을 수 있으며 100~1,500달러의 비용이 부과된다.

각 기관에서 제공되고 있는 센서스 자료는 다양한 형태로 제공되고 있는데, 2000년 센서스 자료의 경우는 1% PUMS(Public Use MicrodataSample)자료, 5% PUMS자료, 괌 PUMS 10% 자료, 버진 아일랜드 PUMS 10% 자료, Long Form 전체자료(전체 조사대상의 1/6)가 제공되고 있다. 5% PUMS자료는 주와 10만 명 이상이 있는 카운티의 정보를 제공하며, 1% PUMS자료의 경우에는 표본 수는 작고 지리적(geographic) 변수는 40만 명 이상이 거주하는 지역의 아이디만 제공할 정도로 세분화 정도는 낮지만 지리적 변수를 제외한 다른 변수는 세분화된 자료를 제공한다. 5% PUMS자료가 어떤 인종이든 1만 명 이상의 인종을 대표할 경우 기타로 분류되어 있는 반면 1% PUMS자료는 자세한 인종 카테고리를 모두 공개한다. Long Form 전체자료의 경우는 데이터실험실(Data Laboratories:DL) 형식인 usRDC로만 이용가능하며 다른 방법으로는 접근이 불가능하다. 1990년 자료는 1% PUMS자료, 5% PUMS자료, 3% 노년인구자료, 0.5% 노동시장구역 자료, 푸에르토리코 PUMS 5%가 제공된다.

나. On-site(Data Laboratories)

1) 개요

센서스국 내부의 경제 분석 부문인 CES(Center for Economics Studies) 에서는 데이터 실험실(Data Laboratories) 형식의 usRDC(U.S. Research Data Center)를 운영하고 있는데, 이는 특정한 부지(site) 내에 설치된 특정한 단말기를 이용하여 데이터에 접근하고 거기에서 모든 작업을 완결하는 형식이다. 1994년 1월 보스턴에서 최초 개소한 이래 현재는 미국 전역의 9곳에서 운영하고 있다. 각각의 RDC(Research Data Center)는 센서스국에 있는 Census Bureau RDC, NBER(National Bureau of Economic Reseach)에 위치한 Boston Census RDC, UC Berkeley에 위치한 California Census RDC-Berkeley, UCLA에 위치한 California Census RDC-Los Angeles, FRBC(Federal Reserve Bank of Chicago)에 위치한 Chicago Census RDC, 미시건 대학에 위치한 Michigan Census RDC, Baruch School of Public Affairs에 위치한 New York Census RDC-Baruch, 코넬 대학에 위치한 New York Census RDC-Cornell, 듀크 대학에 위치한 Triangle Census RDC가 있으며 그 위치는 CES 홈페이지에 상세히 나타나 있다.

〈표 1-4〉 주요 지역 RDC 개요(2006년 10월 기준)

설치도시 (설치기관)	Boston (NBER[1])	LA (UCLA)	Chicago (FRBC[2])
설립년	1994년	1999년	2002년
관계 연구기관	NBER	UCLA	FRBC, Chicago Univ., Illinois Univ., Northwestern Univ., Argonne National Laboratory
센서스국 직원 수	1명 (경제학박사)	2명 (경제학박사 1명 포함)	
직원 외 스태프의 수	0명	0명	
부스의 수	8개	단말기 8개	
이용시간	24시간	24시간	
이용요금	$15,000/년	$3,125/월 $15,000/년 소속대학원생 무료	
연간수락 proposal 수	12개		
이용자 중 대학원생 비율	약 50%	약 33%	

설치도시 (설치기관)	Ann Arbor (미시건대)	Ithaca NY (코넬대)	Durham (듀크대)
설립년	2002년	2006년	2000년
관계 연구기관	미시건대, ICPSR[3]	Baruch Coll., NY city Univ., FRB4) NY, Fordham Univ., NBER, NY Univ., Pace Univ., Princeton Univ., Russell Sage Foundation, Rutgers Univ., NY State Univ., Yale Univ.	Duke Univ., M Carolina Univ., NC State Univ.
센서스국 직원 수	1명 (경제학박사)		1명 (경제학박사)
직원 외 스태프의 수	2명 (미시건대 교원)		1명 (듀크대 교수)
부스의 수	8개		8개
이용시간	24시간		24시간
이용요금	$15,000/년 미시건대 교원, 대학원생 무료	비참가기관의 경우 $20,000/년	$15,000/년 관계대학 교원, 대학원생 무료
연간수락 proposal 수	15개		5~6개
이용자 중 대학원생 비율	약 50%		약 42%

주: 1) National Bureau of Economic Research. 2) Federal Reserve Bank of Chicago.
3) Inter-University Consoritum for Political & Social Research.
4) Federal Reserve Bank.

자료: 神林 龍,「北米における政府統計個票調査公開の現状に關する調査報告: 米國勞働統計局, 米國センサス局およびカナダ統計局のオンサイトリサーチを中心に」, 2007을 참조하여 정리.

usRDC에서 제공하는 자료의 목록은 <부록 A>에 첨부하였으며, 주요 지역 RDC의 개요는 <표 1-4>에 요약하였다. 마이크로데이터 이용시 부과되는 비용은 usRDC 관계기관의 이용자에게는 대부분 무료이나 관계기관에 속하지 않는 이용자는 1년에 약 15,000달러를 지불해야 한다.

2) 자료신청 절차

RDC를 통한 자료의 신청까지는 다소 까다로운 절차가 요구된다. 먼저 웹사이트 상에서 이용하고 싶은 데이터를 확인 후 이용자의 계정(account)을 만들고 가장 가까운 usRDC에 체류하고 있는 센서스국 직원과 접촉한다. 이후 제안서(proposal)와 이력서 등의 필요서류를 첨부하여 온라인으로 제출하면 자료의 신청이 시작된다. 이용자의 요건은 국적이나 소속의 제약은 없으나 usRDC를 이용하기 위해서는 SSS(Special Sworn Status)를 얻고, 무급이지만 연방직원으로서의 고용계약을 체결해야 한다.

제안서가 나타내야 할 기준은 ① 연구 내용이 연방법 13조에 규정된 센서스국의 직무내용을 채우는 것, ② 과학적으로 이점이 있는 것, ③ 공공이용파일(Public Use File)의 데이터에서는 분석 불가능한 것, ④ 분석의 성공 전망이 있는 것, ⑤ 비밀보호 정보·성과물의 취급에 관한 제한을 인정하는 것으로 정하고 있다. 제안서는 수시로 접수가 가능하며 제안서의 심사 프로세스는 다음과 같다. 우선 제출된 제안서에 대해 usRDC의 심사위원회 내에서 보통 2명의 담당자가 결정되어 6주일 이내에 각각 2페이지 정도의 심사보고가 작성되는데, 1개의 제안서당 대체로 4~5시간이 소요되며 이 심사보고가 CES(Center for Economics Studies)의 담당자에게 보내져 자료제공여부가 최종 결정된다. 주된 심사내용은 이용하고자 하는 데이터로 신청된 자료의 분석 가능 여부와 익명성 저촉 여부이며 연구 내용이 센서스국의 직무내용에 일치할지에 대해서는 실질적으로 대부분 심사하지 않는다. usRDC의 이용이 허가된 후에는 SSS를 얻기 위한 법적인 수속으로 옮겨가고, 그 후 연방직원으로서의 고용계약이 체결된다. 최근에는 비밀보호의 문제가 중요시되어 이 수속에 많은 서류와 지문 등의 사적정보와 시간이 요구되며 이것들이 끝난 후에 각 usRDC에서 ID 카드가 작성되어 실제의 분석이 가능하게 된다. CES에서는 이러한 절차를 고려하여 제안서의 제출로부터 데이터의 이용 가능까지의 기간을 6~12개월 정도로 어림잡아 연구계획을 세우도록 권장하고 있다.

3) 제공 후 관리

이용자는 usRDC의 데이터 서버에 접속하고 거기에서 모든 작업을 완결해야 한다. 각 단말기에 출력장치는 없으며 도출결과 등을 종이 혹은 전자매체로 usRDC 밖으로 반출할 경우 각 usRDC에 체류하고 있는 센서스국 직원의 승인을 거쳐야 한다. 분석

프로그램 등을 업로드할 경우도 같다. usRDC로의 출입기록과 데이터 서버의 로그는 모두 보존되어 누가, 언제, 어떻게 데이터에 접근하고 어떤 작업을 행했는지를 사후적으로 확인하는 것이 용이하다. 작업 중에 제안서에 제안한 분석 내용을 변경할 필요가 생겼을 경우에는 원칙적으로는 변경 내용과 그 필요성을 제출하고 승인을 받을 필요가 있다. 그러나 이용하는 데이터를 변경할 경우나 근본적으로 다른 연구를 하는 경우 이외는, 사실상 사후적인 승인절차를 거치고 있다. 제공규정을 위반했을 시에는 강한 법적 제재를 받게 되는데 비밀보호에 대한 위반시 징역 5년 또는 250,000달러 이하의 벌금(혹은 그 양쪽을 병과)이 부과된다.

또한 단순한 분석결과뿐만 아니라 데이터를 이용해 나오는 최종 성과물에 대해서도 엄격히 관리하고 있는데 최종 성과물은 비밀보호에 관한 심사위원회에서 승인을 얻은 후, CES의 discussion paper로 공표되며, 학술지에의 투고가 허가된다. 대개의 경우 회귀분석 및 각 변수의 1차·2차의 요약 통계량이 비밀보호에 저촉될 경우는 드물다. 그러나 다중 크로스 표나 산포도를 작성했을 경우는 비밀보호에 저촉될 수 있는 가능성이 많으므로 심사가 엄중하고 따라서 성과물을 공표할 때까지의 시간이 많이 소비된다. 요약 통계량이라 할지라도 중간값이나 최대값, 최소값 등 특정 샘플을 내보일 경우에는 심사의 대상이 된다. 그래서 CES는 논문으로 게재하는 크로스 표는 될 수 있는 한 공표 데이터를 이용하는 것을 권장하고 있다.

2. 미국 노동통계국(Bureau of Labor Statistics:BLS)

미국의 고용・물가 등의 경제통계를 담당하는 연방기관인 노동통계국에서도 마이크로데이터를 제공하고 있는데, 어느 통계가 어느 방법으로 공개될지는 자료공개검토위원회(Disclosure Review Board)에서 조사항목마다 결정한다. 동일한 조사 내에서 공공이용파일(Public Use File)로 제공되는 부분, Off-site Research[7]에서의 사용이 인정되는 부분, On-site에서만 인정되는 부분, 전혀 사용을 인정받지 못하는 부분이 혼재하는 경우도 자주 있으며, 조사시에는 어느 방법으로 어느 항목이 공개될지가 사전에 피조사자에게 설명된다. 사후적으로 공개방법을 변경 하는 것은 어렵지만 자료공개검토위원회(Disclosure Review Board)의 결정에 근거한 변경은 인정하고 있다. 자료공개검토위원회는 내부조직이며 외부자는 관여하지 않는다. 단, 외부전문가가 관여하는 자료공개기술자문위원회(Technical Review Board)가 별도 설치되고 있어, 외부자가 통계조사의 공개방법에 관해서 의견을 진술하는 기회가 확보되어 있다.

---

7) 자료제공 형식 중 Licensed File에 해당하는 경우로 특정한 심층이용자에게 제공되며 제공 전에 법적인 조치(서명 등)가 필요한 형식이다.

가. 공공이용파일(Public Use File)

공공이용파일로 제공되는 자료는 가계조사인 Consumer Expenditure Survey와 노동력 조사인 CPS(Current Population Survey), 패널데이터인 NLS(National Longitudinal Survey)가 있으며, CPS와 NLS는 센서스국과 공동으로 제공되는 자료이다. 이 자료들은 노동통계국(Bureau of Labor Statistics: BLS)에서 전자매체로 구입이 가능하며, 앞의 센서스국의 Public Use File에서 소개한 바와 같이 ICPSR(Inter-University Consoritum for Political & Social Research) 등 각지의 데이터 아카이브에도 기탁되고 있어 교육이나 연구에 빈번하게 이용되고 있다.

나. Off-site Research(Licensed File)

인가파일(Licensed File) 형식으로 제공하는 Off-site Research는 보다 엄격하게 보호할 필요가 있는 조사항목에 대해서 일정 절차를 거쳐 특정 개인에게 데이터를 제공하는 방법으로, BLS에서는 엄격한 절차를 두고 이 방법으로 데이터를 제공하고 있다.

BLS가 관할하는 통계 안에서 이 방법으로 이용 가능한 데이터는 National ongitudinal Survey Geocode Data와 Census of Fatal Occupational Injuries 등이 있는데, National Longitudinal Survey Geocode Data는 NLS79(National Longitudinal Survey of Youth 1979), NLS(National Longitudinal Survey) Young Adult, NLS97(National Longitudinal Survey of Youth 1997)의 지리적 항목을 중심으로 한 조사항목을 말한다. 이 조사 항목은 각 조사에 있어서의 응답자의 거주지정보 및 해당지역의 경제변수, 태어난 시점의 거주지정보, 14세 시점의 거주지 정보, 이혼 후 양육에 관한 합의가 성립한 거주지 정보, 출생・결혼・이혼・입학 등의 연월일, 입학대학의 명칭과 소재지, 박사학위 취득시 전공명, 양친의 거주 지에서의 거리 등이 있다.

자료의 이용 자격 요건으로는 미국 시민일 필요는 없지만 미국 주재로 소속기관이 미국에 있어야 하며, 이용자와 BLS 사이에는 BLS Agency Agreement가 체결되며[8] 동시에 이용자의 소속기관과 BLS에도 합의 계약이 행해진다. 소속기관에는 기관장(대학이면 부학장 이상)의 승인을 얻는 것 이외에 Project Coordinator라고 불리는 실질적 책임자를 지정한 뒤 서버 등 전자환경, 파일이나 CD-ROM 보존을 위한 물리적인 준비를 갖추는 것이 요청된다. 소속기관과 BLS 사이의 계약서에는 계약 위반이 있었을 경우에는 소속기관이 민사상의 손해배상 의무를 지는 것도 명기되어 있다. 따라서 비밀보호와 관련한 위반 등이 있었을 경우, 이용자가 연방정부의 Agent로서 형사・민사소추를 받을 가능성이 생기는 것은 물론 이용자의 소속기관도 민사소추를 받을 수 있게 된다.

8) BLS의 무급 Agent가 되는 것의 합의 계약.

이용 기간은 학생이 1년이며 연구자는 3년 또는 소속기관에 재적 중인 기간이다. 이용 장소는 소속기관 내의 BLS가 지정한 장소이며 이용자의 자택 등은 인정받지 못한다. 이용자가 소속기관을 변경했을 경우 이용 계약은 그 시점으로 끝난다. 제공 후의 관리도 상당히 엄격한데 자료 이용 종료 시에는 소속기관 책임자에 의한 데이터 CD의 반환과 데이터 소거의 확인, BLS에 의한 성과물의 검열이 요청된다. 또한 이용 기간 동안에 BLS 직원에 의한 현장사찰이 행해지며 사찰 시에는 특히 데이터 보존의 물리적 환경, 서버나 main frame의 보안 환경 등이 체크된다. 현재까지 Off-site Research에 있어서의 심각한 계약 위반은 데이터를 자택에 가져간 경우가 1건 있었다. 이 경우 이용 계약이 중단되어 해당 이용자, 해당 소속기관의 재이용이 금지되었으며, 이용자에 대한 형사 소추, 소속기관에 대한 손해배상청구까지는 이르지 않았다. 사찰로 밝혀진 위반의 대다수는 CD- ROM을 보관하고 있는 캐비닛 열쇠를 잘 사용하지 않은 경우와 데이터를 네트워크 상 공유해 다루는 등이었으며, BLS 직원이 그 자리에서 개선을 요구하고 소속기관이 대응・보고하는 것으로 수습되었다.

다. On-site(Data Laboratories)

BLS의 On-site 이용 조건은 첫째 미국 시민일 것, 둘째 자격기관에 학생으로 재적하고 있거나 또는 고용되어 있는 것인데, 첫 번째 조건은 그다지 엄격히 해석되지 않고 두 번째 조건이 엄격하게 해석되고 있다. 두 번째 조건의 자격기관은 미국고등교육기관, NPO(국세청(IRS)이 정하는 기준 501c (1) 또는 501c (3)에 적합한 조직), 지방정부 또는 지방정부의 부속 기관, 지방정부를 대표로 하는 기관, 연방정부 연구기관 등이다. 외국 국적의 객원교수라도 체류 연구기관이 자격기관에 포함될 경우에 한하여 이용 자격 인정을 받는다. 제공되는 자료의 목록은 부록으로 첨부한다.

제안서 제출에서 자료이용까지의 절차를 살펴보면, 먼저 5~10페이지 분량의 제안서를 제출하고 제안서에는 다만 통계적인 이용을 목적으로 하는 것, 연구 내용에 대해 통계 기술적인 메리트가 있는 것, BLS의 흥미가 있는 것, BLS의 정책을 진척시키는 것을 나타낼 필요가 있다. 그밖에 최종 성과물의 공개 방법, 보조금, 공동 연구자의 유무 등을 써 넣는 것이 요구된다. 제안서의 제출 마감은 연 4회(1월, 4월, 7월, 10월의 각 15일)이며 심사에 2개월, 심사 후 법적 계약 등에 2개월을 필요로 한 경우가 대부분이다. 제안서를 접수하는 시기는 자료이용까지 3개월부터 6개월을 계획하도록 한다.

제안서에 대한 심사는 MARB(BLS-wide Microdata Access Review Board)에서 하는데, BLS 내에서 데이터의 보안을 담당하는 직원, 이외 피신청 데이터를 관할하는 그룹의 시니어・이코노미스트가 출석하여 심사를 담당한다. 이용기간은 신청시 최장 2년간이며

한 번 더 연장이 가능하여 최장 4년 동안 이용 가능하다. 이용시간은 월요일부터 금요일까지의 8시부터 17시까지이며, 별도 신청하면 아침 6시부터 이용할 수 있다.

BLS On-Site 서버에는 SAS 및 STATA가 인스톨되어 있으며 랩탑컴퓨터(laptop computer) 등의 반입은 인정하나 인터넷이나 단말기에의 접속 단자는 설치되어 있지 않고 가져온 전자기기를 어떠한 방법으로도 인터넷에 접속하는 것은 금지되어 있다. 단말기에 허용되는 외부출력은 화면만으로 한정되며 종이매체·전자매체에서의 출력은 할 때마다 BLS 담당자의 검열을 받을 필요가 있다. On-site의 서버가 설치된 장소의 출입에는 미리 등록된 ID 카드를 사용한 인증이 필요하며, 이것들을 통해 언제, 누가, 어디에서, 어느 데이터에 접근하여 무엇을 했는지를 알 수 있다.

BLS의 On-site는 1997년 9월 이후 현재의 방법으로 정식으로 운영된 이래 총 120명의 연구자가 이 제도를 이용하고 있으며, 2006년 기준 연간 이용 중인 연구제안서는 36개이며 연구자는 50명이다.

3. 캐나다

캐나다는 통계에 있어 집중형국가로 행정정보에 접근하는 권한도 통계법에 의해 주어져 있는 등, 정부통계에 관한 강력한 권한을 통계청이 가지고 있다. 그에 따라 공개정보의 비밀보호에 관한 요청도 강해서 이와 관련한 위반에 대해서는 형사죄가 성립하는 등 강한 법적구속력을 가진다. 권한 없는 정보의 접근을 시도할 경우 6개월 이하의 징역 또는1,000달러 이하의 벌금, 혹은 그 양쪽을 병과하며, 자산 가격에 영향을 미치게 하거나 투기행위를 행하거나 하는 등 개인의 특정 정보를 경제 거래에 이용하기로 했을 경우는 5년 이하의 징역 또는 5,000달러 이하의 벌금, 혹은 그 양쪽을 병과한다.

가. 공공이용파일(Public Use File)

각 조사의 공공이용파일 작성 시에 마이크로데이터 제공 위원회(Microdata Release Committee)가 제공하는 통계의 범위나 조사항목을 검토한다. 제공되는 데이터로는 센서스 3% 샘플 자료, 가계소비조사자료 등이다. 2002년 4월 2일부터 원격접속파일(Remote Data Access) 방법으로도 제공하고 있으며, 비용은 2001년 센서스 자료가 1,500달러로 개인 파일은 1,000달러이고 가족파일은 300달러, 가구 및 주택파일은 200달러이다. 2002년 가계소비조사의 경우는 3,210달러이다. 자료 이용 전에 마이크로데이터 허가협정(Microdata Licence Agreement)이 필요하다.

통계연구목적을 위해 개별정보들을 연결한 레코드 연결 데이터(Record Linkage Data)도

일부 제공하고 있는데, 이 경우 일반 Public Use File보다 정보공개를 보다 신중하게 처리한다. 1986년부터 레코드 연결에 관하여는 Agency's Policy on Record Linkage가 담당하고 있으며 이 자료의 이용 조건은 단지 통계적인 목적으로 사용하여야 하며, 자료 이용으로 인하여 창출되는 공공의 이익이 개인정보보호의 이익보다 크다고 판단되는 경우이어야 하는데 이는 검토위원회(Review Committee)의 승인 후 통계청장(Chief Statistician)의 승인 및 장관(Minister)의 승인을 얻어야 한다. 통계청장의 승인을 거칠 때는 비밀보호위원회(Office of the Privacy Commissioner)와 공공이익단체(public interest group)의 자문을 구하여 정한다.

캐나다에서는 특징적으로 DLI(Data Liberation Initiative)를 운영하고 있는데, DLI에 참여하는 캐나다 교육기관(대학, 연구소 등)이 연간비용을 부담하고 학생과 임직원들이 무제한으로 마이크로데이터를 사용하도록 하는 제도(대부분 도서관과 정보센터에 위치)를 두고 있다. 이는 1980년대에 마이크로데이터 비용이 높아져 연구자들이 상대적으로 가격이 저렴한 미국, 영국, 중국 자료를 이용하자, 캐나다의 현실에 대한 연구가 이루어지지 않는 문제가 발생하여 1989년에 센서스 가격의 상승을 막고자 공공 데이터 이용자 연합(Canadian Association of Public Data Users)과 연구 도서관 연합(Canadian Association of Research Libraries)에 의해 임시구매컨소시엄(ad hoc buying consortium)이 조직되고, 1993년에 사회과학연맹(Social Science Federation of Canada)의 지원으로 작업반(Working Group)이 구축되어 DLI프로그램을 제안하여 1996년에 재무부 이사회(Treasury Board)의 승인을 거쳐 DLI 5개년 시험 프로젝트가 성공, 2001년 3월 DLI 프로그램이 영구적으로 정착되었다.

또한 DLI 외부자문위원회(DLI-EAC, External Advisory Committee)가 운영되고 있는데 이는 교육기관, 도서관 대표로 구성된다. 조직구성은 캐나다 통계청 외부에서 9명의 투표권자를 선정(지역 대표도서관, 학계, 연구자 대표)하고 캐나다 통계청 내부에서 5명의 투표권자를 선정한다(DLI 담당과장, 자료제공담당 등). 임기는 6년으로 1/3이 2년마다 바뀌며 66개 기관이 참가, 하부위원회를 둔다. DLI 외부자문위원회는 각 교육 기관 및 연구기관의 요구사항을 대표로 캐나다 통계청에 전달하며 마이크로데이터의 제공방법을 제안하거나 새로운 상품 서비스 형식을 제안 한다. 또한 마이크로데이터와 관련한 연간 보고서를 매년 6월 30일까지 작성하여 캐나다 통계청장에게 제출 발표하며 이 자료는 DLI 웹사이트에 게시된다.

나. On-site(Data Laboratories)

캐나다의 On-site 방법은 미국의 방법을 벤치마킹하여 전반적인 부분에서 거의 비슷하게 운영된다. caRDC(Canada Research Data Center)라고 부르며 이용시 데이터 이용자는 선서를 하고 캐나다 통계청과 무급의 고용계약을 체결한다.

2007년 1월 기준, 캐나다 국내에 본소와 출장소를 합해 20개를 운영하고 있으며 이들 시설은 법적으로 캐나다 통계청의 건물로 차용, 각 시설에는 캐나다 통계청 직원이 적어도 1명은 상주하고 있다. 이용자 개인의 PC 지참, PDA나 휴대폰 등의 사용은 원칙적으로 허용되지 않는다. usRDC가 연구목적의 이용에 거의 한정되는 반면, caRDC는 주로 대학원 수준에서의 교육목적 이용도 적극적으로 추진하고 있다.이용자격 요건으로는 첫째 정부기관 또는 RDC 네트워크 참가기관에 소속되어 있을 것, 둘째 연구자(소속이 있으면 국적은 묻지 않는다), 연방정부기관의 피고용자(주지방정부피고용자는 별도상담), 대학원생의 석사논문 또는 박사논문 집필에 일부 이용하는 자, 대학원생의 석사논문 또는 박사논문과는 관련되지 않고 이용하는 자, 학부학생의 최종학년에 속하는 자 중 어느 하나에 포함될 것이다. usRDC와 비교하여 대학원생의 이용이 특히 명확히 규정화되어 있는 점과 학부학생에게도 제공하고 있는 점이 특징이다. 소속이나 국적이 캐나다와 관련되지 않는 연구자도 전혀 이용할 수 없는 것은 아니고 별도 상담으로 결정되고 있어, 상당히 광범위하게 이용자를 상정하고 있다고 말할 수 있다. BLS나 usRDC와 다르게 통계청 업무와의 관련을 기록할 필요는 없다.

이용 절차는 웹사이트 상으로 사용자 계정을 작성하고 연구계획서와 이용자 본인 및 공동 이용자의 이력서 등을 첨부하여 신청한다. 신청된 연구계획서는 선별 검토 위원회(Selection and Review Committee)에 회부되고 2명의 연구자와 캐나다 통계청 직원이 심사한다. 연구적인 부분과 비밀보호적인 부분을 나누어 심사하는 것이 특징인데, 심사 담당 연구자는 관련 분야에서 뽑혀 연구목적과 데이터분석 수법의 합리성을 주로 심사하며, 통계청 직원은 On-site로의 접근 필요성이나 분석 결과가 비밀보호에 영향을 미치는 부분을 주로 심사한다. 제공되는 자료의 목록은 <부록 C>로 첨부한다.

제공 후의 관리 부분을 살펴보면, 연구 수행 중 반출하는 중간성과물에 대해서는 캐나다 통계청의 승인을 얻을 필요는 있지만 캐나다 통계청에 제출할 필요는 없다. 그러나 최종성과물은 제출해야 한다. 어떤 형태로 제출할 것인가는 계약시에 미리 지정하며 RDC working paper, Peer-reviewed journal article, Book or Book chapter, Thesis or Dissertation, Commissioned Report 중 선택한다.

2006년 776개의 프로젝트, 이용자 853명(516명의 연구자, 337명의 학생), 1,225개 데이터(NLSCY[9]: 245개, CCHS[10]: 223개, SLID[11]: 143개의 순)가 신청되어 활용되고 있다.

9) National Longitudinal Survey of Children and Youth.
10) Canadian Community Health Survey.
11) Survey of Labor and Income Dynamics.

4. 호주

특정 개인이나 기업에 대한 정보를 식별해 낼 수 없도록 비밀보호방법[12]을 적용한 Public Use File을 CURF(Confidentialised Unit Record File)라 부른다. CURF는 세분화 정도(levels of detail)에 따라서 3가지 종류로 분류되며 그에 따라 제공방법을 달리하고 있다.

1) 제공방법

CURF의 세분화 정도는 기본형(Basic), 확장형(Expanded), 전문형(Specialist)으로 분류되며 전문형이 가장 세분화된 자료이다. 기본형(Basic)은 CD-ROM, RADL(Remote Access Data Laboratory), ABSDL(Australian Bureau of Statistics Site Data Laboratory)로 접근 가능하며, 확장형(Expanded)은 RADL, ABSDL로 접근 가능하며, 전문형(Specialist)은 ABSDL로만 접근 가능하다. RADL은 허가 받은 이용자가 호주 통계청(Australian Bureau of Statistic: ABS)이 제공하는 마이크로데이터를 인터넷을 통해 제공받을 수 있는 방법을 말하며, 매년 사용자의 비밀번호가 변경되며 승인된 데이터에만 접근이 가능하다. ABSDL은 ABS의 감독 하에 ABS 내에서 마이크로데이터를 이용하는 방법으로, 향후 먼 지역의 사용자를 위해 ABS 외부 데이터실험실(Data Laboratories: DL)을 개발할 예정이다. 전문파일의 경우 신청하기 전에 호주 통계청(ABS)의 마이크로데이터 담당자와 접촉해야 하며 각각의 경우에 따라서 CURF의 제공 여부를 결정한다.

호주 통계청에서는 확실한 수요(demonstrated user demand)가 있을 때 생산하고 제공되고 있는 CURF 리스트를 홈페이지에 게시하며 향후 서비스될 CURF 목록과 일정을 게시한다. CURF 사용자 훈련 매뉴얼(Training Manual), 사용자 가이드(User Guide)를 제공하며, 마이크로데이터 관련 뉴스레터를 분기별로 발간하고 홈페이지 상의 FAQ도 활성화하는 등 마이크로데이터 제공 관련 홍보도 중요시하고 있다. 외국인에 자료 제공은 경우에 따라 다르지만 외국인이라고 해서 기본적인 자료만 접근할 수 있는 것이 아니라 전문(Specialist)파일에도 접근이 가능하다. 호주 통계청에서 제공하는 전체 자료 목록은 <부록 D>를 참조하기 바란다.

12) 이름, 주소 등 제거, 값 제거(이상치, 작은 값 등), 셀 변경, 통합(Randomly, Swap, Drop, Substitute)기법 등 적용.

〈표 1-5〉 호주의 제공방법별 이점 및 제약사항

	CD-ROM (기본형)	RADL (기본&확장형)	ABSDL (전문형)
데이터 세분화 정도	세분화 정도가 약함	다소 세분화됨	상당히 세분화됨
이점	· 이용자가 자신의 컴퓨터에서 제약없이 사용 가능	· 인터넷이 가능한 모든 컴퓨터에서 접근 가능 · 좀 더 세분화된 데이터 분석이 가능 · 시스템 내에서 SAS, Stata와 SPSS 소프트웨어가 제공됨	· ABS내 관리감독 하에 지정된 컴퓨터에서 접근 가능 · 높은 수준의 데이터 분석이 가능 · 시스템 내에서 SAS, Stata와 SPSS 소프트웨어가 제공됨 · 시스템에서 다른 데이터 세트들과의 결합도 허용될 수 있음
제약사항	· 법적 서약서(legal undertaking)에 의해 구속됨	· 법적 서약서에 의해 구속됨 · 개별 레코드는 ABS IT 환경 내에 남아 있음 · 쿼리의 속성과 결과물의 성격과 크기에 따라 제약이 있음 · 이용자의 활동이 모니터됨	· 법적 서약서에 의해 구속됨 · 개별 레코드는 ABS IT 환경 내에 남아 있음 · 쿼리의 속성과 결과물의 성격과 크기에 따라 제약이 있음 · 이용자의 활동이 모니터됨 · ABS 내에서 ABS의 감독을 받아야 함

자료: 통계청 국제통계협력과, "외국인에 대한 마이크로데이터 제공검토", 2004에서 인용.

2) 가격책정

CURF의 가격책정은 세분화 정도와 자료 제공방법에 따라 <표 1-6> 과 같이 차등 책정되고 있으며, 이용자별 가격책정이 아닌 기관별 책정을 하고 있다. 호주 통계청 Microdata Access Strategies Section의 디렉터인 Kim Farley-Larmour에 따르면, 호주의 마이크로데이터 가격책정은 호주정부 가이드라인에 따라 원가보상원칙(cost recovery ABS Pricing Policy)을 기준으로 책정하고 있다고 한다. 한 기관에서 다수의 이용자가 해당 마이크로자료를 함께 사용한다고 해서 가격이 올라가는 것은 아니지만 이용자 수를 제한할 수는 있다.

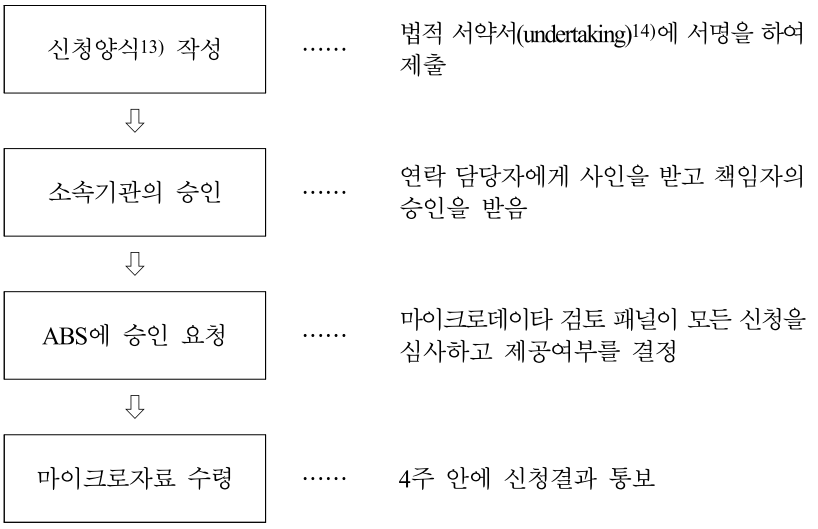
〈표 1-6〉 가격책정 기준

가격책정 기준	가격
기본형(CD-ROM)	$1,320
기본형(RADL)	
확장형(RADL)	
기본형(CD-ROM/RADL)과 확장형(RADL) 묶음자료	$1,980
컨설턴트(프로젝트당)	$435
전문형(ABSDL)	용도에 따라 다름

자료: 호주통계청(ABS) 홈페이지.

3) 자료신청절차

사용자 개인, 연락 담당자(contact officer), 책임자(responsible officer), 컨설턴트, 그리고 컨설턴트 지원기관별로 구분하여 자료 신청절차 및 역할과 의무사항 등에 대한 일련의 과정들을 홈페이지를 통해 상세히 안내하고 있다. 대부분 승인된 기관에서 자료를 사용하고 있으며, 승인된 기관과 전혀 연관이 없는 개인의 CURF 접속은 불가능하다.



13) 신청하는 CURF 개수에 따라 신청 양식이 다름
14) 개인이나 기관의 책임자별로 서약서 양식이 다름

4) 관계자별 역할과 의무

호주 통계청의 경우는 통계청에서 바로 개인으로 자료를 제공하는 것이 아니라 기관을 통해 제공하므로 관계자별 역할과 의무를 가진다. 관계자는 사용자 개인과 연락담당자, 책임자, 컨설턴트로 구분할 수 있는데, 먼저 사용자 개인의 경우는 신청서와 서약서를 작성하기 전에 반드시 CURF 훈련 매뉴얼(Training Manual)을 읽어야 하며, 신청서와 서약서는 소속기관의 연락 담당자의 서명을 받아 ABS에 제출해야 한다.

사용자의 의무는 그들이 서명한 서약서에 의해 통제되며, CURF는 반드시 명시된 통계적 목적으로만 사용해야 한다. 특정 개인이나 기관을 식별하려는 시도를 해서는 안 되며, 다른 사람이나 기관에 자료를 넘겨서는 안 된다. 승인된 기간이 만료된 후이거나 소속기관으로부터 접근권이 박탈된 후 소속기관을 그만두었을 시에는 자료 사용을 금지하고 있다.

연락 담당자는 ABS 담당자와의 연락을 담당하며 사용자 개인의 CURF에 대한 접근을 도와주는 역할을 한다. 또한 소속기관에 대한 책임자의 서약서(ROU[15])를 조정하고, 사용자의 서약서(IUU1[16])와 신청서(REQ) 및 컨설턴트 서약서(CON[17]) 작성을 도와야 한다. 그리고 기관이 보유한 CURF 목록을 관리하며, 원본 CD-ROM를 안전하게 보관하고 복사본을 개인 사용자에 제공하여야 한다. 기간 만료 후 원본 CD-ROM을 ABS에 반납하고, 개인에게 제공되었던 복사본을 파기하여야 한다. 사용자가 소속기관을 떠나게 되면 사용자의 모든 CURF에 대한 접근권이 박탈되고 CURF를 반납하여야 하며 떠나기 최소 10일 전에는 ABS에 알려야 한다.

책임자의 경우는, 마이크로자료 이용 신청 시 책임소재를 명확히 하기 위해 반드시 각 기관의 장 또는 허가된 개인을 책임자로 지정하여 접촉 해야 하며, 일반적으로 대학의 부총장, 정부의 기관장, 사기업의 CEO가 책임자로 지정된다. 컨설턴트의 경우, 각 기관이 기관에서 보유하고 있는 CURF를 사용하는 연구 프로젝트를 수행하기 위해 외부로부터 컨설턴트를 고용하는 경우 컨설턴트용 신청서와 서약서에 서명하면 ABS에서 컨설턴트 접근권을 제공한다. 사용자와 동일한 의무를 부여받으며, 컨설팅(용역 사업)이 완료되면 모든 CURF 접근권이 박탈되고 CURF를 파기하여야 한다. 컨설팅(용역 사업)이 완료되면 고용된 기관으로 CURF와 관련된 모든 분석 결과물들을 넘겨야 하며 이에 대한 소유권은 고용된 기관에 있다.

---

15) Responsible Officer Undertaking.
16) Individual User Undertaking.
17) Consultant Undertaking.

호주 통계청에서는 3명으로 구성된 Data Access & Confidentiality Methodology Unit에서 CURF를 생산하며, 개인의 정보를 보호하기 위하여 CURF를 어떻게 생성하는 것이 효율적인가를 연구하고 있다.

또한 4명으로 구성된 CURF Management(sub-section)에서는 생성된 CURF를 이용자에게 제공하고, 관계자들을 총괄하며 전반적인 관리를 하는 역할을한다.

5) 제공 후 관리

제공 후 CURF를 이용하여 진행된 연구 결과물들을 홈페이지에 공개하고 Auditing Program을 통해서 결과물을 검토하고 있다. Auditing Program은 ABS의 코드와 결과물을 감사(Audit)하여 개별 정보를 노출시키지 않도록 하며, RADL로 이용시 매년 어떤 자료를 이용하였는지, 접근횟수는 얼마인지 등을 체크하여 문제가 발생하면 이용 중단시키는 역할을 하고, 또한 결과물을 체크하여 정보를 지나치게 많이 담는 경우에는 다시 ABS로 회수하고 향후 사용중단 조치를 취한다.

자료 제공에 관한 의무 위반시 CURF 접근권한을 박탈[18]하고 이를 어길 시 13,200달러 이내의 벌금 또는 2년 이하의 징역에 처한다. CDROM의 수거는 잘 이루어지고 있으며 간혹 분실된 경우에는 차후 이용을 못 하게 한다.

호주에는 ABS-AVCC(Australian Vice Chancellors Committee) CURF협약[19]을 두어 학문적 연구와 대학수업을 지원하기 위하여 대학, 정부와 통계청이 협약을 맺어 CURF 이용을 활성화하고 있다. 해당기관의 연락 담당자(Contact Officer)는 CD-ROM 원본을 보유하고 사용자에게 복사본을 제공하고 수거하며 주로 교수, 사서, 관리자 등으로 지정된다. CD-ROM과 RADL에의 접근권이 포함되어 있으나 ABSDL 접근권은 포함되어 있지 않다.

5. 영 국

영국의 마이크로데이터 제공현황에 대해 각 기관별로 알아보고자 한다. 영국은 영국 통계청(Office for National Statistics: ONS)의 주관하에 크게 4개의 기관 CCSR(Centre for Census and Survey Research), UKDA(UK Data Archive), CeLSIUS(Centre for Longitudinal

18) 이전에 일부 위반사례가 발견되어 개인(대학 교수)으로부터 향후 규정을 준수하겠다는 서약을 받은 적이 있으나 중대한 위반사례는 아직 없었음.

19) 대학교에서 CURF를 상업적 목적으로 사용하려는 경우에는 이 협약이 적용되지 않음. 또한 이 협약을 맺지 않은 대학교는 현행 ABS의 CURF 가격을 지불하고 구매할 수 있으며, 학문적 연구 목적으로 CURF를 이용하여 그 결과물을 발간할 경우 이를 인정함.

Study Information and User Support), ONS-IDBR(Inter-Departmental Business Register)에서 마이크로데이터를 제공하고 있으며, 기관에 따라 주로 다루는 자료의 종류가 다르다. 각 기관에서 제공하는 자료는 모두 비밀보호기법이 적용된 마이크로데이터이며, 상세하고 개인 식별이 되는 자료는 ONS의 담당자와 연락하여 제공여부를 결정한다.

가. CCSR(Centre for Census and Survey Research)

ESDS(Economic and Social Data Service) Government의 주도적인 역할을 하는 CCSR은 1992년 4월에 설립되어 맨체스터(Manchester) 대학 내에 위치하고 있으며 주로 인구센서스 자료를 제공한다. 1991년, 2001년의 Census로부터 작성된 SAR[20]을 개인과 세대의 2종류로 제공하고 있는데 제공되는 마이크로데이터로는 2% 개인 SAR(약 110만 레코드), 3%개인 SAR(약 175만 레코드), 1% 세대 SAR(약 20만 레코드, 세대원수 약 50만 명), 5% SAM[21](약 300만 명의 개인 레코드)이며, 보다 상세한 개인 데이터인 CAM(Controlled Access Microdata Samples)은 ONS 내에 On-site로만 이용이 가능하다. On-site로 이용시 좀 더 까다로운 절차를 가지는데, 구체적으로 설명하면 개인 CAM은 CRAB(Census Research Access Board)의 심사를 받은 후, ONS의 마이크로데이터공개위원회(Microdata Release Panel: MRP), GROS(General Register Office for Scotland), NISRA(Northern Ireland Statistics and Research Agency)의 심의 후 이용 가능하다. 영국의 국가특성상 잉글랜드, 스코틀랜드, 북아일랜드 각각의 심의기구를 모두 거쳐야 한다. 외국인에 대해서는 3% 개인 SAR과 5% SAM 데이터에 한정해 이용을 승인하고 있다. 개인 CAM자료를 제외한 1991 SAR, 2001 SAR, 2001 SAM의 이용시에 학술목적으로 이용하는 경우는 아이디와 패스워드를 취득한 다음 CRS(Census Registration Service)에 등록하여 이용 가능하며, 비학술 목적으로 이용하는 경우는 End user Licence Agreement에 서명하여야 한다.

데이터 수집, 제공, 분석에 있어서의 비밀 보호에 관한 연구를 실시하는 CAPRI (Confidentiality and Privacy)라는 팀을 별도 운영하고 있는데, 현재 Dr. Kingsley Purdam이 주도하며 맨체스터 대학에 소속되어 있는 사람들로 구성되어 있다.

마이크로데이터나 소지역 집계 데이터의 리스크 레벨을 입증하는 혁신적인 방법을 개발해 ONS의 센서스나 표본조사의 리스크 평가에 막대한 공헌을 하였으며, 국제 심포지엄도 개최하는 등 활발히 활동하고 있다.

---

20) Sample of Anonymised Record: 비밀보호기법이 적용된 센서스 마이크로데이터를 SAR이라 부른다.
21) Small Area Microdata file.

이용요금은 학술목적과 비학술 목적으로 크게 나누어지는데, 학술목적의 경우 무료로 제공되며 CD-ROM으로 구입 시에는 30파운드가 부과된다. 비학술목적의 경우에는 데이터 세트당 1,000파운드의 높은 가격이 책정되며, 중앙정부, 지방정부, 보건기관, 자선단체의 경우에는 500파운드가 부과된다. 각각의 경우에 17.5%의 부가가치세가 별도로 부과된다.

나. UKDA(UK Data Archive)

1967년에 설립되어 에섹스(ESSEX) 대학 내에 위치한 UKDA는 주로 각종 사회, 인문에 관한 서베이 데이터를 제공하고 있다. 제공되는 자료는 사회.인문 전반의 각종 서베이 데이터 및 2001 세대 SAR(Sample of Anonymised Record)이다. ARC(Acquisitions Review Committee)가 격주마다 개최되어 데이터 제안서를 심사한다.

마이크로데이터의 이용요금은 학술목적인 경우 무료이며, 비학술목적은 500파운드가 부과된다. CD-ROM으로 요청할 경우 기본료 7.5파운드에 수수료(handling fee) 2.5파운드가 부과되며, 우송비가 따로 추가되는데 영국 내에서는 3파운드, 영국 이외 유럽은 4파운드, 그 외 지역은 5파운드이다. 역시 17.5%의 부가가치세가 부과되며, 교육목적으로 사용하는 연구생과 학생, EU의 이용자, 각 학부 내의 목적(departmental purpose)으로 이용하는 에섹스 대학의 대학 관계자는 세금이 면제된다.

다. CeLSIUS(Centre for Longitudinal Study Informationand User Support)

CeLSIUS는 2001년 12월에 설립되어 런던시립대학(London City University) 내에 설치되어 있으며 LS데이터(Longitudinal Study Data)를 제공하고 있다. LS데이터는 1974년부터 작업을 시작하였으며 직업별 사망데이터의 타당성이나 장기적 출생데이터의 필요를 시작으로 작성되었다. 이 데이터는 링키지 된 1%(약 50만)의 잉글랜드와 웨일스의 센서스 개표데이터이다. 특정 4개의 탄생 월에 태어난 사람을 LS샘플로 지정하여 센서스정보와 다른 등록 정보를 연결하고 사망, 출산, 이주, 암환자 등의 다른 정보들을 추가하여 패널 데이터 형식으로 구축한 자료이다. 잉글랜드와 웨일스의 데이터를 제공하고 있으며 스코틀랜드는 현재작성 중이다.

자료의 이용에 관해서는 LSRB(Longitudinal Study Research Board)에서 심의한다. 영국내 학술기관 이용자 이외에는 ONS의 LS담당자와 직접 접촉하여야 한다. 자료이용 신청시 연구의 논점이 명확해야 하며, CeLSIUS의 직원과 연구과제 및 연구기간에 대해 상담하고 온라인 트레이닝을 실시한다. 비밀보호(Confidentiality)에 관한 계약서(LS License Agreement)에 서명을 한 후 연구계획서(form A)를 제출하고 CeLSIUS 직원의 지시에

따라 Data Dictionary를 이용해 변수를 추출하고 추출된 변수를 바탕으로 신청서(form B)를 작성해 ONS LSRB에 제출한다. ONSLSRB의 이용 승인 후 계약서를 제출하고 가공 데이터를 수령한다. 데이터와 관련한 자료를 발표할 때는 사전 final outputs clearance를 취득해야 한다. 이용요금은 영국내 학술기관의 교직원 및 학생은 무료이며 그 외에는 LSRB에서 결정한다.

라. ONS-IDBR(Inter-Departmental Business Register)

ONS 내부의 ONS-IDBR에서는 사업체자료(Business Data)를 제공하고 있다. 제공되는 데이터는 기업 활동 분석의 주된 자료로 이용되고 있으며 기업을 대상으로 하는 표본조사의 표본추출틀(sample frame)도 제공하고 있다. 자료의 제공은 MRP(Microdata Release Panel)를 운영하여 심의를 거쳐 제공한다. 중앙정부의 직원 및 중앙정부와 계약한 사람은 조사 및 분석을 위한 목적으로는 개인식별가능 자료(Disclosive Data)의 이용이 가능하며, 지방정부 및 다른 정부단체의 직원은 일부의 개인식별가능 자료(Disclosive Data) 이용이 가능하고, 일반인의 경우는 개인식별불가능 자료(Non-disclosive Data)의 이용이 가능하다.

이용요금은 개인식별가능 자료와 개인식별불가능 자료, 그리고 개인식별가능 자료의 표본자료를 사용할 경우에 1시간당 60파운드가 부과되며, 개인식별가능 자료의 구역 자료를 사용할 경우는 구역당 150파운드가 부과된다.

6. 일 본

일본의 마이크로데이터 제공은 정부기관에 대해서만 지극히 제한적으로 이루어졌다. 이에 따라 마이크로데이터 제공에 관한 시스템이 다소 미약하다. 최근 마이크로데이터에 대한 요청이 증가함에 따라 2009년 제공을 목표로 '통계데이터의 2차 이용 촉진에 관한 연구회'를 개최하는 등 마이크로데이터 제공 서비스를 준비하는 데 많은 노력을 하고 있다. 현재 히토쓰바시대학 경제연구소 부속 사회과학통계정보연구센터에서 정부통계 마이크로데이터 제공 시범시스템을 구축하여 운영하고 있는데 이를 간략히 소개하는 것으로 일본의 사례를 대신한다.

2004년 11월에 시작하여 2008년 9월까지의 계획으로 마이크로데이터 제공에 관해 시범 운영하고 있다. 이용 장소는 히토쓰바시대학 사회과학통계정보연구센터이며 이용 조건은 학술연구에 한정하고 교육목적 및 영리기업의 위탁연구는 허가하지 않는다. 신청자는 전국 대학의 강사 이상의 전임교원으로 하며 연 4회(1월, 4월, 7월, 9월)

이용자를 모집한다. 이용자는 신청서를 다운로드 하고 센터에 신청한다. 심사 후 센터에서 마이크로데이터를 작성하고 설명회를 개최한 후 이용자는 데이터를 이용하게 되며, 이용기간 종료 후 PC 내의 데이터를 소거하고 CD는 반환한다. 데이터의 사용기간은 6개월이며 이용자는 마이크로데이터를 이용해 작성한 연구논문을 센터에 2부 제출해야 한다. 2007년 7월 기준현재 이용건수는 66건, 이용자 수는 92명이다.

제4절 시사점

앞에서 우리나라 통계청의 제공현황과 해외 여러 나라의 마이크로데이터의 제공현황에 대해 살펴보았다. 이를 바탕으로 다음과 같은 시사점을 도출하였다.

첫째, 해외의 마이크로데이터 제공방법에 있어 공공이용파일(Public Use File)로 제공 가능한 자료는 간단한 절차로 제공하고 있으며, 특별한 학술연구를 목적으로 하는 자료는 On-site방법을 통해 제공하는 것이 추세로 보인다. 또한 '기관 대 개인'이 아닌 '기관 대 기관'의 형태로 제공되고 있음을 눈여겨볼 필요가 있다. 기관을 상대로 하여 제공하는 형태는 개인이 자료 이용시 비용의 부담이 적을 뿐만 아니라 자료의 제공을 담당하는 기관에서도 관리적인 측면에서 더 효율적일 것이라 판단된다.

둘째, 제공절차 면에서 On-site방법으로 자료를 제공할 경우에는 제안서(proposal)의 심사뿐 아니라 최종 성과물 관리까지 엄격하게 관리하고 있는데, 우리나라에서도 이용 중의 상황이나 최종성과물에 대한 보다 철저한 관리감독 체계가 요구된다. 또한 자료제공 위원회를 운영함에 있어 통계작성자의 입장과 이용자의 입장을 원활하게 조율하기 위해 외부전문가를 통한 자료제공기준이나 제공방법, 그리고 새로운 상품이나 서비스 방식 등에 대한 의견수렴의 체계적인 절차가 필요하다.

셋째, 마이크로데이터 제공비용 산정시 이용목적별 또는 기관단위별로 가격을 차등화할 필요가 있다. 연구, 자금지원연구, 발간, 소득창출 등의 각각의 경우에 따라 가격을 다르게 산정하거나, 대학이나 연구소 등의 기관단위로 일정기간을 정해 비용을 산정하는 방식의 검토가 필요하다.

넷째, 각국에서는 다양한 마이크로데이터의 제공과 그에 따른 비밀보호방법의 개발에 많은 노력을 기울이는 점에도 주목할 필요가 있다. 이용자의 데이터 이용 목적 및 수요를 파악하여 그에 따른 다양한 상품을 개발해야 하며, 응답자의 개인정보 보호와 응답거부 방지, 응답자와 통계작성기관 간의 신뢰관계 유지를 위해 제공되는 마이크로데이터에 대해 비밀보호방법을 확대 적용하여야 할 것이다. 한편 이용자를 고려하여 자료 정보의 손실을 최소화하는 비밀보호방법을 연구하여 적용해야 할 것이다.

다섯째, 홍보적 측면으로 보면 이용자의 편의를 위해 마이크로데이터 신청방법, 자료제공규정, 이용절차 등을 수록한 상세한 지침 안내서 등의 마련과 제공에 노력을 기울일 필요가 있을 것이다. 이상에서 외국사례를 통하여 제공방법, 제공절차, 비용 산정, 사후관리, 이용자 홍보에 대한 시사점을 도출하였으나, 무엇보다도 자료의 최대 이용자인 연구자들과 긴밀한 협조체제를 구축하여 단순한 데이터 제공자와 이용자의 관계에서 더 나아가 상호 발전할 수 있는 방향을 모색해야 할 것이다.

효율적인 마이크로데이터의 제공은 연구자들로 하여금 좀 더 복잡한 경제·사회 문제들에 대하여 깊이 있는 연구를 가능하게 하며, 연구자들로부터의 피드백을 통하여 자료의 품질향상을 꾀할 수 있고, 이를 통하여 국가통계의 신뢰성을 구축할 수 있게 할 것이라 기대한다.

참고문헌

박원환・황조연(2004), 「통계자료의 특성과 비밀보호 방법에 관한 연구」, 통계청.
김창환(2007), 「IPUMS-Integrated Public Use Microdata Series」, Minnesota Population Center.
통계청, “통계자료제공규정”.
통계청 국제통계협력과(2004), “외국인에 대한 마이크로데이터 제공검토”.
통계청 통계정보국(2006), “통계 마이크로데이터 제공현황 및 개선방안”.
_____(2006), “마이크로데이터 서비스 업무편람”.
_____(2006), “마이크로데이터 활용도 제고를 위한 심층의견조사결과 보고서”.
_____(2006), “OECD 마이크로데이터 관련회의 참가결과 보고”.
_____(2007), “ON-SITE ACCESS 운영지침”.
_____(2007), “On-Site Access 자료유출 방지 계획(안)”.
_____(2007), “마이크로데이터 제공비용 산정기준 개정방안”.

神林 龍(2007), 「北米における政府統計個票調査公開の現狀に關する調査報告 : 米國勞働統計局, 米國センサス局およびカナダ統計局のオンサイトリサーチを中心に」, Institute of Economic Research Hitotsubashi University.
統計データの二次利用促進に關する研究會(2007), “제1회 회의보고서”.
_____(2007), “諸外國の統計データの二次的利用の狀況”.

ABS(2005), “Responsible Access to ABS CURFs Training Manual”.
____(2006), “Remote Access Data Laboratory User Guide”.
____(2007), “Managing ABS Confidential Unit Record File(CURFs) A Step By Step Guide”.
London School of Economics(2003), “Providing remote access to data : the academic perspective”.
ONS(2004), “ONS Micro-data Access and Release Policy”, Statistical Policy Committee.
____(2005), “Providing access to data and making microdata safe, experiences of the ONS”.
____(2005), “Access to business microdata in the UK : dealing with the irreducible risks”.
____(2007), “MRP Guide - ONS MICRODATA RELEASE PROCEDURE”.

http://www.nso.go.kr/shop/
http://mdss.nso.go.kr/
http://www.census.gov/
http://www.bls.gov/
http://www.thedataweb.org/
http://ces.census.gov/

http://factfinder.census.gov/
http://dataferrett.census.gov/
http://www.ipums.org/
http://www.nber.org/data/
http://www.ceprdata.org/
http://www.unicon.com/
http://www.abs.gov.au/
http://www.statistics.gov.uk/
http://www.data-archive.ac.uk/
http://www.ccsr.ac.uk/
http://www.celsius.lshtm.ac.uk/

<부 록>

A. 미국 usRDC 제공자료 목록(2006년 10월 기준)

1	American Community Survey
2	American Housing Survey
3	Annual Capital Expenditures Survey
4	Annual Survey of Manufactures
5	Auxiliary Establishment - ES9200
6	Business Expenditures Survey
7	Census of Construction Industries
8	Census of Finance, Insurance, and Real Estate
9	Census of Manufactures
10	Census of Mining
11	Census of Retail Trade
12	Census of Services
13	Census of Transportation, Communications, and Utilities
14	Census of Wholesale Trade
15	Commodity Flow Survey
16	Compustat-SSEL Bridge
17	CPS Crosswalk
18	Current Population Survey March Supplement
19	Decennial Census Long Form Sample
20	Decennial Employer-Employee Database
21	Economic Census of Puerto Rico
22	Enterprise Summary Report - ES9100 (large company)
23	LEHD Business Register Bridge
24	LEHD Employer Characteristics File
25	LEHD Employer Quarterly Workforce Indicators
26	Longitudinal Business Database
27	Manufacturing Energy Consumption Survey
28	Medicaid Eligibility Database System - California
29	Medical Expenditure Panel Survey - Insurance Component
30	National Employer Survey

(계속)

31	National Longitudinal Survey
32	Ownership Change Database
33	Quarterly Financial Report
34	SIPP Public-Use Crosswalk
35	Standard Statistical Establishment List - non Name and Address File
36	Standard Statistical Establishment List - Name & Address File
37	Survey of Income and Program Participation (SIPP) Panels
38	Survey of Income and Program Participation - Longitudinal
39	Survey of Industrial Research and Development
40	Survey of Manufacturing Technology
41	Survey of Plant Capacity Utilization
42	Survey of Pollution Abatement Costs and Expenditures
43	Unemployment Insurance-Base Wage File
44	Insurance Programs
45	Occupational Employment Statistics Survey
46	Current Employment Statistics
47	Job Openings and Labor Turnover Survey
48	Mass Layoff Statistics
49	National Compensation Survey and its predecessors, including:
50	Employment Cost Index
51	Employer Costs for Employee Compensation
52	Occupational Compensation Survey
53	Industry Wage Surveys
54	Area Wage Surveys
55	Employee Benefits Surveys
56	Survey of Occupational Injuries and Illnesses
57	Consumer Expenditure Surveys
58	Producer Price Indexes
59	Consumer Price Indexes
60	International Price Program
61	Quarterly Census of Employment and Wages Program, and the associated Longitudinal Database of Establishments covered by State Unemployment

B. BLS On-site 제공자료 목록(2007년 9월 기준)

Data on employment and unemployment :

- National Longitudinal Survey of Youth 1979 (NLS79) Zip Code and Census Tract Files
- National Longitudinal Survey of Youth 1997 (NLS97) Zip Code and Census Tract File
- NLS79 School Surveys
- NLS97 School Surveys
- National Longitudinal Surveys Original Cohorts: Mature and Young Women, and Older and Young Men files with Geographic Variables (2)
- 1993 and 1995 Surveys of Employer Provided Training
- Quarterly Census of Employment and Wages Program, and the associated Longitudinal Database of Establishments covered by State Unemployment Insurance Programs (3)
- Occupational Employment Statistics Survey
- Current Employment Statistics
- Job Openings and Labor Turnover Survey
- Mass Layoff Statistics

Data on prices and living conditions :

- Consumer Expenditure Surveys
- Producer Price Indexes
- Consumer Price Indexes
- International Price Program

Data on compensation and working conditions :

- National Compensation Survey and its predecessors, including:
  - Employment Cost Index
  - Employer Costs for Employee Compensation
  - Occupational Compensation Survey
  - Industry Wage Surveys
  - Area Wage Surveys
  - Employee Benefits Surveys
- Survey of Occupational Injuries and Illnesses

C. 캐나다 caRDC 제공자료 목록(2007년 1월 기준)

1	Aboriginal Peoples Survey (APS)
2	Canadian Community Health Survey (CCHS)
3	Canada Survey of Giving, Volunteering, and Participating (CSGVP)
4	Ethnic Diversity Survey (EDS)
5	General Social Survey (GSS)
6	Health Services Access Survey (HSAS)
7	Labour Force Survey (LFS)
8	Longitudinal Survey of Immigrants to Canada (LSIC)
9	National Longitudinal Survey of Children and Youth (NLSCY)
10	National Population Health Survey (NPHS)
11	National Graduates Survey (NGS)
12	Participation and Activity Limitations Survey (PALS)
13	Public Service Employee Survey (PSES)
14	Survey of Approaches to Educational Planning (SAEP)
15	Survey of Labour and Income Dynamics (SLID)
16	Workplace and Employee Survey (WES)
17	Youth in Transition Survey (YITS)

D. 호주 통계청 제공자료 목록(2007년 9월 기준)

조사명	기준연도	세분화 정도
Aspects of Literacy	1996	기본형*
Australian Housing Survey	1994, 1999	기본형*
Australians' Employment and Unemployment Patterns	1994-97	기본형*, 확장형
Business Longitudinal Survey	1994-98	기본형*
Census - Household Sample File	1981, 1986, 1991, 1996	기본형*
	2001	기본형*, 확장형
Child Care Arrangements Survey	1984	기본형
Child Care Survey	1993,1996	기본형*
	1999	기본형*, 확장형
	2002, 2005	확장형
Disability, Ageing and Carers	1993, 1998, 2003	기본형*
Domiciliary Services, Victoria	1986	기본형*
Education and Training Experience	1989, 1993, 1997	기본형
Education and Training	2001	기본형
Education and Training	2005	기본형*, 확장형
Education and Work	2001, 2003, 2005	기본형*
Families in Australia	1992	기본형
Family Characteristics Survey	2003	확장형
Forms of Employment	1998	기본형*
General Social Survey	2002, 2006	기본형*, 확장형
Giving to Charities Population Survey Monitor	1997	기본형
Household Expenditure Survey	1975-76, 1984, 1988-89, 1993-94	기본형*

(계속)

조사명	기준연도	세분화 정도
Household Expenditure Survey and the Survey of Income and Housing (HES/SIH) Second Edition	2003-04	기본형*, 확장형
Income and Housing Survey	1981-82	기본형*
Income Distribution Survey	1986	기본형*
Income and Housing Costs Survey	1990	기본형*
Income and Housing Costs Survey	1994-98, 1999-2000	기본형*
	2000-01	기본형*, 확장형
Income and Housing Survey	2002-03, 2005-06	기본형*, 확장형
Income and Housing Survey and the Household Expenditure Survey (HES/SIH) Second Edition	2003-04	기본형*, 확장형
Innovation in Australian Business	2003	확장형
Labour Force Survey and Employee Earnings, Benefits and Trade Union Membership	2004	기본형*
	2006	기본형*, 확장형
Labour Mobility	1991, 1994	기본형*
Labour Force Survey and Labour Mobility Survey	2006	기본형*
Mental Health and Wellbeing of Adults, Australia, Second Revised	1997	기본형*
Multi Purpose Household Survey	2004-05, 2005-06	확장형
National Aboriginal and Torres Strait Islander Health Survey	2004-05	기본형*, 확장형
National Aboriginal and Torres Strait Islander Survey	1994	기본형*, 확장형

(계속)

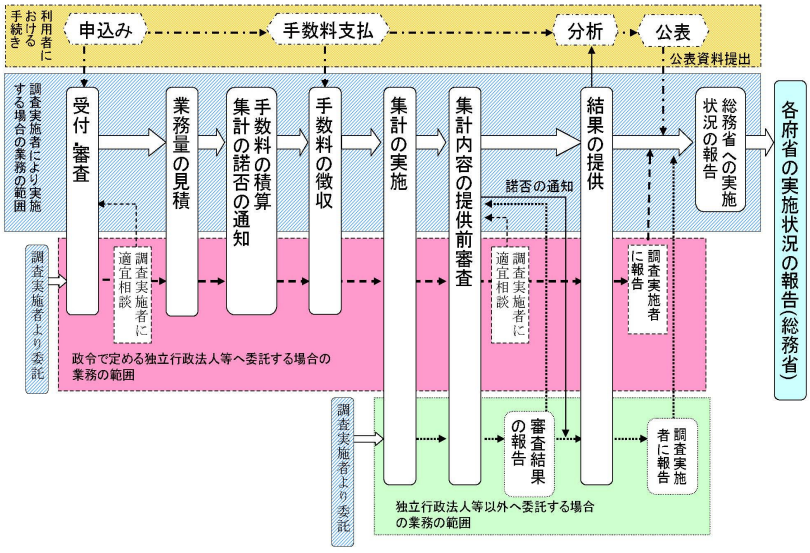
조사명	기준연도	세분화 정도
National Aboriginal and Torres Strait Islander Social Survey	2002	확장형
National Crime and Safety Survey	2002, 2005	확장형
National Health Survey	1977-78, 1983, 1989-90, 1995	기본형*
	2001	기본형*, 확장형
National Health Survey, Indigenous	2001	확장형
National Nutrition Survey	1995	기본형*
Personal Safety Survey	2005	확장형
Pregnancy and Employment Transitions	2005	확장형
Rental Investors' Survey	1993, 1997	기본형*
Rental Tenants Survey	1994	기본형*
Superannuation and Employment Arrangements Survey	2000	확장형
Time Use Survey	1992	기본형*
	1997	기본형*, 확장형
Voluntary Work Survey	2000	기본형*
Women's Safety Survey	1996	기본형*

E. 일본의 자료 이용 흐름도(안)

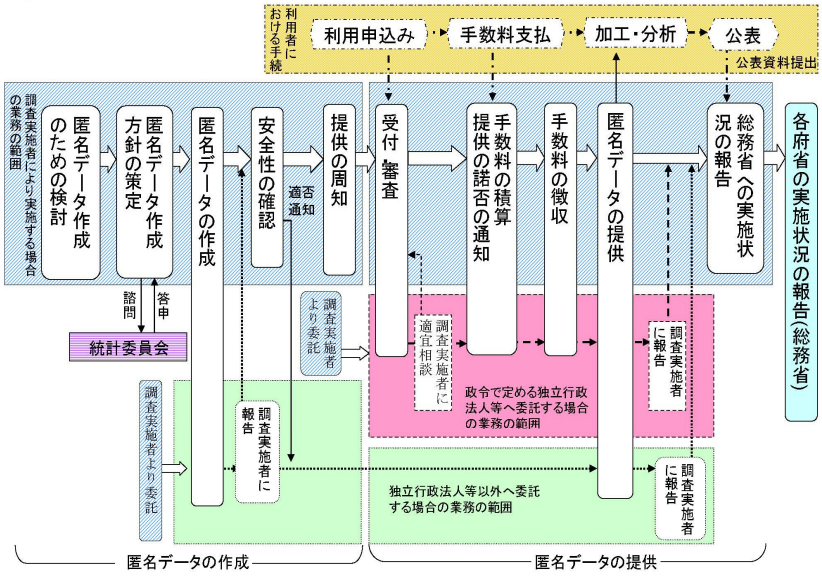
資料10

利用手続きに関する事務の流れ（案）
（※統計データ利用促進ワーキンググループにおいて検討中のもの）

①委託による統計の作成等の場合



②匿名データの作成・提供の場合



F. 각국의 마이크로데이터 제공비용

미 국	캐나다
・Public Use File : 무료 ・RDC : 약 $15,000/년 (기관 vs 기관) ※ ICPSR[1] (가맹자격에 따른 연회비) ○ Category A : $12,730 ○ Category B : $7,930 ○ Category S : $3,980 ○ Category C : $2,000 - Category A : 사회과학 및 관련 분야의 대학원이 있는 대학 및 연구기관 (연구원 60명 이상) - Category B : 사회과학 및 관련분야의 대학원이 없는 대학 (학생수 2,500명 이상) - Category S : 사회과학 및 관련 분야의 대학원이 없는 대학 (학생수 2,500명 미만) - Category C : 개발도상국의 기관	・Public Use File : 파일마다 요금을 징수 ($1,000 ~ $3,000 정도) 일괄구입시 할인이 있기도 함 ・RDC - Direct Access : $3,500/월 - Indirect Access : $5,500/월 ※ DLI(Data Liberation Initiative) - 규모가 큰 기관(27개 기관) : $12,000/년 (CARL, Canadian Association of Research Libraries 등) - 그 외의 기관 : $3,000/년 (CASUL, Canadian Association of Small University Libraries 등)

주: 1) the Inter-University Consoritum for Political & Social Research.

(계속)

호주	영국	그 외 유럽
· CURF : 데이터 세트당 약 $1,000 ~ $2,000 · 컨설팅비용 : 프로젝트당 $435	· CCSR[1] (1991, 2001 SAR) - 학술목적 : 무료 - CD-ROM 구입 : £30 - 상업목적: 데이터 세트당 £1,000 - 지방정부, 보건기관, 중앙정부, 자선단체 : £500 · UKDA[2] (사회, 인문 관련 서베이 데이터) - 학술목적 : 무료 - 상업목적: 데이터 세트당 £500 - CD-ROM: 기본 £7.5, handling fee £2.5, 우송비 £3(영국내), £4(유럽), £5(그 외) - 부가가치세 : 17.5% · CeLSIUS[3] - 국내 학술기관의 교직원, 학생에게 무료 제공 · ONS-IDBR[4] - 약 £60/hour	<이탈리아> - 데이터 세트당 €90 (센서스자료 : €150) 대학, 공공학회는 20% 할인 <독 일> - €65 ~ €95 <네덜란드> - 원칙: 무료 - 실비보상

주: 1) the Centre for Census and Survey Research.
2) the United Kingdom Data Archive.
3) Centre for Longitudinal Study Information and User Support.
4) Inter-Departmental Business Register.

제2장

마이크로자료의 활용도제고를 위한 비밀보호방법

정동명 · 강동환

제1절 서 론

대부분의 통계작성기관에서는 각종 통계조사를 통해 다양한 자료들을 수집하고 있는데, 이렇게 수집된 통계자료는 원자료(raw data)와 마이크로자료(microdata), 그리고 매크로자료(macrodata)로 구분할 수 있다. 원자료란 통계조사 자료에서 최초 입력한 전산화일 자료로서 입력오류, 조사오류 등이 걸러지기 이전 단계의 자료를 말한다. 마이크로자료란 원자료에서 입력오류 등을 제거하여 공표 통계표 작성 등 자료가공의 기초자료로 사용되는 자료를 말하며 통계원시자료라고도 한다. 매크로자료란 마이크로자료를 임의의 기준에 따라 집계한 자료로서, 집계의 정도에 따라 세분화된 자료에서 통합된 자료까지 다양하게 제공할 수 있다. 따라서 통계작성기관에서 자료를 제공할 경우에는 주로 매크로자료를 제공하고 있으며, 일부 제한적인 경우에만 마이크로자료를 제공하고 있다. 그러나 매크로자료를 이용해서 갈수록 복잡하고 다양해지는 경제 · 사회현상을 심층적으로 분석하기에는 어느 정도 한계가 있다. 그러므로 여러 분야의 통계이용자들은 통계를 작성하는 기관에서 매크로자료뿐만 아니라 마이크로자료도 제공할 것을 요구하고 있으며, 이러한 요구는 점차 증가하고 있는 추세이다.

마이크로자료에는 응답자의 개인 사생활에 대한 정보들이 모두 수록되어 있기 때문에 마이크로자료를 그대로 제공할 경우 응답한 개인의 특성을 나타내는 정보들이 그대로 노출될 뿐만 아니라, 심지어는 특정 응답자에 대한 식별도 가능하게 되어 심각한 개인

✍ 이 연구는 2006년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

사생활의 침해가 발생할 위험이 있다. 그러므로 마이크로자료를 제공할 경우에는 사전에 미리 응답자의 개인정보가 노출되지 않도록 통계적으로 보호방법을 적용하는 등 각별한 노력을 기울여야만 한다.

외국의 통계작성기관에서는 1970년대부터 마이크로자료의 제공에 대한 요구와 개인정보의 비밀보호라는 두 가지 상반된 요인을 동시에 고려하여 적정한 수준의 자료를 제공할 수 있는 방법을 찾기 위해 노력해 왔으며, 1980 년대에는 이에 대한 다양한 연구논문들이 발표되었다. 미국 상무부 센서서국에서는 마이크로자료를 제공할 때 통계적 비밀보호방법을 활용하여 개인의 정보를 최대한 보호할 있는 방법들을 연구하였는데, 특히 Cox 등(1986)은 미국의 개인비밀보호와 관련된 자료제공에 대해 언급하였으며, Kim(1986)은 통계적 기법을 적용한 비밀보호방법을 연구하였다. 네덜란드 중앙통계국(1987)에서는 스웨덴과 네덜란드간의 협력세미나를 통해 개인사생활의 개방과 보호에 대해 언급하였다. Bethelehem 등(1990)은 자료파일에서 개인의 식별(identification)과 노출(disclosure)의 문제에 대해 언급하고, 이를 해결하기위한 방법으로 모집단 uniqueness의 추정을 위한 통계적 모형을 설정하고 예제를 통해 제시하였다. Marsh 등(1991)은 영국의 1991년 센서스에서 개인비밀보호를 위해 익명화된 마이크로자료 파일작성의 필요성과 그 효과 등을설명하였다. 이외에도 Dalenius(1977), Fuller(1993) 등 여러 학자들이 노출의 위험성과 비밀보호에 대한 방법을 연구하였다.

우리나라의 경우 2006년 10월 1일 현재 통계청을 포함해서 153개의 통계작성기관에서 673종의 국가통계가 작성되고 있다. 이 중 통계청에서 실시하는 인구주택총조사(census)는 0과 5로 끝나는 연도를 기준으로 매 5년마다 전국을 대상으로 실시하고 있다. 인구주택총조사에서 수집된 결과에는 우리 나라의 인구·사회현상을 심층적으로 파악하고 분석할 수 있는 좋은 정보들이 많이 있어, 학계나 연구기관 등 여러 분야의 통계이용자들이 마이크로자료의 제공을 요구하고 있는 실정이다. 이에 부응코자 통계청에서는 자료이용자가 인터넷을 통해 시스템에 접속하여 원하는 자료를 직접 가공 및 이용할 수 있도록 지원하는 one-stop 서비스 시스템인 MDSS(Micro Data Service System)를 2006년부터 운영하고 있다(http://mdss.nso.go.kr). MDSS는 기존의 주문형 위탁처리 방식에서 탈피하여 마이크로자료를 이용하는 이용자가 직접 자료를 가공할 수 있도록 지원함으로써 보다 빠르고 간편한 절차에 따라 자료를 제공받을 수 있도록 하고 있다. 가령, 자료이용자가 인구주택총조사의 마이크로자료를 원할 경우 MDSS에 접속하면 10% 표본조사 자료 중 2%의 표본자료를 일정한 절차에 따라 제공받을 수 있다.

그러나 개인의 정보가 식별되고 노출되는 민감한 항목들은 일부 제거한 후 제공하고 있기 때문에 자료의 분석 및 활용측면에서 다소 미흡한 점이 있는 것이 사실이다. 이런 점을 고려하여 통계청에서는 개인의 비밀을 보호하면서 마이크로자료를 제공할 수 있는

방법을 개발코자 수년전부터 통계적 보호기법에 대한 연구에 노력하였다. 통계청 통계연구과에서 박원환 등(2004)이 통계자료의 비밀보호에 대해 연구하였다.

본 논문에서는 개인의 비밀을 보호하는 방법들을 살펴보고, 이러한 방법을 적용하여 mask된 마이크로자료 파일을 직접 작성하고자 한다. 2장에서는 노출과 식별에 대해 알아보고, 식별의 측도로 사용되는 uniqueness의 개념을 설명한다. 또한 지금까지 연구되어진 비밀보호의 방법을 이산형(discrete) 변수인 경우와 연속형(continuous) 변수인 경우로 나누어 소개한다. 3장에서는 2005 인구주택총조사 일부 지역의 표본조사 결과자료에 대해 grouping 방법을 적용하여 마이크로 자료파일을 작성해 보고, 이에 대한 내용을 살펴보고자 한다. 마지막으로 4장에서는 본 연구의 최종적인 결론과 이를 토대로 앞으로 고려해야 할 사항들을 언급하고 있다.

제2절 비밀보호방법

1. 노출과 식별

노출(disclosure)이란 통계작성기관이 자료를 수집·정리하여 다양한 형태의 통계정보(통계간행물, 테이프, 디스크 등)로 제공할 경우, 이를 통해서 응답자의 특성이 파악되는 경우를 말한다. 따라서 노출에 대한 문제는 통계작성기관이 통계목적으로 응답자에게 자료를 수집한다는 가정하에서 발생하게 되는데, 어떤 경우에도 노출은 발생되지 않도록 하는 것이 바람직하다. 이에대해 Bethelehem 등(1990)은 노출이 바람직하지 않은 것은 개인의 사생활이 보호되어야하는 법적이고 윤리적인 이유뿐만 아니라 조사의 응답률에 영향을 줄 수 있다는 실질적인 이유가 있기 때문이며, 가능하면 마이크로자료에서 노출을 제한해야 한다고 하였다. 노출은 개인의 식별(identification)과 서로 깊은 연관성을 가지고 있다. 즉, 개인의 민감한 정보가 노출된다면 이는 개인에 대한 식별이 가능하게 되고, 반대로 주어진 통계정보에서 개인의 식별이 가능하다면 노출의 가능성도 더 커지게 된다.

노출을 제한하는 가장 기본적이며 간단한 방법은 많은 마이크로자료가 공표되지 않도록 하는 것이지만, 이는 자료수집의 목적에 위배되는 것이며 현실적으로도 마이크로자료의 제공에 대한 요구가 날로 증가하는 추세이므로 이 방법은 적절하지 않다. 노출과 식별에 영향을 주는 핵심요소는 자료이용자들의 사전 정보이다. 즉, 만약 누군가가 특정 개인에 대한 정보를 알 수 없다면 노출과 식별은 불가능하다. 따라서 노출은 자료이용자들이 사전에 가지고 있는 유용한 정보의 성격과 양에 따라 좌우된다고 할 수 있다. 그러므로

노출을 제한하기 위해서는 제공되는 마이크로자료의 정보와 자료이용자들이 가지고 있는 사전정보가 서로 일치하지 않도록 자료파일을 구성하면 될 것이다.

2. 비밀보호방법

마이크로자료에서 개인의 비밀을 보호하는 방법은 자료의 형태가 이산형인 경우와 연속형인 경우에 따라 달리 적용된다. 이산형 자료인 경우에는 Data swapping, Coding approach, Grouping 등의 방법이 널리 활용되고 있고, 연속형 자료인 경우에는 Rounding, Grouping into Intervals, Micro aggregation 등의 방법이 활용되고 있다.

가. 이산형 변수

이산형 변수의 경우 성별 등과 같이 단 2개의 범주로 구분되는 변수를 Dichotomous variable이라 하고, 교육정도나 혼인상태 등과 같이 3개 이상의 범주로 구분되는 변수를 Polychotomous variable이라 하는데, 이들 특성에 따라 다음과 같은 방법들이 있다.

1) Data Swapping

Data Swapping은 Dalenius(1979)가 제안한 방법으로서, 마이크로자료내의 민감한 항목에 대한 자료노출을 방지하기 위하여 동일한 Key 변수 조합을 갖는 record간에 자료값을 상호 교환하는 방법이다. 가령, HIV 상태 등과 같은 민감한 항목에 대하여 인종이나 성별, 나이, 수입 등과 같은 항목을 Key 변수로 사용하여 값을 적절하게 교환하는 것이다. 이 방법은 masking 후에도 각 Key 변수별로 민감한 항목의 빈도수는 swapping 하기 전의 빈도수와 동일하게 되어 자료의 분석이 용이하고, Dichotomous variable과 Polychotomous variable인 경우에도 모두 적용이 가능하다.

2) Coding Approach

이 방법은 원 자료에 적절한 값(noise)을 추가하여 다른 자료값으로 변형하는 방법이다. 가령, X_i와 Y_i, 그리고 e_i를 0 또는 1을 갖는 Dichotomous variable라고 가정하자. 만약 X_i를 원자료, e_i를 X_i와 서로 독립이고 동일한 분포를 갖는 noise, 그리고 Y_i를 mask된 자료라고 한다면, Y_i는 다음과 같이 나타낼 수 있으며,

$$Y_i = (X_i + e_i)_{\mathrm{mod}2}$$

Y_i가 가질 수 있는 값은 다음과 같다.

X_i	e_i	Y_i
1	1	0 = (1+1)mod2
1	0	1 = (1+0)mod2
0	1	1 = (0+1)mod2
0	0	0 = (0+0)mod2

3) Grouping

자료파일에서 어떤 변수들은 특성상 노출되기가 쉬운 범주로 구성된 경우가 있다. 가령, 나이 변수에서 100세인 사람이 단 1명뿐이라면 누구인지 쉽게 식별할 수가 있을 것이다. 이러한 경우에 인근의 다른 범주들과 통합함으로써 쉽게 식별이 되지 않도록 할 수 있는데, 이러한 방법을 Grouping이라고 한다. 이 방법은 Polychotomous variable인 경우에 적용하면 매우 효과적이다.

나. 연속형 변수

1) Rounding

주어진 자료를 적당한 몫과 나머지의 형태로 나타내는 방법을 Rounding이라고 한다. 예를 들어, 주어진 자료를 적당한 몫과 나머지의 형태로 나타내는 방법을 Rounding이라고 한다. 예를 들어, X를 정수, B를 base, q를 몫, 그리고 r을 나머지라고 하면, $X = qB + r$의 형태로 나타낼 수 있다. 만약, $r \sim U(0, B-1)$이라고 한다면, r의 확률분포와 기댓값 $E(r)$ 및 분산 $Var(r)$은 다음과 같다.

r	0	1	2	…	B-1	계
P(r)	1/B	1/B	1/B	1/B	1/B	1

$$E(r) = \sum_{r=0}^{B-1} r P(r) = \frac{B-1}{2}, \quad Var(r) = \frac{B^2-1}{12}$$

Rounding 방법에서 나머지 r을 처리하는 방법은 크게 4가지로 구분이 되는데, 이에 대한 자세한 내용은 Kim(2003)을 참고하기 바란다.

2) Grouping into Intervals

이 방법은 연속형 자료를 적당한 구간(interval)으로 그룹화하여 각 구간을 대표하는 code값으로 원 자료를 대체하는 방법으로서, 구간을 넓게 하면 자료의 노출 방지에는 용이하다. 그러나 회귀모형 또는 상관계수 등의 자료 분석시에 다소 제약이 있으며, 특히 구간평균이 제공되지 않기 때문에 구간의 중앙값(median)을 사용하는 것이 적절하다.

3) Micro aggregation

자료 파일내에 많은 record들은 선정된 하나의 기준변수에 따라 정렬될 수 있으며, 이를 이용하여 보통 3~4개의 관측값을 그룹화하고 그룹내의 각 자료 값은 그룹평균으로 대체할 수 있는데, 이러한 방법으로 노출을 방지하는 것을 Micro aggregation이라고 한다.

4) 기타

위에서 언급된 방법 이외에도 이산형 변수의 경우에 적용되는 Data swapping 방법을 동일하게 적용할 수 있으며, Kim(1986, 2001)에 의해 체계화된 Adding noise 방법과 Multiplicative noise 방법 등도 있다.

3. Uniqueness

가. Uniqueness

제공되는 마이크로자료 파일에서 식별이 어떻게 이루어지는지를 파악할 수 있다면 어느 정도 노출을 제한할 수가 있는데, 이것은 uniqueness를 통해서 가능하다. Uniqueness란 전체 자료파일에서 조사단위의 특성이 유일하게 존재하는 것을 말하며 어떤 조사단위가 식별될 가능성을 나타내는 측도로 사용된다. 가령, 100명으로 구성된 모집단에서 만약 나이가 100세인 사람이 단 한명 있다면 그 사람은 모집단에서 unique하다고 하며, 그 사람은 다른 사람들에 비해 자료에서 식별될 가능성이 매우 높게 된다. 주어진 자료에서 uniqueness는 하나의 변수만으로도 파악할 수 있고 여러 개의 변수들을 조합하여 파악할 수도 있다. 가령, 위의 예에서 100명 중 나이가 60세인 사람이 3명 있다고 하더라도 직업이라는 변수를 포함하여 사원이면서 나이가 60세인 사람은 단 한명일 수도 있다. 따라서 고려할 변수가 많아질수록 자료의 uniqueness는 점점 더 커지게 된다.

나. 노출위험

노출위험(disclosure risk)은 모집단에서 unique한 경우로 한정되는지 아닌 지에 따라 좁은 의미와 넓은 의미로 정의할 수 있는데, 여기에서는 좁은 의미에서의 노출위험에 대해서 다루기로 한다. 일반적으로 노출은 다음의 3가지 조건을 만족하는 경우에 발생한다.

〈 노출의 발생조건〉

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

따라서 어떤 사람이 위의 조건에서 언급한 자료파일들 중 어느 하나에도 나타나지 않는다면 노출은 일어나지 않는다.

한편, 위에서 언급한 노출의 발생조건 하에서 노출위험은 확률적 모형으로 표현될 수 있으며, 이를 위해 다음과 같이 기호를 정의하기로 한다.

- A : 관심의 대상인 사람
- S_1 : 통계작성기관의 마이크로자료로 구성된 파일 1
- S_2 : 외부인(intruder)에 의해 구성된 파일 2
- U_P : 모집단의 unique class
- U_S : 표본의 unique class

만약, 조사과정에서 측정오차가 없다고 가정하고 어떤 사람(A)이 모집단에서 unique하다면, 그 사람은 또한 표본에서도 unique할 것이다. 즉, 이것을 확률적으로 표현하면 다음과 같이 나타낼 수 있다.

$$\Pr[(A\in U_S)\cap(A\in U_P)] = \Pr(A\in U_P)\Pr[(A\in U_S)\,|\,(A\in U_P)]$$
$$= \Pr(A\in U_P)$$

1) 외부인이 S_2파일에 특정인이 포함된 것을 모르고 있을 경우

만약 금융기관이나 이웃주민 등과 같은 외부인이 그들이 작성한 파일(S_2)에 관심있는 특정 사람이 포함되어 있다는 것을 모르고 있다면, 위의 관계식으로부터 A 의 노출위험 $DR(A)$은 다음과 같이 정의할 수 있다.

$$DR(A) = \Pr[(A \in S_1) \cap (A \in S_2) \cap (A \in U_P)] \tag{1}$$

그리고 조건부 확률공식과 A 가 모집단에서 unique한 사건은 A 가 표본에 추출되거나 추출되지 않거나 독립이라는 사실을 이용하여 위의 식 (1)은 다시 다음과 같이 표현할 수 있다.

$$\begin{aligned} DR(A) &= \Pr[(A \in S_1) \cap (A \in S_2) \mid (A \in U_P)] \Pr(A \in U_P) \\ &= \Pr[(A \in S_1) \cap (A \in S_2)] \Pr(A \in U_P) \end{aligned}$$

반적으로 A 가 파일 1(S_1)에 있다는 사건은 A 가 파일 2(S_2)에 있다는 사건과 서로 독립이 되므로, $\Pr[(A \in S_1) \cap (A \in S_2)] = \Pr[(A \in S_1)] \times \Pr[(A \in S_2)]$이 된다. 그러나 어떤 경우에는 두 사건이 서로 독립이 되지 않는 경우도 있다. 가령, 통계청에서 실시하는 가계조사는 경제활동인구조사의 표본 중 일부만을 표본으로 사용하기 때문에 서로 독립이라고 할 수 없다. 만약 S_1이 경제활동인구조사에서 작성된 파일이고 S_2가 가계조사에서 작성된 파일이라고 한다면 $\Pr[(A \in S_1) \cap (A \in S_2)] = \Pr[(A \in S_2)]$가 될 것이다. 따라서 노출위험은 두 사건이 서로 독립인지 아닌지에 따라 다음과 같이 나타낼 수 있다.

$$DR(A) = \begin{cases} \Pr(A \in S_1) \Pr(A \in S_2) \Pr(A \in U_P), & \text{독립인 경우} \\ \Pr(A \in S_2) \Pr(A \in U_P), & \text{종속인 경우} \end{cases} \tag{2}$$

2) 외부인이 S_2파일에 특정인이 포함된 것을 알고 있을 경우

만약 외부인이 파일(S_2)에 관심있는 특정 사람이 포함되어 있다는 것을 이미 알고 있다고 한다면, $\Pr(A \in S_2) = 1$ 이 될 것이다. 따라서 노출위험은 다음과 같이 된다.

$$DR(A) = \Pr(A \in S_1) \Pr(A \in U_P) \tag{3}$$

실제로 은행이나 대출 회사에서는 어떤 채무자에 대한 자세한 정보를 가지고 있다고 하고, 회사의 고용인이 마이크로자료 파일에서 그 채무자를 식별하기를 시도한다고 가정하면, $\Pr(A \in S_2) = 1$ 이기 때문에 그 채무자의 노출위험은 위의 식 (3)을 사용하여 계산할 수 있다.

특정 사람이 전체 모집단에서 unique하다는 것을 외부인이 알고 있다면 $\Pr(A \in U_P) = 1$ 이 될 것이므로, 노출위험은 다음과 같이 나타낼 수 있다.

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

제3절 마이크로데이터의 Masking

1. 분석대상 자료

본 연구에서 마이크로자료 파일을 작성하기 위해 사용한 자료는 지난 2005년 11월에 통계청에서 실시한 "2005 인구주택총조사"의 표본조사 결과의 충청남도 자료이다. 인구주택총조사의 조사표는 크게 전수조사표와 표본 조사표로 구분하여 가구원과 가구, 그리고 주택에 관한 사항들을 질의하도록 구성되어 있다. 전수조사표는 기본적인 특성 파악을 위해 가구원에 관한 사항 8개, 가구에 관한 사항 7개, 주택에 관한 사항 6개 등 총 21개 항목으로 구성되어 있다. 표본조사표는 전수조사항목을 포함하여 보다 세부적인 특성을 파악코자 가구원에 관한 사항 24개, 가구에 관한 사항 11개, 주택에 관한 사항 6개 등 총 41개 항목으로 구성되어 있으며, 이외에 추가로 16개 시·도별로 각각 서로 다른 조사항목 3개가 포함되어 전체적으로는 44개 조사항목으로 구성되어 있다. 이에 대한 자세한 내용은 2005 인구주택총조사 조사지침서(2005)를 참고하기 바란다.

전수조사에 대한 집계결과는 각 분야별로 올해 세 번(5월, 6월, 7월)에 걸쳐 공표되었으며, 표본조사에 대한 결과는 1차적으로 통근 및 통학, 인구이동에 대한 사항들이 9월에 공표되었고, 나머지 분야에 대한 집계결과는 올 연말까지 순차적으로 공표될 예정이다. 따라서 본 연구의 시작시점이 8월인 것을 감안하여 분석에서는 전수조사 결과와 아울러 표본조사의 잠정결과를 사용하였으며, 대상지역은 우선적으로 충청남도 지역만을 고려하였다. 또한 조사의 편의를 위해 설정된 6종류의 조사구들 중 기숙시설 조사구, 특수사회 시설조사구, 관광호텔 및 외국인거주지역 조사구는 제외하고 아파트 조사구, 보통조사구, 섬조사구를 대상으로 분석하였는데, 이는 자료의 활용측면에서

자료이용자들이 집단가구보다는 개별가구의 특성파악에 더 관심이 많을 것으로 판단하였기 때문이다.

2005 인구주택총조사 결과에 의하면 <표 2-1>에 나타난 바와 같이 충청남도는 총 2,460개 일반 및 아파트조사구에서 가구원 1,798,397명과 660,526가구 및 586,757호의 주택이 있는 것으로 나타났다. 10% 표본의 경우 조사구는 745개이고 가구원은 189,505명, 가구는 71,091가구, 주택은 65,398호인 것으로 각각 나타났다.

〈표 2-1〉 2005 인구주택총조사 결과(충청남도)

(단위 : 개, 명, 가구, 호)

조사	조사구	가구원	가 구	주 택
전수	2,460	1,798,397	660,526	586,757
10% 표본	745	189,505	71,091	65,398

2. 마이크로자료 파일의 작성

가. Key 변수의 선정

비밀보호방법을 적용하여 마이크로자료 파일을 작성하기 위해서는 반드시 주요 Key 변수를 선정하여야 한다. 이에 따라 Key 변수의 선정을 위해 2005 인구주택총조사의 조사항목을 검토하였는바, 조사항목의 수가 다소 많은 편이고 특히 개인의 특성에 대한 민감한 질문들이 포함되어 있어 다소 어려움이 예상되었다. 따라서 전체 표본조사의 항목들 중 지역관련 조사항목 3개와 표본조사 세부항목 19개는 제외하고 전수조사항목 21개만을 우선적으로 검토하였다.

주요 Key 변수는 다음과 같은 3단계의 과정을 거쳐 선정하고자 하였다. 먼저 1단계로는 통계청 이외의 외부기관에서 보유하고 있는 자료들 중 인구주택총조사의 조사항목과 중복되어 식별될 가능성이 높다고 판단되는 항목을 우선 선정하였다. 여기서 외부기관이란 통계를 작성하는 기관이나 중앙행정기관, 지방자치단체, 민간기관, 그리고 이웃주민 등이 될 수 있으며 자금을 대출해주는 은행 등의 금융기관도 포함된다. 2단계에서는 1단계에서 식별가능성이 높다고 판단되는 항목에 대해서 빈도분석(frequency analysis)을 실시하여 각 항목의 빈도수와 분포형태 등 보다 자세한 자료의 특성을 파악하였다. 마지막 3단계에서는 각 항목별로 상대적 정보의 양을 나타내는 측도인 엔트로피(entropy)를 다양한 조합별로 각각 계산한 후, 최종적으로 Key 변수를 선정하였다.

엔트로피란 원래 물리학에서 불확실성의 정도를 나타내는데 사용되고 있으나, Shannon(1948)이 엔트로피를 확률적 방식으로 일반화시켜 정보의 양을 나타내는 측도로 활용하도록 하였다. 가령, 변수 X가 주어져 있을때, X의 엔트로피 $H(X)$는 다음과 같이 정의된다.

$$H(X) = -\sum_{x} P(x)\log_2[P(x)]$$

$$= \sum_{x} P(x)\log_2\left[\frac{1}{P(x)}\right]$$

단, $P(x) = \Pr(X=x)$이다. 또한 변수 $X_1, X_2, \cdots, X_n$의 joint 엔트로피는 다음과 같이 계산할 수 있다.

$$H(X_1,\cdots,X_n) = -\sum_{x_1}\cdots\sum_{x_n} P(x_1,\cdots,x_n)\log_2[P(x_1,\cdots,x_n)]$$

$$= \sum_{x_1}\cdots\sum_{x_n} P(x_1,\cdots,x_n)\log_2\left[\frac{1}{P(x_1,\cdots,x_n)}\right]$$

단, $P(x_1,\cdots,x_n) = \Pr(X_1=x_1,\cdots,X_n=x_n)$이다.

인구주택총조사의 조사항목을 검토한 후, 위에서 언급한 3단계의 과정을 거쳐 Key 변수를 선정한 결과는 다음과 같다.

1) 가구원에 관한 사항

가구원에 관한 사항에서는 총 8개의 전수조사 항목 중 5개 항목(성별, 나이, 가구주와의 관계, 교육정도, 혼인상태)을 외부에서 식별가능성이 높은 항목으로 선정하였다. 이렇게 선정된 5개 항목에 대해 10% 표본조사의 결과를 바탕으로 빈도분석을 실시하여 자료의 특성을 파악하였는데, 성별은 남녀별로 서로 비슷하게 분포되어 있다. 그러나 나이 항목의 경우 30~40대가 전체의 약 30%를 차지하고 있으며, 100세 이상인 고령자도 24명이 있는 것으로 나타났다. 가구주와의 관계 항목에서는 가구주와 배우자, 자녀와 그 배우자가 전체의 90.7% 정도로 대부분을 차지하고 있다. 교육정도 항목에서는 고등학교 졸업이 전체의 약 22.6%로 가장 많았으며, 혼인상태 항목의 경우에는 미혼이 약 20.7%이고 사별과 이혼을 포함해 약 13.7%가 배우자가 없는 것으로 나타났다.

각 항목의 다양한 조합별로 엔트로피를 계산한 결과는 부록 B의 <부표 B-1>에 주어져 있다. 먼저 개별 항목을 살펴보면, 성별의 경우 카테고리가 2개로 엔트로피가 약 0.999이고, 나이는 총 105개의 카테고리에 약 6.398, 가구주와의 관계는 총 14개 카테고리에 약 2.126, 교육정도는 총 32개 카테고리에 약 3.323, 그리고 혼인상태는 결측값을 제외하고 총 4개 카테고리에 약1.355의 엔트로피 값이 각각 나타났다. 항목들 간의 조합에 의한 경우를 살펴보면, 항목이 점차 추가될수록 엔트로피 값이 더 높아지는데 이는 그만큼 더 정보의 양이 많아짐을 의미한다. 가령, 성별의 경우 엔트로피는 약 0.999이지만 성별과 나이를 조합한 경우의 엔트로피는 약 7.390으로 더 높아진다. 이러한 결과를 바탕으로 가구원에 관한 사항에서는 처음 5개의 항목 중 교육정도를 제외하고 성별과 나이, 가구주와의 관계, 혼인상태 등 4개 항목을 최종적인 Key 변수로 선정하였다.

2) 가구에 관한 사항

가구에 관한 사항에서는 총 7개의 전수조사 항목 중 4개 항목(가구구분, 거주층, 점유형태, 주인가구 및 주택소유여부)을 외부에서 식별가능성이 높은 항목으로 선정하였다. 10% 표본조사의 결과를 바탕으로 4개 항목에 대해 빈도분석을 실시하여 특성을 파악한 결과, 가구구분의 항목에서는 가족으로 이루어진 가구가 전체의 74.7%를 차지하고 있으며, 거주층 항목에서는 지상이 99.6%로 가장 많은 빈도수를 나타내고 있다. 점유형태 항목의 경우 전체의 약 65.6%가 주거전용이면서 자기집인 경우이며, 주인가구 및 주택소유여부를 묻는 항목에서는 약 61.9%가 주인가구이면서 다른 곳에 주택미소유인 것으로 나타났다. 4개 항목에 대한 각 범주별 빈도수와 엔트로피를 계산한 결과가 부록 B의 <부표 B-2>에 주어져 있다. 이 결과를 바탕으로 가구구분, 점유형태, 주인가구 등 3개 항목을 Key 변수로 최종 선정하였는데, 특히 주인가구 및 주택소유여부 항목에서 주택소유여부에 대한 내용은 중요성이 상대적으로 낮아 제외하고 주인가구 여부만 Key 변수로 선정하였다.

3) 주택에 관한 사항

주택에 관한 사항에서는 총 6개 항목 중 거처의 종류, 연건평, 대지면적 등 3개 항목을 외부에서 식별가능성이 높은 항목으로 선정하였다. 따라서 10% 표본조사의 결과를 바탕으로 이들 3개 항목에 대해 빈도분석을 실시하여 특성을 파악한 결과, 거처의 종류 항목에서는 단독주택이 59.9%, 아파트가 32.5%로 대부분을 차지하고 있다. 연건평의 항목에서는 전체의 약 80.9%가 20~30평대이며, 200평대가 넘는 경우도 0.06% 정도가 있다. 대지면적의 경우에는 100평대가 약 13.3%로 가장 많고, 400평대 이상도 약 1.1%가 있는것으로 나타났다.

한편, 연건평과 대지면적의 조사결과를 자세히 살펴보면, 5평이나 10평단위로 조사된 결과가 많이 나타남을 알 수 있다. 이는 응답자들이 자세한 면적을 응답하지 않고 5평 또는 10평단위로 응답한 결과로서, 조사과정에서 발생되는 응답자나 조사원에 의한 측정오차(measurement error)라 할 수 있다. 따라서 연건평과 대지면적에 대한 결과분석은 각 조사단위별로 하지 않고 10평단위로 범주화하여 실시하였으며, 응답자들의 응답속성을 감안하여 연건 평이나 대지면적이 25~34평인 것을 30평대, 35~44평인 것을 40평대라고 범주화하였다. 또한 대지면적의 경우 다가구주택 등으로 인해 20평대 이하의 값들도 있는데, 이에 대한 처리는 추후에 별도로 해야 할 것이다. 각 항목별 엔트로피에 대한 결과는 부록 A의 <부표 3>에 주어져 있으며, 이를 바탕으로 주택에 관한 사항에서는 연건평과 대지면적을 제외하고 거처의 종류만 최종적인 Key 변수로 선정하였다.

나. 표본의 추출

현재 통계청에서 제공하고 있는 인구주택총조사의 마이크로자료는 전체 10% 표본조사 중 2%에 해당하는 표본조사 결과자료이다. 따라서 본 연구에서도 비밀보호방법을 적용한 마이크로자료 파일의 작성을 위해 같은 규모인 2%의 표본을 추출하기로 하였다.

1) 추출틀과 추출단위

2% 표본의 추출을 위한 추출틀(sample frame)은 충청남도 표본조사구내에 있는 모든 가구명부로 하였으며, 추출단위(sampling unit)는 가구로 하였다. 추출단위를 가구로 한 것은 인구주택총조사의 표본조사구 특성을 감안할 때, 조사구보다는 가구가 개인의 식별가능성이 줄이는데 더 바람직 할 것으로 판단하였기 때문이다. 즉, 현행 인구주택총조사에서는 60~80가구를 하나의 조사구로 설정한 후, 이 중 10%를 표본조사구로 추출하고 표본조사구내 모든 가구는 표본조사표를 작성하도록 하고 있다. 이러한 방법은 조사의 편리성은 있으나 응답자의 식별가능성이 높아 노출의 위험이 그만큼 더 커지게 된다는 것을 의미한다. 가령, 아파트 표본조사구내에 포함되어 응답한 어떤 가구는 자신들의 이웃집도 같은 표본조사구로 조사에 응답했다는 사실을 쉽게 알 수 있다. 따라서 마이크로자료가 공표될 경우 이웃집의 정보를 식별할 가능성은 상당히 높게 될 것이다. 이러한 이유로 개인의 식별 가능성을 최대한 줄이기 위해 가구를 추출단위로 하였다.

2) 추출방법과 규모

위에서 언급한 바와 같이 마이크로자료 파일의 작성을 위해 새로 추출하는 2% 표본은 가급적 개인의 식별 가능성을 최소화하기 위해 계통추출법을 적용하여 추출하였다. 즉, 16개 시・군・구별로 정렬한 후 조사구 및 거처순으로 다시 정렬을 하고 최종적으로

가구원들의 성별 및 나이순으로 각각 정렬하여 추출틀을 작성하였다. 충청남도 전체의 2% 표본은 기존의 10% 표본의 20%에 해당되므로 SAS 프로그램을 이용하여 최종적인 2% 표본을 계통추출 하였으며, 그 결과가 <표 2-2>에 주어져 있다. 이렇게 추출한 표본이 모집단을 얼마나 잘 대표하는지를 알아보기 위해 가구원의 성별을 비교한 결과, 모집단(전수조사 결과)과 10% 표본의 구조와 크게 다르지 않음을 알 수 있었다.

〈표 2-2〉 표본추출 현황

(단위 : 개, 명, 가구, 호)

구분	전수(A)	10% 표본(B)		2% 표본(C)		
			구성비(B/A)		구성비(C/A)	구성비(C/B)
가구원	1,798,397	189,505	10.54%	38,027	2.11%	20.07%
남	892,044	93,023	10.43%	18,644	2.09%	20.04%
녀	906,353	96,482	10.65%	19,383	2.14%	20.09%
가 구	660,526	71,091	10.76%	14,218	2.15%	20.00%
주 택	586,757	65,398	11.15%	13,038	2.22%	19.94%

다. 노출위험의 계산

마이크로자료 파일의 작성을 위해 새로 추출한 2% 표본에서 노출과 식별의 가능성이 어느 정도인지를 파악하고자 uniqueness와 노출위험을 계산하기로 하고, 이를 위해 계산에 필요한 기호를 다음과 같이 나타내었다.

- A : 관심의 대상인 사람
- S_1 : 새로 추출한 2% 표본파일
- S_2 : 외부인에 의해 구성된 파일
- U_P : 모집단(전수조사결과)의 unique class
- U_S : 2% 표본의 unique class

Uniqueness와 노출위험의 계산을 위해 먼저, 앞 절에서 최종 선정한 8개의 Key 변수에 대한 다양한 조합을 고려하였다. 8개 변수로 얻을 수 있는 경우의 수는 1개 변수를 이용한 조합부터 8개 변수를 모두 이용한 조합까지 합하여 총 247개가 된다.

충청남도의 전수조사결과를 모집단으로 하고 새로 추출한 2% 표본조사결과를 표본으로 한 후, Key 변수의 각 조합별로 빈도수가 1이 되는 경우를 파악한 결과가 부록 C의

<부표 C-1>과 <부표 C-2>에 주어져 있다. 여러 조합들 중 8개 변수를 모두 고려한 조합이 uniqueness의 발생이 최대가 되는 경우이므로, 결국 uniqueness는 8개 변수를 모두 고려한 조합의 발생 빈도수로 파악할 수가 있는데, 모집단에서는 총 9,664개가 있으며 표본에서는 1,810개가 있는 것으로 각각 나타났다. 따라서 모집단에서 어떤 사람(A)이 unique할 확률은 약 0.54%가 되고,

$$\Pr(A \in U_P) = \frac{9,664}{1,798,397} = 0.00537$$

같은 방법으로 표본에서 어떤 사람(A)이 unique할 확률은 약 4.76%가 된다.

$$\Pr(A \in U_S) = \frac{1,810}{38,027} = 0.04759$$

즉, 표본에서의 uniqueness가 모집단에서의 uniqueness보다 약 8.8배 이상 더 높게 나타나고 있음을 알 수 있다. 이는 모집단보다 표본에서 unique한 사람이 더 많이 발생한다는 것을 의미하지만, 모집단에서 unique하지 않은 사람도 표본에서는 unique할 수가 있기 때문에 표본에서 unique한 사람이 반드시 모집단에서도 unique한 것은 아니다. 만약 모집단에서 unique한 어떤 사람이 표본으로 추출되었다면 이 사람의 노출 가능성은 상대적으로 높을 것이다. 충청남도의 경우 전체 모집단에서 unique한 9,664명 중 216명이 다시 2% 표본으로 추출되어 unique한 것으로 나타났다.

(단위 : 명, %)

조사	모집단	2% 표본	U_P	U_S	$U_P \cap U_S$
개수	1,798,397	38,027	9,664	1,810	216
구성비			0.537	4.760	0.01

이상의 uniqueness에 대한 결과를 이용하여 충청남도의 노출위험을 계산해 보았다. 노출위험의 발생 가능성이 최대가 되는 가장 극단적인 경우를 고려하여 외부인이 직접 작성한 파일(S_2)에 어떤 특정사람이 포함되어 있다는 것을 이미 알고 있다고 가정하였다. 이 가정하에서는 $\Pr(A \in S_2) = 1$ 이 되므로 2장에서 정의한 수식 (3)을 이용하여 노출위험을 계산하면, 어떤 특정 사람(A)이 모집단에서 unique하면서 인구주택총조사의 2% 표본에 포함될 확률은 약 0.01%가 된다. 즉,

$$DR(A) = \Pr(A \in S_1)\Pr(A \in U_P) = 0.02 \times \frac{9,664}{1,798,397} = 0.00011$$

여기서 $\Pr(A \in S_1) = 0.02$ 인데, 이는 제공하는 마이크로자료의 규모가 2%이기 때문이다. 위의 결과는 외부인이 가지고 있는 파일에 어떤 사람이 포함되어 있다는 것을 알고 있다고 가정하고 2%를 마이크로자료 파일로 제공할 경우 10,000명 중 1명 정도가 노출될 가능성이 있다는 것을 의미한다.

라. 자료의 Grouping

위에서 2% 표본의 uniqueness에 대해 살펴본 결과 총 38,027명 중 약4.76%인 1,810명이 unique한 것으로 나타났다. 만약 추출된 2%의 표본에 비밀보호방법을 적용하면 unique한 사람은 더 감소할 것이며, 이는 결국 mask 된 자료파일을 제공할 경우 개인의 식별과 노출에 대한 위험이 훨씬 더 줄어들 것이다. 비밀보호를 위한 방법은 자료의 형태에 따라 다양한 방법이 있는데, 여기서는 grouping방법을 적용하여 mask된 자료파일을 만들고자 하였다. Grouping 방법은 주로 이산형 자료로 구성된 경우 적용이 간편하고 매우 효과적인 방법이다.

본 연구에서 선정한 8개의 Key 변수들은 나이 변수를 제외하고는 모두 이산형 변수들이고 나이변수도 범주형 자료로 변환할 수가 있으므로, 각 변수들의 빈도분석의 결과를 이용하여 적정한 수준의 범주로 grouping하였다. 먼저, 가구원에 관한 사항에서는 4개의 Key 변수가 사용되었는데 성별 항목은 남녀별 범주를 그대로 사용하고, 나이 항목은 99세까지는 5세별로 20개의 범주로 하고, 100세 이상은 하나의 범주로 하여 총 21개 범주로 grouping하였다. 가구주와의 관계 항목은 14개의 범주를 총 4개(가구주, 배우자, 자녀 및 자녀의 배우자, 기타)의 범주로 축소하였다. 혼인상태 항목은 4개 범주 중 사별과 이혼을 합하여 총 3개(미혼, 기혼 및 배우자 있음, 기혼 및 배우자 없음)의 범주로 grouping하였다.

가구에 관한 사항에서는 Key 변수가 3개인데, 가구구분 항목은 5개 범주에서 가족이 아닌 남남끼리 사는 5인 이하 가구와 가족이 아닌 남남끼리 사는 6인 이상 가구를 합쳐 기타 범주로 하여 총 4개 범주로 grouping하였다. 점유형태는 원래 12개의 범주가 있는데, 주거전용과 영업전용은 그대로 두고 4개(자가, 전세, 월세, 무상) 범주로 하여 총 8개의 범주로 grouping하였다. 주인가구 및 주택소유여부 항목은 주인가구와 세들어 사는 가구로 크게 2개 범주로 축소하였다. 한편, 주택에 관한 사항에서는 거처의 종류 항목이 Key 변수로 사용되었는데, 기존의 12개 범주를 4개(단독주택, 아파트, 공동주택,기타)의 범주로 grouping하였다. 각 항목에 대한 grouping의 결과가 <표 2-3>에 주어져 있다.

2% 표본에서 8개 Key 변수를 <표 2-3>에 주어진 범주별로 그룹화한 후 uniqueness를 살펴보면, 총 553명이 unique한 것으로 나타났는데, 이는 그룹화하기 전의 1,810명보다

1,257명이 줄어들어 69.4% 정도 감소하였다. 따라서 그룹화를 한 후 2% 표본을 마이크로자료 파일로 제공해도 노출위험이 적기 때문에 적절하다고 할 수 있을 것이다.

〈표 2-3〉 Key 변수의 grouping 결과

(단위 : 개)

Key 변수	Grouping 전		Grouping 후	
	개수	범주	개수	범주
성별	2	① 남자 ② 여자	2	ⓐ 남자 (①) ⓑ 여자 (②)
나이	105	각 세별	21	ⓐ~ⓣ (각 5세별) ⓤ 100세 이상
가구주와의 관계	14	① 가구주 ② 배우자 ③ 자녀 ④ 자녀의 배우자 ⑤ 가구주의 부모 ⑥ 배우자의 부모 ⑦ 손자녀, 그의 배우자 ⑧ 증손자녀, 그의 배우자 ⑨ 조부모 ⑩ 형제자매, 그 배우자 ⑪ 형제자매의 자녀 ⑫ 부모의 형제자매 ⑬ 기타 친인척 ⑭ 기타동거인	4	ⓐ 가구주 (①) ⓑ 배우자 (②) ⓒ 자녀 및 자녀의 배우자 (③+④) ⓓ 기타 (⑤+ … +⑭)
혼인상태	4	① 미혼 ② 배우자 있음 ③ 사별 ④ 이혼	3	ⓐ 미혼 (①) ⓑ 기혼 및 배우자 있음 (②) ⓒ 기혼 및 배우자 없음 (③+④)
가구구분	5	① 가족으로 이루어진 가구 ② 가족과 가족이외의 사람이 함께사는 가구 ③ 1인 가구 ④ 가족이 아닌 남남끼리 사는 5인 이하 가구 ⑤ 가족이 아닌 남남끼리 사는 6인 이상 가구	4	ⓐ 가족으로 이루어진 가구 (①) ⓑ 가족과 가족이외의 사람이 함께사는 가구 (②) ⓒ 1인 가구 (③) ⓓ 기타 (④+⑤)

〈표 2-3〉 Key 변수의 grouping 결과(계속)

(단위 : 개)

Key 변수	Grouping 전		Grouping 후	
	개수	범주	개수	범주
점유형태	12	① 주거전용 + 자기집 ② 주거전용 + 전세 ③ 주거전용+보증금있는 월세 ④ 주거전용+보증금없는 월세 ⑤ 주거전용 + 사글세 ⑥ 주거전용 + 무상 ⑦ 영업전용 + 자기집 ⑧ 영업전용 + 전세 ⑨ 영업전용+보증금있는 월세 ⑩ 영업전용+보증금없는 월세 ⑪ 영업전용 + 사글세 ⑫ 영업전용 + 무상	8	ⓐ 주거전용 + 자기집 (①) ⓑ 주거전용 + 전세 (②) ⓒ 주거전용 + 월세 (③+④+⑤) ⓓ 주거전용 + 무상 ⑥) ⓔ 영업전용 + 자기집 (⑦) ⓕ 영업전용 + 전세 (⑧) ⓖ 영업전용 + 월세 (⑨+⑩+⑪) ⓗ 영업전용 + 무상 (⑫)
주인가구	6	① 주인가구 + 다른곳에 주택소유 ② 주인가구 + 다른곳에 주택미소유 ③ 대표가구 + 다른곳에 주택소유 ④ 대표가구 + 다른곳에 주택미소유 ⑤ 세들어 사는가구 + 다른곳에 주택소유 ⑥ 세들어 사는가구 + 다른곳에 주택미소유	2	ⓐ 주인가구 (①+②) ⓑ 세들어 사는가구 (③+ … +⑥)
거처의 종류	12	① 단독주택 ㉠ 일반단독주택 ㉡ 다가구단독주택 ㉢ 영업겸용단독주택 ② 아파트 ③ 연립주택 ④ 다세대주택 ⑤ 비거주용 건물 ⑥오피스텔 ⑦ 호텔, 여관내 객실 ⑧ 기숙사 및 특수시설 ⑨ 판잣집 등 ⑩ 기타	4	ⓐ 단독주택 (①) ⓑ 아파트 (②) ⓒ 공동주택 (③+④) ⓓ 기타 (⑤+ … +⑩)

제4절 결 론

앞에서 살펴본 바와 같이 본 연구에서는 마이크로자료 파일을 공표할 경우 개인의 비밀을 보호하는 다양한 형태의 masking방법을 소개하였고, 마이크로자료가 제공될 때 노출과 개인의 식별의 발생원인과 문제점 등을 살펴보았다. 또한 식별의 측도로 사용되는 uniqueness의 개념도 소개하였으며 특히, 노출의 위험을 통계적 확률로 정리하여 계산함으로써 쉽게 이해할 수 있도록 설명하였다.

비밀보호방법을 적용한 마이크로자료 파일을 작성하기 위해 2005 인구주택총조사의 표본조사 결과자료 중 충청남도 지역의 자료를 이용하였다. 먼저 masking에 필요한 Key 변수로 8개 변수를 선정하였다. 즉, 가구원에 관한 사항에서는 4개 항목(성별, 나이, 가구주와의 관계, 혼인상태)을 선정하였고, 가구에 관한 사항에서는 3개 항목(가구구분, 점유형태, 주인가구)을, 그리고 주택에 관한 사항에서는 1개 항목(거처의 종류)을 각각 선정하였다. 통계이용자들에게 제공할 마이크로자료 파일은 현행 통계청에서 제공하는 것과 같은 규모인 2% 표본조사결과로 정하고, 충청남도의 10% 표본조사결과들 중 20%를 계통추출법으로 추출하여 전체적으로 2% 표본을 구성하였다.

한편, 자료파일의 uniqueness와 노출의 위험을 알아보기 위해 전수조사결과와 2% 표본조사 결과에서 8개 Key 변수의 각 조합별로 빈도분석을 통해 빈도수가 1개인 경우를 파악하여 보았다. 그 결과, 모집단(전수조사결과)에서는 1,798,397명 중 9,664명(약 0.54%)이 unique한 것으로 나타났고, 2% 표본에서는 38,027명 중 1,810명(약 4.76%)이 unique한 것으로 나타났다. 그리고 이에 따른 노출 위험은 약 0.01%가 됨을 알 수 있었다. 8개 Key 변수에 대해 grouping방법을 적용하여 2% 표본을 다시 구성한 결과 unique한 경우는 553명으로 grouping하기 전의 1,810명에 비해 약 69.4%가 감소하였다. 따라서 grouping을 한 후 2% 표본을 마이크로자료 파일로 제공해도 노출위험이 적기 때문에 적절하다고 할 수 있다.

일반적으로 Key 변수의 선정은 연구자에 따라 다소 주관적이며, 일정하지 않을 수도 있다. 즉, 본 연구에서는 여러 과정을 거쳐 최종적으로 8개를 선정하였으나, 외부인(금융기관, 이웃주민 등)들이 파악할 수 있는 항목이 8개 이외에도 더 있다면 이 항목을 Key 변수의 선정에 고려해야 할 것이다. 또 한 자료를 grouping할 때에도 자료제공의 범위 등에 따라 다양하게 할 수도 있다. 가령, 나이 항목의 경우 본 연구에서는 5세 간격으로 하고 100세 이상은 하나의 범주로 grouping하였으나, 경우에 따라서는 10세 간격이나 20세 간격으로 할 수도 있을 것이다. 따라서 본 연구에서 선정한 8개 Key 변수와 그에 따른 grouping방법이 최적의 방법이라고 할 수는 없으며, 자료의 특성이나 제공의 범위에 따라 적정한 방법을 적용하면 될 것이다.

본 연구에서는 마이크로자료 파일을 작성할 때 인구주택총조사의 질문항목이 주로 이산형 형태로 구성된 것을 감안하여 이산형 변수의 비밀보호기법인 grouping방법을 적용하였으나, 인구주택총조사의 일부 항목들은 연속형인 경우도 있다. 가령, 주택에 관한 사항에서 연건평이나 대지면적 등의 항목은 연속형 자료로 얻어진다. 따라서 rounding방법 등과 같은 연속형 변수에 대한 기법을 적용하는 것이 더 타당할 것이다. 그러나 여기서는 연속형 변수에 대한 방법은 적용하지 않았기 때문에, 위의 항목에 대한 자료를 제공할 경우에는 원래자료를 제공하는 것보다는 이산형 형태로 범주화시켜 제공하는 것이 더 바람직할 것으로 판단된다.

향후에는 본 연구의 결과를 바탕으로 인구주택총조사의 표본조사결과에 대해 16개 시·도를 모두 포함한 전국 단위의 마이크로자료 파일을 작성할 수 있는 연구가 계속해서 추진되어야 할 것이다. 아울러 여기서 언급한 grouping방법 이외에 다른 방법들을 적용하는 연구뿐만 아니라, 통계청에서 생산하고 있는 각종 통계들 중 마이크로자료의 제공을 많이 요구하는 통계 조사를 대상으로 비밀보호방법을 적용하는 연구도 병행해서 진행되어야 할 것이다. 또한 자료이용자들의 요구가 발생할 경우 마이크로자료의 제공범위를 어느 정도로 할 것인가에 대한 고민과 함께 얼마만큼 신속하게 제공할 수 있는가에 대한 검토도 필요하다. 끝으로 이번 연구를 계기로 노출과 개인 식별에 대한 올바른 이해와 비밀보호방법에 대한 관심이 고조되어 다양한 방법들이 개발되고 활용되어지길 기대해 본다.

참고문헌

[1] 박원환, 황 조연(2004), “통계자료의 특성과 비밀보호방법에 관한 연구”, 통계청, 통계연구보고서.

[2] 통계청(2005), 2005 인구주택총조사 조사지침서.

[3] Bethelehem, J. G., Keller, W. J., and Pannekoek J. (1990), “Disclosure Control of Microdata”, *Journal of the American Statistical Association*, 85, 38-45.

[4] Cox, L. H., McDonald, S., and Nelson, D. (1986), “Confidentiality Issues at the United States Bureau of the Census”, *Journal of OfficialStatistics*, 2, 135-160.

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

[6] Fuller, W. A. (1993), “Masking Procedures for Microdata Disclosure Limitation”, *Journal of Official Statistics,* 9, 383-406.

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

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

[9] Kim, J. (1990), “Subdomain Estimation for the Masked Data”, in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 456-461.

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

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

[12] Kim, J., Katzoff, M., Gonzalez Jr, J., and Williams, P. (2003), “Techniques for Masking Microdata”, *National Center for Health Statistics internal memorandum.*

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

[14] Netherlands Central Bureau of Statistics (1987), *Proceedings of the Seminar on Openness and Protection of Privacy in the Information Society,* Voorburg: Author.

[15] Shannon, C. E. (1948), “A Mathematical Theory of Communication”, *The Bell System Technical Journal*, 27, 379-423, 623-656.

[16] Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994), “Disclosure Control for Census Microdata”, *Journal of Official Statistics*, 10, 31-51.

[17] Skinner, C., and Holmes, D. (1998), “Estimating the Re-identification Risk Per Record in Microdata”, *Journal of Official.*

<부 록>

A. 2005 인구주택총조사 결과(충청남도)

〈부표 A-1〉 전수조사 결과

(단위 : 개, 명, 가구, 호)

지역	조사구	가구원	가구	주택
전체	2,460	1,798,397	660,526	586,757
천안시	281	502,515	174,753	143,630
공주시	173	117,880	44,012	38,141
보령시	79	94,071	35,111	33,033
아산시	148	187,336	68,406	60,845
서산시	105	139,239	50,420	44,565
논산시	151	117,293	44,382	40,733
계룡시	110	30,476	9,768	9,357
금산군	129	50,648	20,057	18,714
연기군	239	75,479	28,369	23,600
부여군	148	70,647	27,732	26,764
서천군	96	55,993	22,892	21,821
청양군	64	30,592	12,430	12,068
홍성군	223	80,165	30,372	27,118
예산군	195	78,911	29,737	28,220
태안군	150	57,492	21,758	20,808
당진군	169	109,660	40,327	37,340

〈부표 A-2〉 10% 표본조사 결과

(단위 : 개, 명, 가구, 호)

지역	조사구	가구원	가구	주택
전체	749	189,505	71,091	65,398
천안시	139	43,233	15,098	12,878
공주시	51	13,554	5,098	4,513
보령시	35	10,598	4,023	3,858
아산시	68	18,324	6,732	6,009
서산시	52	13,300	4,771	4,365
논산시	57	12,777	4,788	4,568
계룡시	16	2,877	955	891
금산군	34	6,866	2,840	2,694
연기군	37	7,418	2,772	2,492
부여군	36	11,203	4,598	4,494
서천군	32	7,277	3,155	3,045
청양군	21	5,917	2,486	2,432
홍성군	44	9,275	3,587	3,468
예산군	42	9,374	3,609	3,505
태안군	34	6,677	2,519	2,476
당진군	51	10,835	4,060	3,710

〈부표 A-3〉 2% 표본조사 결과

(단위 : 개, 명, 가구, 호)

지역	조사구	가구원	가구	주택
전체	749	38,027	14,218	13,038
천안시	139	8,592	3,020	2,569
공주시	51	2,719	1,019	894
보령시	35	2,166	805	773
아산시	68	3,722	1,346	1,195
서산시	52	2,707	954	870
논산시	57	2,544	958	906
계룡시	16	562	191	177
금산군	34	1,384	568	537
연기군	37	1,493	554	495
부여군	36	2,251	920	898
서천군	32	1,422	631	608
청양군	21	1,193	497	487
홍성군	44	1,869	718	693
예산군	42	1,867	721	701
태안군	34	1,371	504	495
당진군	51	2,165	812	740

B. Key 변수의 Entropy

〈부표 B-1〉 성별의 Entropy

(단위 : 명, %)

성별	빈도수	구성비	Entropy
계	**189,505**	**100.0**	**0.99976**
남	93,023	49.1	0.50392
여	96,482	50.9	0.49584

〈부표 B-2〉 나이의 Entropy

(단위 : 세, %)

나이	빈도수	구성비	Entropy	나이	빈도수	구성비	Entropy
계	**189,505**	**100.0**	**6.39830**	19	2,035	1.1	0.07024
1	1,338	0.7	0.05045	20	1,585	0.8	0.05772
2	1,692	0.9	0.06078	21	1,471	0.8	0.05441
3	1,879	1.0	0.06600	22	1,360	0.7	0.05112
4	1,903	1.0	0.06666	23	1,960	1.0	0.06821
5	2,194	1.2	0.07447	24	2,224	1.2	0.07526
6	2,438	1.3	0.08080	25	2,256	1.2	0.07610
7	2,408	1.3	0.08003	26	2,340	1.2	0.07828
8	2,493	1.3	0.08220	27	2,249	1.2	0.07592
9	2,592	1.4	0.08469	28	1,995	1.1	0.06916
10	2,534	1.3	0.08323	29	2,150	1.1	0.07331
11	2,743	1.4	0.08844	30	2,247	1.2	0.07586
12	2,656	1.4	0.08629	31	2,240	1.2	0.07568
13	2,657	1.4	0.08632	32	2,654	1.4	0.08624
14	2,604	1.4	0.08499	33	2,642	1.4	0.08594
15	2,528	1.3	0.08308	34	2,697	1.4	0.08731
16	2,316	1.2	0.07766	35	3,056	1.6	0.09602
17	2,176	1.1	0.07400	36	2,832	1.5	0.09063
18	2,081	1.1	0.07147	37	2,616	1.4	0.08529

(단위 : 세, %)

나이	빈도수	구성비	Entropy	나이	빈도수	구성비	Entropy
38	2,945	1.6	0.09336	70	2,782	1.5	0.08940
39	2,636	1.4	0.08579	71	2,220	1.2	0.07516
40	2,944	1.6	0.09334	72	2,242	1.2	0.07573
41	2,719	1.4	0.08785	73	2,188	1.2	0.07431
42	2,777	1.5	0.08928	74	1,933	1.0	0.06748
43	3,166	1.7	0.09863	75	1,666	0.9	0.06004
44	2,978	1.6	0.09416	76	1,593	0.8	0.05795
45	3,007	1.6	0.09485	77	1,291	0.7	0.04903
46	3,399	1.8	0.10405	78	1,405	0.7	0.05246
47	2,995	1.6	0.09457	79	1,011	0.5	0.04028
48	3,093	1.6	0.09690	80	1,083	0.6	0.04258
49	2,957	1.6	0.09365	81	862	0.5	0.03539
50	2,618	1.4	0.08534	82	759	0.4	0.03190
51	2,644	1.4	0.08599	83	709	0.4	0.03016
52	2,644	1.4	0.08599	84	663	0.3	0.02855
53	1,987	1.0	0.06895	85	516	0.3	0.02320
54	2,658	1.4	0.08634	86	378	0.2	0.01789
55	1,868	1.0	0.06569	87	332	0.2	0.01604
56	2,162	1.1	0.07363	88	275	0.1	0.01368
57	2,316	1.2	0.07766	89	219	0.1	0.01128
58	2,171	1.1	0.07387	90	211	0.1	0.01092
59	2,578	1.4	0.08434	91	162	0.1	0.00871
60	2,041	1.1	0.07040	92	113	0.1	0.00639
61	1,626	0.9	0.05890	93	92	0.0	0.00534
62	1,924	1.0	0.06723	94	69	0.0	0.00416
63	1,927	1.0	0.06731	95	56	0.0	0.00346
64	2,343	1.2	0.07836	96	25	0.0	0.00170
65	2,660	1.4	0.08639	97	33	0.0	0.00217
66	1,909	1.0	0.06682	98	15	0.0	0.00108
67	2,263	1.2	0.07628	99	12	0.0	0.00088
68	2,554	1.3	0.08374	100~	24	0.0	0.00193
69	2,216	1.2	0.07505				

〈부표 B-3〉 가구주와의 관계의 Entropy

(단위 : 명, %)

가구주와의 관계	빈도수	구성비	Entropy
계	189,505	100.0	2.1258
1	71,091	37.5	0.5306
2	45,889	24.2	0.4954
3	53,422	28.2	0.5150
4	1,515	0.8	0.0557
5	6,715	3.5	0.1707
6	491	0.3	0.0223
7	4,624	2.4	0.1307
8	41	0.0	0.0026
9	123	0.1	0.0069
10	1,368	0.7	0.0514
11	394	0.2	0.0185
12	36	0.0	0.0023
13	457	0.2	0.0210
14	3,339	1.8	0.1027

〈부표 B-4〉 혼인상태의 Entropy

(단위 : 명, %)

혼인상태	빈도수	구성비	Entropy
계	154,415	100.0	1.3547
1	32,020	20.7	0.4707
2	101,159	65.5	0.3997
3	17,610	11.4	0.3572
4	3,626	2.3	0.1271

〈부표 B-5〉 교육정도의 Entropy

(단위 : 명, %)

교육정도	빈도수	구성비	Entropy
계	189,505	100.0	3.3231
10	31,976	16.9	0.4332
21	32,768	17.3	0.4378
22	15,464	8.2	0.2950
24	2	0.0	0.0002
25	1,738	0.9	0.0621
31	15,725	8.3	0.2980
32	7,343	3.9	0.1817
34	2	0.0	0.0002
35	698	0.4	0.0298
41	42,892	22.6	0.4851
42	6,251	3.3	0.1624
44	22	0.0	0.0015
45	770	0.4	0.0323
51	9,915	5.2	0.2227
52	1,459	0.8	0.0541
53	232	0.1	0.0118
54	504	0.3	0.0228
55	588	0.3	0.0259
61	12,716	6.7	0.2615
62	4,154	2.2	0.1208
63	144	0.1	0.0079
64	629	0.3	0.0273
65	688	0.4	0.0294
71	1,628	0.9	0.0590
72	352	0.2	0.0169
73	319	0.2	0.0155
74	26	0.0	0.0018
75	23	0.0	0.0016
81	322	0.2	0.0156
82	65	0.0	0.0039
83	81	0.0	0.0048
84	9	0.0	0.0007

〈부표 B-6〉 가구구분의 Entropy

(단위 : 가구, %)

가구구분	빈도수	구성비	Entropy
계	**71,091**	**100.0**	**0.9630**
1	53,085	74.7	0.3146
2	266	0.4	0.0302
3	16,054	22.6	0.4848
4	1,624	2.3	0.1245
5	62	0.1	0.0089

〈부표 B-7〉 거주층의 Entropy

(단위 : 가구, %)

거주층	빈도수	구성비	Entropy
계	**71,091**	**100.0**	**0.0382**
1	230	0.3	0.0268
2	70,822	99.6	0.0054
3	39	0.1	0.0059

〈부표 B-8〉 주인가구여부의 Entropy

(단위 : 가구, %)

주인가구여부	빈도수	구성비	Entropy
계	**71,091**	**100.0**	**1.6045**
11	4,086	5.7	0.2369
12	43,983	61.9	0.4286
21	2,665	3.7	0.1776
22	14,664	20.6	0.4698
30	5,693	8.0	0.2917

〈부표 B-9〉 점유형태의 Entropy

(단위 : 가구, %)

점유형태	빈도수	구성비	Entropy
계	71,091	100.0	1.8059
11	46,661	65.6	0.3987
12	8,095	11.4	0.3569
13	5,851	8.2	0.2965
14	1,984	2.8	0.1441
15	1,467	2.1	0.1155
16	4,463	6.3	0.2507
21	1,408	2.0	0.1121
22	229	0.3	0.0267
23	429	0.6	0.0445
24	165	0.2	0.0203
25	55	0.1	0.0080
26	284	0.4	0.0318

〈부표 B-10〉 거처의 종류의 Entropy

(단위 : 가구, %)

거처종류	빈도수	구성비	Entropy
계	65,398	100.0	1.5397
11	37,928	58.0	0.4558
12	913	1.4	0.0860
13	365	0.6	0.0418
2	21,277	32.5	0.5270
3	1,999	3.1	0.1538
4	1,430	2.2	0.1206
5	1,095	1.7	0.0988
6	123	0.2	0.0170
7	38	0.1	0.0062
8	17	0.0	0.0031
9	50	0.1	0.0079
10	163	0.2	0.0216

〈부표 B-11〉 연건평의 Entropy

(단위 : 평대, %)

연건평	빈도수	구성비	Entropy
계	65,130	100.0	1.8682
10	6,779	10.4	0.3397
20	29,070	44.6	0.5194
30	23,618	36.3	0.5307
40	3,315	5.1	0.2187
50	1,176	1.8	0.1046
60	513	0.8	0.0550
70	201	0.3	0.0257
80	92	0.1	0.0134
90	79	0.1	0.0118
100	55	0.1	0.0086
110	22	0.0	0.0039
120	43	0.1	0.0070
130	23	0.0	0.0040
140	26	0.0	0.0045
150	37	0.1	0.0061
160	7	0.0	0.0014
170	7	0.0	0.0014
180	13	0.0	0.0025
190	13	0.0	0.0025
200	24	0.0	0.0042
210 ~	17	0.0	0.0031

〈부표 B-12〉 대지면적의 Entropy

(단위 : 평대, %)

대지면적	빈도수	구성비	Entropy
계	65,130	100.0	1.8682
10	160	0.4	0.0324
20	955	2.4	0.1305
30	2,530	6.5	0.2551
40	2,799	7.1	0.2719
50	4,830	12.3	0.3722
60	2,663	6.8	0.2635
70	2,729	7.0	0.2676
80	2,879	7.3	0.2767
90	1,383	3.5	0.1702
100	5,220	13.3	0.3873
110	909	2.3	0.1259
120	1,566	4.0	0.1856
130	960	2.4	0.1310
140	523	1.3	0.0831
150	2,339	6.0	0.2426
160	456	1.2	0.0747
170	460	1.2	0.0752
180	665	1.7	0.0998
190	299	0.8	0.0536
200	2,382	6.1	0.2455
210	170	0.4	0.0340
220	150	0.4	0.0307
230	143	0.4	0.0295
240	102	0.3	0.0223
250	337	0.9	0.0590
260	69	0.2	0.0161
270	101	0.3	0.0222
280	110	0.3	0.0238
290	59	0.2	0.0141
300	551	1.4	0.0865
310	25	0.1	0.0068
320	45	0.1	0.0112
330	28	0.1	0.0075
340	22	0.1	0.0061
350	67	0.2	0.0157
360	18	0.0	0.0051
370	27	0.1	0.0072
380	27	0.1	0.0072
390	19	0.0	0.0053
400	186	0.5	0.0366
410 ~	243	0.6	0.0455

C. Key 변수의 Uniqueness

〈부표 C-1〉 모집단(전수조사)의 Uniqueness

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
1개 변수	1	S	-	2	-	2
	2	A	2	111	1	105
	3	R	-	14	-	14
	4	M	-	4	-	4
	5	HT	-	5	-	5
	6	PT	-	12	-	12
	7	HO	-	3	-	3
	8	PJ	-	10	-	10
2개 변수	1	S×A	5	217	3	206
	2	S×R	-	28	-	28
	3	S×M	-	10	-	10
	4	S×HT	-	10	-	10
	5	S×PT	-	24	-	24
	6	S×HO	-	6	-	6
	7	S×PJ	-	20	-	19
	8	A×R	65	1,037	101	916
	9	A×M	11	335	16	303
	10	A×HT	29	414	31	360
	11	A×PT	63	963	77	848
	12	A×HO	10	274	11	253
	13	A×PJ	71	717	77	566
	14	R×M	-	55	-	53
	15	R×HT	-	5	-	5
	16	R×PT	-	12	-	12
	17	R×HO	-	3	-	3
	18	R×PJ	-	10	-	10
	19	M×HT	-	22	-	21
	20	M×PT	1	54	3	51
	21	M×HO	-	15	1	14
	22	M×PJ	2	43	2	39
	23	HT×PT	-	60	4	58
	24	HT×HO	-	15	-	15

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
2개 변수	25	HT×PJ	1	47	-	43
	26	PT×HO	-	22	-	22
	27	PT×PJ	9	116	10	102
	28	HO×PJ	-	20	1	20
3개 변수	1	S×A×R	167	1,917	206	1,631
	2	S×A×M	30	647	34	576
	3	S×A×HT	72	778	67	655
	4	S×A×PT	147	1,841	207	1,551
	5	S×A×HO	22	532	24	489
	6	S×A×PJ	151	1,282	136	996
	7	S×R×M	2	108	5	102
	8	S×R×HT	-	10	-	10
	9	S×R×PT	-	24	-	24
	10	S×R×HO	-	6	-	6
	11	S×R×PJ	-	20	-	19
	12	S×M×HT	-	44	1	39
	13	S×M×PT	6	106	3	99
	14	S×M×HO	2	30	1	26
	15	S×M×PJ	2	84	3	72
	16	S×HT×PT	5	120	6	109
	17	S×HT×HO	-	30	-	29
	18	S×HT×PJ	3	91	5	80
	19	S×PT×HO	-	44	-	44
	20	S×PT×PJ	22	213	19	175
	21	S×HO×PJ	-	40	1	38
	22	A×R×M	349	2,337	379	1,823
	23	A×R×HT	29	414	31	360
	24	A×R×PT	63	963	77	848
	25	A×R×HO	10	274	11	253
	26	A×R×PJ	71	717	77	566
	27	A×M×HT	154	1,170	139	939
	28	A×M×PT	330	2,843	441	2,222

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
3개 변수	29	A×M×HO	57	888	62	782
	30	A×M×PJ	300	1,911	274	1,437
	31	A×HT×PT	544	3,157	514	2,339
	32	A×HT×HO	117	1,069	105	895
	33	A×HT×PJ	473	2,221	387	1,604
	34	A×PT×HO	139	1,667	181	1,412
	35	A×PT×PJ	1,029	4,102	863	2,803
	36	A×HO×PJ	162	1,285	160	988
	37	R×M×HT	-	22	-	21
	38	R×M×PT	1	54	3	51
	39	R×M×HO	-	15	1	14
	40	R×M×PJ	2	43	2	39
	41	R×HT×PT	-	60	4	58
	42	R×HT×HO	-	15	-	15
	43	R×HT×PJ	1	47	-	43
	44	R×PT×HO	-	22	-	22
	45	R×PT×PJ	9	116	10	102
	46	R×HO×PJ	-	20	1	20
	47	M×HT×PT	24	238	22	207
	48	M×HT×HO	3	66	2	59
	49	M×HT×PJ	16	172	12	142
	50	M×PT×HO	1	94	3	91
	51	M×PT×PJ	55	386	50	304
	52	M×HO×PJ	3	83	6	74
	53	HT×PT×HO	3	109	7	101
	54	HT×PT×PJ	64	396	69	317
	55	HT×HO×PJ	6	90	6	80
	56	PT×HO×PJ	9	116	10	102
4개 변수	1	S×A×R×M	713	4,034	685	3,039
	2	S×A×R×HT	72	778	67	655
	3	S×A×R×PT	147	1,841	207	1,551
	4	S×A×R×HO	22	532	24	489

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	5	S×A×R×PJ	151	1,282	136	996
	6	S×A×M×HT	304	2,054	282	1,595
	7	S×A×M×PT	836	4,987	887	3,600
	8	S×A×M×HO	120	1,657	134	1,450
	9	S×A×M×PJ	601	3,204	500	2,343
	10	S×A×PT×PT	1,086	5,370	996	3,818
	11	S×A×PT×HO	248	1,956	248	1,603
	12	S×A×PT×PJ	840	3,646	694	2,581
	13	S×A×PT×HO	338	3,129	439	2,494
	14	S×A×PT×PJ	1,833	6,645	1,450	4,324
	15	S×A×HO×PJ	330	2,259	294	1,727
	16	S×R×M×HT	-	44	1	39
	17	S×R×M×PT	6	106	3	99
	18	S×R×M×HO	2	30	1	26
	19	S×R×M×PJ	2	84	3	72
	20	S×R×HT×PT	5	120	7	107
	21	S×R×HT×HO	-	30	-	29
	22	S×R×HT×PJ	3	91	5	80
	23	S×R×PT×HO	-	44	-	44
	24	S×R×PT×PJ	22	213	19	175
	25	S×R×HO×PJ	-	40	1	38
	26	S×M×HT×PT	54	441	46	367
	27	S×M×HT×HO	9	129	4	110
	28	S×M×HT×PJ	37	312	34	250
	29	S×M×PT×HO	6	186	4	179
	30	S×M×PT×PJ	108	661	103	509
	31	S×M×HO×PJ	14	161	15	138
	32	S×HT×PT×HO	12	212	19	191
	33	S×HT×PT×PJ	119	680	129	524
	34	S×HT×HO×PJ	11	168	17	144
	35	S×PT×HO×PJ	22	213	19	175
	36	A×R×M×HT	154	1,168	139	939

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	37	A×R×M×PT	330	2,843	441	2,222
	38	A×R×M×HO	57	888	62	792
	39	A×R×M×PJ	300	1,911	274	1,437
	40	A×R×HT×PT	544	3,157	514	2,339
	41	A×R×HT×HO	117	1,069	105	895
	42	A×R×HT×PJ	473	2,221	387	1,604
	43	A×R×PT×HO	139	1,667	181	1,412
	44	A×R×PT×PJ	1,029	4,102	863	2,803
	45	A×R×HO×PJ	162	1,285	160	988
	46	A×M×HT×PT	1,766	7,107	1,484	4,706
	47	A×M×HT×HO	480	2,825	454	2,192
	48	A×M×HT×PJ	1,309	4,774	1,012	3,204
	49	A×M×PT×HO	782	4,661	821	3,409
	50	A×M×PT×PJ	2,679	8,747	1,979	5,359
	51	A×M×HO×PJ	592	3,291	519	2,421
	52	A×HT×PT×HO	958	4,864	931	3,523
	53	A×HT×PT×PJ	3,022	9,043	2,209	5,445
	54	A×HT×HO×PJ	853	3,591	673	2,500
	55	A×PT×HO×PJ	1,029	4,102	863	2,803
	56	R×M×HT×PT	24	238	22	207
	57	R×M×HT×HO	3	66	2	59
	58	R×M×HT×PJ	16	172	12	142
	59	R×M×PT×HO	1	94	3	91
	60	R×M×PT×PJ	55	386	50	304
	61	R×M×HO×PJ	3	83	6	74
	62	R×HT×PT×HO	3	109	7	101
	63	R×HT×PT×PJ	64	396	69	317
	64	R×HT×HO×PJ	6	90	6	80
	65	R×PT×HO×PJ	9	116	10	102
	66	M×HT×PT×HO	44	401	49	342
	67	M×HT×PT×PJ	231	1,093	220	797
	68	M×HT×HO×PJ	32	305	33	252

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	69	M×PT×HO×PJ	55	386	50	304
	70	HT×PT×HO×PJ	64	396	69	317
5개 변수	1	S×A×R×M×HT	304	2,054	282	1,595
	2	S×A×R×M×PT	836	4,987	887	3,600
	3	S×A×R×M×HO	120	1,657	134	1,450
	4	S×A×R×M×PJ	601	3,204	500	2,343
	5	S×A×R×HT×PT	1,086	5,370	996	3,818
	6	S×A×R×HT×HO	248	1,956	248	1,603
	7	S×A×R×HT×PJ	840	3,646	694	2,581
	8	S×A×R×PT×HO	338	3,129	439	2,494
	9	S×A×R×PT×PJ	1,833	6,645	1,450	4,324
	10	S×A×R×HO×PJ	330	2,259	294	1,727
	11	S×A×M×HT×PT	3,182	11,114	2,492	7,002
	12	S×A×M×HT×HO	965	4,812	835	3,601
	13	S×A×M×HT×PJ	2,167	7,383	1,705	4,778
	14	S×A×M×PT×HO	1,672	7,789	1,570	5,307
	15	S×A×M×PT×PJ	4,467	12,902	3,095	7,475
	16	S×A×M×HO×PT	1,171	5,419	981	3,831
	17	S×A×HT×PT×HO	1,888	8,112	1,709	5,571
	18	S×A×HT×PT×PJ	4,933	13,421	3,384	7,736
	19	S×A×HT×HO×PJ	1,477	5,804	1,203	3,959
	20	S×A×PT×HO×PJ	1,833	6,645	1,450	4,324
	21	S×R×M×HT×PT	54	441	46	367
	22	S×R×M×HT×HO	9	129	4	110
	23	S×R×M×HT×PJ	37	312	34	250
	24	S×R×M×PT×HO	6	186	4	179
	25	S×R×M×PT×PJ	108	661	103	509
	26	S×R×M×HO×PJ	14	161	15	138
	27	S×R×HT×PT×HO	12	212	19	191
	28	S×R×HT×PT×PJ	119	680	129	524
	29	S×R×HT×HO×PJ	11	168	17	144
	30	S×R×PT×HO×PJ	22	213	19	175

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
5개 변수	31	S×M×HT×PT×HO	92	726	94	599
	32	S×M×HT×PT×PJ	407	1,755	370	1,226
	33	S×M×HT×HO×PJ	71	544	75	432
	34	S×M×PT×HO×PJ	108	661	103	509
	35	S×HT×PT×HO×PJ	119	680	129	524
	36	A×R×M×HT×PT	1,766	7,107	1,484	4,706
	37	A×R×M×HT×HO	480	2,825	454	2,192
	38	A×R×M×HT×PJ	1,309	4,774	1,012	3,204
	39	A×R×M×PT×HO	782	4,661	821	3,409
	40	A×R×M×PT×PJ	2,679	8,747	1,979	5,359
	41	A×R×M×HO×PJ	592	3,291	519	2,421
	42	A×R×HT×PT×HO	958	4,864	931	3,523
	43	A×R×HT×PT×PJ	3,022	9,043	2,209	5,445
	44	A×R×HT×HO×PJ	853	3,591	673	2,500
	45	A×R×PT×HO×PJ	1,029	4,102	863	2,803
	46	A×M×HT×PT×HO	2,925	10,372	2,354	6,624
	47	A×M×HT×PT×PJ	6,455	16,014	4,002	8,663
	48	A×M×HT×HO×PJ	2,178	7,477	1,649	4,818
	49	A×M×PT×HO×PJ	2,679	8,747	1,979	5,359
	50	A×HT×PT×HO×PJ	3,022	9,043	2,209	5,445
	51	R×M×HT×PT×HO	44	401	49	342
	52	R×M×HT×PT×PJ	231	1,093	220	797
	53	R×M×HT×HO×PJ	32	305	33	252
	54	R×M×PT×HO×PJ	55	386	50	304
	55	R×HT×PT×HO×PJ	64	396	69	317
	56	M×HT×PT×HO×PJ	231	1,093	220	797
6개 변수	1	S×A×R×M×HT×PT	3,182	11,114	2,492	7,002
	2	S×A×R×M×HT×HO	965	4,812	835	3,601
	3	S×A×R×M×HT×PJ	2,167	7,383	1,705	4,778
	4	S×A×R×M×PT×HO	1,672	7,789	1,570	5,307
	5	S×A×R×M×PT×PJ	4,467	12,902	3,095	7,475
	6	S×A×R×M×HO×PJ	1,171	5,419	981	3,831

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
6개 변수	7	S×A×R×HT×PT×HO	1,888	8,112	1,709	5,571
	8	S×A×R×HT×PT×PJ	4,933	13,421	3,384	7,736
	9	S×A×R×HT×HO×PJ	1,477	5,804	1,203	3,959
	10	S×A×R×PT×HO×PJ	1,833	6,645	1,450	4,324
	11	S×A×M×HT×PT×HO	5,099	15,718	3,773	9,521
	12	S×A×M×HT×PT×PJ	9,664	22,110	5,719	11,461
	13	S×A×M×HT×HO×PJ	3,667	11,339	2,709	7,009
	14	S×A×M×PT×HO×PJ	4,467	2,902	3,095	7,475
	15	S×A×HT×PT×HO×PJ	4,933	13,421	3,384	7,736
	16	S×R×M×HT×PT×HO	92	726	94	599
	17	S×R×M×HT×PT×PJ	407	1,755	370	1,226
	18	S×R×M×HT×HO×PJ	71	544	74	432
	19	S×R×M×PT×HO×PJ	108	661	103	509
	20	S×R×HT×PT×HO×PJ	119	680	129	524
	21	S×M×HT×PT×HO×PJ	407	1,755	370	1,226
	22	A×R×M×HT×PT×HO	2,925	10,372	2,354	6,624
	23	A×R×M×HT×PT×PJ	6,455	16,014	4,002	8,663
	24	A×R×M×HT×HO×PJ	2,178	7,477	1,649	4,818
	25	A×R×M×PT×HO×PJ	2,579	8,747	1,979	5,359
	26	A×R×HT×PT×HO×PJ	3,022	9,043	2,209	5,445
	27	A×M×HT×PT×HO×PJ	6,455	16,014	4,002	8,663
	28	R×M×HT×PT×HO×PJ	231	1,093	220	797
7개 변수	1	S×A×R×M×HT×PT×HO	5,099	15,718	3,773	9,521
	2	S×A×R×M×HT×PT×PJ	9,664	22,110	5,719	11,461
	3	S×A×R×M×HT×HO×PJ	3,667	11,339	2,708	7,009
	4	S×A×R×M×PT×HO×PJ	4,467	12,902	3,095	7,475
	5	S×A×R×HT×PT×HO×PJ	4,933	13,421	3,384	7,736
	6	S×A×M×HT×PT×HO×PJ	9,664	22,110	5,719	11,461
	7	S×R×M×HT×PT×HO×PJ	407	1,755	370	1,226
	8	A×R×M×HT×PT×HO×PJ	6,455	16,014	4,002	8,663
8개 변수	1	S×A×R×M×HT×PT×HO×PJ	9,664	22,069	5,719	11,461

주) S(성별), A(나이), R(가구주와의 관계), M(결혼상태), HT(가구구분), PT(점유형태), HO(주택소유), PJ(거처종류)

〈부표 C-2〉 2% 표본자료의 Uniqueness

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
1개 변수	1	S	-	2	-	2
	2	A	1	101	3	96
	3	R	-	14	1	14
	4	M	-	4	-	4
	5	HT	-	5	-	5
	6	PT	-	12	1	12
	7	HO	-	3	-	3
	8	PJ	-	10	-	6
2개 변수	1	S×A	1	196	4	183
	2	S×R	1	28	2	28
	3	S×M	-	8	-	8
	4	S×HT	-	10	2	10
	5	S×PT	1	24	2	21
	6	S×HO	-	6	-	6
	7	S×PJ	1	20	-	12
	8	A×R	123	630	120	426
	9	A×M	20	259	35	219
	10	A×HT	49	262	30	184
	11	A×PT	102	582	115	349
	12	A×HO	12	223	24	189
	13	A×PJ	79	377	69	266
	14	R×M	6	46	3	34
	15	R×HT	-	5	-	5
	16	R×PT	-	12	1	12
	17	R×HO	-	3	-	3
	18	R×PJ	-	10	-	6
	19	M×HT	2	20	3	17
	20	M×PT	2	46	8	40
	21	M×HO	-	12	-	12
	22	M×PJ	6	37	1	23
	23	HT×PT	5	49	6	36
	24	HT×HO	-	14	1	13

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
2개 변수	25	HT×PJ	6	38	3	21
	26	PT×HO	1	22	2	21
	27	PT×PJ	12	65	9	42
	28	HO×PJ	-	17	-	11
3개 변수	1	S×A×R	242	1,012	217	647
	2	S×A×M	41	481	76	371
	3	S×A×HT	78	439	67	308
	4	S×A×PT	250	974	199	528
	5	S×A×HO	34	422	62	331
	6	S×A×PJ	145	605	154	411
	7	S×R×M	11	82	10	62
	8	S×R×HT	-	10	2	10
	9	S×R×PT	1	24	2	21
	10	S×R×HO	-	6	-	6
	11	S×R×PJ	1	20	-	12
	12	S×M×HT	4	36	7	31
	13	S×M×PT	7	87	12	65
	14	S×M×HO	-	24	-	24
	15	S×M×PJ	10	60	6	45
	16	S×HT×PT	12	86	13	60
	17	S×HT×HO	1	27	6	25
	18	S×HT×PJ	15	64	5	37
	19	S×PT×HO	2	43	4	35
	20	S×PT×PJ	24	110	17	70
	21	S×HO×PJ	1	33	-	22
	22	A×R×M	293	991	224	590
	23	A×R×HT	49	262	30	184
	24	A×R×PT	102	582	115	349
	25	A×R×HO	12	223	24	187
	26	A×R×PJ	79	377	69	266
	27	A×M×HT	124	568	140	365
	28	A×M×PT	398	1,198	283	603

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
3개 변수	29	A×M×HO	89	595	139	422
	30	A×M×PJ	225	772	206	472
	31	A×HT×PT	388	1,158	259	579
	32	A×HT×HO	128	579	114	389
	33	A×HT×PJ	269	748	182	430
	34	A×PT×HO	220	833	163	440
	35	A×PT×PJ	474	1,196	272	551
	36	A×HO×PJ	133	607	127	391
	37	R×M×HT	2	20	3	17
	38	R×M×PT	2	46	8	40
	39	R×M×HO	-	12	-	12
	40	R×M×PJ	6	37	1	23
	41	R×HT×PT	5	49	6	36
	42	R×HT×HO	-	14	1	13
	43	R×HT×PJ	6	38	3	21
	44	R×PT×HO	1	22	2	21
	45	R×PT×PJ	12	65	9	42
	46	R×HO×PJ	-	17	-	11
	47	M×HT×PT	29	142	30	88
	48	M×HT×HO	10	51	8	39
	49	M×HT×PJ	24	95	16	60
	50	M×PT×HO	6	80	14	62
	51	M×PT×PJ	40	176	34	102
	52	M×HO×PJ	6	61	5	43
	53	HT×PT×HO	10	80	11	56
	54	HT×PT×PJ	53	166	31	89
	55	HT×HO×PJ	8	58	7	34
	56	PT×HO×PJ	12	65	9	42
4개 변수	1	S×A×R×M	473	1,478	355	819
	2	S×A×R×HT	78	439	67	307
	3	S×A×R×PT	250	974	199	528
	4	S×A×R×HO	34	422	62	330

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	5	S×A×R×PJ	145	605	154	410
	6	S×A×M×HT	242	903	237	513
	7	S×A×M×PT	699	1,669	425	785
	8	S×A×M×HO	274	979	265	592
	9	S×A×M×PJ	381	1,094	326	599
	10	S×A×PT×PT	675	1,673	419	791
	11	S×A×PT×HO	243	949	241	587
	12	S×A×PT×PJ	387	1,061	296	578
	13	S×A×PT×HO	464	1,299	279	643
	14	S×A×PT×PJ	774	1,657	402	730
	15	S×A×HO×PJ	260	947	237	560
	16	S×R×M×HT	4	36	7	30
	17	S×R×M×PT	7	87	12	65
	18	S×R×M×HO	-	24	-	24
	19	S×R×M×PJ	10	60	6	45
	20	S×R×HT×PT	12	86	13	60
	21	S×R×HT×HO	1	27	6	25
	22	S×R×HT×PJ	15	64	5	37
	23	S×R×PT×HO	2	43	4	35
	24	S×R×PT×PJ	24	110	17	70
	25	S×R×HO×PJ	1	33	-	22
	26	S×M×HT×PT	59	229	40	133
	27	S×M×HT×HO	15	87	12	68
	28	S×M×HT×PJ	36	151	33	99
	29	S×M×PT×HO	20	144	22	100
	30	S×M×PT×PJ	75	269	60	155
	31	S×M×HO×PJ	12	100	10	75
	32	S×HT×PT×HO	26	136	19	90
	33	S×HT×PT×PJ	81	254	48	131
	34	S×HT×HO×PJ	21	99	13	59
	35	S×PT×HO×PJ	24	110	17	70
	36	A×R×M×HT	124	568	140	365

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	37	A×R×M×PT	398	1,198	283	603
	38	A×R×M×HO	89	595	139	422
	39	A×R×M×PJ	225	772	206	472
	40	A×R×HT×PT	388	1,158	259	579
	41	A×R×HT×HO	128	579	114	389
	42	A×R×HT×PJ	269	748	182	430
	43	A×R×PT×HO	220	833	164	440
	44	A×R×PT×PJ	474	1,196	272	551
	45	A×R×HO×PJ	133	607	127	391
	46	A×M×HT×PT	862	1,861	464	799
	47	A×M×HT×HO	369	1,143	307	620
	48	A×M×HT×PJ	515	1,251	347	613
	49	A×M×PT×HO	624	1,507	363	708
	50	A×M×PT×PJ	939	1,911	459	782
	51	A×M×HO×PJ	395	1,194	307	621
	52	A×HT×PT×HO	609	1,492	343	695
	53	A×HT×PT×PJ	922	1,804	429	739
	54	A×HT×HO×PJ	404	1,096	285	584
	55	A×PT×HO×PJ	474	1,196	272	551
	56	R×M×HT×PT	29	142	30	88
	57	R×M×HT×HO	10	51	8	39
	58	R×M×HT×PJ	24	95	16	60
	59	R×M×PT×HO	6	80	14	62
	60	R×M×PT×PJ	40	176	34	102
	61	R×M×HO×PJ	6	61	5	43
	62	R×HT×PT×HO	10	80	11	56
	63	R×HT×PT×PJ	53	166	31	89
	64	R×HT×HO×PJ	8	58	6	28
	65	R×PT×HO×PJ	12	65	9	42
	66	M×HT×PT×HO	52	213	40	126
	67	M×HT×PT×PJ	131	353	77	174
	68	M×HT×HO×PJ	30	151	27	92

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	69	M×PT×HO×PJ	40	176	34	102
	70	HT×PT×HO×PJ	53	166	31	89
5개 변수	1	S×A×R×M×HT	242	903	237	513
	2	S×A×R×M×PT	699	1,669	425	785
	3	S×A×R×M×HO	274	979	265	592
	4	S×A×R×M×PJ	381	1,094	326	599
	5	S×A×R×HT×PT	675	1,673	419	791
	6	S×A×R×HT×HO	243	949	241	587
	7	S×A×R×HT×PJ	387	1,061	296	578
	8	S×A×R×PT×HO	464	1,299	279	643
	9	S×A×R×PT×PJ	774	1,657	402	730
	10	S×A×R×HO×PJ	260	947	237	560
	11	S×A×M×HT×PT	1,262	2,386	617	961
	12	S×A×M×HT×HO	689	1,643	460	785
	13	S×A×M×HT×PJ	799	1,661	467	728
	14	S×A×M×PT×HO	988	2,001	526	896
	15	S×A×M×PT×PJ	1,280	2,323	588	904
	16	S×A×M×HO×PT	678	1,600	435	749
	17	S×A×HT×PT×HO	1,003	2,034	536	911
	18	S×A×HT×PT×PJ	1,256	2,274	576	893
	19	S×A×HT×HO×PJ	630	1,534	411	737
	20	S×A×PT×HO×PJ	774	1,657	402	730
	21	S×R×M×HT×PT	59	229	40	133
	22	S×R×M×HT×HO	15	87	12	68
	23	S×R×M×HT×PJ	36	151	33	99
	24	S×R×M×PT×HO	20	144	22	100
	25	S×R×M×PT×PJ	75	269	60	155
	26	S×R×M×HO×PJ	12	100	10	75
	27	S×R×HT×PT×HO	26	136	19	90
	28	S×R×HT×PT×PJ	81	254	48	131
	29	S×R×HT×HO×PJ	21	99	13	59
	30	S×R×PT×HO×PJ	24	110	17	70

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
5개 변수	31	S×M×HT×PT×HO	98	334	62	190
	32	S×M×HT×PT×PJ	213	497	112	241
	33	S×M×HT×HO×PJ	52	237	44	142
	34	S×M×PT×HO×PJ	75	269	60	155
	35	S×HT×PT×HO×PJ	81	254	48	131
	36	A×R×M×HT×PT	862	1,861	464	799
	37	A×R×M×HT×HO	369	1,143	307	620
	38	A×R×M×HT×PJ	515	1,251	347	613
	39	A×R×M×PT×HO	624	1,496	363	708
	40	A×R×M×PT×PJ	939	1,911	459	782
	41	A×R×M×HO×PJ	395	1,194	308	621
	42	A×R×HT×PT×HO	609	1,492	343	695
	43	A×R×HT×PT×PJ	922	1,804	429	739
	44	A×R×HT×HO×PJ	404	1,096	285	584
	45	A×R×PT×HO×PJ	474	1,196	272	551
	46	A×M×HT×PT×HO	1,153	2,184	566	906
	47	A×M×HT×PT×PJ	1,442	2,465	609	900
	48	A×M×HT×HO×PJ	836	1,755	464	757
	49	A×M×PT×HO×PJ	939	1,911	459	782
	50	A×HT×PT×HO×PJ	922	1,804	429	739
	51	R×M×HT×PT×HO	52	213	40	126
	52	R×M×HT×PT×PJ	131	353	77	174
	53	R×M×HT×HO×PJ	30	151	27	92
	54	R×M×PT×HO×PJ	40	176	34	102
	55	R×HT×PT×HO×PJ	53	166	31	89
	56	M×HT×PT×HO×PJ	131	353	77	174
6개 변수	1	S×A×R×M×HT×PT	1,262	2,386	617	961
	2	S×A×R×M×HT×HO	689	1,643	460	785
	3	S×A×R×M×HT×PJ	799	1,661	467	728
	4	S×A×R×M×PT×HO	988	2,001	526	896
	5	S×A×R×M×PT×PJ	1,280	2,323	588	904
	6	S×A×R×M×HO×PJ	678	1,600	435	749

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
6개 변수	7	S×A×R×HT×PT×HO	1,003	2,034	536	911
	8	S×A×R×HT×PT×PJ	1,256	2,274	576	893
	9	S×A×R×HT×HO×PJ	630	1,534	411	737
	10	S×A×R×PT×HO×PJ	774	1,657	402	730
	11	S×A×M×HT×PT×HO	1,563	2,693	576	893
	12	S×A×M×HT×PT×PJ	1,810	2,873	709	995
	13	S×A×M×HT×HO×PJ	1,196	2,194	570	863
	14	S×A×M×PT×HO×PJ	1,280	2,323	588	904
	15	S×A×HT×PT×HO×PJ	1,256	2,274	726	1,063
	16	S×R×M×HT×PT×HO	98	334	62	190
	17	S×R×M×HT×PT×PJ	213	497	112	241
	18	S×R×M×HT×HO×PJ	52	237	44	142
	19	S×R×M×PT×HO×PJ	75	269	60	155
	20	S×R×HT×PT×HO×PJ	81	254	48	131
	21	S×M×HT×PT×HO×PJ	213	497	112	241
	22	A×R×M×HT×PT×HO	1,153	2,184	566	906
	23	A×R×M×HT×PT×PJ	1,442	2,465	609	900
	24	A×R×M×HT×HO×PJ	836	1,755	464	757
	25	A×R×M×PT×HO×PJ	939	1,911	459	782
	26	A×R×HT×PT×HO×PJ	922	1,804	429	739
	27	A×M×HT×PT×HO×PJ	1,442	2,465	609	900
	28	R×M×HT×PT×HO×PJ	131	353	77	174
7개 변수	1	S×A×R×M×HT×PT×HO	1,563	2,693	726	1,063
	2	S×A×R×M×HT×PT×PJ	1,810	2,873	709	995
	3	S×A×R×M×HT×HO×PJ	1,196	2,194	570	863
	4	S×A×R×M×PT×HO×PJ	1,280	2,323	588	904
	5	S×A×R×HT×PT×HO×PJ	1,256	2,274	576	893
	6	S×A×M×HT×PT×HO×PJ	1,810	2,873	709	995
	7	S×R×M×HT×PT×HO×PJ	213	497	112	241
	8	A×R×M×HT×PT×HO×PJ	1,442	2,465	609	900
8개 변수	1	S×A×R×M×HT×PT×HO×PJ	1,810	2,873	709	995

주) S(성별), A(나이), R(가구주와의 관계), M(결혼상태), HT(가구구분), PT(점유형태), HO(주택소유), PJ(거처종류)

〈부표 C-3〉 Grouping한 2% 표본의 Uniqueness

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
1개 변수	1	S	-	2	-	2
	2	A	-	21	-	20
	3	R	-	4	-	4
	4	M	-	4	-	4
	5	HT	-	4	-	4
	6	PT	-	8	-	8
	7	HO	-	2	-	2
	8	PJ	-	8	-	4
2개 변수	1	S×A	1	42	2	40
	2	S×R	-	8	-	8
	3	S×M	-	8	-	8
	4	S×HT	-	8	-	8
	5	S×PT	-	16	-	14
	6	S×HO	-	4	-	4
	7	S×PJ	-	8	-	8
	8	A×R	3	70	5	66
	9	A×M	1	46	7	46
	10	A×HT	2	59	5	49
	11	A×PT	6	110	20	90
	12	A×HO	1	34	1	31
	13	A×PJ	2	64	5	58
	14	R×M	-	12	1	12
	15	R×HT	-	4	-	4
	16	R×PT	-	8	-	8
	17	R×HO	-	2	-	2
	18	R×PJ	-	4	-	4
	19	M×HT	-	12	-	12
	20	M×PT	-	24	2	23
	21	M×HO	-	6	-	6
	22	M×PJ	-	12	-	12
	23	HT×PT	2	32	3	26
	24	HT×HO	-	8	-	8

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
2개 변수	25	HT×PJ	-	15	1	16
	26	PT×HO	-	8	-	8
	27	PT×PJ	2	29	4	27
	28	HO×PJ	-	8	-	8
3개 변수	1	S×A×R	5	133	10	122
	2	S×A×M	3	90	10	84
	3	S×A×HT	9	109	17	89
	4	S×A×PT	25	204	37	145
	5	S×A×HO	3	67	4	61
	6	S×A×PJ	6	122	17	105
	7	S×R×M	-	24	1	23
	8	S×R×HT	-	8	-	8
	9	S×R×PT	-	16	-	14
	10	S×R×HO	-	4	-	4
	11	S×R×PJ	-	8	-	8
	12	S×M×HT	1	23	4	23
	13	S×M×PT	1	48	4	39
	14	S×M×HO	-	12	-	12
	15	S×M×PJ	-	24	-	24
	16	S×HT×PT	3	57	8	43
	17	S×HT×HO	-	16	2	16
	18	S×HT×PJ	1	30	3	29
	19	S×PT×HO	-	16	-	14
	20	S×PT×PJ	8	55	5	44
	21	S×HO×PJ	-	16	-	16
	22	A×R×M	5	134	-	12
	23	A×R×HT	2	59	5	49
	24	A×R×PT	6	110	20	90
	25	A×R×HO	1	34	1	31
	26	A×R×PJ	2	64	5	58
	27	A×M×HT	18	132	22	100
	28	A×M×PT	27	234	48	167

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
3개 변수	29	A×M×HO	3	82	8	77
	30	A×M×PJ	12	145	24	122
	31	A×HT×PT	66	289	47	171
	32	A×HT×HO	9	109	12	80
	33	A×HT×PJ	30	184	31	128
	34	A×PT×HO	6	110	20	90
	35	A×PT×PJ	52	285	58	186
	36	A×HO×PJ	11	120	18	102
	37	R×M×HT	-	12	-	12
	38	R×M×PT	-	24	2	23
	39	R×M×HO	-	6	-	6
	40	R×M×PJ	-	12	-	12
	41	R×HT×PT	2	32	3	26
	42	R×HT×HO	-	8	-	8
	43	R×HT×PJ	-	15	1	16
	44	R×PT×HO	-	8	-	8
	45	R×PT×PJ	2	29	4	27
	46	R×HO×PJ	-	8	-	8
	47	M×HT×PT	13	83	16	57
	48	M×HT×HO	2	24	2	21
	49	M×HT×PJ	5	44	9	41
	50	M×PT×HO	-	24	2	23
	51	M×PT×PJ	9	76	14	61
	52	M×HO×PJ	-	24	-	24
	53	HT×PT×HO	2	32	3	26
	54	HT×PT×PJ	17	89	20	65
	55	HT×HO×PJ	1	30	5	27
	56	PT×HO×PJ	2	29	4	27
4개 변수	1	S×A×R×M	17	253	33	206
	2	S×A×R×HT	9	109	17	89
	3	S×A×R×PT	25	204	37	145
	4	S×A×R×HO	3	67	4	61

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	5	S×A×R×PJ	6	122	17	105
	6	S×A×M×HT	34	224	41	162
	7	S×A×M×PT	75	389	75	250
	8	S×A×M×HO	11	154	25	134
	9	S×A×M×PJ	35	251	58	196
	10	S×A×PT×PT	128	460	80	265
	11	S×A×PT×HO	29	193	33	143
	12	S×A×PT×PJ	70	306	57	201
	13	S×A×PT×HO	25	204	37	145
	14	S×A×PT×PJ	106	473	97	277
	15	S×A×HO×PJ	23	221	40	172
	16	S×R×M×HT	1	23	4	23
	17	S×R×M×PT	1	48	4	39
	18	S×R×M×HO	-	12	-	12
	19	S×R×M×PJ	-	24	-	24
	20	S×R×HT×PT	3	57	8	43
	21	S×R×HT×HO	-	16	2	16
	22	S×R×HT×PJ	1	30	3	29
	23	S×R×PT×HO	-	16	-	14
	24	S×R×PT×PJ	8	55	5	44
	25	S×R×HO×PJ	-	16	-	16
	26	S×M×HT×PT	31	141	22	89
	27	S×M×HT×HO	5	45	6	38
	28	S×M×HT×PJ	8	79	18	69
	29	S×M×PT×HO	1	48	4	39
	30	S×M×PT×PJ	18	132	20	98
	31	S×M×HO×PJ	-	47	-	44
	32	S×HT×PT×HO	3	57	8	43
	33	S×HT×PT×PJ	25	148	28	98
	34	S×HT×HO×PJ	4	56	9	47
	35	S×PT×HO×PJ	8	55	5	44
	36	A×R×M×HT	18	132	22	100

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	37	A×R×M×PT	27	234	48	167
	38	A×R×M×HO	3	82	8	77
	39	A×R×M×PJ	12	145	24	122
	40	A×R×HT×PT	66	289	47	171
	41	A×R×HT×HO	9	109	12	80
	42	A×R×HT×PJ	30	184	31	128
	43	A×R×PT×HO	6	110	20	90
	44	A×R×PT×PJ	52	285	58	186
	45	A×R×HO×PJ	11	120	18	102
	46	A×M×HT×PT	141	495	100	282
	47	A×M×HT×HO	42	226	40	159
	48	A×M×HT×PJ	74	337	82	217
	49	A×M×PT×HO	27	234	48	167
	50	A×M×PT×PJ	129	532	111	298
	51	A×M×HO×PJ	35	262	47	191
	52	A×HT×PT×HO	66	289	47	171
	53	A×HT×PT×PJ	197	600	113	296
	54	A×HT×HO×PJ	67	305	60	193
	55	A×PT×HO×PJ	52	285	58	186
	56	R×M×HT×PT	13	83	16	57
	57	R×M×HT×HO	2	24	2	21
	58	R×M×HT×PJ	5	44	9	41
	59	R×M×PT×HO	-	24	2	23
	60	R×M×PT×PJ	9	76	14	61
	61	R×M×HO×PJ	-	24	-	24
	62	R×HT×PT×HO	2	32	3	26
	63	R×HT×PT×PJ	17	89	20	65
	64	R×HT×HO×PJ	1	30	5	27
	65	R×PT×HO×PJ	2	29	14	61
	66	M×HT×PT×HO	13	83	16	57
	67	M×HT×PT×PJ	40	191	43	122
	68	M×HT×HO×PJ	7	78	17	64

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
4개 변수	69	M×PT×HO×PJ	2	29	14	61
	70	HT×PT×HO×PJ	17	89	20	65
5개 변수	1	S×A×R×M×HT	34	224	41	162
	2	S×A×R×M×PT	75	389	75	250
	3	S×A×R×M×HO	11	154	25	134
	4	S×A×R×M×PJ	35	251	58	196
	5	S×A×R×HT×PT	128	460	80	265
	6	S×A×R×HT×HO	29	193	33	143
	7	S×A×R×HT×PJ	70	306	57	201
	8	S×A×R×PT×HO	25	204	37	145
	9	S×A×R×PT×PJ	106	473	97	277
	10	S×A×R×HO×PJ	23	221	40	172
	11	S×A×M×HT×PT	255	742	170	405
	12	S×A×M×HT×HO	85	373	83	252
	13	S×A×M×HT×PJ	165	525	139	312
	14	S×A×M×PT×HO	75	389	75	250
	15	S×A×M×PT×PJ	252	775	174	404
	16	S×A×M×HO×PT	86	428	101	290
	17	S×A×HT×PT×HO	128	460	80	265
	18	S×A×HT×PT×PJ	335	855	182	415
	19	S×A×HT×HO×PJ	131	485	99	288
	20	S×A×PT×HO×PJ	106	473	97	277
	21	S×R×M×HT×PT	31	141	22	89
	22	S×R×M×HT×HO	5	45	6	38
	23	S×R×M×HT×PJ	8	79	18	69
	24	S×R×M×PT×HO	1	48	4	39
	25	S×R×M×PT×PJ	18	132	20	98
	26	S×R×M×HO×PJ	-	47	-	44
	27	S×R×HT×PT×HO	3	57	8	43
	28	S×R×HT×PT×PJ	25	148	28	98
	29	S×R×HT×HO×PJ	4	56	9	47
	30	S×R×PT×HO×PJ	8	55	5	44

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
5개 변수	31	S×M×HT×PT×HO	31	141	22	89
	32	S×M×HT×PT×PJ	85	306	65	178
	33	S×M×HT×HO×PJ	15	138	27	104
	34	S×M×PT×HO×PJ	18	132	20	98
	35	S×HT×PT×HO×PJ	25	148	28	98
	36	A×R×M×HT×PT	141	495	100	282
	37	A×R×M×HT×HO	42	226	40	159
	38	A×R×M×HT×PJ	74	337	82	217
	39	A×R×M×PT×HO	27	234	48	167
	40	A×R×M×PT×PJ	129	532	111	298
	41	A×R×M×HO×PJ	35	262	47	191
	42	A×R×HT×PT×HO	66	289	47	171
	43	A×R×HT×PT×PJ	197	600	113	296
	44	A×R×HT×HO×PJ	67	305	60	193
	45	A×R×PT×HO×PJ	52	285	58	186
	46	A×M×HT×PT×HO	141	495	100	282
	47	A×M×HT×PT×PJ	363	903	200	419
	48	A×M×HT×HO×PJ	161	536	123	303
	49	A×M×PT×HO×PJ	129	532	111	298
	50	A×HT×PT×HO×PJ	197	600	113	296
	51	R×M×HT×PT×HO	13	83	16	57
	52	R×M×HT×PT×PJ	40	191	43	122
	53	R×M×HT×HO×PJ	7	78	17	64
	54	R×M×PT×HO×PJ	9	76	14	61
	55	R×HT×PT×HO×PJ	17	89	20	65
	56	M×HT×PT×HO×PJ	40	191	43	122
6개 변수	1	S×A×R×M×HT×PT	255	742	170	405
	2	S×A×R×M×HT×HO	85	373	83	252
	3	S×A×R×M×HT×PJ	165	525	139	312
	4	S×A×R×M×PT×HO	75	389	75	250
	5	S×A×R×M×PT×PJ	252	775	174	404
	6	S×A×R×M×HO×PJ	86	428	101	290

Key 변수의 조합			충청남도		천안시	
			1인 경우	유효조합 수	1인 경우	유효조합 수
6개 변수	7	S×A×R×HT×PT×HO	128	460	80	265
	8	S×A×R×HT×PT×PJ	335	855	182	415
	9	S×A×R×HT×HO×PJ	131	485	99	288
	10	S×A×R×PT×HO×PJ	106	473	97	277
	11	S×A×M×HT×PT×HO	255	742	170	405
	12	S×A×M×HT×PT×PJ	553	1,197	298	534
	13	S×A×M×HT×HO×PJ	286	779	208	415
	14	S×A×M×PT×HO×PJ	252	775	174	404
	15	S×A×HT×PT×HO×PJ	335	855	182	415
	16	S×R×M×HT×PT×HO	31	141	22	89
	17	S×R×M×HT×PT×PJ	85	306	65	178
	18	S×R×M×HT×HO×PJ	15	138	27	104
	19	S×R×M×PT×HO×PJ	18	132	20	98
	20	S×R×HT×PT×HO×PJ	25	148	28	98
	21	S×M×HT×PT×HO×PJ	85	306	65	178
	22	A×R×M×HT×PT×HO	141	495	100	282
	23	A×R×M×HT×PT×PJ	363	903	200	419
	24	A×R×M×HT×HO×PJ	161	536	123	303
	25	A×R×M×PT×HO×PJ	129	532	111	298
	26	A×R×HT×PT×HO×PJ	197	600	113	296
	27	A×M×HT×PT×HO×PJ	363	903	200	419
	28	R×M×HT×PT×HO×PJ	40	191	43	122
7개 변수	1	S×A×R×M×HT×PT×HO	255	742	170	405
	2	S×A×R×M×HT×PT×PJ	553	1,197	298	534
	3	S×A×R×M×HT×HO×PJ	287	779	208	415
	4	S×A×R×M×PT×HO×PJ	252	775	174	404
	5	S×A×R×HT×PT×HO×PJ	335	855	182	415
	6	S×A×M×HT×PT×HO×PJ	553	1,197	298	534
	7	S×R×M×HT×PT×HO×PJ	85	306	65	178
	8	A×R×M×HT×PT×HO×PJ	363	903	200	419
8개 변수	1	S×A×R×M×HT×PT×HO×PJ	553	1,197	298	534

주) S(성별), A(나이), R(가구주와의 관계), M(결혼상태), HT(가구구분), PT(점유형태), HO(주택소유), PJ(거처종류)

2부

매스킹 기법의 적용과 평가

3. 2005 인구주택총조사자료의 개인정보 노출제한방법

정동명 · 정미옥

4. 가계조사 마이크로데이터의 비밀보호

정동명 · 정남수 · 한승훈

5. 승법잡음모형을 이용한 가계조사 자료의 비밀보호

정동명 · 김경미

6. 가계금융 · 복지조사 마이크로데이터 제공을 위한 매스킹 방안

박민정 · 권순필 · 심규호

7. 가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가

박민정

제3장

2005 인구주택총조사자료의 개인정보 노출제한방법

정동명 · 정미옥

제1절 서 론

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

통계작성기관이 마이크로자료 파일을 제공할 때는 외부 이용자가 그 파일로부터 특정 개인을 연결시킬 수 없다는 것을 확실하게 보장해야만 한다. 마이크로자료에서 개인정보의 노출위험을 제거하는 가장 안전한 방법은 아예 마이크로자료를 제공하지 않는 것이지만, 이는 자료수집의 근본취지에 위배되는 행위이므로 통계작성기관은 이러한 노출위험을 최소화하면서 최대한 유용한 정보가 들어 있는 자료를 제공하기 위해 노력하고 있다. 미국 · 호주 등 외국의 통계작성기관에서는 이미 오래 전부터 노출의

✍ 이 연구는 2007년 상반기 통계개발원 연구보고서에 수록된 자료입니다.

위험성을 깊이 인식하고 개인정보의 비밀보호를 위해 다양한 방법들을 적용하여 마이크로자료를 제공하고 있다. 미국 센서스국에서는 2000년 인구센서스 결과에 다양한 기법들을 적용시킨 후 마이크로자료를 제공하고 있으며, 호주의 경우에도 인구주택총조사 마이크로자료의 비밀보호를 위해 이름이나 주소와 같이 직접적으로 개인을 식별 가능하게 하는 변수를 삭제하고, 상세한 수준의 범주를 제거하는 방법을 사용하였다. 이 밖에 네덜란드와 스웨덴, 영국 등에서도 노출위험의 심각성과 비밀보호방법의 필요성 등에 대해 관심을 가지고 많은 연구가 진행되었으며, Dalenius & Reiss(1982), Kim(1986), Bethelehem et al.(1990), Marsh et al.(1991), Fuller(1993) 등 여러 학자들도 노출의 위험성과 비밀보호에 대한 여러 방법을 제시하였다.

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

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

제2절 노출과 노출위험

1. 노출과 유일성

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

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

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

2. 노출위험의 확률모형

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

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

만약 금융기관이나 이웃주민 등과 같은 외부인이 자신들이 직접 작성한 파일(S_2)에 관심 있는 어떤 사람(A)이 포함되어 있다는 것을 모르고 있다면, 특정인의 노출위험 $DR(A)$은 다음과 같이 정의할 수 있다.

$$DR(A) = \Pr[(A \in S_1) \cap (A \in S_2) \cap (A \in U_P)]$$

그러나 외부인이 자신들의 파일에 관심 있는 특정인(A)이 포함되어 있다는 것을 이미 알고 있다고 한다면 $\Pr(A \in S_2) = 1$이 될 것이며, 이때 노출위험은 다음과 같이 된다.

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

제3절 인구주택총조사 자료의 비밀보호

1. 인구주택총조사 자료

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

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

2005 인구주택총조사 결과에 의하면, <표 3-1>에 나타난 바와 같이 총 조사구 규모는 265,298개이고 인구는 45,772,054명, 가구는 15,895,481가구, 주택은 12,693,578호가 있는 것으로 나타났다. 10% 표본결과의 경우 조사구는 26,713개이고 인구는 4,455,527명, 가구는 1,582,681가구, 주택은 1,295,389호인 것으로 각각 나타났다. 이에 대한 자세한 결과는 <부록 1>을 참조하기 바란다.

〈표 3-1〉 2005 인구주택총조사 결과

(단위: 개, 명, 가구, 호)

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

2. 마이크로자료 파일의 작성

가. Key 변수의 선정

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

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

1) 인구에 관한 사항

가구원에 관한 사항에는 개인에 대한 민감한 정보들이 많이 들어 있어 다른 분야에 비해 key 변수를 더 많이 고려하였으며, 외부에서 식별가능성이 높다고 판단되는 8개 항목(성별, 나이, 가구주와의 관계, 교육정도, 혼인상태, 활동제약, 산업, 직업)을 key 변수로 선정하였다. 선정된 key 변수의 빈도분석결과를 살펴보면, 성별의 경우 남녀가 비슷한 비율을 차지하면서 지역별로도 거의 차이가 없었다. 나이는 만 100세 이상의 고령자가 총 24명으로 각 시도에 적게는 1명에서 많게는 6명 정도가 분포하고 있으며, 특·광역시보다 도지역에서 평균연령이 다소 높게 나타났다. 가구주와의 관계 항목에서는 가구주, 자녀, 가구주의 배우자의 비중은 매우 높은 반면, 그 이외의 범주에 대해서는 상대적으로 낮게 나타났다. 교육정도는 고등학교 졸업이 약 19~31.9% 정도로 지역별로 다소차이를 보이고 있지만 가장 큰 비중을 차지하고 있었다. 다음으로 활동제약 항목은 육체적·정신적 제약과 활동제약 부분의 세부항목이 있는데, 대부분의 지역에서 제약 없음이 압도적으로 나타났다. 하지만 활동제약이 있는 인구가 매우 적더라도 활동제약이 있는 경우 외부에서 쉽게 식별 가능하기 때문에 key 변수로 선정하였다. 취업인구에 대해 직장의 산업과 직업을 묻는 항목에서는 표준산업분류 및 표준직업분류를 기준으로 대분류에 대하여 분석하였으며, 산업항목의 경우 지역별 산업의 구성비에는 다소 차이가 있었으나 대부분의 지역에서 도매 및 소매업과 제조업, 농업이 큰 비중을 차지하고 있었다. 직업항목의 경우 군인이 강원도 1.89%를 제외하고 대부분의 지역에서 약 0.5% 이하로 매우 적게 분포하였다. 마지막으로 혼인상태는 전국적으로 약 26.6%, 사별과 이혼이 각각 8.5%, 3%인 것으로 나타났다.

2) 가구에 관한 사항

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

3) 주택에 관한 사항

주택에 대한 정보는 개인을 식별해 내는 데 크게 영향을 미칠 만한 변수는 많지 않기 때문에 거처종류 항목만을 key 변수로 선정하였다. 거처종류 항목의 단독주택 비율이 가장 낮은 지역은 울산으로 약 29.8%였고, 전남은 73.2%로 가장 높게 나타났다. 또한 아파트는 전국적으로 약40%의 비율을 차지하고 있었다.

나. 표본의 추출

현재 통계청에서 제공하고 있는 인구주택총조사의 마이크로자료는 전체 규모의 2%에 해당하는 표본조사 결과자료이다. 이는 전체 규모로는 2%이지만 표본조사 결과만으로 보면 약 20%에 해당된다. 따라서 전체 10% 표본조사구 내에 있는 모든 가구의 명부를 추출틀(sample frame)로 사용하여 표본가구를 추출하였다. 이때 추출방법은 개인의 식별 가능성의 최소화에 적합한 계통추출법(systematic sampling)을 적용하였으며, 이렇게 추출된 결과가 <표 3-2>에 주어져 있다.

〈표 3-2〉 2% 표본현황

(단위: 개, 명, 가구, 호)

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

다. 노출위험

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

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

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

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

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

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

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

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

여기서 $\Pr(A \in S_1) = 0.2$ 인데, 이는 제공하는 마이크로자료의 규모가 2%이므로 10% 표본조사결과에서 20%를 표본으로 추출하였기 때문이다. 위의 결과는, 외부인이 가지고

있는 파일에 어떤 사람이 포함되어 있다는 것을 알고 있다고 가정하고 2%를 마이크로자료 파일로 제공할 경우, 100명 중 약 8명 정도가 노출될 가능성이 있다는 것을 의미한다.

〈표 3-3〉 노출제한방법 사용 전 유일성

(단위: 명, %)

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

라. 노출위험의 제한방법

위에서 2% 표본의 유일성에 대해 살펴본 결과, 총 892,024명 중 약51.4%인 458,160명이 유일한 것으로 나타났다. 만약 추출된 2%의 표본에 노출제한방법을 적용하면 유일성은 더 감소하게 될 것이며, 이는 결국 개인의 식별과 노출에 대한 위험이 훨씬 더 줄어들게 됨을 의미한다. 노출을 제한하는 방법은 자료의 종류와 형태에 따라 다양한 방법이 있는데, 본 연구에서 자료의 제공범위를 제한하는 방법과 자료의 정보를 축소하는 방법을 이용하였다.

1) 제공범위의 제한

마이크로자료의 제공범위를 제한하기 위해 조사구의 경우 아파트조사구와 보통조사구, 섬조사구만을 대상으로 하였다. 가구의 경우 가족으로 이루어진 가구와 가족과 가족 이외의 사람이 함께 사는 가구, 1인가구, 가족이 아닌 남남끼리 함께 사는 5인 이하의 가구만

대상으로 하였으며, 지역자료도 시·도 단위까지만 제공하기로 하였다. 그리고 남북이산가족, 임차료, 대지면적 등 3개 항목은 조사의 특성상 공표대상에서 제외하기로 하였다.

2) 자료의 축소

인구주택총조사의 특성상 이산형 변수가 대부분이고 연속형 변수인 나이, 연건평 등도 구간으로 변환이 가능하기 때문에 본 연구에서는 주로 이산형 변수에 효과적인 그룹화(grouping)와 코딩(top-coding, bottomcoding), 하위 세부항목의 통합 및 제거 등의 방법을 이용하여 자료의 정보량을 다소 축소하였다. 각 분야별 key 변수에 대해 살펴보면, 인구에 관한 항목의 경우 성별은 남녀범주를 그대로 사용하였고, 나이는 각 세별로 되어 있던 코드를 0~84세까지는 그대로 두고, 85세 이상은 topcoding하여 하나의 범주로 처리를 하였다. 가구주와의 관계 항목은 14개로 구분되어 있던 범를 범주간 관련성(부모세대 혹은 자녀세대, 혈연관계 혹은 비혈연관계)과 각 범주의 빈도수를 고려하여 8개 범주로 그룹화하였다. 교육정도는 학력에 대한 세부항목에서 4년제 미만과 4년제 이상으로 구분되어 있던 대학범주를 대학교라는 하나의 범주로, 석사과정과 박사과정을 대학원범주로 통합하였다. 또한 졸업, 재학, 수료, 휴학, 중퇴라는 범주로 세분화되어 있던 교육상태 항목을 졸업과 졸업이 아닌 상태로 축소하였다. 활동제약 항목의 육체적·정신적 제약 부분은민감하면서 그 빈도가 매우 낮은 치매와 중풍을 각각 정신적 제약과 육체적 제약으로 묶어 주었다. 다음으로 종사산업은 표준산업분류 대분류 기준으로 20개의 대분류 산업을 16개로 축소하였고, 직업 항목에서는 표준직업분류 대분류 기준 10개의 범주 중에 군인은 그 빈도가 매우 낮아 직업미상과 묶어 주었다. 마지막으로 혼인상태는 기존의 범주를 그대로 사용하였다. 가구에 관한 사항에서 가구구분은 가족과 가족 이외의 사람이 함께 사는 가구와 가족이 아닌 남남끼리 함께 사는 5인 이하의 가구를 기타의 범주로 묶어 기존 4개의 범주를 3개로 조정하였다. 점유형태 항목에서는 매월 주거비용을 내는지의 개념을 적용시켜 보증금 있는 월세와 보증금 없는 월세, 사글세를 월세라는 하나의 범주로 그룹화하였다. 주택소유여부 항목에서 주인가구여부 세부항목은 현재 주택을 소유하고 있는지의 개념을 적용하여 주인가구와 주인 아닌 가구로 범주를 축소하였고, 주택소유여부 세부항목은 기존 범주를 그대로 사용하였다. 마지막으로 주택에 관한 사항에서 거처종류 항목은 기존의 10개의 범주를 5개(단독주택, 아파트, 연립 및 다세대 주택, 비거주용 건물내 주택, 기타)로 그룹화하였다. 이에 대한 결과가 <표 3-4>에 나타나 있으며, 각 항목에 대한 자세한 내용은 <부록 3>을 참고하기 바란다.

한편, 위에서 언급한 노출제한방법을 적용한 후 모집단과 2% 표본의 유일성을 파악한 결과, <표 3-5>에 나타난 바와 같이 모집단의 경우 총4,455,527명 중 약 24.4%인 1,089,142명이 유일한 것으로 나타났으며, 2% 표본의 경우 총 892,024명 중 38.2%인 341,168명이 유일한 것으로 나타났다. 이는 노출제한방법을 적용하기 전보다 상당히

줄어들었음을 보여주고 있다. 이를 바탕으로 외부인이 직접 작성한 파일에 어떤 특정인(A)이 포함되어 있다는 것을 이미 알고 있다는 가정 하에서 노출위험을 계산하면 약 4.9%가 되어, 노출제한방법을 적용하기 전의 노출위험(8.1%)보다 약 40% 정도 감소한 것을 알 수 있다.

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

〈표 3-4〉 노출제한방법

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

〈표 3-5〉 노출제한방법 사용 후 유일성

(단위: 명, %)

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

제4절 결 론

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

한편, 본 연구에서는 마이크로자료 파일을 작성할 때 인구주택총조사의 질문항목이 대부분 이산형 형태로 구성된 것을 감안하여 이산형변수의 노출제한방법을 주로 적용하였으나, 나이 등 일부 항목들은 연속형 변수이므로 자료를 범주화하여 제공하는 것이 더 바람직할 것이다. 또한 10% 표본조사결과를 모집단으로 하여 자료 분석 등을 하였으나, 만약 전수조사결과를 모집단으로 하고 유일성 등을 파악한다면 자료의 수가 많을수록 유일성이 감소하는 특성으로 인해 본 연구에서 계산한 것보다 훨씬 더 노출위험이 줄어들 것이다. 그러나 활동제약이나 산업 등의 key 변수는 10% 표본조사결과에만 있고 전수조사결과에는 없기 때문에 전수조사결과에서의 노출위험을 계산하지는 못하였다. 향후 이에 대한 추가검토와 함께 연속형 자료의 노출제한방법에 대한 연구도 지속할 예정이다. 아울러 본 연구의 결과를 바탕으로 통계청에서 생산하고 있는 각종 통계들 중 마이크로자료의 제공을 많이 요구하는 통계조사를 대상으로 다양한 노출제한방법을 적용하는 연구가 활발히 진행되길 기대해 본다.

참고문헌

박원환 · 황조연(2004), 「통계자료의 특성과 비밀보호방법에 관한 연구」, 통계연구결과 보고서, 통계청.

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

통계청(2005), 「2005 인구주택총조사 조사지침서」.

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

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

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

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

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

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

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

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

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

Marsh, C., C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D.

Lievesley, and N. Walford(1991), "The Case for Samples of Anonymized Records from the 1991 Census", *Journal of the Royal Statistical Society A*, 154, pp.305-340.

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

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

<부 록>

1. 2005 인구주택총조사 결과

〈부표 3-1〉 전수조사 결과

(단위: 개, 명, 가구, 호)

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

〈부표 3-2〉 10% 표본조사 결과

(단위: 개, 명, 가구, 호)

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

2. 변수별 노출제한방법

〈부표 3-3〉 인구에 관한 사항

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

〈부표 3-3〉 인구에 관한 사항 (계속)

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

〈부표 3-4〉 가구에 관한 사항

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

〈부표 3-5〉 주택에 관한 사항

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

제4장

가계조사 마이크로데이터의 비밀보호

정동명 · 정남수 · 한승훈

제1절 서 론

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

그러나 갈수록 급변하는 경제 · 사회현상에 대한 심층적 자료분석에는 매크로데이터의 한계가 있기 때문에 여러 분야의 통계이용자들은 마이크로데이터의 제공에 대한 요구가 커지고 있다. 따라서 최근 들어 통계청을 비롯한 여러 통계작성기관들은 그들이 작성하고 있는 통계에 대한 마이크로데이터를 이용자들에게 제공하는 것을 법적으로 제도화하고 있는 실정이다. 하지만, 마이크로데이터를 그대로 제공할 경우 응답한 개인의 특성을 나타내는 정보들이 그대로 노출되어 심각한 개인 사생활의 침해가 발생할 가능성이 높다. 그러므로 마이크로데이터를 제공 할 경우에는 사전에 미리 응답자의 개인정보가 노출되지 않도록 통계적으로 보호방법을 적용하는 등 각별한 노력을 기울여야 한다.

외국의 통계작성기관에서는 1970년대부터 마이크로데이터의 제공에 대한 요구와

✍ 이 연구는 2007년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

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

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

본 연구에서는 개인정보의 노출위험성과 비밀보호방법들을 살펴보고, 몇 가지 방법을 실제 자료에 적용하여 비밀이 보호된 마이크로데이터를 작성하고자 한다. 제2절에서는 개인정보의 노출과 노출의 위험성, 그리고 비밀을 보호하는 다양한 방법을 소개한다. 제3절에서는 다양한 비밀보호방법들 중 연속형자료의 비밀보호방법을 가계조사 결과에 적용하여 비밀보호된 마이크로데이터를 작성하는 과정을 설명하고, 이에 대한 결과를 분석한다. 마지막으로 제4절에서는 본 연구의 결론과 향후 고려할 사항들을 간략히 언급하고자 한다.

제2절 노출과 유일성

1. 노출과 노출위험

가. 노출과 유일성

통계작성기관에서 응답자로부터 수집·정리된 자료를 다양한 형태의 통계정보로

제공할 경우 이를 통해서 응답자의 특성이 파악되는 것을 노출(disclosure)이라 하는데, 만약 개인의 민감한 정보가 노출된다면 이는 개인에 대한 식별(identification)이 가능하게 된다는 것을 의미한다. 이러한 노출은 외부이용자(intruder)가 사전에 가지고 있는 유용한 정보의 성격과 양에 따라 좌우되므로, 그들이 가지고 있는 사전정보와 통계작성기관에서 제공하는 마이크로데이터의 정보가 서로 일치하지 않도록 한다면 노출이 발생될 가능성은 아주 낮게 될 것이다.

유일성(uniqueness)이란 전체 자료파일에서 조사단위의 특성이 유일하게 존재하는 것을 말하며, 어떤 조사단위가 식별될 가능성을 나타내는 측도로 사용된다. 가령, 100명으로 구성된 모집단에서 나이가 100세인 사람이 단 1명 있다면 그 사람은 모집단에서 유일하다고 하는데, 이렇게 유일한 사람은 다른 사람들에 비해 자료에서 식별될 가능성이 매우 높다. 주어진 자료에서 유일성은 하나의 변수만으로도 파악할 수 있고 여러 개의 변수들을 조합하여 파악할 수도 있으며, 고려되는 변수가 많아질수록 유일성은 점점 더 커지게 된다. 가령, 100명 중 나이가 60세인 사람이 3명 있다고 하더라도 직업이라는 변수를 포함하여 회사원이 면서 나이가 60세인 사람은 단 1명일 수도 있다.

나. 노출위험의 확률모형

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

〈 노출의 발생조건 〉

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

노출의 발생조건하에서 노출위험(disclosure risk)은 통계적 확률모형으로 표현할 수 있다. 먼저, A를 관심의 대상인 사람, S_1을 통계작성기관의 마이크로데이터로 구성된 파일 1, S_2를 외부인(intruder)에 의해 구성된 파일 2, U_P를 모집단의 유일성집단, 그리고 U_S를 표본의 유일성집단이라고 각각 정의하자. 만약 금융기관이나 이웃주민 등과 같은 외부인이 자신들이 직접 작성한 파일(S_2)에 관심 있는 어떤 사람 A가 포함되어 있다는 것을 모르고 있다면, 특정사람의 노출위험 $DR(A)$은 다음과 같이 정의할 수 있다.

$$DR(A) = \Pr[(A \in S_1) \cap (A \in S_2) \cap (A \in U_P)]$$

그러나 외부인이 자신들의 파일에 관심있는 특정사람 A 가 포함되어 있다는 것을 이미 알고 있다고 한다면 $\Pr(A \in S_2) = 1$ 이 될 것이며, 이때 노출위험은 다음과 같이 된다.

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

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

2. 비밀보호방법

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

가. 자료교환

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

나. 코딩접근법

이 방법은 원자료에 적절한 값(noise)을 추가하여 다른 자료값으로 변형하는 방법이다. 가령, X_i 와 Y_i, 그리고 e_i 를 0 또는 1을 갖는 이산형 변수라고 가정하자. 만약 X_i 를

원자료, e_i를 X_i와 서로 독립이고 동일한 분포를 갖는 값, 그리고 Y_i를 변형된 자료라고 한다면 $Y_i = (X_i + e_i)_{\mathrm{mod}2}$이 되고, X_i와 e_i가 서로 같은 값이면 0, 아니면 1의 값을 각각 가진다.

다. 그룹화

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

라. 반올림

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

마. 구간그룹화

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

바. 가법 및 승법잡음

가법잡음은 원자료에 적절한 잡음(noise)을 더해 주어 다른 자료값으로 변형하는방법으로서 X_i를 원자료, e_i를 잡음, 그리고 Y_i를 변형된 자료라 하면 Y_i는 다음과 같이 나타낼 수 있다.

$$Y_i = X_i + e_i$$

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

$$Y_i = X_i\ e_i$$

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

제3절 비밀보호된 마이크로자료 작성

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

1. 분석대상

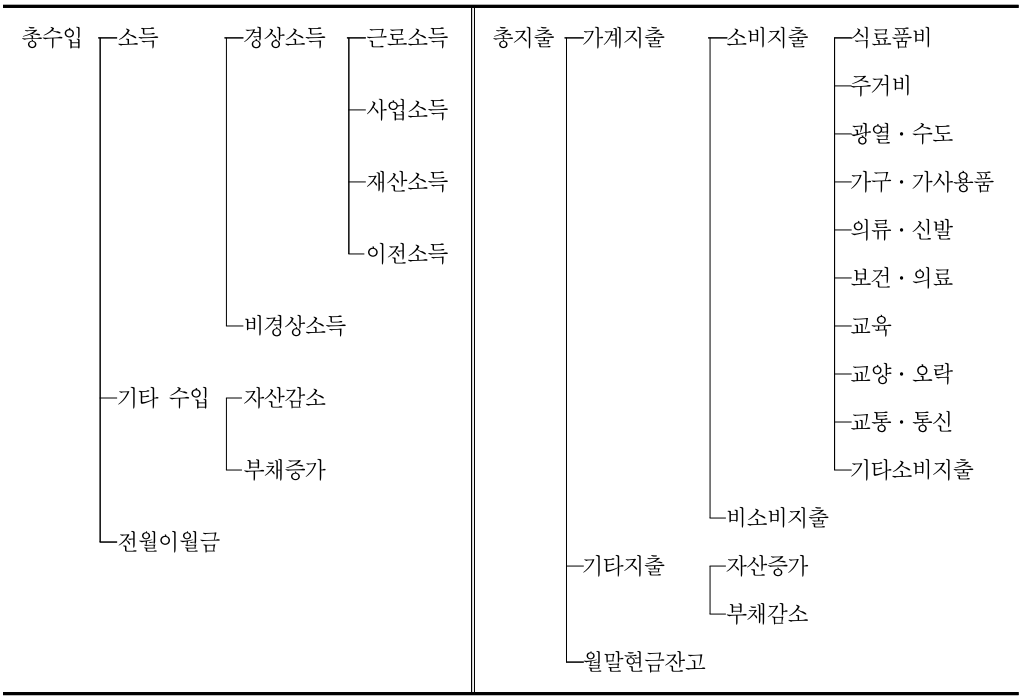
본 연구에서는 최근에 작성된 2006년 가계조사결과를 대상으로 비밀 보호된 마이크로데이터를 작성코자 한다. 통계청에서 실시하는 가계조사는 우리나라 가구에 대한 가계수지의 실태를 파악하여 국민의 소득과 소비수준 변화의 측정 및 분석 등에 필요한 자료를 제공하기 위해 매월 전국의 표본가구를 대상으로 한다. 2005 인구주택총조사의 10% 표본조사구를 모집단으로 하여 전국을 25개로 층화한 후, 확률비례계통추출법으로 표본

조사구를 우선 추출하고, 조사구당 10가구 추출을 기준으로 부적격 가구를 제외한 약 8,000가구를 최종 표본가구로 선정하였다. 이렇게 선정된 전국의 표본가구를 대상으로 가구의 수입과 지출관련 항목들을 응답가구에서 직접 가계부를 매일 작성하도록 하고 있으며, 가계부의 결과는 매월단위로 집계·정리한 후 분기별 및 연도별로 공표하고 있다.

가계부를 통해 조사되는 항목은 <표 4-1>에 주어져 있는 바와 같이 가구사항을 제외하면 총수입 항목과 총지출 항목으로 구분되는데, 총수입은 다시 소득과 기타수입, 그리고 전월이월금으로 나누어지고, 총지출은 가계지출과 기타지출, 그리고 월말 현금잔고로 구분된다. 조사항목들을 세분화하여 분류한 후 이를 다시 품목별로 정리하면 수입관련 품목이 47개이고 지출관련 품목이 510개로서 가계조사에서 조사되는 품목은 모두 557개가 된다. 이에 대한 자세한 내용은 가계조사 지침서(2006)를 참고하기 바란다.

한편, 2006년 가계조사의 조사대상은 <표 4-2>에 나타난 바와 같이 동일한 가구의 중복을 제외하면 총 12,458가구이며, 이 중 서울이 1,638가구이고 서울 이외 지역이 10,820가구이다. 이 결과는 분기별 및 연별로 조사결과가 집계되어 공표되는 가구 수와 다른데, 이는 1년간 매월 조사에 응답한 동일가구는 12가구가 아닌 1가구로 처리하여 중복성을 제외하였기 때문이다.

〈표 4-1〉 가계수지 항목분류



〈표 4-2〉 2006년도 가계조사의 표본규모

(단위: 가구)

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

2. 식별정보의 비밀보호

가. key변수의 선정

가계조사의 조사항목에는 월소득이나 지출규모 등 개인 및 가구의 특성에 관한 민감한 정보들이 많이 있기 때문에 key변수의 선정에 신중을 기하였다. key변수의 선정을 위해 외부기관에서 보유하고 있는 자료들 중 가계조사의 조사항목과 중복되어 식별될 가능성이 높다고 판단되는 항목을 우선 선정한 후, 빈도분석(frequency analysis) 등을 실시하여 각 항목의 빈도수와 분포형태 등 보다 자세한 자료의 특성을 지역별 및 분기별로 파악하였다. 이는 지역별 및 분기별 분포를 비교해 보고, 항목의 각 범주별로 최소 응답수를 확인함으로써 특정 지역이나 항목에 있어서 특징적으로 나타날 수 있는 특성들을 살펴보기 위함이었다. 이러한 면밀한 검토 작업을 거친 후 가구유형, 거처구분, 가구원수, 배우자 유무, 가구주성별, 가구주연령, 가구주직업 등 모두 7개 항목을 key변수 로 최종 선정하였다.

나. key변수의 노출위험

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

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

2005 인구주택총조사의 10% 표본조사결과를 모집단으로 하고 2006년 가계조사의 마이크로데이터를 표본으로 한 후, 7개 key변수의 각 조합별로 빈도수가 1이 되는

가구를 지역별 및 분기별로 파악한 결과가 <표 4-3>에 주어져 있다. 결과해석의 편의를 위해 여기서는 연간자료인 경우에 대해서만 설명하기로 한다. 여러 조합들 중 7개 변수를 모두 고려한 조합이 유일성의 발생이 최대가 되는 경우인데, 모집단에서는 총 21,728가구가 있으며 표본에서는 4,267가구가 있는 것으로 각각 나타났다. 따라서 어떤 가구(A)가 모집단에서 유일하게 될 확률은 약 1.52%가 되고, 표본에서 유일하게 될 확률은 약 34.25%가 된다.

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

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

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

〈표 4-3〉 key변수의 유일성(그룹화 前)

(단위: 가구, %)

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

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

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

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

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

다. key변수의 그룹화

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

key변수로 선정된 7개 변수의 경우 나이를 제외하고는 모두 이산형 변수이고 나이 항목도 구간으로 변환이 가능하다. 따라서 key변수의 비밀보호를 위해 주로 이산형 변수에 효과적인 그룹화와 상한그룹화(topcoding)등의 방법을 이용하였는데, 이렇게 할 경우 자료의 정보량은 다소 축소된다. 7개 key변수 중 가구유형(4개), 거처구분(4개), 배우자유무(3개), 가구주성별(2개) 등 4개 변수는 원래 범주를 축소하지 않고 그대로 사용하기로 하였으며, 나머지 3개 변수는 범주의 특성을 고려하여 적정규모의 범주로 축소하였다. 먼저 가구원수의 경우 5인 이하는 그대로 두고 6인 이상은 상한그룹으로 하나의 범주로 하여 총 6개 범주로 구분하였다. 가구주연령 변수의 경우, 각 세별로 되어 있던 것을 15~79세까지는 5세 간격(13개)으로 그룹화하고 80세 이상은 상한그룹으로 1개 범주로 처리하여 총 14개 범주로 통합하였다. 끝으로 가구주직업 변수의 경우 직업 대분류에 의한 12개 범주로 구분된 것을 그대로 유지하되 범주의 빈도수를 고려하여

직업군인은 기타에 분류하여 총 11개 범주로 구분하였다. 이렇게 선정된 key변수에 대해 그룹화한 결과를 정리하여 <표 4-4>에 나타내었다.

〈표 4-4〉 key변수의 그룹화

(단위: 개)

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

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

이러한 결과를 바탕으로 외부인이 직접 작성한 파일에 어떤 특정가구(A)가 포함되어있다는 것을 이미 알고 있다는 가정하에서 가계조사의 노출위험을 계산하면

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

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

〈표 4-5〉 key변수의 유일성(그룹화 後)

(단위: 가구, %)

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

3. 민감정보의 비밀보호

가. 민감변수의 선정

앞에서 언급한 바와 같이 민감변수란 주어진 자료에서 개인정보의 노출가능성이 높거나 노출을 제한해야 할 필요성이 있는 변수를 말하는데, 자료제공시 이러한 민감변수의 자료값을 적절한 비밀보호방법으로 변환시켜 줌으로써 정보의 노출을 제한

시킬 수 있다. 가계조사의 경우 앞의 <표 4-1>에 나타난 바와 같이 조사항목은 크게 가구의 수입과 지출로 구분되는데, 다시 세분화하면 47개의 수입관련 항목과 510개의 지출관련 항목으로 나눌 수 있다. 그러나 이 항목들은 가계수지의 현황을 파악하기에는 매우 유용하지만 어떤 특정가구의 식별을 위한 특성을 나타내는 데 모든 항목들이 사용되지는 않는다. 가령, 수입부문의 가구주 급여소득 등과 같은 항목은 어떤 특정 가구만이 해당될 수 있어 그 가구의 식별이 가능하지만, 지출부문의 식료품비를 나타내는 일부 항목들은 대부분의 가구에서 쉽게 파악될 수 있어 이 항목만으로는 가구의 식별이 용이하지 않다.

따라서 본 연구에서는 가계수지 항목별로 그 특성을 파악하여 이들 중 응답가구의 식별에 영향을 줄 것으로 판단되는 82개 항목들을 민감변수로 선정하였는데 수입관련 항목이 31개, 지출관련 항목이 51개이다. 이에 대한 결과는 <표 4-6>에 주어져 있다.

〈표 4-6〉 민감변수

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

나. 민감변수의 비밀보호

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

1) 반올림

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

<표 4-6>에 주어진 82개 민감변수는 자료특성에 따라 조사된 자료값의 단위가 매우 다양하다. 가령, 저금항목의 경우 연중 최소값은 백단위지만 최대값은 억단위가 된다. 따라서 본 연구에서는 이러한 자료의 특성을 고려하여 원자료를 백단위와 천단위, 그리고 만단위에서 반올림하는 3가지 방법을 고려하였다. 그리고 응답가구에서 가계부를 작성할 때 정확한 값이 아닌 일정단위 이상으로 반올림하여 응답한(즉, 반올림된 응답자료) 경우가 있는데, 이런 경우에는 반올림하고자 하는 일정단위 이하에서 반올림된 응답자료는 일정단위로 반올림하고, 일정단위 이상에서 반올림된 응답자료는 별도의 반올림 없이 그 값을 그대로 사용하기로 하였다. 가령, 어떤 변수의 자료값을 천단위에서 반올림한다고 할 때 응답자료가 13,000원이면 반올림하여 10,000원으로 변환하고, 응답자료가 130,000원이면 별도의 반올림 없이 130,000원을 그대로 사용하는 것이다. 가계조사의 민감변수에 대해 백단위와 천단위, 그리고 만단위로 각각 반올림하여 자료를 변환해서 시기별 및 지역별로 평균과 표준편차를 구하였으며, 이 중 수입관련 항목과 지출관련 항목 1개에 대한 결과를 <표 4-7>에 나타내었다. 여기서 원자료를 백단위에서 반올림한 것을 방법 1, 천단위로 반올림한 것을 방법 2, 그리고 만단위로 반올림한 것을 방법 3으로 나타내었다. 이 표의 결과에 의하면, 가구주급여소득 항목이나 식사대 항목에서 시기별 및 지역별로 방법 1이 원자료와 가장 유사한 결과를 보여주고 있는데, 이는 낮은 단위에서의 반올림이 높은 단위에서의 반올림보다 원자료와 더 유사하기 때문이다. 일반적으로 자료분석 측면에서는 높은 단위보다 낮은 단위에서 반올림하는 것이 더 효과적일 수 있지만, 비밀보호의 측면에서는 반올림의 단위를 낮게 할수록 정보노출의 가능성이 더 높게 된다. 따라서 반올림을 이용한 자료변환의 경우, 무조건 낮은 단위에서 반올림하는 것보다는 가급적 원자료의 특성을 고려하여 적정 수준의 단위에서 반올림하도록 하는 것이 더 바람직할 것이다.

2) 구간그룹화

구간그룹화란 연속형 자료를 적당한 구간으로 그룹화하여 각 구간을 대표하는 계급값으로 자료를 변환하는 방법을 말한다. 자료를 구간으로 변그룹화하기 위해서는 계급의 수와 계급간격, 그리고 계급의 한계를 정해 주어야 하는데 이들을 정하는 과정은 다음과 같다.

〈표 4-7〉 반올림 적용결과

(단위: 가구, %)

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

〈표 4-7〉 반올림 적용결과 (계속)

(단위: 가구, %)

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

먼저 주어진 자료의 수를 n, 계급의 수를 k라 하면 적정규모의 계급 수는 $2^k \geq n$을 만족시키는 최소의 정수값이 된다. 가령, $n = 50$ 이면, k는 6이 되어 적정규모의

계급 수는 6개가 된다. 다음으로 계급 수가 결정되면 각 계급별 간격을 결정해야 하는데, 간격을 동일하게 할 경우 계급간격은 자료의 범위를 계급 수로 나누어 주면 된다. 즉,

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

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

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

가) 균등분포인 경우

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

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

나) 파레토분포인 경우

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

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

다) log-mormal분포인 경우

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

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

이상의 결과를 정리하여 구간별 분포와 계급값에 따른 6가지의 구간그룹화 방법을 <표 4-8>에 나타내었다.

〈표 4-8〉 구간별 분포와 계급값

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

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

〈표 4-9〉 구간그룹화 적용결과

(단위: 원)

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

〈표 4-9〉 구간그룹화 적용결과 (계속)

(단위: 원)

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

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

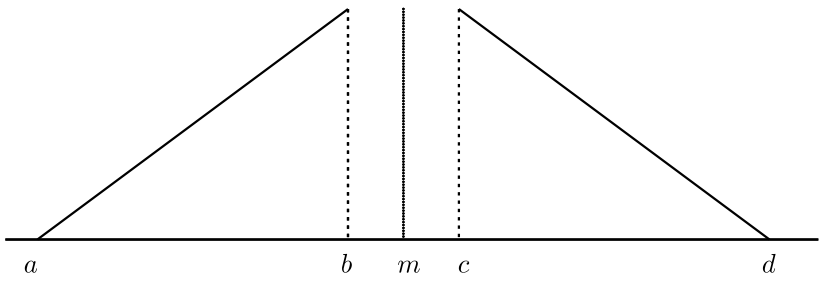
3) 승법잡음

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

일반적으로 잡음 e의 값은 e에 대한 적절한 분포를 가정하고 이에 따라 컴퓨터로 난수를 발생시키는 방법으로 구한다. 잡음 e의 분포로는 다양한 형태가 적용될 수 있는데, Kim & Winkler(2001)는 평균 μ를 중심으로 좌우가 절단된 정규분포(truncated normal distribution)를 적용한 결과를 소개하였다. Massell & Russell(2006)은 절단된 삼각분포(truncated triangular distribution)를 제시하였으며, Kim(2007)은 절단된 삼각분포를 변형한 다양한 형태의 분포를 제시하고 그 특성을 분석하였다. 본 연구에서는 e가 절단된 삼각분포를 따른다고 가정하고 승법잡음모형을 적용하여 선정된 민감변수의 자료를 변환하고 그 결과를 분석하였다.

가) 절단된 삼각분포

삼각분포는 자료의 최빈값을 중심으로 삼각형 모양으로 이루어진 분포를 말하고, 절단된 삼각분포는 최빈값을 중심으로 좌우로 일정부분 절단시킨 분포를 말한다. 일반적으로 삼각분포나 절단된 삼각분포는 좌우가 비대칭인 경우도 있으나, 여기서는 편의상 좌우대칭인 경우만 고려하였다. 만약 자료의 최소값을 a, 최대값을 d라 하고 최빈값 m을 중심으로 좌우대칭이면서 또한 절사점 b와 $c\,(>b)$도 m에서 좌우대칭이 되는 절단된 삼각분포는 [그림 4-1]에 주어진 것과 같은 형태가 된다.



[그림 4-1] 좌우대칭의 절단된 삼각분포

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

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

또한 e의 기대값과 분산을 계산하면 다음과 같이 된다.

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

나) 난수의 발생

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

〈표 4-10〉 승법잡음의 유형

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

<표 4-10>에 주어진 방법에 따라 난수를 발생시킨 결과가 실제 이론에 의해 절단된 삼각분포와 어느 정도 차이가 나는지를 검토하기 위해 기초통계량을 계산해 보았다.

〈표 4-11〉 생성난수의 결과비교

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

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

다) 자료의 변환

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

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

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

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

한편, 승법잡음모형으로 자료를 변환한 후 통계이용자들에게 변환된 자료를 제공할 경우, 이용자들은 원자료의 평균과 분산은 모르고 단지 변환된 변환된 자료 Y_i와 잡음 e_i의 평균과 분산만 알 수 있다. 따라서 원자료 X_i의 평균과 분산을 구할 수 있는 계산식이 필요한데, 이들의 평균과 분산은 위의 식을 이용하여 다음과 같이 얻을 수 있다.

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

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

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

〈표 4-12〉 승법잡음 적용결과

(단위: 원)

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

〈표 4-12〉 승법잡음 적용결과 (계속)

(단위: 원)

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

다. 최적의 비밀보호방법

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

먼저, 반올림을 적용한 경우에는 백단위에서 반올림한 방법 1을 선정하였고, 구간그룹화를 적용한 경우에는 log-normal 분포를 가정한 방법 3을 선정하였다. 그리고 승법잡음을 적용한 방법에서는 최빈값을 중심으로 가까운 값을 절사점으로 사용한 방법 1을 선정하여 이들 3가지 방법에 의한 평균과 표준편차를 시기별 및 지역별로 구한 후, 그 결과를<표 4-13>에 수록하였다. 이 표에 의하면 거의 모든 경우에 반올림으로 변환한 평균값이 원자료의 평균값과 차이가 가장 적다는 것을 알 수 있다. 즉, 백단위에서 반올림한 결과가 구간그룹화나 승법잡음에 의한 결과보다 거의 대부분의 경우에 더 우수한 것으로 나타났다. 또한 반올림을 제외하고 구간그룹화와 승법잡음의 두 방법만을 비교한다면 거의 모든 경우에 승법잡음에 의한 변환이 더 우수함을 알 수 있다. 이러한 결과를 정리하면, 가계조사자료의 비밀보호를 위해서는 백단위에서 반올림하는 것이 원자료와 가장 유사하다고 할 수 있으나, 자료제공 및 정보손실 등의 측면에서 보면 승법잡음을 이용한 방법도 효과적이라고 할 수 있다.

〈표 4-13〉 3가지 방법별 결과비교

(단위: 원)

변수	시기	지역	평 균			
			원자료	반올림 (방법 1)	구간그룹화 (방법 3)	승법잡음 (방법 1)
가구주 급여소득	연간	전국	1,967,254	1,967,268	1,961,531	1,966,919
		서울	2,133,225	2,133,231	2,122,965	2,123,645
		기타	1,941,942	1,941,958	1,936,911	1,943,018

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

〈표 4-13〉 3가지 방법별 결과비교 (계속)

(단위: 원)

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

제4절 결 론

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

한편, 민감정보의 노출제한을 위해 전체 조사항목들 중 수입관련 항목 31개와 지출관련 항목 51개를 합해 총 82개 항목을 민감변수로 선정하였으며, 이들 민감변수는 모두 연속형 자료값을 가지기 때문에 반올림과 구간그룹화, 승법잡음 등 연속형 자료의 비밀보호방법을 적용하여 분석하였다. 먼저 원자료를 백단위와 천단위, 그리고 만단위에서 반올림 한 후 변환된 자료의 평균을 계산한 결과, 낮은 단위에서 반올림한 결과가 더 우수한 것으로 나타났다. 구간그룹화에서는 적절한 계급수와 계급간격 및 계급의 한계 등을 정하고 각 구간별로 균등분포와 파레토분포, 그리고 log-normal분포를 가정하여 6가지의 방법에 따라 자료를 변환한 후 원자료의 평균값과 비교한 결과, log-normal로 가정한 경우가 가장 우수한 것으로 나타났다. 마지막으로 승법잡음의 모형에서 잡음의 분포를 절단된 삼각분포로 가정하여 4가지 방법으로 난수를 생성한 후 원자료에 곱하여 자료를 변환시킨 결과, 자료의 최빈값을 중심으로 가장 가까운 값을 절사점으로 사용한 방법이 가장 우수한 것으로 나타났다.

이상에서 살펴본 바와 같이, 가계조사의 경우 연속형 자료가 대부분이므로 본 연구에서 검토한 세 가지 방법에 의한 자료변환이 비밀보호 된 마이크로데이터를 작성하는데 상당히 효과적임을 알 수 있다. 다만 가계조사의 특성상 고려해야 할 변수들이 너무 많고 변수별로도 적합한 방법이 다르기 때문에 어떤 하나의 방법을 최적이라고 하기에는

다소무리가 있다. 그렇지만 자료제공과 정보노출 제한의 측면을 동시에 고려한다면 승법잡음을 이용한 비밀보호가 매우 유용한 방법이 될 수 있을 것으로 판단된다. 물론 잡음의 분포와 난수생성, 절사점 선정 등의 다소 복잡한 과정이 필요하지만, 모의실험 등을 통한 후속 연구가 활발히 이루어지고 이에 대한 알고리즘이 정립된다면 승법잡음 모형을 이용한 방법이 연속형 자료의 비밀보호에 매우 효과적일 수 있을 것이라 사료된다. 따라서 향후에도 승법잡음모형에 대한 이론적 연구가 꾸준히 진행되어 연속형 자료의 마이크로데이터 작성에 많이 활용되길 기대해본다.

참고문헌

정동명 · 강동환(2007), “마이크로자료의 활용도제고를 위한 비밀보호방법”, 통계연구결과 보고서, 통계청.

정동명 · 김종익 · 강동환(2007), “인구센서스자료의 비밀보호방법”, 「통계연구」, 제12권제1호.

통계청(2006), 「가계조사 조사지침서」. Bethelehem, J. G., W. J. Keller, and J. Pannekoek(1990), “Disclosure

Control of Microdata”, *Journal of the American Statistical Association*, 85, pp.38-45.

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

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

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

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

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

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

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

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

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

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

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

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

제5장

승법잡음모형을 이용한 가계조사 자료의 비밀보호

정동명 · 김경미

제1절 서 론

최근 들어 개인정보의 비밀보호가 사회적으로 이슈화되면서 통계청을 비롯한 국가통계 작성기관에서는 마이크로자료(micro data)나 매크로자료(macro data)를 외부에 제공할 때, 사전에 미리 응답자의 개인정보가 노출되지 않도록 적절한 비밀보호방법을 적용하고 있다. 예를 들어, 통계청에서는 통계이용자들을 위해 5년마다 실시하는 인구주택총조사 결과 중 2%를 마이크로자료 파일로 작성하여 그대로 제공하였으나, 지난 2005 인구주택총조사 결과부터는 일부 민감한 변수에 대해 비밀보호방법이 적용된 2% 마이크로자료 파일을 작성하여 제공하였다.

이러한 비밀보호방법은 1970년대 이후부터 영국이나 네덜란드, 미국 등 여러 통계선진국에서 꾸준히 연구되고 있는 분야이며, 자료의 특성에 따라 자료교환(data swapping)이나 그룹화(grouping), 반올림(rounding), 구간그룹화(grouping into intervals) 등 다양한 방법이 개발되어 현재 널리 활용되고 있다. 우리나라에서도 최근에 통계작성기관과 학계에서 이에 대한 연구가 점차 활성화되고 있는데, 특히 통계청에서는 2007년 통계법 개정으로 인한 마이크로자료 제공의무에 따라 응답자의 비밀보호를 위한 연구를 체계적으로 진행하고 있다. 정동명(2007) 등은 2005 인구주택총조사 결과에 그룹화 등의 방법을 적용하여 비밀보호된 마이크로자료 파일을 작성하는 연구를 하였다. 또한 정동명(2008) 등은 우리나라 가구의 소득과 소비지출의 특성을 파악하는 가계조사의 결과에도 반올림과 구간그룹화, 승법잡음 등의 방법을 적용하여 분석하고 서로 비교해 보았다.

✍ 이 연구는 2008년 상반기 통계개발원 연구보고서에 수록된 자료입니다.

본 연구에서는 연속형 자료의 비밀보호방법에 효과적으로 적용할 수 있는 승법잡음모형에 대해 살펴보고, 이 모형을 가계조사에 적용하여 비밀보호된 마이크로자료를 작성하는 과정을 소개하고자 한다. 2절에서는 원자료의 변환에 이용되는 잡음의 모형을 소개하고, 특히 승법잡음모형에서 활용할 수 있는 4가지 분포와 이들의 특성을 제시한다. 3절에서는 제시된 분포별로 난수프로그램을 작성하여 잡음을 생성하고 그 결과를 비교해본다. 또한 가계조사에서 민감변수로 선정된 항목을 대상으로 자료를 변환하는 과정을 설명하고, 원자료와 승법잡음모형에 의해 변환된 자료를 분석하여 서로 비교해 본다. 마지막으로 본 연구의 결론은 4절에서 살펴보도록 한다.

제2절 잡음을 이용한 비밀보호

1. 변환모형

가. 가법잡음모형

주어진 원자료를 X라 하고, 이에 대응하는 잡음을 e라고 하자. 만약 원자료 X에 적절한 수준의 잡음을 더해주면 변환된 자료 Y가 얻어지는데, 이러한 변환모형을 가법잡음(additive noise)모형이라 하고 다음과 같이 나타낸다.

$$Y = X + e$$

여기서 잡음 e의 분산과 공분산은 원자료 X에 의존된다. 가법잡음모형은 연속형 자료(continuous data)인 경우 효과적으로 활용할 수 있으나, 원자료가 0인 경우 잡음 e의 값에 따라 변환된 Y값이 음수가 될 수 있기 때문에 비밀보호된 자료값이 항상 0이나 양수가 되어야 할 경우에는 적용하기가 곤란하다. 가법잡음모형의 특성 등에 대한 자세한 내용은 Kim(1986)과 Fuller(1993) 등을 참고하기 바란다.

나. 승법잡음모형

가법잡음모형과 달리 원자료 X에 대응하는 잡음 e를 곱해주어 자료를 변환하는 것을 승법잡음(multiplicative noise)모형이라 하고, 이 때 변환된 자료 Y는 다음과 같이 나타낸다.

$$Y = X \cdot e$$

잡음 e 는 일반적으로 1이 아니면서 1을 중심으로 양의 값을 가지는 것이 바람직하다. 이것은 e 가 만약 음수가 되면 변환된 변수도 음의 값이 되고, e 가 1이 되면 Y 는 X 와 같게 되어 변환의 의미가 없기 때문이다. 또한 e 가 1에서 멀리 떨어진 값을 가지면, 원자료 X 와 변환된 자료 Y 가 너무 차이가 나서 정보의 손실이 크게 된다. 따라서 e 는 1이 아니면서 1과 가까운 값을 가지는 것이 바람직하다.

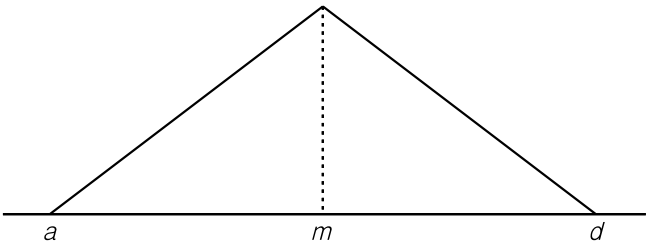
2. 잡음의 분포와 그 특성

잡음 e 를 이용하여 자료를 변환시키고자 할 경우, e 를 구하는 것이 매우 중요한 요인이 된다. 일반적으로 e 를 구하기 위해서는 먼저 e 에 대한 적절한 분포를 가정한 후, 컴퓨터를 이용해서 난수(random number)를 발생시키는 프로그램을 작성하여 e 의 값을 구한다. 이러한 e 에 대한 분포로는 모형에 따라 다양한 형태가 적용될 수 있으나, 이에 대한 연구는 많지 않다. 다만, 승법잡음모형인 경우, Kim & Winkler(2001)는 평균을 중심으로 좌우가 절단된 정규분포(truncated normal distribution)를 적용한 결과를 소개하였다. 또한 Kim(2003) 등은 절단된 삼각분포를 이용하여 승법잡음모형에 의한 자료변환과 이에 대한 특성을 분석하였다.

본 절에서는 승법잡음모형에서 잡음의 분포로 이용할 수 있는 삼각분포와 절단된 삼각분포, 사다리꼴분포, 이중삼각분포 등 모두 4가지 유형의 분포를 소개하고, 이들의 특성을 각각 살펴보고자 한다.

가. 삼각분포

삼각분포(triangular distribution)는 [그림 5-1]에 나타난바와 같이 m 을 중심으로 삼각형 모양으로 이루어진 분포를 말한다. 여기서 a 는 자료의 최소값, m 은 최빈값(mode), d 는 최대값이 된다.



[그림 5-1] 삼각분포

잡음 e 가 최빈값 m 을 중심으로 삼각분포를 따른다고 하면, 확률밀도함수(probability density function; pdf), $f^t(e)$ 는 다음과 같이 나타낼 수 있다.

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

그리고 $f^t(e)$을 이용하여 잡음 e의 기대값 $E^t(e)$와 분산 $Var^t(e)$을 구하면 다음과 같이 나타낼 수 있다.

$$E^t(e) = \frac{(m+a+d)}{3} \tag{1}$$

$$Var^t(e) = \frac{[d^2 + a^2 + m^2 - m(a+d) - ad]}{18}$$

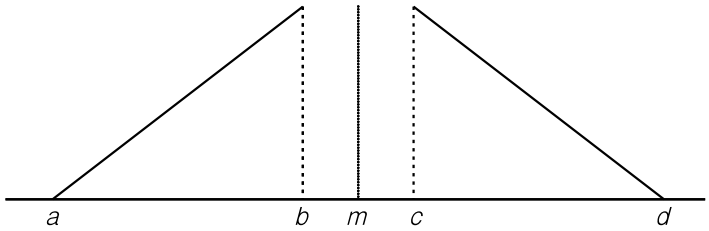
만약 잡음 e가 최빈값 m을 중심으로 좌우대칭인 삼각분포를 따른다면 $m-a = d-m$이고 $a+d = 2m$이 되므로, 위 식 (1)에서 $E^t(e)$와 $Var^t(e)$는 다음과 같이 간단하게 표현할 수 있다.

$$E^t(e) = m$$

$$Var^t(e) = \frac{(d-m)^2}{6}$$

나. 절단된 삼각분포

절단된 삼각분포(truncated triangular distribution)는 [그림 5-1]의 삼각분포에서 최빈값 m을 중심으로 좌우로 일부분이 절단된 형태로 이루어진 분포를 말한다. 즉, 절단된 삼각분포는 [그림 5-2]에 나타난 바와 같이 m을 중심으로 절사점 b와 c $(c > b)$에서 각각 절단된 형태를 하고 있어 최빈값을 포함하지 않는다.



[그림 5-2] 절단된 삼각분포

잡음 e가 최빈값 m을 중심으로 절단된 삼각분포를 따른다면 확률밀도함수(pdf), $f^{tt}(e)$는 다음과 같이 나타낼 수 있다.

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

그리고 위의 식 $f^{tt}(e)$을 이용하여 잡음 e의 기대값 $E^{tt}(e)$와 분산 $Var^{tt}(e)$은 다음과 같이 구할 수 있다.

$$E^{tt}(e) = \frac{[(d-m)(b-a)^2(2b+a)+(m-a)(d-c)^2(2c+d)]}{3[(d-m)(b-a)^2+(m-a)(d-c)^2]} \tag{2}$$

$$\begin{aligned} Var^{tt}(e) = & \frac{1}{18[(b-a)^2(d-m)+(d-c)^2(m-a)]^2}\{(b-a)^6(d-m)^2 \\ & + (d-c)^6(m-a)^2 + (b-a)^2(d-m)(d-c)^2(m-a) \\ & [(3b+a)^2+(3c+d)^2+2(a^2+d^2)-4(2b+a)(2c+d)]\} \end{aligned}$$

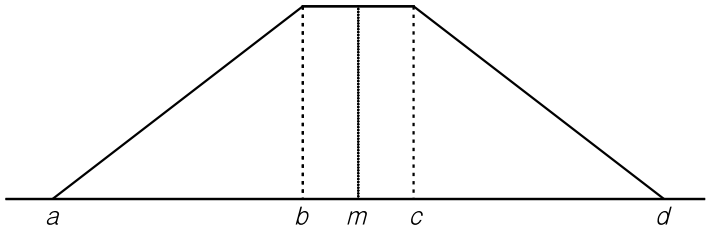
만약 최빈값 m을 중심으로 좌우대칭이면, $b-a = d-c$가 되므로 위의 식 (2)에서 $E^{tt}(e)$와 $Var^{tt}(e)$는 다음과 같이 간단히 나타낼 수 있다.

$$\begin{aligned} E^{tt}(e) &= m \\ Var^{tt}(e) &= \frac{(d+c)^2+2[(2m-c)^2-m(m+2d)]}{6} \end{aligned}$$

절단된 삼각분포는 절단되지 않은 삼각분포와 기대값은 같지만 분산은 서로 다르다. 즉, 절단된 경우의 분산이 더 커지게 되는데, 이는 평균 주위의 자료들이 절단됨으로써 그만큼 더 나머지 자료들이 평균에서 멀리 떨어지기 때문일 것이다. 예를 들어, $a=0.5$, $b=0.75$, $m=1$, $c=1.25$, $d=1.5$라고 하면 절단된 삼각분포의 기대값은 1로서 절단되지 않은 삼각분포와 같다. 즉, $E^{t}(e) = 1 = E^{tt}(e)$이 된다. 그러나 절단된 삼각분포의 분산은 $Var^{t}(e) = 0.5/18 = 0.028$이고 절단되지 않은 삼각분포의 분산은 $Var^{tt}(e) = 8.25/72 = 0.115$이 된다. 따라서 이 예제에서는 절단된 경우가 절단되지 않은 경우에 비해 약 4배 정도 더 큰 것으로 나타났다.

다. 사다리꼴분포

사다리꼴분포(trapezoidal distribution)는 자료의 최빈값 m을 중심으로 [그림 5-3]과 같은 모양으로 이루어진 분포를 말한다. 즉, 사다리꼴분포는 [그림 5-2]의 절단된 삼각분포에서 두 절사점 b와 $c\,(c>b)$ 사이에 일정한 값을 가지는 형태로서, 절단된 삼각분포와는 달리 최빈값을 포함하고 있다.



[그림 5-3] 사다리꼴분포

만약 잡음 e가 최빈값 m을 중심으로 사다리꼴분포를 따른다고 한다면, 확률밀도함수(pdf), $f^{tr}(e)$는 다음과 같다.

$$f^{tr}(e) = \begin{cases} \dfrac{2k_1}{(d-a)(b-a)}(e-a), & a \le e < b \\ \dfrac{2k_2}{(d-a)}, & b \le e < c \\ \dfrac{2k_3}{(d-a)(d-c)}(d-e), & c \le e < d \end{cases}$$

여기서 k_1, k_2, k_3는 분포의 높이에 의존하는 값으로, 만약 높이가 같다면 이들 값은 모두 같게 된다. 즉, 두 절사점 b와 c에서 $f(b)=f(c)$가 된다면 다음과 같이 된다.

$$k_1 = k_2 = k_3 = \frac{(d-a)}{(d+c-b-a)}$$

따라서 사다리꼴분포에서 높이가 동일하다면 $f^{tr}(e)$는 다음과 같이 나타낼 수 있다.

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

위의 식 $f_{eq}^{tr}(e)$ 을 이용하여 잡음 e 의 기대값 $E^{tr}(e)$ 와 분산 $Var^{tr}(e)$ 을 구하면 다음과 같이 구할 수 있다.

$$E^{tr}(e) = \frac{(d^2+c^2-b^2-a^2+cd-ab)}{3(d+c-b-a)} \tag{3}$$

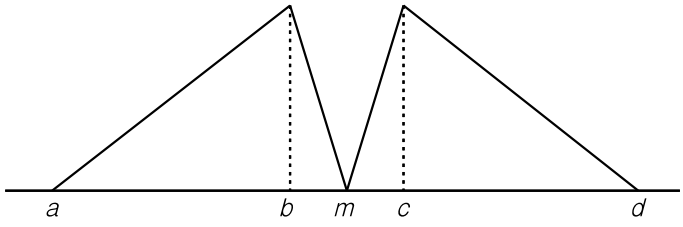
$$\begin{aligned} Var^{tr}(e) &= \frac{1}{18(d+c-b-a)} \\ &\quad \times [3(d^3+c^3-b^3-a^3-a^2b-ab^2+c^2d+cd^2) \\ &\quad - \frac{2[(d+c-b-a)(d+c+b+a)+ab-cd]^2}{(d+c-b-a)}] \end{aligned}$$

만약 잡음 e 가 최빈값 m 을 중심으로 좌우대칭이라고 한다면 $b+c = d+a = 2m$ 이 되므로, $a = 2m-d$ 와 $b = 2m-c$ 를 위 식 (3)에 각각 대입하면 $E^{tr}(e)$ 와 $Var^{tr}(e)$ 은 다음과 같이 간단하게 나타낼 수 있다.

$$\begin{aligned} E^{tr}(e) &= m \\ Var^{tr}(e) &= \frac{[d^3+c^3-4m^3+(6m^2+cd)(d+c)-4m(d^2+c^2+cd)]}{6(d+c-2m)} \end{aligned}$$

라. 이중삼각분포

이중삼각분포(double triangular distribution)는 [그림 5-4]에 나타난바와 같이 최빈값 m 을 중심으로 좌우가 각각 삼각형 모양으로 이루어진 분포를 말한다. 즉, 최소값 a 에서 절사점 b 를 중심으로 최빈값 m까지 하나의 삼각형 모양이 되고, 최빈값 m에서 절사점 c 를 중심으로 최대값 d 까지 또 하나의 삼각형 모양이 되는 분포이다. 이 분포는 두 절사점 b 와 c 사이에서 적절한 값을 갖지만 사다리꼴분포와는 달리 최빈값 m 을 포함하지 않는다.



[그림 5-4] 이중삼각분포

잡음 e가 최빈값 m을 중심으로 이중삼각분포를 따른다면, 확률밀도함수(pdf), $f^{dt}(e)$는 다음과 같다.

$$f^{dt}(e) = \begin{cases} \frac{1}{(m-a)(b-a)}(e-a) & , \quad a \le e < b \\ \frac{1}{(m-a)(m-b)}(m-e) & , \quad b \le e < m \\ \frac{1}{(d-m)(c-m)}(e-m) & , \quad m \le e < c \\ \frac{1}{(d-m)(d-c)}(d-e) & , \quad c \le e < d \end{cases}$$

또한 $f^{dt}(e)$을 이용하여 잡음 e의 기대값 $E^{dt}(e)$와 분산 $Var^{dt}(e)$을 계산하면 다음과 같이 된다.

$$E^{dt}(e) = \frac{(a+b+c+d+2m)}{6} \tag{4}$$

$$Var^{dt}(e) = \frac{1}{36}[\,2a^2+2b^2+2c^2+2d^2+2m^2-m(a+b+c+d) - 2(a+b)(c+d)+ab+cd\,]$$

만약 잡음 e가 최빈값 m을 중심으로 좌우대칭이 된다면 $b+c = d+a = 2m$이 되므로, $a = 2m-d$와 $b = 2m-c$를 위 식 (4)에 각각 대입하면 $E^{dt}(e)$와 $Var^{dt}(e)$은 다음과 같이 간단히 나타낼 수 있다.

$$E^{dt}(e) = m$$

$$Var^{dt}(e) = \frac{[\,(d-m)^2+(c-m)^2-m(d+c-m)+cd\,]}{6}$$

제3절 적용사례

1. 분석대상

가. 가계조사

앞 절에서 소개한 4개의 잡음분포를 이용한 승법잡음모형의 특성을 파악하기 위해 통계청에서 실시하는 가계조사에 적용해 보았다. 가계조사는 우리나라 가구에 대한 가계수지의 실태를 파악하여 국민의 소득과 소비수준 변화의 측정 및 분석 등에 필요한 자료를 제공하기 위한 조사이다. 2006년 가계조사의 대상가구는 약 8,000가구이며, 2005 인구주택총조사의 10% 표본조사구를 모집단으로 하여 전국을 25개로 층화한 후 확률비례추출법을 이용하여 선정하였다. 이렇게 선정된 전국의 표본가구를 대상으로 가구의 수입과 소비지출 관련 항목들을 응답가구에서 직접 가계부에 매일 작성하도록 하고 있으며, 매월단위로 가계부의 내용을 집계, 정리한 후 그 결과를 분기별 및 연도별로 공표하고 있다.

2006년 가계조사의 대상가구는 <표 5-1>에 나타난바와 같이 동일한 가구의 중복을 제외하면 총 12,458가구이며, 이 중 서울이 1,638가구이고 서울이외지역이 10,820가구이다. 이 결과는 분기별 및 연별로 조사결과가 집계되어 공표되는 가구 수와 다른데, 이는 1년간 매월 조사에 응답한 동일가구는 12가구가 아닌 1가구로 처리하여 중복성을 제외하였기 때문이다.

〈표 5-1〉 2006년 가계조사의 표본규모

(단위: 가구)

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

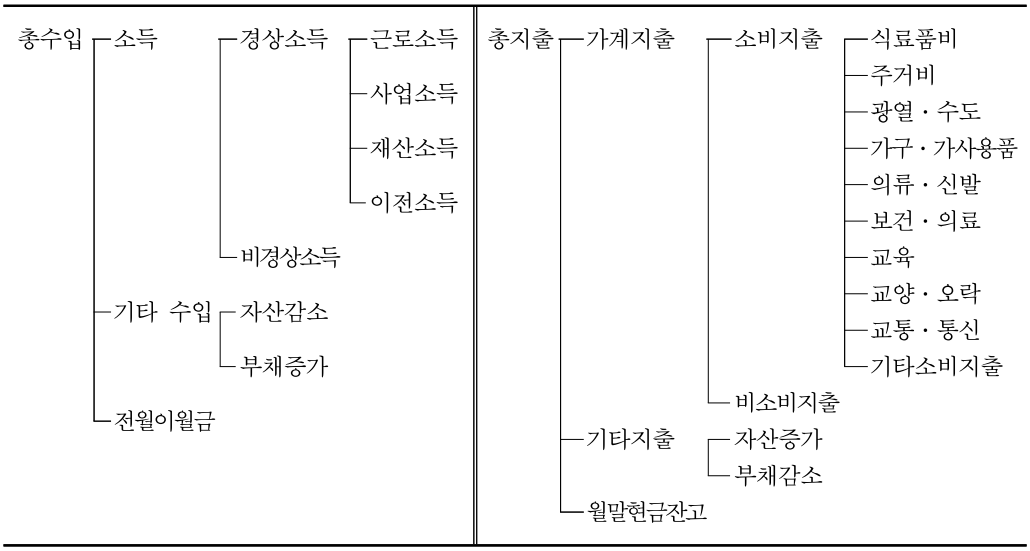
나. 분석항목

가계조사에서 가계부를 통해 조사되는 항목은 <표 5-2>에 주어져 있는 바와 같이 가구사항을 제외하면 총수입 항목과 총지출 항목으로 구분되는데, 총수입은 다시 소득과 기타수입, 그리고 전월이월금으로 나누어지고 총지출은 가계지출과 기타지출, 그리고

월말 현금잔고로 구분된다. 조사항목들을 세분화하여 분류한 후 이를 다시 품목별로 정리하면, 수입관련 품목이 47개이고 지출관련 품목이 510개로 모두 557개가 된다. 따라서 가계조사에서 조사되는 항목은 가구특성 62개와 조사품목 557개를 합해 총 619개 항목이 된다. 이에 대한 자세한 내용은 가계조사 지침서(2006)를 참고하기 바란다.

한편, 정동명(2008) 등은가계조사 항목을식별정보(identifying information)와 민감정보(sensitive information)로 각각 구분한 후, 이에 따라 비밀보호방법을 달리 적용하였다. 그들은 62개의 가구특성 항목 중 응답자의 식별가능성이 높다고 판단되는 7개 항목을 식별변수(또는 key변수)로 선정하였다. 그리고 수입관련 품목 47개 중 31개, 지출관련 품목 510개 중 51개 등 총 82개 품목을 응답자들이 노출을 꺼려하는 민감변수로 각각 선정하였는데, 선정된 민감변수는 모두 최하단계의 품목들로 구성되어 있다.

〈표 5-2〉 가계수지 항목분류



그러나 본 연구에서는 최하단계 품목이 아닌 보다 상위단계에 있는 품목들을 민감변수로 새로 선정하였다. 이것은 최하단위의 품목이 너무 세분화 되어 있어 이용자들에게 제공할 경우 통계정보로서의 가치가 낮거나 전혀 의미없는 정보가 될 수도 있으며, 오히려 이들의 합의 형태로 구성된 상위단계의 품목이 통계정보로서 더 효용성이 높을 것이라고 판단했기 때문이다. 가령, 수입관련 품목 중 가구주급여소득과 가구주상여금을 합하면 가구주소득이 얻어지는데, 이용자들에게는 가구주급여소득이나 가구주상여금에 대한 것보다는 가구주소득에 대한 정보가 더 유용할 수 있을 것이다.

이러한 기준에 따라 최하단계의 품목은 이용자들에게 제공하지 않는 것으로 가정하고, 외부인(intruder)이 다른 경로를 통해 정보를 얻을 수 있을 것으로 판단되는 품목들을 분석한 후, 최종적으로 수입관련 9개품목과 지출관련 21개 품목 등 총 30개 품목을 민감변수로 선정하였다. 이에 대한 결과는 <표 5-3>에 주어져 있다.

〈표 5-3〉 민감변수

구분	품 목		
총수입 (9개)	· 가구주소득	· 배우자소득	· 기타 가구원소득
	· 사업소득	· 재산소득	· 이전소득
	· 비경상소득	· 자산감소로인한수입	· 부채증가로인한수입
총지출 (21개)	· 월세	· 주택설비및수선비	· 기타주거
	· 수도료	· 전기료	· 연료
	· 공동주택난방비	· 의약품	· 보건의료용품기구
	· 보건의료서비스	· 납입금	· 교재비
	· 보충교육비	· 문구류	· 교통비
	· 통신비	· 조세	· 공적연금
	· 사회보험	· 기타비소비지출	· 월말현금잔고

2. 승법잡음모형에 의한 비밀보호

가. 잡음의 생성

자료의 비밀보호를 위해 승법잡음모형을 적용하기 위해서는 우선 잡음을 생성해야만 하는데, 여기서는 앞 절에서 소개한 4가지 분포에 따라 잡음을 생성하는 과정을 소개하고자 한다. 먼저 잡음 ε가 각각의 분포를 따른다는 가정하에 최빈값(m)을 1로 하고 자료의 범위가 $0.01 \leq |\varepsilon - 1| \leq 0.4$ 을 만족하도록 최소값(a)과 최대값(d), 두 절사점 (b, c) 등을 정하였다. 이것은 두 절사점이 최빈값 1을 중심으로 너무 멀리 떨어지지 않도록 하기 위해서이다. 만약 절사점이 최빈값 1에서 너무 멀리 떨어지면 편향이 발생하고 분산이 커지게 된다. 이에 대한 자세한 내용은 정동명(2008) 등을 참고하기 바란다.

위의 조건이 만족되면 $a = 0.6$, $d = 1.4$, $b = 0.99$, $c = 1.01$ 이 되는데, 이 값들을 이용하여 각 분포별로 난수발생 프로그램을 SAS로 작성하였다. 그리고 2006년도 가계조사의 월별 표본규모를 고려하여 약 90,000개의 난수를 생성하였다.

각 분포별로 난수를 발생시킨 결과가 실제 이론에 의한 값과 어느정도 차이가 나는지를 검토하기 위해 기초통계량을 계산하였으며, 그결과를 <표 5-4>에 나타내었다. 이 표에 따르면 대부분의 경우에 큰 차이가 나지 않음을 알 수 있는데, 이는 각 분포별로 난수가 잘 생성되었음을 보여주고 있다. 가령, 각 분포별로 난수의 평균값은 모두 1.0002로 나타나 이론적 평균값인 1과 거의 차이가 나지 않으며, 표준편차의 경우에도 거의 차이가 없음을 알 수 있다.

다만, 각 분포별로 이론적인 표준편차가 약간씩 다르게 나타나는데, 삼각분포가 가장 작고 이어서 사다리꼴분포와 이중삼각분포, 절단된 삼각분포의 순으로 나타났다. 이것은 두 절사점 사이의 정보량이 얼마나 더 많은지에 따른 차이로서, 두 절사점 사이에서 난수가 많이 발생되는 분포가 그렇지 못한 분포에 비해 정보량이 더 많아 표준편차가 적어지기 때문인 것 같다. 그러므로 두 절사점 사이에서 난수가 발생되지 않는 절단된 삼각분포의 표준편차가 가장 크게 나타난다고 할 수 있을 것이다.

〈표 5-4〉 생성된 난수의 결과비교

분포	구분	통계량			
		평균	표준편차	최소값	최대값
삼각	이론적(A)	1.0000	0.1633	0.6000	1.4000
	난수결과(B)	1.0002	0.1624	0.6041	1.3916
	차이(B-A)	0.0002	-0.0009	0.0041	-0.0084
절단된 삼각	이론적(A)	1.0000	0.1675	0.6000	1.4000
	난수결과(B)	1.0002	0.1666	0.6040	1.3918
	차이(B-A)	0.0002	-0.0009	0.0040	-0.0082
사다리꼴	이론적(A)	1.0000	0.1634	0.6000	1.4000
	난수결과(B)	1.0002	0.1624	0.6041	1.3916
	차이(B-A)	0.0002	0.0010	0.0041	-0.0084
이중삼각	이론적(A)	1.0000	0.1654	0.6000	1.4000
	난수결과(B)	1.0002	0.1645	0.6041	1.3917
	차이(B-A)	0.0002	0.0009	0.0041	-0.0083

나. 자료의 변환

원자료를 X_i, 잡음을 e_i 라 하면 승법잡음모형에 따라 변환된 자료 Y_i 는 다음과 같이 나타낼 수 있다.

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

여기서 X_i와 e_i 는 서로 독립이기 때문에 변환된 Y_i의 기대값$E(Y_i)$ 와 분산 $Var(Y_i)$ 은 다음과 같이 구할 수 있다.

$$E(Y_i) = E(X_i) \cdot E(e_i) \tag{5}$$

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

한편, 통계이용자들에게 변환된 자료를 제공할 경우 이용자들은 원자료 X_i의 평균과 분산은 모르고 단지 변환된 자료 Y_i 와 잡음 e_i 의 평균과 분산만 알 수 있게 된다. 따라서 이용자들 중에는 추가적인 자료분석을 위해 원자료의 평균과 분산을 요구할 수도 있다. 이를 위해 원자료 X_i의 평균과 분산을 구하는 추정식이 필요한데, 위의 식 (5)을 이용하여 X_i의 기대값 $E(X_i)$ 와 분산 $Var(X_i)$ 은 다음과 같이 얻을 수 있다.

$$\hat{E}(X_i) = \hat{E}(Y_i) / E(e_i) = \hat{E}(Y_i)$$

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

다. 소지역자료의 변환

소지역(domain) 자료의 변환을 위해 소지역을 나타내는 첨자를 s 라 하면, 승법잡음모형에 의해 변환된 소지역자료 Y_i^s 는 다음과 같이 나타낼 수 있다.

$$Y_i^s = X_i^s \cdot e_i, \quad i = 1, \cdots, n_s$$

여기서 n_s 는 소지역의 크기를 나타내며, 소지역의 원자료 X_i^s 와 e_i 는 서로 독립이기 때문에 변환된 Y_i^s 의 기대값 $E(Y_i^s)$ 와 분산 $Var(Y_i^s)$ 은 다음과 같이 구할 수 있다.

$$E(Y_i^s) = E(X_i^s) \cdot E(e_i) \tag{6}$$

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

따라서 위의 식 (6)을 이용하여 소지역 s 에서의 원자료 X_i^s 의 분산 $Var(X_i^s)$ 은 다음과 같이 얻을 수 있다.

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

라. 변환결과의 비교 · 분석

승법잡음모형의 자료변환에 대한 분석을 위해 2006년 가계조사 결과에 직접 적용해 보았다. 먼저 30개 민감변수 중 수입관련 품목의 가구주소득과 지출관련 품목의 월세를 대상으로 앞 절에서 언급한 4가지 잡음분포에 따라 난수를 생성한 후, 이 잡음을 가계조사자료에 곱하여 자료를 변환해 보았다. 즉, 가계조사의 연간자료를 원자료로 한 후 전국 및 지역별로 원자료와 변환된 자료의 평균과 표준편차를 각각 계산하고, 또한 원자료와 변환된 자료의 차이를 비교하기 위해 상대차이(relative difference)도 계산하였다. 이에 대한 결과가 <표 5-5>와 <표 5-6>에 각각 주어져 있다.

가구주소득에 대한 결과를 살펴보면, 먼저 전국단위의 경우 원자료와 변환된 자료의 평균값 차이는 약 2,200~2,300원 정도이고, 표준편차의 차이는 1,000원 내외인 것으로 나타나 이들의 차이가 크지 않음을 알 수 있다. 또한 원자료와 변환된 자료의 상대차이도 약 0.1% 정도로 나타나 각 분포별로 큰 차이가 없는 것으로 나타났다. 분포별로는 삼각분포를 이용하여 변환한 결과가 원자료와 가장 차이가 적고 이어서 사다리꼴분포와 이중삼각분포, 그리고 절단된 삼각분포의 순으로 나타났다. 이는 삼각분포나 사다리꼴분포에서 생성된 난수가 1을 포함하므로 자료변환이 이루어지지 않은 경우가 있기 때문인 것 같다.

〈표 5-5〉 품목별 변환결과

(단위 : 원)

품목	통계량	지역	원자료	변환된 자료			
				삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
가구주 소득	평균	전국	2,258,159	2,255,944	2,255,851	2,255,942	2,255,894
		서울	2,295,665	2,299,081	2,298,817	2,299,078	2,298,942
		기타지역	2,252,439	2,249,366	2,249,299	2,249,364	2,249,330
	표준편차	전국	1,636,192	1,635,247	1,635,132	1,635,209	1,635,181
		서울	1,559,763	1,574,642	1,574,811	1,574,607	1,574,723
		기타지역	1,647,478	1,644,212	1,644,055	1,644,173	1,644,124

품목	통계량	지역	원자료	변환된 자료			
				삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
월세	평균	전국	181,064	180,943	180,975	180,943	180,958
		서울	246,382	247,216	247,280	247,216	247,247
		기타지역	171,099	170,833	170,860	170,833	170,846
	표준편차	전국	133,014	132,038	132,168	132,037	132,102
		서울	164,236	165,359	165,639	165,358	165,499
		기타지역	124,615	123,095	123,195	123,094	123,144

〈표 5-6〉 품목별 변환자료의 상대차이

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
가구주 소득	평균	전국	0.00098	0.00102	0.00098	0.00100
		서울	0.00149	0.00137	0.00149	0.00143
		기타지역	0.00136	0.00139	0.00137	0.00138
	표준편차	전국	0.00058	0.00065	0.00060	0.00062
		서울	0.00954	0.00965	0.00952	0.00959
		기타지역	0.00198	0.00208	0.00201	0.00204

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
월세	평균	전국	0.00067	0.00049	0.00067	0.00058
		서울	0.00338	0.00365	0.00339	0.00351
		기타지역	0.00156	0.00140	0.00156	0.00148
	표준편차	전국	0.00733	0.00636	0.00734	0.00686
		서울	0.00684	0.00855	0.00684	0.00770
		기타지역	0.01220	0.01140	0.01221	0.01181

* 상대차이 = |원자료-변환자료|/원자료

비밀보호를 위해 실제 자료에 적용할 경우에는 잡음이 1이 포함되지 않는 것이 더 바람직하므로 절단된 삼각분포와 이중삼각분포만을 비교해 보면, 이중삼각분포를 이용하여 변환한 경우가 원자료와의 차이가 약간 더 적다. 이러한 결과는 절단된 삼각분포의 경우 두 절사점 b와 c사이에서 난수가 전혀 생성되지 않지만, 이중삼각분포의 경우 최빈값을 중심으로 어느 정도 난수가 생성되어 그만큼 더 많은 정보가 이용되기 때문인 것 같다.

지역단위의 경우에도 원자료와 변환된 자료의 차이가 크지 않은 것으로 나타났으나, 서울지역은 절단된 삼각분포인 경우가 가장 차이가 적어 전국단위나 기타지역의 결과와 다르게 나타났다. 이것은 두 절사점 사이에서 난수가 발생하지 않아 정보량이 적으면 상대적으로 변환된 자료는 원자료와 차이가 크지만, 원자료에 어떤 값이 곱해지는가에 따라 차이가 크지 않을 수도 있다는 것이다. 다시 말하면, 난수발생을 위한 잡음의 분포가 자료의 변환에 중요한 요인이 될 수 있지만, 무엇보다도 원자료에 곱해지는 값이 가장 절대적인 요인이 된다는 것을 보여주고 있다.

월세에 대한 결과를 살펴보면, 전국단위의 경우 원자료와 변환된 자료의 평균값 차이는 100원 내외이고 표준편차의 차이는 840~980원 사이인 것으로 나타나 차이가 크지 않음을 알 수 있다. 그러나 분포별 변환 결과는 가구주소득의 경우와 반대로 나타나는데, 전국단위와 기타지역의 경우에는 절단된 삼각분포를 이용하여 변환한 결과가 원자료와 가장 차이가 적고, 서울지역의 경우에는 삼각분포를 이용하여 변환한 결과가 가장 원자료와 차이가 적은 것으로 나타났다. 이러한 결과는 가구주소득에서 설명한 것과 유사하게 해석하면 될 것이다.

위에서 언급한 방법과 동일하게 모든 민감변수에 대해 자료를 변환하고 이들의 평균과 표준편차를 구한 후, 원자료와 어느 정도 차이가 나는지를 알아보기 위해 상대차이를 계산하여 그 결과를 <부록>에 수록하였다. 이 결과에 따르면 일부 품목을 제외한 대부분의 경우 삼각분포를 이용하여 변환한 결과가 원자료와 가장 차이가 적은 것으로 나타났다. 또한 잡음 e가 1을 포함하지 않는 절단된 삼각분포와 이중삼각분포만을 비교해 보면, 두 절사점 사이의 정보를 어느 정도 이용하는 이중삼각분포가 더 효과적이라고 할 수 있다.

이상의 분석결과를 정리하면, 승법잡음모형을 이용하여 자료를 변환할 경우 원자료에 곱해지는 잡음의 값이 가장 중요한 요인이 되지만, 어떤 잡음분포를 이용해서 잡음을 생성해야 하는지도 중요한 요인이 될 수 있음을 보여주고 있다. 따라서 만약 승법잡음모형으로 자료를 변환하고자 할 경우 본 연구에서 제시한 이중삼각분포를 잡음의 분포로 이용하면 상당히 효과적이라고 할 수 있을 것이다.

제4절 결 론

본 연구에서는 최근 들어 연속형 자료의 비밀보호방법에서 널리 활용되고 있는 승법잡음모형을 소개하고, 2006년 가계조사 자료에 이 모형을 적용하여 비밀보호된 마이크로자료를 작성하는 과정을 설명하고 분석하였다. 이를 위해 먼저 승법잡음모형에서 요구되는 잡음의 분포에 대해 4종류의 분포를 제시하고, 이들 분포의 특성을 살펴보았다. 먼저 삼각분포를 기본으로 하여 자료의 최빈값을 중심으로 좌우가 절단된 삼각분포와 두 절사점간에 일정한 값을 갖는 사다리꼴분포, 그리고 두 삼각형의 형태로 이루어진 이중삼각분포에 대해 확률밀도함수를 각각 구하고 이를 이용하여 평균과 분산의 계산공식을 유도하였다.

가계조사 자료의 비밀보호를 위해 가구특성항목을 제외한 557개 조사품목 중 수입관련 품목 9개와 지출관련 품목 21개 등 총 30개 품목을 민감변수로 선정한 후, 이들 변수별로 제시된 잡음분포에 따라 승법잡음모형을 적용하였다. 그 결과, 대부분의 항목에서 삼각분포를 가정하고 변환한 자료가 원자료와 가장 유사한 것으로 나타났는데, 이는 삼각분포를 가정할 경우 잡음 e가 1을 포함하므로 원자료가 변환되지 않을 수도 있기 때문이다.

따라서 잡음 e가 1이 포함하지 않는 절단된 삼각분포와 이중삼각분포만을 비교해 보면, 대부분의 변수에서 이중삼각분포를 가정한 경우가 절단된 삼각분포를 가정한 경우보다 약간 더 효과적인 것으로 나타났다. 이러한 결과는 이중삼각분포가 두 절사점 사이의 정보를 이용하기 때문이라 할 수 있다. 이상에서 살펴본바와 같이 가계조사의 경우 응답가구에서 노출을 꺼려하는 민감한 정보들이 대부분 연속형 자료이므로 이중삼각분포를 가정한 승법잡음모형을 적용하는 것이 비밀보호에 상당히 효과적이라고 판단된다.

한편, 본 연구에서는 최하단계의 품목들을 제공하지 않는다는 가정하에, 상위단계의 품목들로 민감변수를 구성하여 분석하였으나 이에 대한 다른 견해도 있을 것이다. 하지만 가계조사의 경우 너무 세분화된 자료가 이용자들에게 제공됨으로써 오히려 정보의 질을 저하시킬 가능성도 있으므로, 수집된 자료를 일방적으로 모두 제공하기보다는 자료제공심의 등의 절차를 거쳐 자료제공의 범위와 수준을 정하고, 이에 따라 이용자들에게 제공하는 것이 더 바람직할 수도 있을 것이다.

향후에 본 연구의 결과를 실무에 적용하여 활용성을 높이기 위해서는 품목간의 연관성에 대한 분석 등 추가적인 연구가 필요하다. 외국의 경우에도 새로운 비밀보호 방법의 실무적용을 위해서 연구자가 지속적인 연구가 수행되도록 다양한 지원을 하고 있다. 따라서 본 연구에서 제안한 승법잡음모형을 이용하여 비밀보호된 가계조사의

마이크로자료를 작성하고, 이용자들에게 제공하기 위해서는 체계적이고 지속적인 연구와 더불어 이와 관련된 부서에서도 꾸준한 관심과 적극적인 지원이 요구된다. 끝으로 본 연구를 계기로 응답자의 정보를 보호할 수 있는 다양한 연구가 활발히 진행되고, 그 결과를 바탕으로 자료제공의 신뢰성 확보와 함께 궁극적으로 조사의 응답률도 제고될 수 있기를 기대해 본다.

참고문헌

정동명, 김종익, 강동환(2007), “인구센서스자료의 비밀보호방법”, 「통계연구」, 제12권 제1호, 95-120.

정동명, 정미옥(2008), “인구주택총조사 마이크로자료의 개인정보 노출제한방법”, 「응용통계연구」, 제21권 제2호, 313-325.

정동명, 정남수, 한승훈(2008), “가계조사 마이크로데이터의 비밀보호”, 연구보고서, 통계청.

통계청(2006), 「가계조사 조사지침서」.

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

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

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

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

Kim, J. and Jeong, D. M. (2007), “Application of the Concept of Uniqueness for Creating Public Use Microdata Files”, *in Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, UNECE, CD-ROM.

<부록> 품목별 변환결과의 상대차이

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
가구주소득	평균	전국	0.00098	0.00102	0.00098	0.00100
		서울	0.00149	0.00137	0.00149	0.00143
		기타	0.00136	0.00139	0.00137	0.00138
	표준편차	전국	0.00058	0.00065	0.00060	0.00062
		서울	0.00954	0.00965	0.00952	0.00959
		기타	0.00198	0.00208	0.00201	0.00204
배우자소득	평균	전국	0.00116	0.00127	0.00116	0.00122
		서울	0.00053	0.00058	0.00053	0.00055
		기타	0.00145	0.00159	0.00145	0.00152
	표준편차	전국	0.00143	0.00125	0.00141	0.00134
		서울	0.00693	0.00692	0.00691	0.00694
		기타	0.00048	0.00026	0.00046	0.00037
기타 가구원소득	평균	전국	0.00242	0.00261	0.00242	0.00252
		서울	0.00251	0.00300	0.00251	0.00276
		기타	0.00240	0.00252	0.00240	0.00246
	표준편차	전국	0.00417	0.00451	0.00419	0.00435
		서울	0.00382	0.00396	0.00385	0.00390
		기타	0.00425	0.00463	0.00428	0.00444
사업소득	평균	전국	0.00286	0.00297	0.00286	0.00292
		서울	0.00095	0.00101	0.00095	0.00099
		기타	0.00313	0.00326	0.00314	0.00320
	표준편차	전국	0.01582	0.01628	0.01584	0.01605
		서울	0.00773	0.00745	0.00775	0.00761
		기타	0.01673	0.01726	0.01675	0.01700
재산소득	평균	전국	0.00892	0.00921	0.00893	0.00907
		서울	0.01429	0.01504	0.01430	0.01468
		기타	0.00764	0.00782	0.00765	0.00774
	표준편차	전국	0.05780	0.05887	0.05782	0.05835
		서울	0.06435	0.06696	0.06440	0.06571
		기타	0.05674	0.05739	0.05675	0.05707

* 상대차이= |(원자료-변환자료)|/원자료

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
이전소득	평균	전국	0.00066	0.00058	0.00066	0.00062
		서울	0.01365	0.01419	0.01366	0.01393
		기타	0.00054	0.00069	0.00055	0.00062
	표준편차	전국	0.00271	0.00180	0.00271	0.00225
		서울	0.06537	0.06977	0.06544	0.06759
		기타	0.00672	0.00839	0.00673	0.00756
비경상소득	평균	전국	0.00077	0.00107	0.00077	0.00092
		서울	0.00863	0.00950	0.00864	0.00907
		기타	0.00185	0.00229	0.00186	0.00207
	표준편차	전국	0.00170	0.00455	0.00173	0.00314
		서울	0.06663	0.07272	0.06671	0.06970
		기타	0.00766	0.01125	0.00769	0.00947
자산감소로 인한 수입	평균	전국	0.00288	0.00301	0.00288	0.00295
		서울	0.00244	0.00279	0.00245	0.00262
		기타	0.00295	0.00304	0.00295	0.00300
	표준편차	전국	0.02813	0.02881	0.02815	0.02848
		서울	0.00280	0.00035	0.00275	0.00121
		기타	0.03255	0.03287	0.03256	0.03272
부채증가로 인한 수입	평균	전국	0.00032	0.00043	0.00032	0.00037
		서울	0.01471	0.01490	0.01472	0.01482
		기타	0.00231	0.00221	0.00231	0.00227
	표준편차	전국	0.01729	0.01809	0.01729	0.01768
		서울	0.07107	0.07284	0.07109	0.07201
		기타	0.00129	0.00083	0.00130	0.00109
월세	평균	전국	0.00067	0.00049	0.00067	0.00058
		서울	0.00338	0.00365	0.00339	0.00351
		기타	0.00156	0.00140	0.00156	0.00148
	표준편차	전국	0.00733	0.00636	0.00734	0.00686
		서울	0.00684	0.00855	0.00684	0.00770
		기타	0.01220	0.01140	0.01221	0.01181

* 상대차이= |(원자료-변환자료)|/원자료

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
주택설비 및 수선비	평균	전국	0.00606	0.00607	0.00606	0.00607
		서울	0.02614	0.02782	0.02616	0.02699
		기타	0.00385	0.00368	0.00385	0.00377
	표준편차	전국	0.00305	0.00225	0.00304	0.00265
		서울	0.05203	0.05632	0.05209	0.05419
		기타	0.00046	0.00161	0.00047	0.00104
기타주거	평균	전국	0.00221	0.00229	0.00221	0.00225
		서울	0.00550	0.00591	0.00550	0.00570
		기타	0.00165	0.00167	0.00165	0.00166
	표준편차	전국	0.01778	0.01924	0.01781	0.01852
		서울	0.04892	0.05154	0.04896	0.05024
		기타	0.00078	0.00000	0.00076	0.00037
수도료	평균	전국	0.00054	0.00056	0.00054	0.00055
		서울	0.00117	0.00113	0.00117	0.00115
		기타	0.00082	0.00083	0.00082	0.00083
	표준편차	전국	0.00185	0.00192	0.00187	0.00189
		서울	0.01250	0.01273	0.01253	0.01262
		기타	0.00021	0.00025	0.00023	0.00024
전기료	평균	전국	0.00135	0.00139	0.00135	0.00138
		서울	0.00087	0.00104	0.00087	0.00096
		기타	0.00142	0.00144	0.00142	0.00143
	표준편차	전국	0.00486	0.00521	0.00488	0.00504
		서울	0.00852	0.00883	0.00854	0.00870
		기타	0.00449	0.00485	0.00452	0.00467
연료	평균	전국	0.00191	0.00199	0.00191	0.00195
		서울	0.00206	0.00229	0.00206	0.00218
		기타	0.00189	0.00194	0.00189	0.00191
	표준편차	전국	0.01155	0.01185	0.01157	0.01171
		서울	0.00833	0.00852	0.00835	0.00844
		기타	0.01184	0.01215	0.01185	0.01199

* 상대차이= |(원자료-변환자료)|/원자료

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
공동주택 난방비	평균	전국	0.00131	0.00130	0.00131	0.00130
		서울	0.00034	0.00018	0.00034	0.00027
		기타	0.00172	0.00166	0.00172	0.00169
	표준 편차	전국	0.00577	0.00602	0.00578	0.00590
		서울	0.00086	0.00161	0.00089	0.00123
		기타	0.00733	0.00743	0.00735	0.00739
의약품	평균	전국	0.00186	0.00195	0.00187	0.00191
		서울	0.00152	0.00172	0.00153	0.00163
		기타	0.00192	0.00198	0.00192	0.00195
	표준 편차	전국	0.00825	0.00845	0.00826	0.00835
		서울	0.00159	0.00320	0.00160	0.00238
		기타	0.01079	0.01147	0.01081	0.01113
보건의료 용품기구	평균	전국	0.00159	0.00161	0.00159	0.00160
		서울	0.01345	0.01391	0.01346	0.01369
		기타	0.00018	0.00023	0.00018	0.00021
	표준 편차	전국	0.00354	0.00354	0.00355	0.00353
		서울	0.10388	0.10784	0.10394	0.10588
		기타	0.01006	0.01056	0.01006	0.01032
보건의료 서비스	평균	전국	0.00097	0.00101	0.00097	0.00099
		서울	0.00502	0.00513	0.00502	0.00507
		기타	0.00183	0.00190	0.00183	0.00187
	표준 편차	전국	0.01457	0.01579	0.01459	0.01519
		서울	0.01331	0.01387	0.01331	0.01359
		기타	0.01818	0.01964	0.01821	0.01892
납입금	평균	전국	0.00181	0.00160	0.00180	0.00170
		서울	0.01621	0.01673	0.01621	0.01648
		기타	0.00517	0.00501	0.00516	0.00509
	표준 편차	전국	0.00072	0.00118	0.00071	0.00094
		서울	0.03409	0.03481	0.03409	0.03445
		기타	0.00576	0.00535	0.00576	0.00557

* 상대차이= |(원자료-변환자료)|/원자료

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
교재비	평균	전국	0.00427	0.00441	0.00427	0.00434
		서울	0.01103	0.01158	0.01103	0.01130
		기타	0.00334	0.00342	0.00334	0.00338
	표준편차	전국	0.02348	0.02379	0.02349	0.02364
		서울	0.03212	0.03304	0.03214	0.03257
		기타	0.02283	0.02310	0.02285	0.02297
보충교육비	평균	전국	0.00041	0.00040	0.00041	0.00040
		서울	0.00052	0.00060	0.00052	0.00057
		기타	0.00058	0.00059	0.00058	0.00058
	표준편차	전국	0.00595	0.00702	0.00598	0.00650
		서울	0.02414	0.02760	0.02420	0.02589
		기타	0.00038	0.00013	0.00036	0.00025
문구류	평균	전국	0.00279	0.00289	0.00279	0.00284
		서울	0.00791	0.00830	0.00792	0.00810
		기타	0.00185	0.00190	0.00185	0.00188
	표준편차	전국	0.00395	0.00442	0.00394	0.00419
		서울	0.00826	0.00916	0.00827	0.00871
		기타	0.00227	0.00259	0.00227	0.00244
교통	평균	전국	0.00263	0.00264	0.00263	0.00264
		서울	0.00346	0.00333	0.00346	0.00339
		기타	0.00343	0.00342	0.00343	0.00343
	표준편차	전국	0.01030	0.01043	0.01031	0.01040
		서울	0.00722	0.00568	0.00719	0.00644
		기타	0.01304	0.01295	0.01305	0.01303
통신	평균	전국	0.00162	0.00166	0.00162	0.00164
		서울	0.00052	0.00060	0.00052	0.00056
		기타	0.00179	0.00182	0.00179	0.00181
	표준편차	전국	0.00390	0.00388	0.00392	0.00390
		서울	0.00143	0.00188	0.00141	0.00165
		기타	0.00476	0.00480	0.00478	0.00478

* 상대차이= |(원자료-변환자료)|/원자료

품목	통계량	지역	상 대 차 이			
			삼각분포	절단된 삼각분포	사다리꼴분포	이중삼각분포
조세	평균	전국	0.00530	0.00549	0.00530	0.00540
		서울	0.00203	0.00182	0.00203	0.00193
		기타	0.00645	0.00664	0.00645	0.00655
	표준편차	전국	0.01810	0.01864	0.01812	0.01837
		서울	0.01608	0.01602	0.01607	0.01605
		기타	0.02429	0.02492	0.02431	0.02461
공적연금	평균	전국	0.00146	0.00154	0.00146	0.00150
		서울	0.00164	0.00183	0.00164	0.00174
		기타	0.00143	0.00150	0.00143	0.00147
	표준편차	전국	0.00044	0.00005	0.00046	0.00025
		서울	0.00833	0.00881	0.00837	0.00857
		기타	0.00036	0.00084	0.00035	0.00060
사회보험	평균	전국	0.00167	0.00174	0.00167	0.00171
		서울	0.00062	0.00075	0.00062	0.00069
		기타	0.00185	0.00190	0.00185	0.00187
	표준편차	전국	0.00598	0.00612	0.00601	0.00606
		서울	0.00426	0.00412	0.00428	0.00419
		기타	0.00635	0.00654	0.00637	0.00645
기타 비소비지출	평균	전국	0.00400	0.00420	0.00400	0.00409
		서울	0.00354	0.00426	0.00355	0.00390
		기타	0.00406	0.00419	0.00406	0.00411
	표준편차	전국	0.02099	0.02320	0.02100	0.02188
		서울	0.03830	0.04431	0.03836	0.04133
		기타	0.01758	0.01902	0.01757	0.01804
월말 현금잔고	평균	전국	0.00140	0.00146	0.00140	0.00143
		서울	0.00040	0.00056	0.00041	0.00049
		기타	0.00154	0.00158	0.00154	0.00156
	표준편차	전국	0.01401	0.01338	0.01403	0.01377
		서울	0.00086	0.00093	0.00088	0.00090
		기타	0.01460	0.01393	0.01461	0.01434

* 상대차이= |(원자료-변환자료)|/원자료

제6장

가계금융 · 복지조사 마이크로데이터 제공을 위한 매스킹 방안

박민정·권순필·심규호

제1절 서론

1. 연구배경 및 목적

세계 각국의 통계청은 각종 조사통계에 대한 마이크로데이터를 공표하고 있는데, 이는 집계 자료가 아닌 조사에 응답한 개별 단위의 자료가 통계적 목적을 위하여 활용되도록 하는 것을 의미한다. 때문에 각 국가통계기관은 마이크로자료를 공표할 때 응답자가 식별되어 민감한 정보들이 노출되지 않도록 다양한 노력을 기울이고 있다. 우리나라 통계청도 응답자 정보에 대한 비밀보호[1] 노력의 일환으로 마이크로데이터에 대한 각종 노출제한 기법을 연구 및 적용하여 왔다.

마이크로데이터는 응답한 개별 자료(record)들을 쌓아놓은 집합이며, 각 레코드는 보통 응답자 특성에 관한 변수들과 조사 내용에 관한 변수들로 구성되어 있다. 변수들 중 주민번호나 주소와 같은 응답자의 직접적인 식별정보들을 제거한다 하더라도, 그 외의 공표 대상이 되는 응답자의 특성을 나타내는 변수들의 조합을 통해서 응답자의 식별이 가능할 수 있다. 예를 들어 공표 대상인 특성 변수에 「지역·연령·성별·직업·주택 유형」이 포함되고, 「A지역·50대·여성·군인·연립거주자」라는 특성 변수 값의 조합을 가진 사람이 모집단에서 유일하고 조사에 응했다면 그 레코드는 노출위험이 매우 크다고 할

✍ 이 연구는 2013년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

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

수 있다. 이러한 간접적인 식별력을 가지는 응답자 특성을 나타내는 변수들을 키변수라 한다.

키변수뿐만 아니라 조사 내용에 관한 변수들 중 노출에 민감한 변수들도 비밀보호의 대상이 될 수 있다. 예를 들면 고액 자산가의 총자산이나 소득 등은 응답자가 노출을 극도로 꺼릴 것이다. 이러한 민감변수들이 응답자 특성에 따라 식별되지 않도록 노출 위험을 제어하는 것은 마이크로데이터 공표에 있어서 필수적 조치라 할 수 있다.

공공재로서 마이크로데이터는 각종 연구와 정책개발에 중요한 기반이며 활용 가치도 크다. 하지만 개인의 사생활이 중시되고 국가통계 작성을 위한 각종 조사의 응답률이 나날이 떨어지고 있는 상황 속에서 마이크로데이터를 수집하고 공표하는 것은 점점 어려운 일이 되어가고 있다. 어렵게 수집한 자료에서 생성한 마이크로데이터를 적절한 비밀보호 처리를 하는 것은 사생활의 보호일 뿐만 아니라 응답자의 응답을 장려하는 가장 기본적인 작업이기도 하다. 이러한 비밀보호 처리의 문제를 검토하고자 본 연구에서는 특별히 가계금융·복지조사[2]의 마이이크로데이터를 연구 대상으로 하였다.

가계금융·복지조사의 마이크로데이터는 2012년부터 바로 공표되기 시작하였는데, 노출 위험으로 인해 시도변수는 공표에서 제외되었다. 그러나 시도별 특성에 관한 통계 공급을 위해서 시도별 자산·부채·소득의 집계자료는 보도자료 형태로 별도 공표되어 있다. 따라서 본 연구에서는 공표된 집계자료의 범위에 포함된 시도변수를 마이크로데이터에서도 공표하기 위해 각종 매스킹 기법들을 중심으로 비밀보호 처리 방안을 모색해 보고자 한다.

시도변수를 제외한 마이크로데이터(S1)가 시도변수를 포함할 경우(S2)보다 변수의 개수 자체가 작으므로, 작은 노출위험과 큰 정보손실(작은 자료 유용성)을 가지는 것은 자명하다. 본 연구를 통해 얻고자 하는 것은 S2와 같이 시도변수를 포함하면서 매스킹 처리된 마이크로데이터(M2)이며, 이는 S2보다는 작은 노출위험을 유지하면서 자료 유용성을 확보하는 것을 추구하는 것이다.

시도변수가 포함된 원래 자료인 S2는 노출위험으로 공표될 수 없으며, 이론적으로 M2는 현재 공표 중인 S1을 대체하기 위해 생성되어야 한다. 그러나 S1은 시도변수를 제외한다면 최대한의 자료 유용성을 이용자에게 제공하고 있어 정책적으로 S1의 공표는 유지되어야 하는 상황이다. 만약 S1은 계속 공표하면서 별도로 M2를 작성하기 위한 매스킹 방안을 모색하고자 한다면, M2는 S1과 공통변수를 가지므로 둘을 연계할 가능성이

2) 가계금융·복지조사는 가계부문의 미시적 재무 건전성을 주기적으로 파악하여 정부의 정책이나 금융시스템 발전 등을 위한 기초자료로 활용하기 위해 만들어졌다. 또한, 조사된 가구특성별 자산과 부채에 관한 정보는 각종 경제, 사회, 금융 관련 학문연구의 기초자료로 사용될 수 있다. 가계금융·복지조사는 연간조사이며 2012년에 첫 번째로 실시되었고, 전국을 대표하는 2만 가구를 표본으로 추출하여 횡단 및 패널 분석이 가능하도록 표본설계를 하였다. 참고로, 가계금융·복지조사의 전신은 2010년부터 2회 실시되었던 가계금융조사이며, 가계금융조사의 표본규모와 조사내용을 금융 및 복지부문으로 확대 개편하여 가계금융·복지조사가 만들어졌다(통계청, 2012).

있어 S1을 대체하여 M2만 공표하는 경우보다는 공표변수 범위 선택이나 노출위험 감소에서 한계를 가짐을 밝혀 두고자 한다.

2. 국내 연구 현황

이제 현재까지 우리나라 마이크로데이터 비밀보호의 연구 현황을 소개하며 서론을 마치고자 한다. 우리나라 마이크로데이터 비밀보호 연구는 주로 통계청을 중심으로 이루어져 왔다. 최초의 연구는 인구주택총조사 마이크로데이터의 노출제한을 위한 것으로 2006년과 2007년에 이루어졌다. 인총자료의 노출제한 연구에서는 특성 변수들을 중심으로 유일성에 의한 식별이나 노출위험에 관한 개념이 소개되었고, 유일성에 근거한 노출위험 측정이 이루어졌다(정동명과 강동환, 2006; 정동명과 정미옥, 2007).

한편, 같은 시기에 가계조사의 마이크로데이터 비밀보호를 위한 연구도 이루어졌는데 주로 민감정보에 대한 개념 소개와 잡음을 이용한 노출제한 기법들을 적용한 결과를 보여주고 있다(정동명 외, 2007; 정동명과 김경미, 2008). 이는 응답자 특성의 유일성에 국한되었던 연구 내용을 조사 내용의 자료 구조를 보존하면서 노출을 제한하는 영역으로 확장시킨 의의를 가진다. 즉 노출위험에 국한되었던 국내 비밀보호 연구를 자료 유용성 관점까지 다루도록 확장하였다고 할 수 있다. 그 외에 각종 해외사례에 관해 정리한 보고서들을 통해 어떠한 마이크로데이터의 비밀보호 기법들이나 행정적 절차가 해외에서 사용되고 있는지 참고할 수 있다(김경미 외, 2007; 윤연옥, 2010).

한편, 본 연구의 대상인 가계금융·복지조사에 관해서는 2012년의 연구를 통해서 특성 변수들의 유일성에 근거해 노출위험을 측정하여 시도변수 제공의 위험성이 확인된 바 있다(김경미와 임경은, 2012). 다음 <표 6-1>은 이상의 연구 내용을 정리한 것이다. 표의 R-U map이란 노출위험-자료 유용성 지도로 노출위험과 자료 유용성(정보손실)에 관련된 각 수치를 그림으로 표현한 것을 말한다.

〈표 6-1〉 국내 마이크로데이터 비밀보호 연구 현황

연구 시기	대상 자료	신개념 소개	주요 방법론	측정 기준
2006년	인총 2%표본(충남)	식별, 유용성	그룹화	노출위험
2007년	인총 2%표본(전국)	노출위험	재코딩	노출위험
2007년	가계조사	민감정보	그룹화, 반올림	자료 유용성
2008년	가계조사	민감정보	승법잡음	자료 유용성
2012년	가계금융·복지조사	R-U map	재코딩	노출위험

이러한 연구 흐름에 이어 본 연구에서는 가계금융·복지조사 마이크로데이터의 비밀보호를 위해, 먼저 2절에서 전반적인 마이크로데이터 노출제어 방향들과 각종 매스킹 기법들을 개괄한다. 이어 3절과 4절에서는 가계금융·복지조사에 매스킹 기법들을 적용하여 비밀보호가 이루어진 정도를 노출위험과 자료 유용성에 근거하여 평가한다. 마지막으로 5절에서는 향후 연구방향을 논하며 결론을 맺는다. 각종 매스킹 기법 적용을 위한 프로그램으로는 R 패키지 sdcMicro(Templ, 2008)를 사용하였다.

제2절 통계적 노출제어와 매스킹

1. 노출제어의 종류

각종 마이크로데이터의 노출을 제어하는 것은 두 방향에서 가능하다. 다음 <표 6-2>에서는 최종 제공하는 자료 형식에 따라 노출제어 방법들을 분류하였다.

〈표 6-2〉 노출제어의 종류와 제공되는 자료의 형식

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

SDL: Statistical Disclosure Limitation

만약 이용자들이 자료를 통계적 목적으로만 이용하였고 생산된 결과물에 응답자의 정보 노출위험이 없다면, 자료는 노출에서 안전하다고 할 수 있다. 이러한 제어는 물리적으로도 가능한데 크게 결과물을 통제하거나 이용자의 접근을 제어하는 방법들이 있을 수 있다. 우선 결과물을 통제하는 방법으로는 이용자의 요구에 따라 통계기관이 직접 통계표를 작성하여 서비스하거나 이용자의 접속을 원격으로 제어하고 그 결과물을 검열 후 제공하는 것이 있다. 한편 이용자의 접근을 물리적으로 제어하는 비밀보호 방법으로는 인가파일을 제공하거나 데이터 실험실을 운영하는 것을 들 수 있다. 인가파일이나 데이터 실험실의 두 가지 예를 소개하자면 다음과 같다(김경미 외, 2007).

먼저 미국 노동통계국은 노동통계국과 자료 이용자의 소속기관 사이의 기관 대 기관 계약을 체결한 후 자료를 제공한다. 계약 후 자료 이용자 소속기관 책임자의 데이터 반환 및 소거 확인 절차, 자료 제공자인 노동통계국에 의한 성과물 검열 절차, 이용 기간

동안의 노동통계국 직원에 의한 현장 사찰 등의 과정을 거친다. 이렇게 제공하는 파일을 인가파일(licensed file)이라고 한다.

다음으로 미국 센서스국에서 운영하는 데이터 실험실을 살펴보자. 미국 센서스국은 9개의 센터를 운영하고 있고, 자료 이용자는 가까운 데이터 실험실로 찾아가서 센터의 서버에 접속해 모든 필요한 작업을 완결한다. 이용자가 자료의 추가 분석을 필요로 할 경우 다시 이용자가 센터로 이동하는 것이 필요하다. 이렇게 얻은 최종 성과물은 심사위원회의 승인 후 학술지 투고가 허용되며, 전체적으로 자료 이용 제안서 제출부터 데이터 이용까지 총 6~12개월이 소요될 것을 예상해야 한다.

이상의 물리적 제어 방법들은 이용자와 자료 제공자 모두에게 복잡한 절차를 요구한다. 자료 제공자가 이용자의 결과물을 대신 생산하거나 혹은 일일이 검열하는 것이나, 계약 등의 행정 절차들이 자료 요구 건마다 수반되는 것, 그리고 자료 이용자가 특정 장소로 이동하는 것 등은 모두 쉽지 않으며 많은 비용이 발생하는 일들이다. 이에 따라 각국 통계기관은 불특정 다수가 통계적으로 이용할 수 있으나 응답자의 정보가 식별되지 않는 공공이용파일(Public Use Microdata File)을 제공하여 마이크로데이터에 대한 수요를 좀 더 많이 충족시키고자 노력하고 있다. 마이크로자료를 공공이용파일로 공표할 때 추구되는 목표는 크게 다음 세 가지가 있다(Reiter, 2004).

- 응답자의 식별정보나 민감정보를 알려는 외부 시도에 대한 **안전성**
- 광범위한 통계적 분석을 뒷받침할 수 있는 **정보의 충분성**
- 보편적인 통계적 방법론들을 사용하는 **이용자의 편의성**

위의 세 목표를 만족시키며 공공이용파일을 작성하기 위해 사용되는 방법들로 크게 매스킹 처리와 인위자료의 활용을 들 수 있다. 매스킹(masking)이란 원래 자료에 적절한 변환을 통해 간접적인 식별 정보를 가리는 것으로 매스킹된 자료란 변환된 자료와 변환에 관한 정보 모두를 의미한다. 이때 변환에 관한 정보는 자료 이용자가 변환된 자료를 적절히 분석하는데 필요하다. 매스킹 기법들을 분류해 보면 부분적으로 자료값을 제공하거나(sampling), 자료값을 가리거나(suppression), 구분할 수 없는 그룹으로 값을 섞는 것(aggregation), 그리고 잡음 추가(noise addition)나 자료값 간 교환(swapping) 등을 통해 변조적인 값(perturbative data)을 제공하는 것이 있다(Duncan 외, 2011).

이와 같은 여러 가지 매스킹 기법들을 자료에 적용했다면 그 중 어느 방법이 효율적인지를 판단하기 위한 기준이 필요하다. 보통은 노출위험과 정보손실(자료 유용성) 정도를 계산하여 실제로 사용할 매스킹 기법을 선택하는데 이용한다. 노출위험은 주로 응답자의 특성을 나타내는 키변수 조합의 유일성이나 응답자가 응답한 민감변수들의 매스킹 전후

거리에 근거해 계산한다. 정보손실은 매스킹 전후의 자료 간의 거리나 고유벡터 값들 사이의 거리를 이용해 계산한다. 이때 노출위험은 매스킹 전후의 거리가 짧을수록 정보손실은 그 거리가 길수록 커져서 서로 상충적인(trade-off) 관계가 있으며, 노출위험과 정보손실 각 측도에서 동시에 좋은 결과를 얻을 수는 없다. 또한, 매스킹 기법에 대한 이용자의 숙지도 필요하여 이용자의 편의성에서도 부족한 점을 가진다. 그러나 현실적으로 최근까지 다른 적절한 제안이 충분한 실용성을 가지고 제시되지는 않아, 자료의 비밀보호를 위해 매스킹 기법이 널리 이용되어 왔다.

매스킹 기법은 자료마다 적합한 기법이 다르고 위에서 언급한 노출위험과 정보손실 사이의 상충관계가 언제나 존재하여 어떠한 매스킹 기법을 적용한다고 해도 공공이용파일이 추구하는 세 가지 목표를 동시에 달성하기 어려운 한계를 가지고 있다. 또한 시간에 따라 차원이 증가하게 되는 종단자료에 대해서는 매스킹 적용이 사실상 불가능한 단점도 가지고 있다. 이러한 한계를 근본적으로 뛰어넘기 위해서 인위자료(synthetic data) 접근법이 최근 발전하였다.

인위자료 접근법은 ①공표 자료 내 어떤 개체도 원자료 내 민감변수의 값을 가지지 않고, ②원자료에서 가능한 통계 분석/추론이 공표 자료에서 가능할 것을 추구한다(Rubin, 1993). 이것은 노출위험을 0으로 둔 상태에서 자료 구조의 궁극적인 보존을 추구하는 것을 의미한다고 할 수 있다. 인위자료 접근법의 기본 아이디어는 표본 조사 등을 통해 모집단을 대표하는 다수의 표본 추출이 가능한 것처럼 조사된 표본과 구조가 유사한 인위자료 세트들을 통계적으로 생성하자는 것이며, 이는 베이지안적 사고를 기반으로 하는 것이다.

인위자료 관련 초기 연구로는 2001년 프랑스의 종단 연계 자료에 관한 사례가 있다. 이는 자료 수집 기관 및 표본 추출 방법이 서로 다른 여러 자료들을 종단적으로 연계하여 공표하고자 기존의 매스킹 기법이 아닌 인위자료 방법론 도입을 처음 시도한 것으로(Abowd와 Woodcock, 2001) 이러한 접근법을 완전 인위자료 생성 방법이라 한다. 보통 인위자료를 활용하는 과정은 ①모집단에서 개체를 추출해 인위자료 세트(들)을 구성하고 ②원래의 조사 자료를 기반으로 만든 모형을 가지고 다중대체(multiple imputation) 기법(Rubin, 1993)을 사용해 위의 인위자료 세트를 채우며 ③여러 개의 인위자료 세트들을 공표하고 ④각 세트들에서 얻은 통계량을 결합하는 방법(combining rule)을 제시하는 단계를 가진다.

완전 인위자료 이용은 통계적 노출제어에서 획기적인 아이디어로 간주되며 매스킹 기법만으로는 성취하기 어려웠던 공공이용파일 생산의 궁극적인 목표들을 추구할 수 있는 길을 열어주었다. 일반적으로 사용자 모형의 정확도가 높을 때 노출위험이 높아지는데 인위적으로 덜 정확한 인위자료 생성 모형 사용을 통해 노출제한이 가능하며, 조사된 자료의 분포와 통계적으로 유사한 분포에서 인위자료를 생성하여 다양한 추론에 대해

통계적 타당성을 제공하고, 또한 소지역을 포함한 지리적 정보 공표 가능성도 가지는 장점이 있다. 나아가 비밀보호기법과 관련된 새로운 통계 이론이나 프로그램 습득이 불필요하고 원래 자료의 표본추출법에 관한 고려 없이 단순 추출을 가정하고 통계 분석이 가능하여 이용자의 편의성을 크게 높이기도 한다. 단점은 인위자료를 만드는 모형 및 가정에 지나치게 의존적이며 추정량의 편의가 큰 것으로 어떤 통계기관도 완전 인위자료 접근법을 전적으로 받아들이지 않았다. 완전 인위자료 접근법은 준모수적 및 비모수적 모형 탐구, 다양한 구조를 가지는 모집단에 대한 추론 타당성 연구 및 부분 인위자료 활용 연구로 발전하는 모습을 보인다.

최근 실제 자료에 적용되기 시작한 부분 인위자료 접근법은 완전 인위자료 접근법과 기본적으로 유사하나, 조사 자료 중 노출위험이 큰 정보만을 선택해 인위자료로 변환하여 일부 정보에 대해서만 자료를 대체함으로써 모형 의존성을 낮추는 효과를 가진다는 점이 다르다. 연구 사례로는 독일 IAB 기관 패널 조사(Drechsler와 Reiter, 2009) 및 미국 사업체 종단자료(Kinney 외, 2011)의 비밀보호가 있다. 이상 인위자료의 이용을 통한 비밀보호는 본 연구의 범위를 넘어 향후 연구방향으로(박민정과 김경미, 2013) 고려하기로 하며 본 연구는 매스킹 기법들을 이용한 비밀보호에 초점을 맞추어 진행되었음을 밝혀둔다.

2. 매스킹 기법의 종류

앞에서 언급한 대로 마이크로데이터의 비밀보호는 물리적 접근 및 통계적 접근을 통해서 이루어질 수 있고, 통계적 접근의 비밀보호 방안으로는 매스킹 기법과 인위자료의 활용을 들 수 있다. 이러한 비밀보호 방안들 중, 본 연구는 매스킹 기법을 우리나라 공표 자료 중 하나인 가계금융·복지조사 마이크로데이터에 적용하는 것에 초점을 맞추어 진행되었다. 수많은 매스킹 기법들 중 세계적으로 가장 보편적으로 활용되고 있는 것들을 중심으로 우리나라 자료에 적용한 후 비밀보호 정도를 살펴보았다.

매스킹이란 원자료에 적절한 변환을 통해 간접적인 식별정보를 가리는 것이다. 매스킹된 자료란 변환된 자료와 변환에 관한 정보 모두를 의미한다. 변환에 관한 정보는 이용자의 분석이 올바르게 진행되는데 필요하므로 제공되어야 한다. 매스킹 기법들은 크게 네 가지 범주인 ①부분적으로 자료값을 제공하거나(sampling) ②자료값을 가리거나(suppression) ③구분할 수 없는 그룹으로 값을 섞어 제공하거나(aggregation) ④자료값을 변조하는 것으로(perturbative data) 나누거나(Duncan 외, 2011), 연속형 혹은 범주형 등 변수의 형식에 따라서도 분류하여 볼 수도 있다.

이 절에서는 비교적 널리 사용되고 있는 매스킹 기법들을 하나씩 간단히 살펴보고자 한다. 매스킹 기법들은 세부 종류가 매우 많고, 각 기법별로 전문적으로 방대하게 연구

되어 있어 자세한 내용을 여기에서 모두 다루기는 불가능하다. 자세한 내용에 관심이 있는 독자들은 필요에 따라 기법별 관련 논문들을 참고하길 권한다. 더 많은 종류의 마이크로데이터의 매스킹 기법들에 대해 좀 더 전반적으로 알기 위해서는 Duncan 외 (2011)의 5장 및 관련 참고문헌들이나 통계자료 비밀보호론(충남대) 등을 참고할 것을 추천한다.

국소 감추기(local suppression)

감추기(suppression)는 보통 레코드나 속성 차원에서 이루어진다. 예를 들면, 특정 금액 이상의 고소득자의 자료를 모두 감추거나(레코드 감추기), 레코드의 식별을 간접적으로 도울 수 있는 지역 변수를 모든 레코드에 대해 감추는 것(속성 감추기)을 생각할 수 있다. 그러나 이러한 종류의 감추기는 너무 많은 정보의 손실을 야기한다.

국소 감추기(local suppression)는 노출위험을 높이는 키변수 조합의 몇 개의 값만을 감추는 것을 말한다. 예를 들면, 한 레코드 값에 대해 변수 조합 지역·연령·성별·직업·주택유형이 A지역·50대·여성·군인·연립거주자일 때 성별이나 직업을 감추는 것을 생각할 수 있다. 만약 자료 이용자가 직업에 관한 정보를 더 필요로 할 경우에는 직업 대신 성별을 감추는 것이 유익할 것이다. 이러한 국소 감추기는 키변수 특정 조합의 레코드 수가 적당히 많도록 익명성을 획득하기 위해 이용할 수 있다. 예를 들면 같은 키조합을 가지는 레코드의 수인 익명성(k-anonymity)이 2 이상이 되도록 국소 감추기 기법을 활용할 수 있다. 이 때 되도록이면 감추어지지 않기를 원하는 변수를 선택하는 알고리즘도 sdcMicro에 구현되어 있다. 본 연구에서는 다른 매스킹 기법들을 모두 적용한 후에 국소 감추기를 키변수에 대해 마지막으로 적용하여 익명성 획득에 활용하고자 하였다.

전반적 재코딩(global recoding)

전반적 재코딩이란 특정 속성의 범주를 더 상위 범주(less-specific categories)로 묶는 것을 말한다. 예를 들면 직업이 의사와 간호사일 경우 모두 의료계 종사자로 묶는 것을 생각할 수 있다. 연령 변수도 마찬가지여서 각 세 연령을 5세나 10세 단위로 묶어 익명성을 확보할 수 있다. 전반적 재코딩 중 특정 값 이상/이하를 묶는 것을 top/bottom 코딩이라 하며 그 그룹에 속하는 값들을 해당 그룹의 평균이나 중앙값으로 대체하여 표시하곤 한다. 본 연구에서는 각 키변수에 대해서 빈도수가 너무 적은 범주의 그룹이 생기지 않도록 전반적 재코딩을 하였다. 이에 대한 자세한 내용은 키변수의 매스킹을 다루는 제4절에서 기술하도록 한다.

국소통합(microaggregation)

국소통합[3]이란 자료를 최소 k개(보통 3개) 이상의 레코드를 가지는 그룹들로 묶고, 각 그룹의 자료 값들을 그 그룹의 평균이나 중앙값 등의 동일한 한 값으로 대체하는 기법을 말한다. 레코드를 묶어 그룹을 만들 때, 단일 변수를 기준으로 비슷한 값을 가지는 그룹들을 만들 수도 있고, 여러 변수를 모두 고려하여 그룹을 정할 수도 있다. 여러 변수를 고려할 때는 주성분을 이용하거나, 네델란드 통계청에서 제안했던 자료 간 거리가 먼 것부터 묶어가는 MADV 알고리즘을 이용할 수도 있다(Statistics Netherlands, 2007). 관심 있는 독자는 다양한 종류의 그룹화 기법이나 그룹 내 자료의 수 k의 결정에 관해 비교한 연구(Mateo-Santz와 Domingo-Ferrer, 1998)도 있으니 참고하기 바란다. 여기에서는 각 방법들을 자세히 서술하는 것은 생략하며, 제3절 민감변수의 매스킹 방안에서 현재 구현되어 있는 그룹화 방법들을 다양하게 적용하고 노출위험-자료 유용성 측도를 기준으로 각 국소통합 기법 혹은 그룹화 방법들을 비교하도록 한다.

잡음 추가(noise addition)

잡음 추가 기법은 연속형 변수의 자료에 적용되는 자료 변조적(perturbative) 기법 중 하나이다. 이는 통계적 잡음을 원래의 자료에 더해서 특정한 변수의 정보가 공유된 다른 자료와의 정확 매칭(exact matching)을 피하기 위해 사용된다. 세부적으로는 먼저 원자료의 평균과 분산을 가지는 정규 분포로부터 무상관(uncorrelated) 잡음을 발생시켜 더하는 방법을 생각할 수 있다[4]. 이 경우 원래 자료의 공분산 구조를 보존하지 않는다. 다음으로, 원래 자료의 공분산 행렬에 비례하는 공분산을 가지는 분포에서 잡음을 발생시켜 원래 자료에 더할 수 있는데 이를 상관(correlated) 잡음 기법이라 한다(Brand, 2004). 상관 잡음의 경우에도 정규성 가정이 성립하지 않는다면 매스킹 작업 후에 자료의 구조가 심각하게 왜곡될 수도 있다. 이에 따라 상관 잡음 기법의 로버스트한 알고리즘이 제시되었고(Templ과 Meindl, 2008), 이 기법 역시 sdcMicro 패키지에 구현되어 있다. 그 외에 자료의 크기를 고려하거나 자료 중에서 이상값들만 골라서 잡음을 더하는 알고리즘들도 존재한다.

3) microaggregation에 관한 정확한 통계용어가 정의되어 있지 않아 여기서는 임의적으로 국소통합이라 번역하였음을 밝혀둔다.

4) 통계개발원의 기존 연구에서는 무상관 가법 잡음 기법을 이용할 때 발생하는 문제점을 극복하기 위해 승법 잡음 모형이 연구되었으나(정동명과 김경미, 2008) 본 연구에서는 승법 잡음 모형은 배제하였고, 대신 무상관 가법 잡음 기법을 뛰어 넘기 위해 제시된 상관 가법 잡음 기법을 받아들여 연구를 진행하였음을 밝혀 둔다.

자료순위 교환(rank swapping)

일반적으로 자료 교환(data swapping)은 범주형 변수를 대상으로 한다. 자료순위 교환 기법은 자료 교환 기법의 확장으로 이를 연속형 변수에도 적용할 수 있는 것으로 알려져 있다. 즉 자료순위 교환이란 자료를 정렬하고 제한된 범위 내에서 자료를 서로 교환하는 것으로 본 연구에서는 위아래 1%를 제한 범위로 정하여 자료순위 교환 기법을 적용하였다. 자세한 내용은 Moore(1996)를 참고하면 된다.

자료섞기(shuffling)

자료섞기[5] 기법은 연속형(numerical) 자료를 교환(data swapping)하는 방법을 확장시킨 것으로 매스킹된 자료를 생성하기 위해 조건부 분포[6]를 이용한다(Muralidhar와 Sarathy, 2006). 조건부 분포를 이용한다는 것은 표본으로부터 모집단의 분포를 모수적으로 추정하여 새로운 자료 세트를 생성하고자 하는 접근으로 설명될 수 있다. 이와 관련하여 정규분포 가정을 확장하여 치우친 t분포(skewed-t distribution)를 가정하여 새로운 자료 세트를 생성하는 기법이 연구되기도 하였다(Lee 외, 2010). 이러한 연구 내용들은 기법의 특성상 매스킹 기법과 인위자료 활용 사이에 위치하고 있다고도 판단될 수 있으며, 이들에 대한 좀 더 엄밀한 검토가 향후 필요할 것으로 생각되나 여기서는 연구의 범위를 넘는 것으로 간주하였다. 본 연구에서는 자료섞기 세부 기법들 중 일반적 가법 자료 변조(GADP[7])의 결과물을 이용할 경우와 정규 코풀라(normal copula)를 이용할 경우의 알고리즘을 마이크로데이터 매스킹을 위해 이용하여 보았다.

이상 본 연구에서 설명하는 대부분의 매스킹 기법들은 통계 소프트웨어 R의 sdcMicro라는 패키지에 구현되어 있다(Templ, 2008). 이 sdcMicro 패키지에는 네덜란드 통계청에서 개발한 선도적인 마이크로데이터의 비밀보호를 위한 프로그램인 μ-Argus에서 구현된 기법들이 최신의 내용까지 반영되어 모두 담겨있을 뿐만 아니라, 그 외의 다른 최근 매스킹 기법들까지 추가로 구현되어 있다. 사용자의 편의성을 위해 μ-Argus와 같은 메뉴 형식의 그래픽 인터페이스(GUI)도 구성되어 있으며, 명령어 형식일 때는 입력 자료를 행렬 형태나 패키지 내에서 정의하는 sdc 개체(object)로 처리하기도 하므로 마이크로데이터 비밀보호 담당자들이 필요에 따라 적절히 사용할 수 있다.

---

5) shuffling에 관한 정확한 통계용어가 정의되어 있지 않아 여기서는 임의적으로 자료섞기라 번역하였음을 밝혀둔다.

6) 자료 교환 기법은 자료 변조적(perturbative) 기법으로 주로 언급되나 자료섞기 기법은 자료의 분포를 이용하므로 자료 변조(perturbation) 기법으로 구분되기도 한다.

7) General Additive Data Perturbation

제3절 민감변수의 매스킹 방안

가계금융·복지조사의 마이크로데이터는 응답자 특성을 나타내는 성별, 연령 등의 키변수와 민감한 응답 내용으로 구성된 자산, 부채, 소득과 같은 민감변수로 구성되어 있다. 이 절에서는 민감변수들의 매스킹 방안을 다루고자 한다. 고려한 매스킹 기법은 국소통합, 잡음 추가, 자료순위 교환, 자료섞기이며, 각 방법을 적용한 후 노출위험 및 정보손실 측도를 산출하여 비교하였다. 이러한 방법들 중 국소통합과 잡음 추가 기법들이 노출위험과 정보손실 측면에서 절대적 우위를 보여 그 결과를 상세히 살펴보기로 한다. 또한 노출위험과 정보손실 사이의 상충관계를 감안하여 정보손실 증가와 노출위험 감소 효과를 가지는 국소통합과 잡음 추가의 결합안도 함께 검토하고자 한다.

본 연구에서 고려한 민감변수의 종류는 <표 6-3>과 같으며 이는 보도자료 형태로 공표된 2012년 집계자료의 민감변수와 동일한 것들이다. 각 상위변수는 하위변수들의 합으로 구성되므로 (예. 자산총액 = 실물자산 + 금융자산) 본 연구에서는 <표 6-3>의 변수명이 굵게 표시된 하위분류 12개에 대해 매스킹 작업을 하며, 상위변수의 매스킹된 값들은 매스킹된 하위변수들 값의 합으로 결정되도록 하였다. 즉, 본 연구에서는 가계금융·복지조사 마이크로데이터 12개 민감변수로 구성된 2만여 개 레코드에 대해 관련 매스킹 기법들을 적용하여 노출위험과 정보손실 측도를 계산하였다.

〈표 6-3〉 매스킹에 고려된 가계금융·복지조사 마이크로데이터의 민감변수

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**				

1. 매스킹 기법별 노출위험과 정보손실 측정

민감변수의 매스킹 기법들 중 잡음 추가 기법을 적용할 때, 음이 아닌 값을 가지는 변수들이 잡음 추가에 의해 음의 값을 가지는 경우가 있게 된다. 예를 들면, 실제로 소득이 없거나 무응답에 의해 소득을 0이라고 응답한 값들이 잡음 추가 매스킹 기법에

의해 음의 값의 소득을 가질 수 있게 된다. 따라서 매스킹 기법별로 노출위험과 정보손실을 측정하기 이전에 음의 값 발생 문제를 어떻게 처리할지 고민해야 한다. 본 연구에서는 음의 값 처리에 대해 다음의 세 가지 시나리오를 고려하였다.

첫 번째는 잡음 추가 후, 발생한 음의 값들을 모두 절대값으로 바꾸어주는 것이다. 이 경우, 0으로 응답한 경우 또는 무응답[8]에 대한 모든 정보가 손실되는 단점이 있다. 즉, 이렇게 매스킹 처리된 자료에서는 0인 값이 거의 존재하지 않을 수 있다. 이는 원래의 마이크로데이터와 매우 큰 차이이며 노출위험은 적겠지만 정보손실이 매우 크다고 할 수 있다.

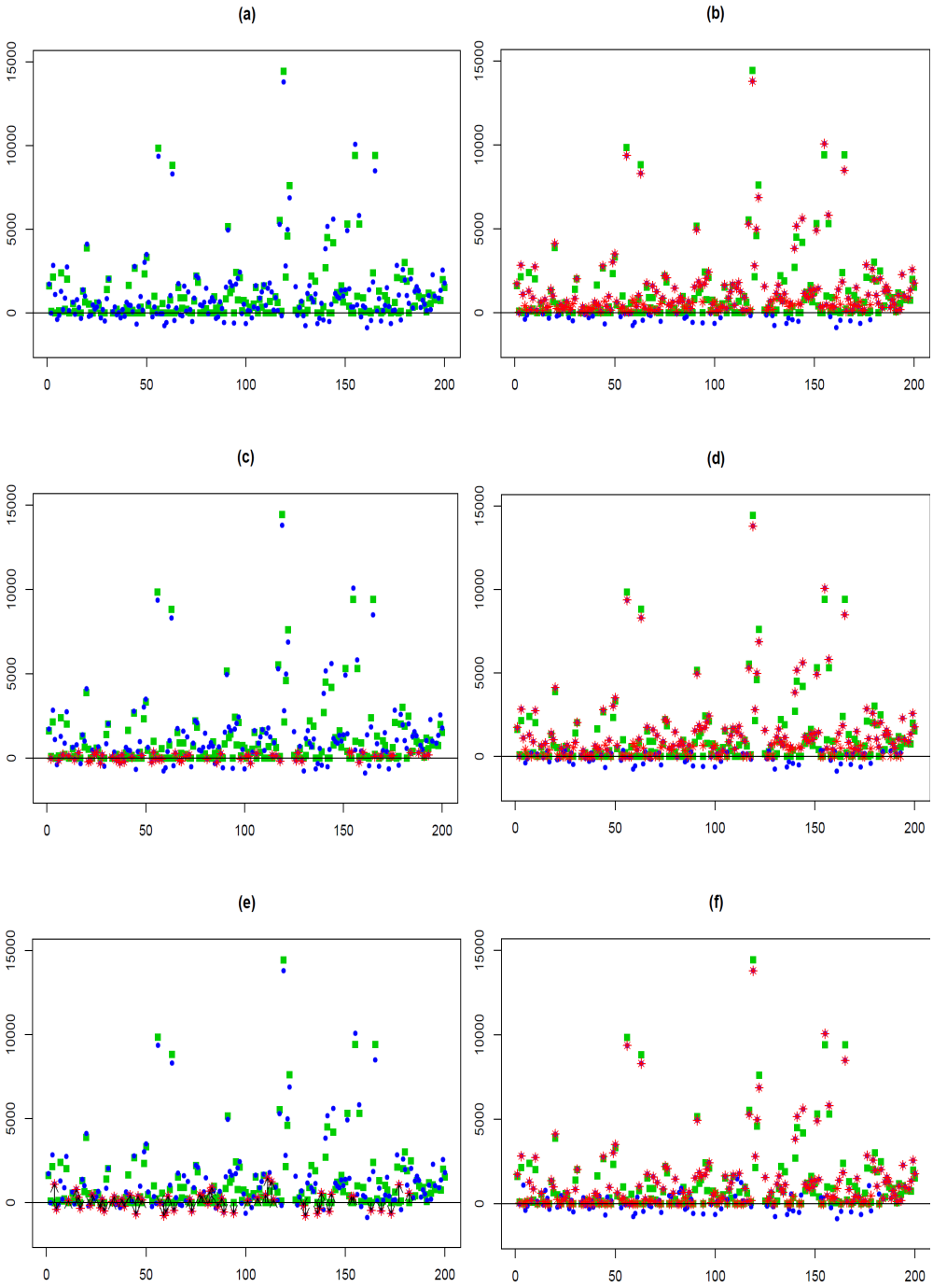
두 번째 시나리오는 원래 마이크로데이터의 무응답 자료의 개수를 유지하는 것으로 잡음 추가 후에 절대값이 작은 순서대로 원래 무응답 개수만큼 0으로 바꾸어 주는 것이다. 이 경우 원래 마이크로데이터 무응답 자료의 개수라는 정보가 보존된다. 마지막 세 번째 시나리오는 원래 무응답 정보를 모두 그대로 보존하는 것이다. 이 경우 원래 마이크로데이터에서 0인 값들은 모두 0으로 환원시키고 나머지 잡음 추가로 인해 음의 값이 발생한 경우 절대값을 취하도록 한다. 세 번째 시나리오에 의해서는 정보손실은 최소이겠으나 노출위험은 그만큼 커질 수밖에 없게 된다.

다음 [그림 6-1]은 이러한 세 가지 시나리오를 그림으로 표현한 것이다. 이해를 돕기 위해 한 변수에 대해서 시나리오별 매스킹 결과를 200개의 표본[9]을 이용해 나타내었다. [그림 6-1]의 (a)에서는 원래 마이크로데이터를 초록색 네모로, 잡음을 추가하여 매스킹한 결과는 파란색 원으로 표현하였다. [그림 6-1]의 (a)~(f) 모두 한 번의 잡음 추가된 자료를 이용해 그려 초록색 네모와 파란색 원은 모두 동일함을 밝혀 둔다. 한편, 잡음 추가 후 절대값을 취한 것은 [그림 6-1]의 (b)에서 붉은색 별로 표시하였다. 따라서 잡음 추가를 하여도 음의 값이 아닌 자료들은 원위에 별모양이 덮여 그려졌으며, 잡음 추가를 통해 음의 값으로 매스킹된 것들은 원으로 그림에 남아있고, 이들의 절대값들은 0보다 큰 영역에 별모양으로 표현되어 있다. 결과적으로 첫 번째 시나리오에 의해서는 네모로 표현된 원래 자료들이 원으로 표시된 잡음 추가된 자료들이 아니라, 잡음 추가 후에 절대값을 취한 결과인 별 모양의 자료들로 변환됨을 알 수 있다.

---

8) 이하 0으로 응답한 경우 또는 무응답을 묶어서 무응답 자료라 부르도록 하겠다.

9) 이 200개의 표본은 임의로 추출된 것이며, 무응답 자료 처리에 대한 시나리오를 설명하기 위해 이 그림에서만 사용하였다.



[그림 6-1] 잡음 추가 기법 적용과 무응답 정보 처리에 대한 세 가지 시나리오

두 번째 시나리오는 [그림 6-1]의 (c)와 (d)에 나타나있다. 원래 무응답의 개수를 k개라고 하면, 잡음 추가 후에 절대값이 작은 순서대로 k개의 자료가 별 모양으로 [그림 6-1]의 (c)에 그려져 있다. 이들을 모두 0으로 처리하고, 나머지 자료에 대해서 매스킹 처리 후의 절대값을 별 모양으로 표현한 것이 [그림 6-1]의 (d)이다. 즉, [그림 6-1]의 (d)가 두 번째 시나리오의 결과를 보여준다고 할 수 있다. 이 경우 최종 매스킹된 자료들을 크기 순서대로 정렬하면 0인 값과 다음 값 사이에 일정한 간격이 존재하게 되는 특징이 생기게 된다.

세 번째 시나리오는 무응답 정보를 그대로 보존하는 것이었다. [그림 6-1]의 (e)는 화살표와 별 모양으로 원래 0인 자료들에 대하여 잡음 추가 후 변환된 결과를 보여주고 있다. 이들을 원래 값인 0으로 환원시키고, 나머지 값들의 잡음 추가 이후 절대값을 별 모양으로 그린 것이 [그림 6-1]의 (f)에 나타나있다. 이러한 세 번째 시나리오로 자료를 매스킹 처리할 경우, 무응답 정보가 그대로 보존되어 노출위험이 높아지나 무응답과 관련된 정보손실은 작아지게 된다.

이상으로 잡음 추가 기법을 사용할 경우 발생할 수 있는 음의 값 처리에 대한 세 시나리오를 살펴보았다. 이러한 잡음 추가 기법을 포함하여 민감변수의 비밀보호를 위해 본 연구에서 고려한 각 매스킹 기법들을 세부 알고리즘별로 정리하면 다음 <표 6-4>와 같다. 국소통합은 같은 값을 대체할 그룹을 결정하는 알고리즘별로, 잡음 추가는 공분산 정보 보존 여부나 자료 크기 정보 반영(restr) 등에 따른 알고리즘별로 나열되어 있다. 또한, 자료순위 교환은 상하 0.1 퍼센트 기준으로, 자료섞기는 일반적 가법 자료 변조(GADP[10])의 결과물을 이용할 경우(sh)와 정규 코풀라(normal copula)를 이용할 경우(sh.mvn)에 따른 각각의 세부 기법들을 나타내고 있다. 표에 나열된 것들 이외의 알고리즘도 다수 존재하나 안정성의 문제로 본 연구에서는 고려하지 않았다.

〈표 6-4〉 민감변수를 위한 매스킹 기법 및 세부 알고리즘

국소통합 (microaggregation)	잡음 추가 (noise addition)	자료순위 교환 (rank swapping)	자료섞기 (shuffling)
single	additive	0.1 percent	sh
simple	correlated		sh.mvn
onedims	correlated2		
mdav	restr		
influence			
pca			
clustpca			

10) General Additive Data Perturbation

이들 세부 알고리즘별로 노출위험과 정보손실 측도 값들을 살펴보면 다음 <표 6-5>와 같다. 앞에서 언급했던 것처럼 정보손실1(il1)은 자료들의 물리적 거리에 근거하여 얻어진 값인 반면에 정보손실2(il.eigen)는 고유벡터를 이용한 거리에 근거하여 얻어진 값을 나타낸다. 본 연구에서는 이상의 노출위험과 정보손실 측도 결과 값들에 대해 값 자체에 의미를 부여하기보다는 각 기법들의 상대적 비교를 위해서만 사용함을 강조하여 밝혀 둔다.

〈표 6-5〉 세부 알고리즘별 노출위험과 정보손실 측도 결과

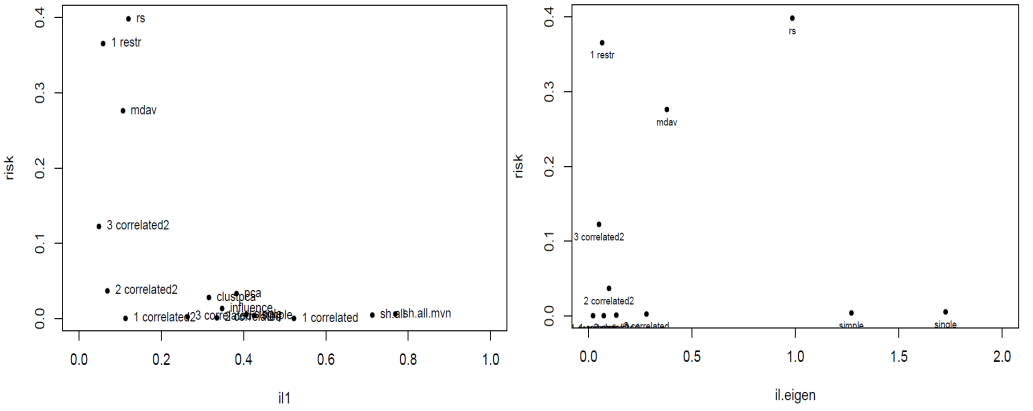
매스킹 기법	세부 알고리즘	노출위험 (risk)	정보손실1 (il1)	정보손실2 (il.eigen)
국소통합 (microaggregation)	single	**0.01**	0.41	1.73
	simple	**0.00**	0.43	1.27
	onedims	0.99	**0.00**	0.12
	mdav	**0.28**	**0.11**	0.38
	influence	**0.01**	0.35	2.18
	pca	**0.03**	0.38	3.75
	clustpca	**0.03**	0.32	3.01
잡음 추가1 (noise addition1)	1 additive	**0.00**	5.64	6.81
	1 correlated	**0.00**	0.52	0.02
	1 correlated2	**0.00**	**0.11**	0.07
	1 restr	0.37	**0.06**	0.07
잡음 추가2 (noise addition2)	2 additive	**0.00**	3.91	6.86
	2 correlated	**0.00**	0.34	0.13
	2 correlated2	**0.04**	**0.07**	0.10
	2 restr	0.57	**0.04**	0.05
잡음 추가3 (noise addition3)	3 additive	**0.00**	2.28	3.69
	3 correlated	**0.00**	0.26	0.28
	3 correlated2	**0.12**	**0.05**	0.05
	3 restr	0.78	**0.02**	0.04
자료순위 교환 (rank swapping)	rs	0.40	**0.12**	0.99
자료섞기 (shuffling)	sh	**0.00**	0.71	3.16
	sh.mvn	**0.01**	0.77	6.15

한편, <표 6-5>에 나타난 매스킹 기법의 잡음 추가 1, 2, 3은 앞에서 설명한 무응답 자료의 처리에 관한 시나리오 1, 2, 3에 해당하는 것이다. 표에서는 노출위험이 0.3이하이고 정보손실1 측도[11] 값이 0.15보다 작은 것들에 대해 해당 수치를 굵게 표시하고, 두 측도 모두 이러한 값들보다 작은 경우는 세부 알고리즘명을 굵게 표시하였다.

11) 정보손실2 측도는 다변량 분석에 유용한 마이크로데이터 생산을 위해 중요한 의미를 가지나 정보손실1의 값과 비례하는 경향을 보이므로, 분석의 단순성을 위해 여기에서는 고려하지 않고 참고만 하기로 한다.

자료 비밀보호 목적은 노출위험도 작으면서 정보손실도 동시에 최소화해야 하므로 두 측도 중 하나만 작게 만드는 것은 의미가 있다고 할 수 없다. 따라서 민감변수들에 매스킹 기법들을 적용하여 노출위험과 정보손실 측도를 살펴본 결과, 두 측도 값들 모두가 작은 기법은 국소통합(mdav) 및 잡음 추가(correlated2)라고 할 수 있다. 국소통합 세부 기법 mdav는 자료 공간에서 가장 멀리 있는 것들부터 그룹화해가는 알고리즘을(Statistics Netherlands, 2007) 나타내며, 잡음 추가 세부 기법 correlated2는 자료의 공분산 구조를 고려하는 상관 잡음 기법의 로버스트한 알고리즘(Templ과 Meindl, 2008)을 나타내고 있다.

이러한 노출위험-정보손실 측도 결과 값들을 평면에 표현한 것을 위험-유용성 지도(R-U map)라고 하며, [그림 6-2]는 위의 <표 6-5>의 결과에 대한 위험-유용성 지도에 해당된다. 한편, 유용성은 정보손실에 의해서 측정이 되므로 이제 위험-유용성 지도를 위험-정보손실 지도라고 부르도록 하겠다. [그림 6-2]의 왼쪽은 노출위험 대비 정보손실1(il1) 측도, 오른쪽은 정보손실2(il.eigen) 측도에 대한 결과이다. 이러한 위험-정보손실 지도에서 좌표가 원점에 가까울수록 매스킹 처리가 잘 된 것이라고 할 수 있는데, 그림에서 보아도 잡음 추가 기법들의 매스킹 결과가 좋은 것을 볼 수 있다. 국소통합 기법에 대해서도 노출위험과 정보손실1 측도 모두 상대적으로 작은 범위에 있다고 할 수 있다.



[그림 6-2] 세부 기법별 노출위험-정보손실 지도

참고로 노출위험과 정보손실의 개념 이해에 도움을 주고자 <부록>의 <부표 6-1>과 <부표 6-2>를 통해 특정 표본에 대하여[12] 각 민감변수별로 매스킹 세부 기법별 평균과 표준편차를 나타내었다. <부록>의 이 표들을 살펴보면 <표 6-5>에서 나타난 노출위험이

12) 노출위험-정보손실 관련 수치들은 전체 마이크로데이터를 이용하여 계산하였으나 노출위험-정보손실 수치 이외의 이하 본문 및 부록의 평균 및 표준편차 관련 표들과 상자그림, 각종 산점도, 상관계수 값 등은 가계금융·복지조사 마이크로데이터의 비밀보호를 위하여 자료 내의 특정한 하나의 표본을 이용하여 얻었음을 밝혀 둔다. 이를 표본 S라고 부르도록 하겠다.

작고 정보손실이 큰 경우[13] 상대적으로 평균이나 표준편차가 원래 마이크로데이터의 경우와 차이가 크게 나는 것을 볼 수 있다. 반면에 노출위험이 크고 정보손실이 작은 기법을 사용할 경우[14] 상대적으로 그 차이가 작은 경향이 있는 것 또한 확인할 수 있다. 평균과 표준편차는 자료의 유용성 관점에서 검토할 수 있으며 노출위험과 정보손실 관련 수치가 모두 상대적으로 작은 국소통합(mdav) 및 잡음 추가(correlated2) 기법의 결과에 대해서는 다음 절에서 좀 더 자세히 검토하기로 한다.

한편, 노출위험과 정보손실 사이의 상충(trade-off)관계가 있음을 감안한다면 원점에 대해 볼록한 동등한(equivalent) 함수가 존재하여 전체적인 효용이 동일한 조합이 있을 수 있다. 예를 들면 [그림 6-2]의 왼쪽 위험-정보손실 지도에서 보이지 않는 동등한 효용 곡선이 잡음 추가 기법(correlated2)의 시나리오별 결과값들(3 correlated2, 2 correlated2, 1 correlated2)을 지나고 있을 수 있다. 그러나 이러한 동등한 효용 함수의 존재를 확인할 수 없으므로 본 연구에서는 시나리오별 잡음 추가 기법과 국소통합(mdav)의 결과를 심층적으로 살펴보는 동시에, 또한 이들의 결합안도 함께 검토하여 민감변수의 매스킹 방안들을 도출해 보고자 한다.

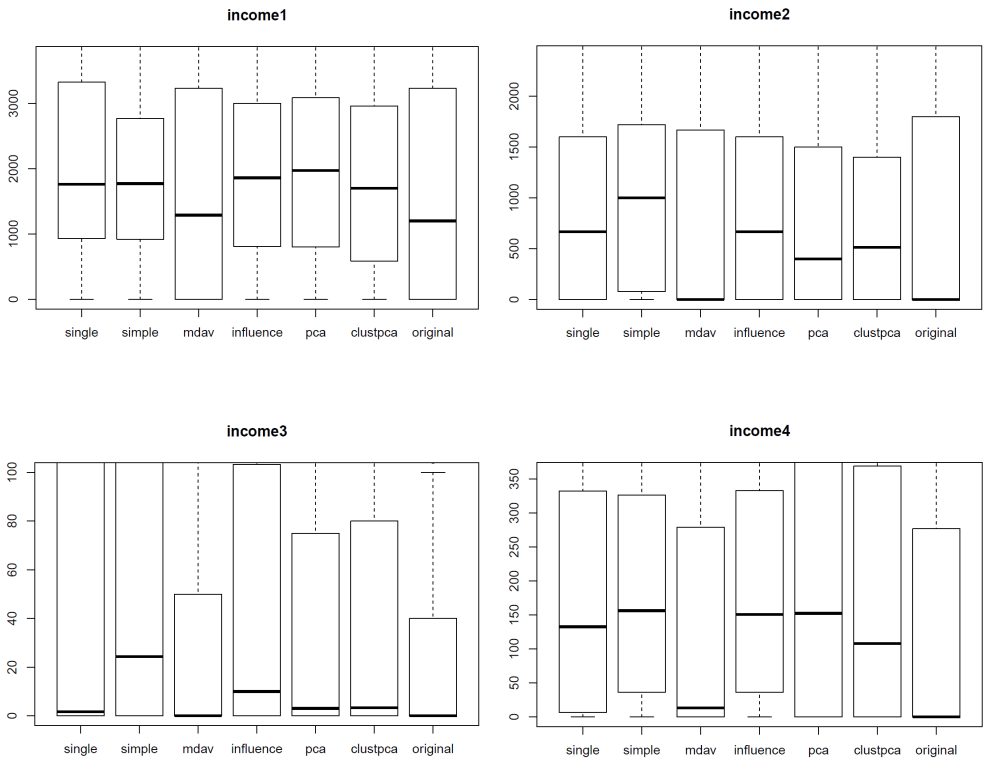
2. 국소통합 기법 적용 결과

앞에서는 민감변수를 위한 매스킹 방법들 중 현실적으로 이용 가능한 대부분의 기법들을 세부 알고리즘별로 적용하여 노출위험 및 정보손실 측도 값들을 살펴보았다. 이들 중에서 위험-정보손실 측도 값들이 모두 상대적으로 수용할 만한 범위에 있는 방법들은 국소통합과 잡음 추가 기법이라 할 수 있었다. 이제 그 중에서 먼저 국소통합 기법의 적용 결과를 자세히 살펴보도록 하겠다. 참고로 가계금융·복지조사 마이크로데이터에 극단 이상점들이 많으므로 표현의 효율성을 위해 본문 및 부록의 여러 그림들은 이상점들을 배제하고 주요 분포가 드러나는데 중심을 두고 그려졌음을[15] 밝혀 둔다.

먼저, 각 국소통합 세부 기법별로 상자그림을 그려보면 소득 변수에 대하여 [그림 6-3]과 같이 나타난다. 원래 자료에 대한 상자그림은 각 그림들 오른쪽 끝에 original로 표현되어 있다. 자산 및 부채 변수에 대한 상자그림은 <부록>의 [부그림 6-1] 및 [부그림 6-3]에 수록하였다. 노출위험-정보손실 측도에서도 mdav[16] 기법이 가장 좋은 결과를 보였었는데

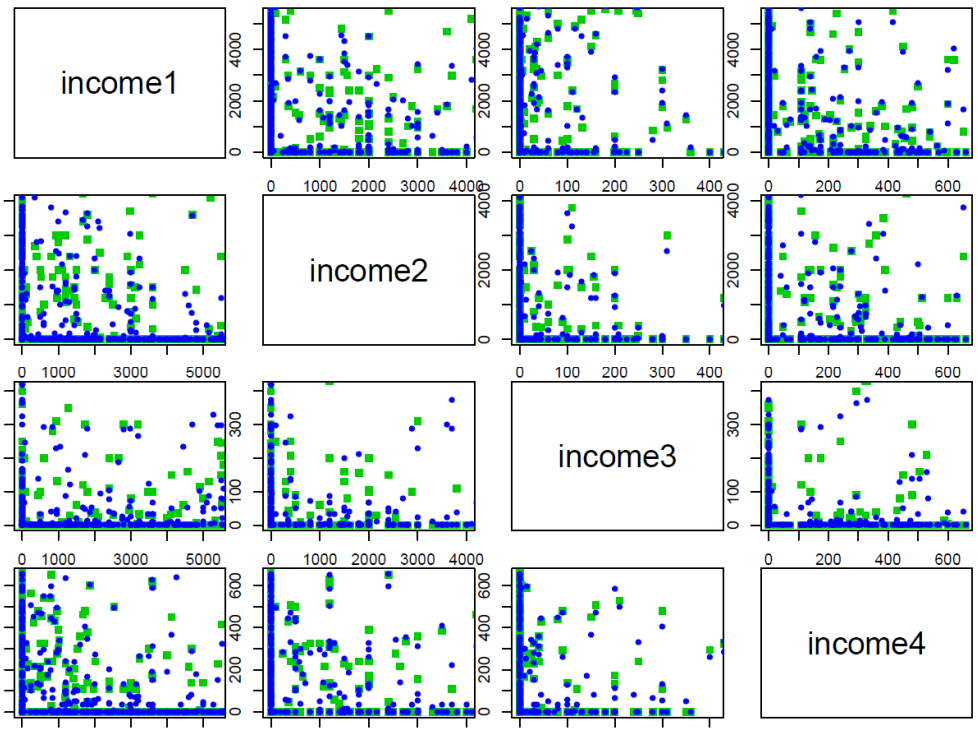
13) 세부 기법 single, simple, influence, pca, clustpca, additive, correlated, sh, sh.mvn 등
14) 세부 기법 onedims 등
15) 각종 상자그림, 산점도 등은 부록의 <부표 6-1> 및 <부표 6-2>의 평균과 표준편차를 구한 동일한 표본 S를 사용하였으나, 극단 이상점들로 인해 부록 표의 수치들과 큰 차이가 있어 보임을 참고하기 바란다.
16) 국소통합 세부 기법 mdav는 자료 공간에서 가장 멀리 있는 것들부터 그룹화해가는 알고리즘(Statistics Netherlands, 2007)을 나타낸다.

상자그림을 점검해도 이상점들을 배제하면 전체적인 자료의 분포를 보존하는데 있어서 mdav 기법이 가장 나은 것을 볼 수 있다.



[그림 6-3] 국소통합 세부 기법별 소득 변수들의 상자그림

참고로 mdav에 의한 매스킹 결과를 보기 위해 하위변수별 산점도를 비교하면 [그림 6-4]와 같다. [그림 6-4]는 소득변수에 대해 표본 S를 이용한 결과를 보인 것이다. 자산 및 부채 변수에 대한 하위변수별 산점도는 <부록>의 [부그림 6-2] 및 [부그림 6-4]에 수록하였다. 소득 하위변수들 사이의 상관계수를 구해보면 <표 6-6>과 같이 나타난다. 하위변수 income2와 income3 사이의 상관계수에 비교적 큰 변화가 있음을 알 수 있다. 부채 하위변수들 중에서 debt12와 debt02 사이의 상관계수에도 비교적 큰 변화가 있었다.



[그림 6-4] 국소통합(mdav) 기법 적용 전후 소득 변수들의 산점도

〈표 6-6〉 국소통합(mdav) 기법 적용 전후 소득 변수들의 상관계수

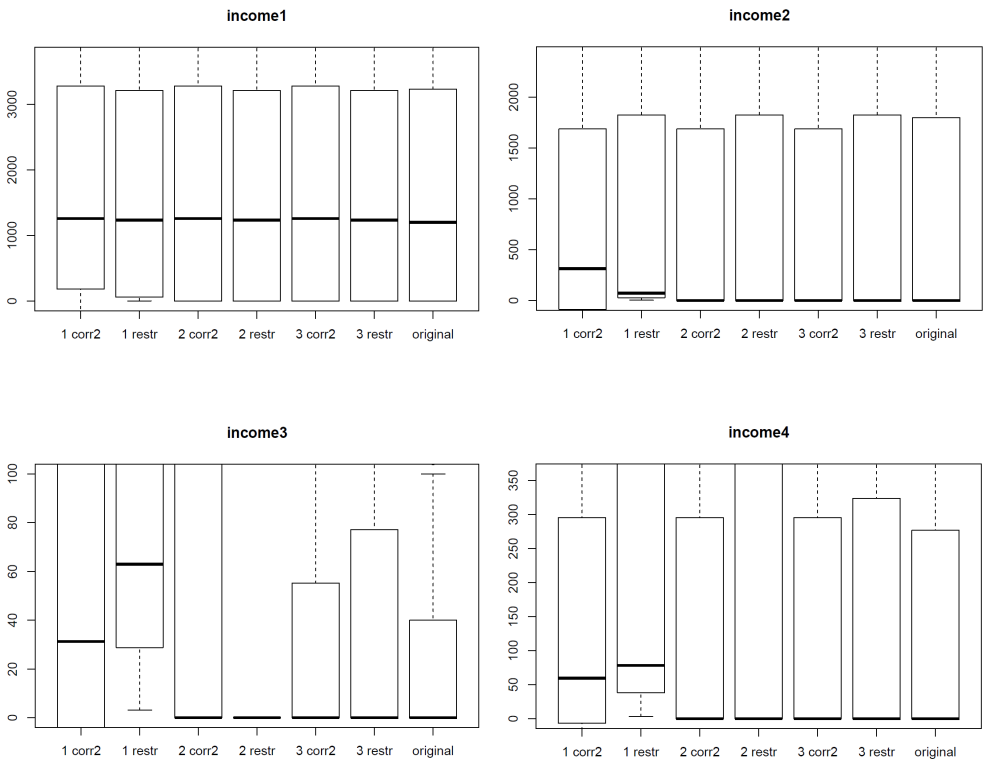
	original				국소통합(mdav)			
	income1	income2	income3	income4	income1	income2	income3	income4
income1	1.00	-0.23	0.02	-0.21	1.00	-0.22	0.05	-0.20
income2	-0.23	1.00	0.43	-0.10	-0.22	1.00	0.19	-0.12
income3	0.02	0.43	1.00	0.04	0.05	0.19	1.00	0.02
income4	-0.21	-0.10	0.04	1.00	-0.20	-0.12	0.02	1.00

3. 잡음 추가 기법 적용 결과

잡음 추가 기법은 위험-정보손실 측도를 기준으로 세부 기법 중[17] correlated2가 가장 좋은 결과를 보였다. 비교를 위해 각 시나리오별로 잡음 추가 세부 기법 correlated2 및

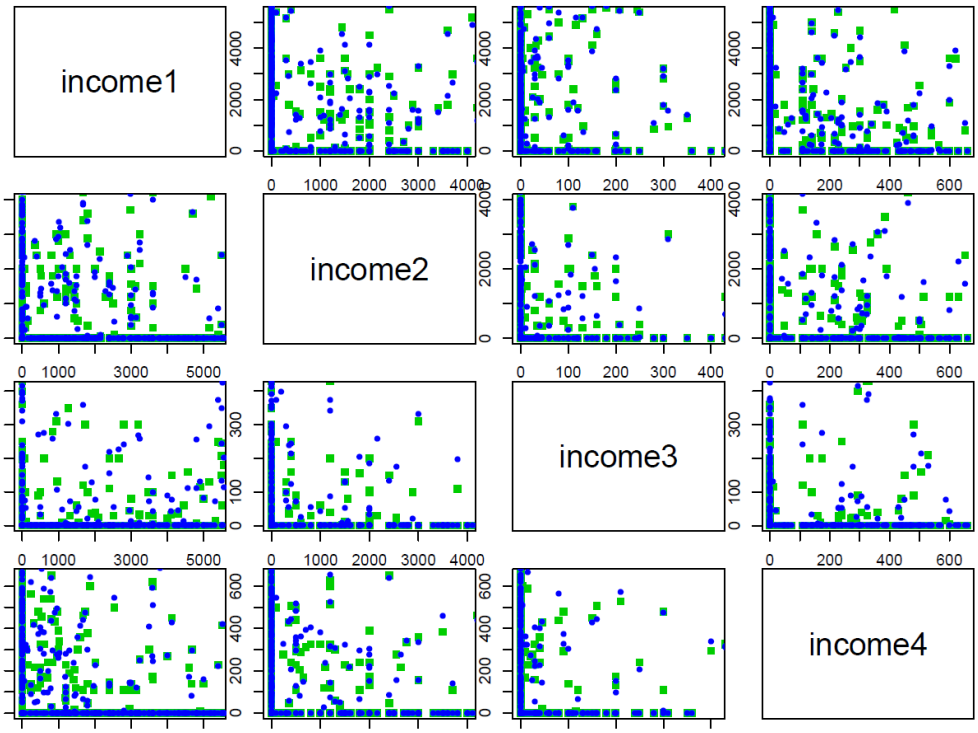
17) 잡음 추가 세부 기법 correlated2는 자료의 공분산 구조를 고려하는 상관 잡음 기법의 로버스트한 알고리즘(Templ과 Meindl, 2008)을, 세부 기법 restr은 자료의 크기를 고려하는 알고리즘을 나타낸다.

restr에 대해 상자그림을 그려보면 소득 변수에 대하여 [그림 6-5]와 같이 나타난다. 자산 및 부채 변수에 대한 상자그림은 <부록>의 [부그림 6-5] 및 [부그림 6-7]에 수록하였다. 노출위험은 다소 증가하나 앞의 정보손실 측도에서도 시나리오 3의 correlated2 기법([그림 6-5]에서는 3 corr2)이 가장 좋은 결과를 보였는데 상자그림을 점검해도 전반적으로 시나리오 3의 correlated2 기법이 원래 자료의 구조를 가장 잘 보존하는 것을 볼 수 있다.



[그림 6-5] 잡음 추가 시나리오별 소득 변수들의 상자그림

시나리오 3의 correlated2에 의한 매스킹 결과를 보기 위해 하위변수별 산점도를 비교하면 [그림 6-6]과 같다. [그림 6-6]은 소득변수에 대해 표본 S를 이용한 결과를 보인 것이다. 자산 및 부채 변수에 대한 하위변수별 산점도는 <부록>의 [부그림 6-6] 및 [부그림 6-8]에 수록하였다. 소득 하위변수들 사이의 상관계수를 구해보면 <표 6-7>과 같이 나타난다. 소득의 하위변수들 사이의 상관계수에 비교적 변화가 없음을 알 수 있다. 자산과 부채의 하위변수들 사이의 상관계수에도 비교적 변화가 없었다.



[그림 6-6] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 소득 변수들의 산점도

〈표 6-7〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 소득 변수들의 상관계수

	original				잡음 추가3(correlated2)			
	income1	income2	income3	income4	income1	income2	income3	income4
income1	1.00	-0.23	0.02	-0.21	1.00	-0.23	0.02	-0.22
income2	-0.23	1.00	0.43	-0.10	-0.23	1.00	0.44	-0.10
income3	0.02	0.43	1.00	0.04	0.02	0.44	1.00	0.04
income4	-0.21	-0.10	0.04	1.00	-0.22	-0.10	0.04	1.00

4. 결합안 적용 결과 및 비교

노출위험과 정보손실 사이에는 상충관계가 있으므로 정보손실을 낮추는 것을 일정 부분 포기하면서 노출위험을 낮추는 결합안을 도출해 볼 수 있다. 먼저 앞의 국소통합 기법과 시나리오별 잡음 추가 기법들의 노출위험과 정보손실 측도를 비교하면 <표 6-8>과 같다. <표 6-8>을 보면 국소통합 기법만을 사용할 경우에는 노출위험과 정보손실 측도 모두 잡음 추가 기법만을 사용할 때보다 그 결과가 좋지 않다. 하지만 국소통합을 사용한 후에 잡음 추가를 하여 잡음 추가만 할 경우보다 노출위험을 훨씬 낮출 수 있다면 전체적인 효용이 동등한 새로운 안이 도출될 수도 있다. <표 6-8>의 결합안 1, 2, 3은 각각 국소통합 결과에 잡음 추가 세부 기법 correlated2를 이용해 얻어진 잡음을 더하고 나서 시나리오 1, 2, 3 각각을 따라 무응답 자료를 처리한 것이다. 이렇게 얻어진 매스킹 처리된 전체 자료에 대하여 노출위험과 정보손실1 및 2 측도를 계산하면 <표 6-8>의 결합안 1, 2, 3의 결과를 얻게 된다. 시나리오별로 잡음 추가에 비해 노출위험 측도는 낮아지고 정보손실 측도는 높아졌음을 알 수 있다.

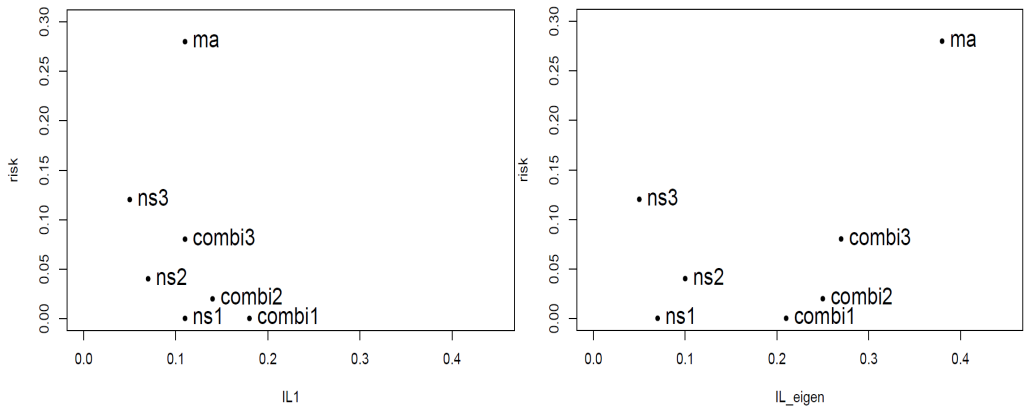
〈표 6-8〉 국소통합, 잡음 추가 및 결합안의 노출위험과 정보손실 측도 결과

매스킹 기법	세부 알고리즘	노출위험 (risk)	정보손실1 (il1)	정보손실2 (il.eigen)
국소통합 (microaggregation)	mdav	0.28	0.11	0.38
잡음 추가1 (noise addition1)	1 correlated2	0.00	0.11	0.07
잡음 추가2 (noise addition2)	2 correlated2	0.04	0.07	0.10
잡음 추가3 (noise addition3)	3 correlated2	0.12	0.05	0.05
결합안1 (combination1)		0.00	0.18	0.21
결합안2 (combination2)		0.02	0.14	0.25
결합안3 (combination3)		0.08	0.11	0.27

다음의 [그림 6-7]은 <표 6-8>의 결과를 노출위험-정보손실 지도로 표현한 것이다. 국소통합은 ma로, 잡음 추가는 시나리오별로 ns1, ns2, ns3으로, 각 결합안은 combi1, combi2, combi3으로 표현하였다. 잡음 추가 기법들과 결합안들이 국소통합 결과에 비해

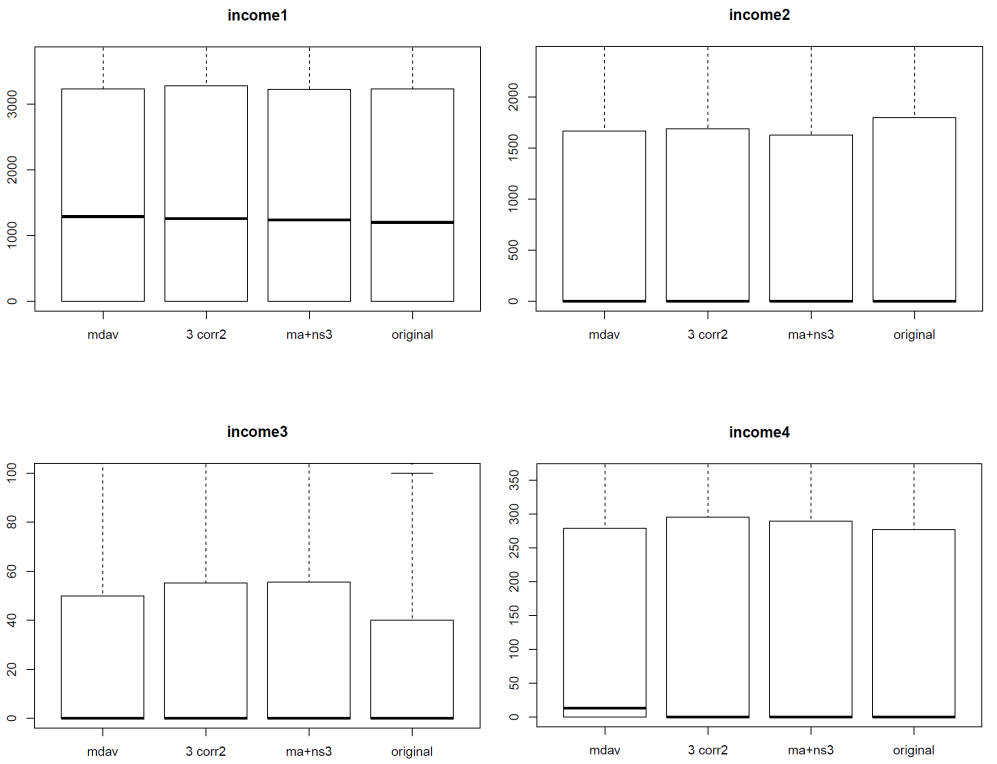
절대적으로 원점에 가깝게 위치해 있다. 그러나 시나리오별로 잡음 추가 기법과 결합안은 각각 보이지 않는 효용이 동등한 곡선 위에 있을 수 있다. 가령 ns2와 combi2 혹은 ns3과 combi3이 각각 노출위험-정보손실 측면에서 동등할 수 있다는 것이다. 따라서 각 시나리오별로 잡음 추가 기법과 결합안은 어느 것이 더 효율적일지 추가적인 엄밀한 검토가 필요하다고 할 수 있다.

한편, 무응답 자료 처리 시나리오 2번의 잡음 추가 기법(ns2)과 시나리오 3번을 반영한 결합안(combi3)을 보면 ns2가 원점에 가까우므로 절대적으로 우위에 있다고 할 수 있다. 그러나 무응답 정보를 어느 수준까지 제공할지에 대해서 정책적인 접근이 필요하므로 이러한 노출위험-정보손실 측도만을 가지고 시나리오 2번을 채택하는 것은 적절하다고 하기 어렵다. 즉, 정책적인 측면과 노출위험-정보손실 측도, 실제 자료 분포의 변화 등을 더욱 면밀히 검토할 필요가 있다고 볼 수 있다.



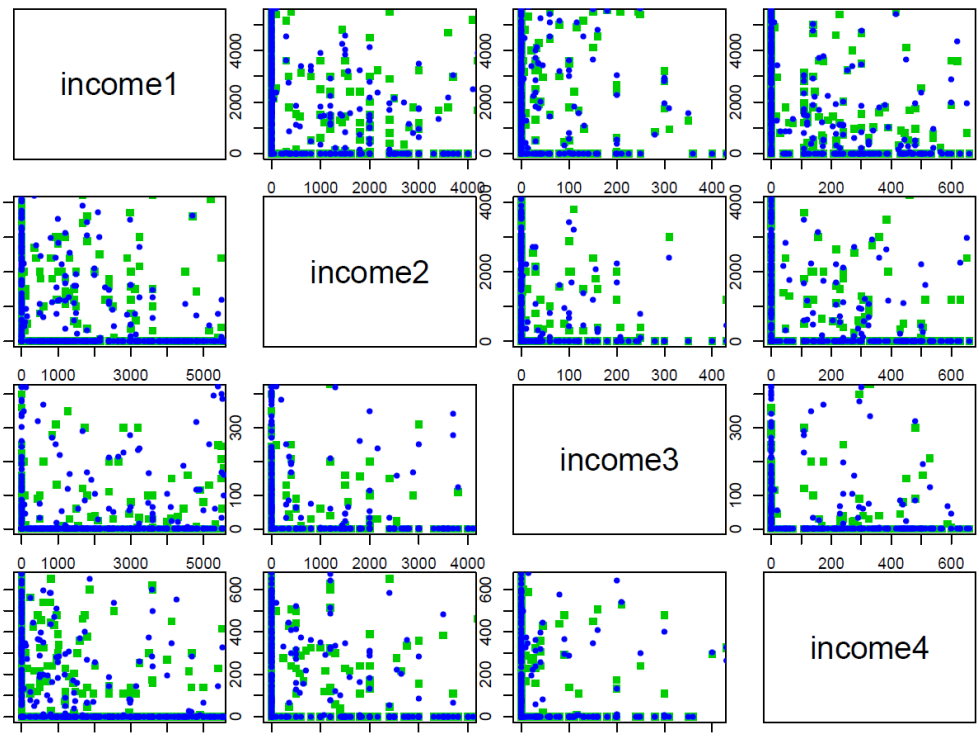
[그림 6-7] 국소통합, 잡음 추가 및 결합안의 노출위험-정보손실 지도

이제 세 가지 시나리오 중 무응답 정보를 가장 많이 제공하는 세 번째 시나리오를 대상으로 국소통합, 잡음 추가 및 결합안을 비교하도록 하겠다. 각각에 대하여 상자그림을 그려보면 소득 변수에 대하여 [그림 6-8]과 같이 나타난다. 자산 및 부채 변수에 대한 상자그림은 <부록>의 [부그림 6-9] 및 [부그림 6-11]에 수록하였다. 잡음 추가 기법과 결합안은 노출위험과 정보손실에서 서로 엇갈린 결과를 보였는데 상자그림에서도 분포 보존에 있어 서로 다른 결과를 보이며 전체적으로 확연한 차이를 찾기는 어려운 것을 알 수 있다.



[그림 6-8] 국소통합, 잡음 추가3 및 결합안3에 대한 소득 변수들의 상자그림

시나리오 3의 결합안 의한 매스킹 결과를 보기 위해 하위변수별 산점도를 비교하면 [그림 6-9]와 같다. [그림 6-9]는 소득변수에 대해 표본 S를 이용한 결과를 보인 것이다. 자산 및 부채 변수에 대한 하위변수별 산점도는 <부록>의 [부그림 6-10] 및 [부그림 6-12]에 수록하였다. 소득 하위변수들 사이의 상관계수를 구해보면 <표 6-9>와 같이 나타난다. 매스킹 전후 상관계수 변화는 국소통합의 결과와 같으며 소득의 하위변수 중 income2와 income3 사이의 상관계수에 비교적 큰 변화가 있음을 알 수 있다. 자산과 부채의 하위변수들 중에서도 국소통합 결과와 마찬가지로 debt12와 debt02 사이의 상관계수에서 비교적 큰 변화가 있었다. 즉, 결합안은 국소통합 기법을 활용하여 잡음 추가 기법의 노출위험을 더 낮추기 위해 자료 유용성을 어느 정도 포기하는 방안이라 할 수 있겠다.



[그림 6-9] 결합안3 적용 전후 소득 변수들의 산점도

〈표 6-9〉 결합안3 적용 전후 소득 변수들의 상관계수

	original				결합안			
	income1	income2	income3	income4	income1	income2	income3	income4
income1	1.00	-0.23	0.02	-0.21	1.00	-0.22	0.05	-0.20
income2	-0.23	1.00	0.43	-0.10	-0.22	1.00	0.19	-0.12
income3	0.02	0.43	1.00	0.04	0.05	0.19	1.00	0.02
income4	-0.21	-0.10	0.04	1.00	-0.20	-0.12	0.02	1.00

제4절 키변수의 공표 범위와 노출위험

서론에서 언급했던 것처럼 기존 연구에서 가계금융·복지조사 마이크로데이터를 대상으로 키변수의 유일성에 근거해 노출위험을 측정하여 시도변수 제공의 위험성이 확인된 바 있다(김경미와 임경은, 2012). 본 연구에서는 민감변수에 대한 매스킹 처리와는 별개로 시도변수를 포함시키기 위한 키변수의 매스킹 방안을 모색하고자 한다. 고려한 키변수들은 시도, 성별, 연령, 교육정도, 종사상지위, 동거여부, 혼인상태, 직업, 가구원수, 주택유형, 주거면적, 소유형태 등의 총 12개이며, 사용하는 매스킹 기법은 재코딩과 국소 감추기이다. 다만, 국소 감추기는 다른 매스킹 기법들을 적용한 후에 익명성(k anonymity)을 획득하기 위해 이용하므로 구체적인 내용과 결과는 여기서 언급하지 않기로 한다. 재코딩 기법을 적용한 키변수와 빈도수 결과는 다음 <표 6-10>과 같다. 빈도수가 적은 구간이 발생하지 않도록 재코딩 기준을 결정하였다.

〈표 6-10〉 키변수의 재코딩과 빈도수

나이	~30	~35	~40	~45	~50	~55	~60
빈도수	977	1563	2212	2552	2587	2573	1859
	~65	~70	~75	~80	80~		
	1557	1337	1195	805	527		
동거유형	1	2	3				
빈도수	3578	15914	252				
종사상지위	1	2	3	4	기타		
빈도수	7941	2830	1167	4045	3761		
직업	1	2	3	4	5	6	7
빈도수	573	2493	2448	1236	1675	1353	1792
	8	9					
	2236	2085					
가구원수	1	2	3	4	5~		
빈도수	3578	5027	3968	5348	1823		
집넓이	~60	~85	~110	~135	~160	~185	185~
빈도수	8623	6200	2060	1448	491	451	471

이러한 재코딩 작업 후에 시도변수를 포함하여 공표 범위를 결정하기 위해 변수들을 추가하면서 유일성에 근거한 노출위험을 측정하였다. 비교를 위하여 시도변수를 포함하지 않은 경우에 대해서도 노출위험을 측정하였다. 다음 <표 6-11>은 공표 범위에

따른 노출위험 측도 결과를 나타내고 있다. <표 6-11>의 변수 이름은 각각 시도(V1), 성별(V2), 연령(V3), 교육정도(V4), 종사상지위(V5), 동거여부(V6), 혼인상태(V7), 직업(V8), 가구원수(V9), 주택유형(V10), 주거면적(V11), 주택소유형태(V12)이며, 각 변수를 포함할 때는 1, 포함하지 않을 때는 0으로 표시하였다. 변수들 중 성별(V2), 연령(V3), 교육정도(V4), 종사상지위(V5)는 반드시 포함하도록 하고, 나머지 7개 변수의 추가 여부에 따라 총 128개의 경우 유일성과 노출위험을 시도변수 포함 여부에 따라 계산하고 일부 결과를 정리하였다.

다음 <표 6-11>에서 U1은 동일한 키조합을 가지는 레코드가 1개인 경우들을 센 것이고, U2는 동일한 키조합을 2개 이하로 가지는 레코드의 개수를 센 결과 이다. 또한 risk는 유일성에 근거한 노출위험을 측정한 것이다. 키변수의 노출위험은 자료의 가중값을 반영하여 계산하므로 가중값이 달라지면 노출위험도 달라진다. 본 연구에서는 키변수의 노출위험 역시 그 수치를 절대적으로 받아들이기보다는 키변수 공표 범위에 따른 노출위험 변화를 상대적으로 비교하기 위해 이용하고자 한다.

이제 <표 6-11>의 노출위험을 자세히 살펴보자. 시나리오 S1을 살펴보면 변수들 중에서 성별(V2), 연령(V3), 교육정도(V4), 종사상지위(V5)만을 포함하고 시도변수를 포함하지 않을 경우를 나타내며 이때 유일성과 노출위험 모두 낮은 수치를 보인다. 그러나 시도변수를 포함할 경우 유일성과 노출위험이 상대적으로 매우 증가함을 볼 수 있다. 만약 11개 변수를 모두 포함할 경우(S128) 시도변수를 제외하면 유일성 U1은 11829, 노출위험은 0.29인 반면, 시도변수를 포함하면 U1은 17138, 노출위험은 0.42로 약 50%가 증가하는 것을 볼 수 있다. 즉 <표3-11>을 통해 모든 변수의 포함 여부별로 시도변수의 영향을 살펴 볼 수 있다.

한편, 좀 더 일반적으로 접근하여 시도(V1), 성별(V2), 연령(V3), 교육정도(V4), 종사상지위(V5)를 모두 포함하고 나서 공표 변수를 하나 더 늘리는 경우를 생각한다면, S2는 주택소유형태(V12)를 추가하고 S5는 주택유형(V10)을 추가한다. 이때, 노출위험은 S2의 경우 0.14, S5의 경우 0.13으로 주택유형(V10)를 추가하는 것이 노출위험이 더 낮다. 이런 식으로 공표 변수를 하나씩 추가할 경우 어느 변수가 노출위험을 가장 적게 증가시키는지 파악하는데 <표 6-11> 키변수의 노출위험 측도 결과를 활용할 수 있다.

키변수의 공표 범위 선택은 노출위험 수치에만 의존하기보다는 각 변수들의 중요도를 감안하는 것이 필요하다. 각 변수들의 중요도에 대한 결정을 한 후에 키변수 조합에 따른 노출위험을 참고하여 마이크로데이터의 매스킹 방향을 결정하는 것이 바람직할 것이다.

〈표 6-11〉 공표 범위에 따른 유일성과 노출위험

	키변수 포함 여부												시도변수 제외			시도변수 포함		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	U1	U2	risk	U1	U2	risk
S1		1	1	1	1	0	0	0	0	0	0	0	80	126	0	1938	1740	0.07
S2		1	1	1	1	0	0	0	0	0	0	1	586	578	0.02	4854	2826	0.14
S3		1	1	1	1	0	0	0	0	0	1	0	722	780	0.03	5746	3178	0.17
S4		1	1	1	1	0	0	0	0	0	1	1	1962	1472	0.06	9001	3626	0.25
S5		1	1	1	1	0	0	0	0	1	0	0	436	420	0.01	4369	2748	0.13
S21		1	1	1	1	0	0	1	0	1	0	0	1800	1342	0.05	8245	3655	0.22
S22		1	1	1	1	0	0	1	0	1	0	1	4083	2343	0.11	11626	3166	0.30
S23		1	1	1	1	0	0	1	0	1	1	0	4523	2467	0.13	12440	3426	0.32
S24		1	1	1	1	0	0	1	0	1	1	1	7266	3074	0.19	14856	2623	0.37
S25		1	1	1	1	0	0	1	1	0	0	0	2232	1616	0.06	9630	3724	0.26
S56		1	1	1	1	0	1	1	0	1	1	1	8837	2908	0.23	15501	2268	0.38
S57		1	1	1	1	0	1	1	1	0	0	0	3320	1924	0.09	10492	3389	0.27
S58		1	1	1	1	0	1	1	1	0	0	1	6104	2480	0.16	13217	2866	0.33
S59		1	1	1	1	0	1	1	1	0	1	0	6422	2872	0.17	14185	2858	0.36
S60		1	1	1	1	0	1	1	1	0	1	1	9329	3104	0.24	16172	2165	0.4
S96		1	1	1	1	1	0	1	1	1	1	1	11202	3015	0.28	16996	1743	0.41
S97		1	1	1	1	1	1	0	0	0	0	0	773	594	0.02	4531	2226	0.13
S98		1	1	1	1	1	1	0	0	0	0	1	2140	1388	0.06	7234	2674	0.19
S99		1	1	1	1	1	1	0	0	0	1	0	2085	1372	0.06	8155	3154	0.22
S100		1	1	1	1	1	1	0	0	0	1	1	4004	2058	0.11	10892	3276	0.28
S124		1	1	1	1	1	1	1	1	0	1	1	9500	3078	0.24	16251	2154	0.40
S125		1	1	1	1	1	1	1	1	1	0	0	5890	2647	0.15	13077	2847	0.33
S126		1	1	1	1	1	1	1	1	1	0	1	8804	2672	0.22	15032	2239	0.37
S127		1	1	1	1	1	1	1	1	1	1	0	9250	3016	0.24	15837	2246	0.39
S128		1	1	1	1	1	1	1	1	1	1	1	11829	2722	0.29	17138	1647	0.42

제5절 결론

이번 연구에서는 먼저 민감변수에 현재 적용 가능한 매스킹 기법들을 대부분 적용하여 노출위험, 유용성(정보손실) 측도를 계산하여 비교한 후, 노출위험 및 정보손실 수치가 작은 기법으로 국소통합(mdav) 및 잡음 추가(correlated2) 기법을 선택하였다. 이때 무응답 등으로 값이 0인 자료들에 대해 세 가지 시나리오를 적용하였다. 또한, 잡음 추가 기법을 적용한 결과보다 노출위험을 추가로 감소시키기 위해 국소통합과 잡음 추가 기법의 결합안을 제시하였다. 결론적으로 총 7가지 방안에 대하여 노출위험 및 유용성 측도를 검토하고, 서로 보이지 않는 효용이 같을 수 있는 방안들을 비교하기 위해 자료의 분포 및 상관계수의 변화를 살펴보았다.

민감변수의 최종 7가지 매스킹 방안 중에서 무엇을 선택할지 결정하기 위해서는 노출위험과 유용성 중에서 어느 것에 더 많은 비중을 둘 것인가를 결정하는 정책적인 측면과, 매스킹 전후의 자료 특성의 변화 및 통계 분석 결과의 차이를 심층적으로 검토하는 과정이 추가로 필요하다. 이는 본 연구의 범위를 넘어서는 문제로 향후 연구 과제가 된다. 또한 가중값을 민감변수로 취급하여 매스킹 처리할 경우의 자료의 유용성 감소 문제 역시 향후 연구 대상에 포함된다.

다음으로 키변수에 관하여는 공표범위에 따른 유일성 기반 노출위험을 제시하였다. 변수의 중요성을 고려하여 반드시 포함되는 5개의 변수를 선택하고 나머지 변수들을 하나씩 추가하면서 노출위험을 살펴보았다. 변수의 중요성에 대한 정책적인 측면과 노출위험 측도를 감안하여 공표할 변수의 범위가 결정이 되면, 국소 감추기 기법을 이용하여 2 이상의 익명성을 획득하는 과정을 수행할 수 있다. 한편 국소 감추기를 이용할 경우 자료 유용성 변화의 세부적인 측면을 검토하는 것은 향후 과제가 된다.

본 연구에서는 민감변수와 키변수 각각에 대하여 가능한 매스킹 방안들을 검토하였다. 여러 가지 매스킹 방안들 중에서 무엇을 선택할지의 문제를 풀기 위해서는 실제로 매스킹된 자료가 공표되었을 때 노출위험과 정보손실이 어떻게 발생할지를 심층적으로 검토하는 것이 필요하다. 이러한 매스킹 전후 자료의 비교는 모든 지역, 모든 변수에 대해서 이루어질 수 있으며 매우 광범위하다. 향후 매스킹 방안을 포함하여 마이크로데이터의 비밀보호 연구가 꾸준히 발전하길 기대한다.

참고문헌

김경미, 이의규, 정미옥 (2007), 마이크로데이터 제공에 관한 해외사례연구, 통계개발원.
김경미, 임경은 (2012), 가계금융・복지조사 자료 비밀보호방법 연구, 통계개발원.
박민정, 김경미 (2013), 종단자료 비밀보호의 국제 연구동향 및 향후 추진방향, 통계개발원.
윤연옥 (2010), 통계자료의 비밀보호기법 연구 -OECD 국가의 마이크로데이터 제공을 중심으로-, 통계청.
정동명, 강동환 (2006), 마이크로자료의 활용도 제고를 위한 비밀보호방법, 통계개발원.
정동명, 김경미 (2008), 승법잡음모형을 이용한 가계조사 자료의 비밀보호, 통계개발원.
정동명, 정남수, 한승훈 (2007), 가계조사 마이크로데이터의 비밀보호, 통계개발원.
정동명, 정미옥 (2007), 2005 인구주택총조사자료의 개인정보 노출제한방법, 통계개발원.
충남대학교 통계학과 대학원, 통계자료 비밀보호론 [매스킹(masking)이론].
통계청 (2012), 2012 가계금융·복지조사 지침서.

Abowd, J. M. and Woodcock, S. D. (2001), Disclosure limitation in longitudial 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*, 215-277, Amsterdam: North-Holland.

Brand, R. (2004), Microdata protection through noise addition, In: *Inference Control in Statistical Databases*, Lecture Notes in Computer Science Volumn 2316, 97-116.

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.

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

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.

Lee, S., Genton, M. G. and Arellano-Valle, R. B. (2010), Perturbation of numerical confidential data via skew- t distributions, *Management Science*, 56(2), 318-333.

Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1998), A comparative study of microaggregation methods, Qüestiió, 22: 511-526.

Moore, Jr. R. (1996), Controlled data-swapping techniques fo masking public use microdata, *U.S. Bureau of Census Statistical Research Division Report Series*, RR 96-04.

Muralidhar, K. and Sarathy, R. (2006), Data shuffling - a new masking approach for numerical data. *Management Science*, 52(5), 658-670.

Reiter, J. P. (2004), New approaches to data dissemination: A glimpse into the future, *Chance*,

17:3 (Summer 2004), 12-16.

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

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

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, In: *Privacy in Statistical Databases*, Lecture Notes in Computer Science Volumn 5262, Springer, 177-189.

<부 록>

3.1 매스킹 기법별 노출위험과 정보손실 측정

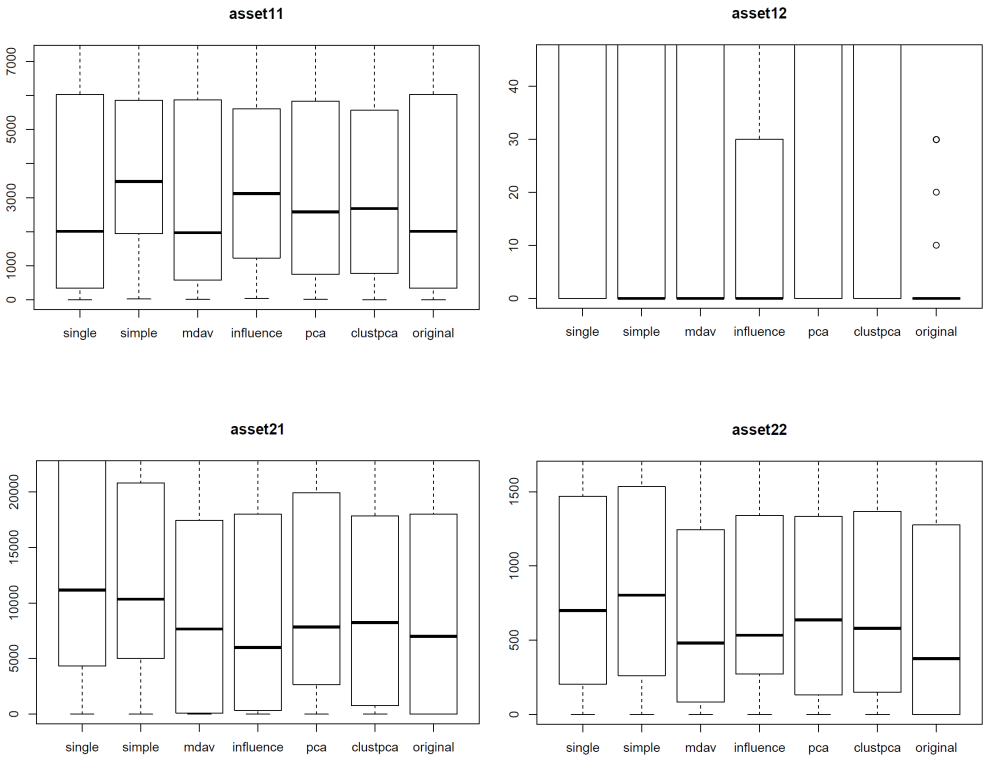
〈부표 6-1〉 매스킹 기법 적용에 따른 표본 S 민감변수들의 평균

평균	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4
original	5095.20	308.10	19943.80	1926.50	2942.10	871.40	178.10	469.60	1990.00	1305.50	183.70	239.00
single	5115.30	1377.60	19649.80	1689.50	2681.50	704.70	134.30	1075.00	2322.20	1165.40	176.40	251.30
simple	5095.80	308.10	19957.60	1927.30	2942.10	871.40	178.10	469.60	1986.40	1306.30	183.70	239.50
onedims	5114.30	304.20	19477.00	1922.70	3101.20	867.80	177.20	465.00	1987.80	1311.40	183.30	238.80
mdav	5102.90	361.10	19961.20	1869.70	2550.90	856.50	187.50	594.90	2005.90	1302.30	171.10	238.60
influence	5396.40	353.10	19724.30	1508.80	2327.70	725.50	155.10	643.30	2137.00	1191.20	183.60	240.80
pca	5219.40	1108.00	19450.50	1588.00	2573.30	697.00	147.70	1010.30	2226.80	1150.50	207.90	282.50
clustpca	5100.40	1001.20	19848.00	1560.50	2799.70	613.10	147.70	709.20	2102.30	1198.40	176.40	287.50
1 additve	5666.60	91.70	46529.90	1836.40	4734.90	1129.20	-11.60	1246.90	2819.50	885.30	61.00	337.90
1 corr	11222.20	2317.80	42211.40	3650.70	5855.70	1464.40	310.50	2005.70	4613.90	2469.40	383.70	518.90
1 corr2	5086.00	355.50	19759.90	1905.00	2984.80	864.80	181.10	501.50	1998.20	1306.00	182.20	243.50
1 restr	5056.40	383.40	19545.80	1964.00	2954.00	933.10	257.80	541.00	2024.70	1358.10	2620	315.90
2 additve	60968.80	14690.20	176063.00	21517.00	29961.90	5660.40	1694.90	7814.60	12722.80	7940.10	2045.90	1666.20
2 corr	11235.90	923.00	35710.80	3495.50	4129.70	1094.00	258.40	724.00	3764.50	1806.70	254.00	374.70
2 corr2	5432.10	433.30	20182.90	2095.00	3071.10	870.70	190.60	560.80	2008.50	1327.70	196.70	244.60
2 restr	5056.40	330.20	19529.70	1954.40	2920.80	889.80	232.90	498.60	2004.90	1317.40	223.90	286.60
3 additve	60969.20	4918.90	122505.80	16626.20	13777.50	4155.30	1030.60	1767.20	9434.30	6395.30	1409.20	1012.50
3 corr	11235.90	782.80	33998.00	3256.80	3806.40	1069.50	234.10	587.30	3705.10	1811.80	249.40	352.20
3 corr2	5432.10	366.40	19833.70	2010.10	2960.00	864.80	182.50	476.60	1995.90	1310.80	189.20	240.90
3 restr	5056.40	316.90	19518.10	1940.50	2893.70	875.90	202.00	462.70	1990.50	1311.70	202.20	267.60
rs	5718.40	329.40	18941.10	1906.50	2544.90	760.10	168.90	379.50	1953.90	1310.30	169.70	227.70
sh	6371.40	905.70	18935.00	2665.50	2177.60	793.80	181.90	1144.50	2468.50	1530.50	222.10	263.70
sh.mvn	5177.90	1735.10	18265.60	1422.20	2347.90	450.70	99.80	1026.10	2388.40	1028.80	156.10	297.30

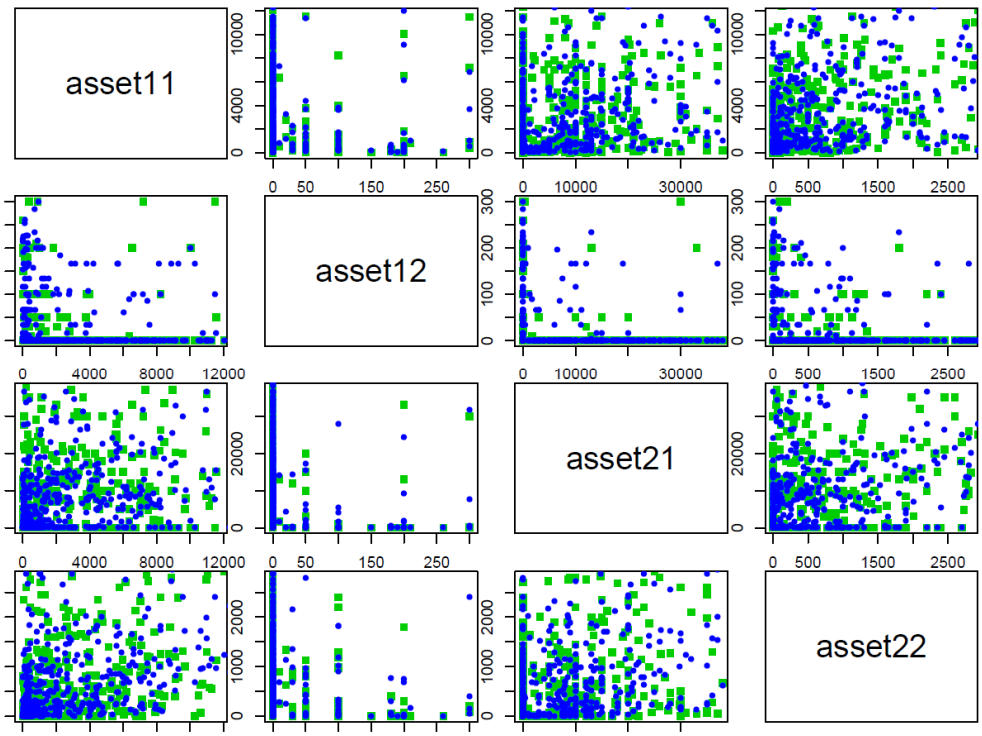
〈부표 6-2〉 매스킹 기법 적용에 따른 표본 S 민감변수들의 표준편차

표준편차	asset11	asset12	asset21	asset22	debt11	debt12	debt134	debt02	income1	income2	income3	income4
original	11778.20	1422.20	79137.40	7924.10	16817.70	2488.20	683.80	3374.70	2383.80	3191.40	740.30	539.70
single	12086.90	2671.80	35049.00	3982.80	7173.00	1266.60	299.90	3092.00	1973.40	1815.40	593.10	343.10
simple	6719.90	837.60	44818.50	4432.80	9499.40	1401.50	381.60	1923.60	1460.50	1789.20	441.50	297.20
onedims	12086.80	1406.90	71880.20	7909.30	19931.30	2478.20	682.30	3346.90	2380.70	3359.10	738.40	539.80
mdav	10308.90	1425.50	65711.30	7670.00	12286.70	2446.80	785.60	3882.10	2367.50	2745.20	607.80	524.70
influence	16229.30	1427.90	74383.40	3541.10	9722.90	1396.50	349.00	2389.30	1797.60	2371.70	892.20	302.70
pca	17026.90	1991.80	60377.80	4540.10	9774.90	1691.90	421.20	3699.30	1853.90	2475.90	1345.40	394.10
clustpca	15254.20	1990.30	70868.40	4789.00	14414.30	1447.60	377.90	3414.40	2164.60	2381.20	919.60	483.80
1 additve	76614.00	28451.40	245009.30	27874.10	57343.30	15428.90	3525.10	31922.60	17374.40	18187.10	52820	2989.80
1 corr	12159.60	1837.70	79918.10	7944.60	17016.80	2610.40	700.90	3742.40	2489.50	3363.00	754.90	547.60
1 corr2	117840	1528.50	78589.00	7875.20	16758.30	2482.30	682.80	3401.10	2407.20	3212.60	754.90	541.60
1 restr	11506.10	1389.00	77227.80	7746.40	16407.80	2427.10	690.90	3290.20	2330.50	3123.60	734.60	544.30
2 additve	46660.40	21474.40	172586.20	17639.60	43543.60	11421.70	2688.30	21597.10	11880.70	13619.50	3925.80	2270.70
2 corr	12146.90	2185.90	82249.90	8002.40	17329.90	2707.30	712.30	3763.00	3338.90	3620.70	778.00	620.90
2 corr2	11628.20	1459.60	78459.30	7826.30	16727.90	2469.50	679.00	3341.10	2393.30	3193.20	748.10	539.80
2 restr	11506.10	1399.70	77231.90	7748.80	16413.60	2441.50	698.70	3296.10	2347.10	3139.50	744.80	558.00
3 additve	46659.90	12370.40	159085.00	17282.10	30216.40	8079.90	1894.10	7968.20	11413.70	10364.70	2882.70	1743.70
3 corr	12146.90	1997.00	82389.90	8049.20	17314.30	2692.80	711.50	3682.20	3348.20	3599.40	771.10	621.60
3 corr2	11628.20	1441.60	78520.90	7843.90	16738.70	2469.40	680.10	3332.50	2401.20	3196.60	748.00	540.80
3 restr	11506.10	1397.70	77234.80	7751.80	16417.90	2444.50	695.50	3298.30	2357.40	3141.10	740.10	556.10
rs	17187.30	1874.40	64890.30	8382.60	14708.40	2031.90	786.10	2350.70	2338.30	3790.10	637.70	500.40
sh	12134.40	2656.40	35013.20	8192.20	7354.90	2660.60	662.40	5256.70	2989.50	3428.20	761.40	583.80
sh.mvn	10213.30	4860.50	28928.70	3641.70	7235.80	1489.40	542.50	3947.90	3041.70	2631.80	656.50	618.60

3.2 국소통합 기법 적용 결과



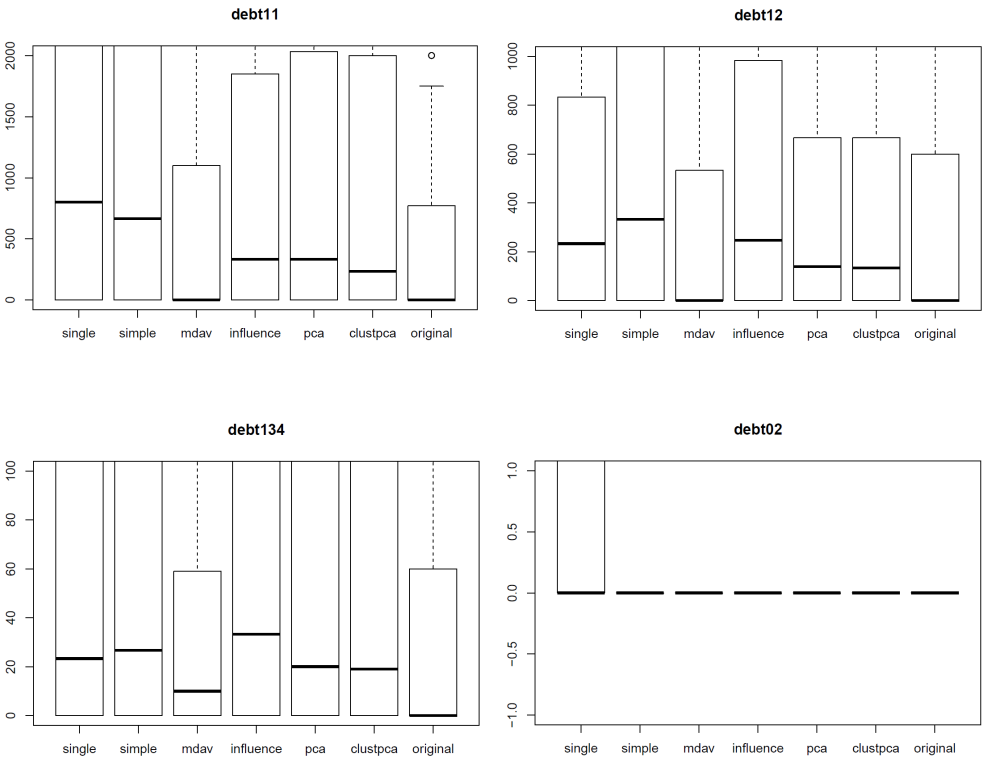
[부그림 6-1] 국소통합 세부 기법별 자산 변수들의 상자그림



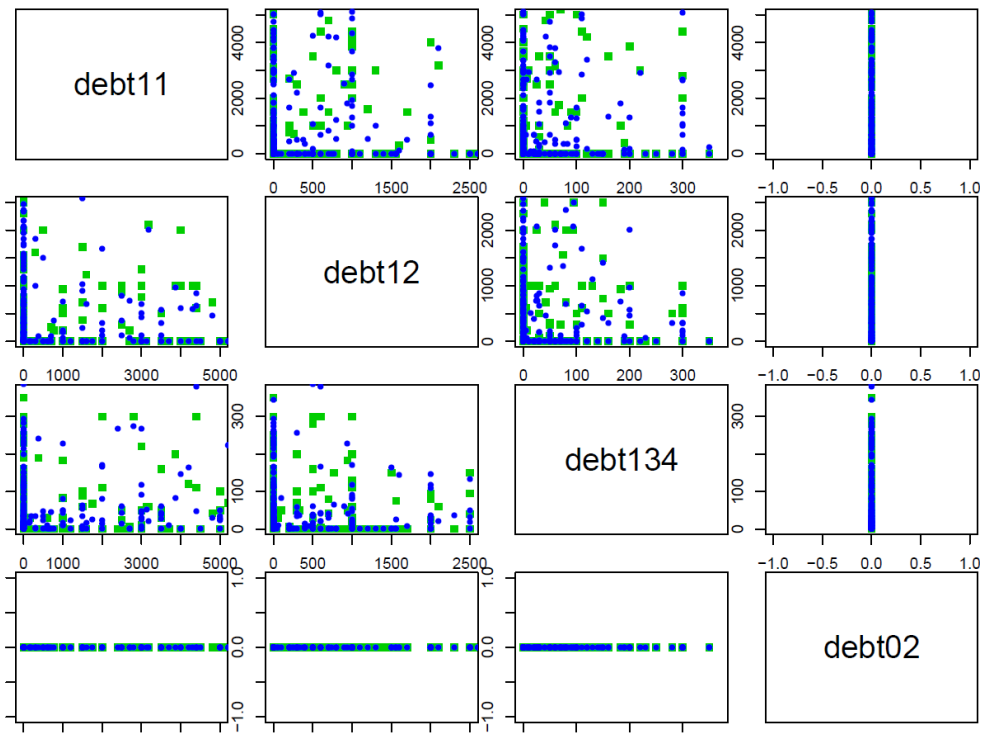
[부그림 6-2] 국소통합(mdav) 기법 적용 전후 자산 변수들의 산점도

〈부표 6-3〉 국소통합(mdav) 기법 적용 전후 자산 변수들의 상관계수

	original				국소통합(mdav)			
	asset11	asset12	asset21	asset22	asset11	asset12	asset21	asset22
asset11	1.00	0.03	0.66	0.36	1.00	0.02	0.62	0.39
asset12	0.03	1.00	-0.02	0.05	0.02	1.00	-0.03	0.00
asset21	0.66	-0.02	1.00	0.43	0.62	-0.03	1.00	0.43
asset22	0.36	0.05	0.43	1.00	0.39	0.00	0.43	1.00



[부그림 6-3] 국소통합 세부 기법별 부채 변수들의 상자그림

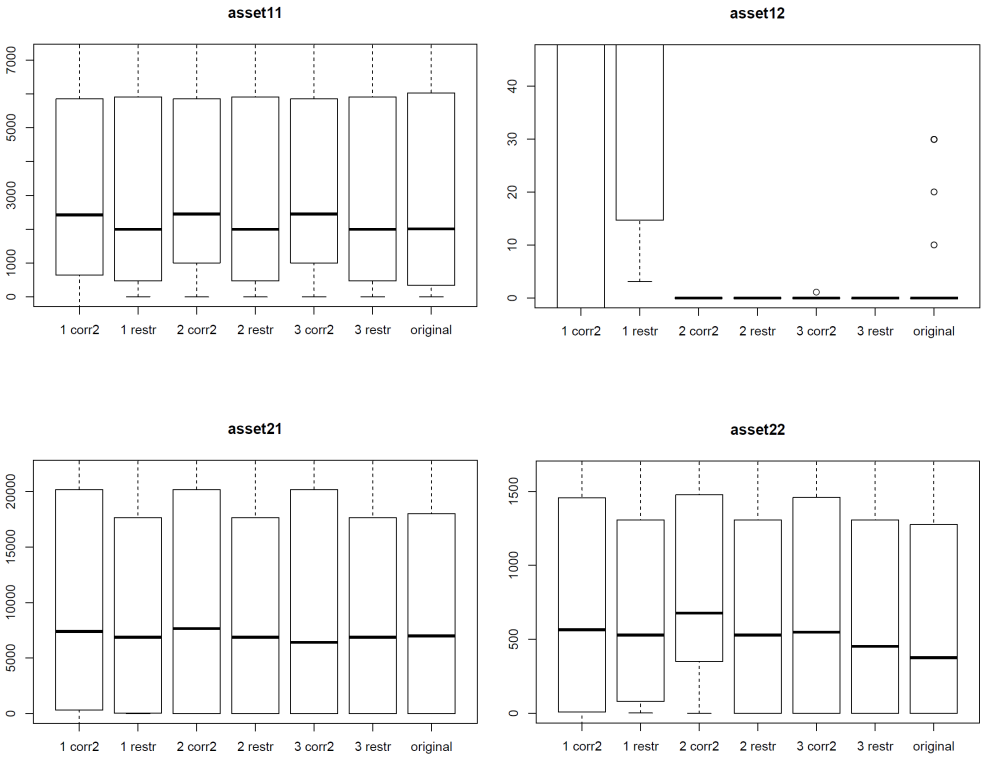


[부그림 6-4] 국소통합(mdav) 기법 적용 전후 부채 변수들의 산점도

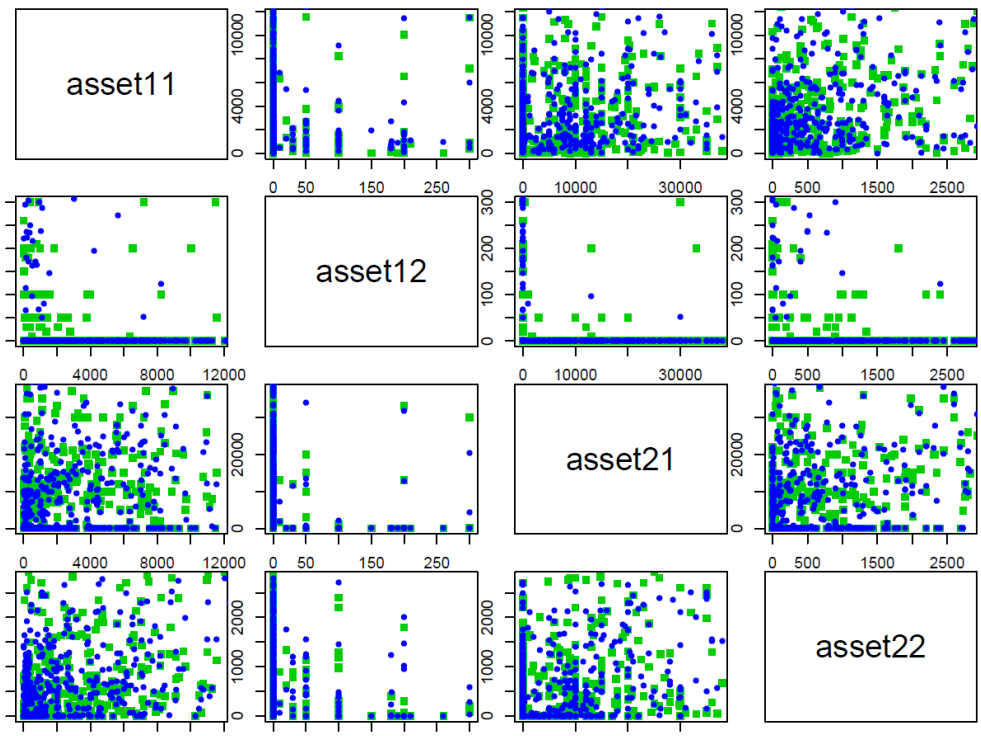
〈부표 6-4〉 국소통합(mdav) 기법 적용 전후 부채 변수들의 상관계수

	original				국소통합(mdav)			
	debt11	debt12	debt134	debt02	debt11	debt12	debt134	debt02
debt11	1.00	0.09	0.03	0.17	1.00	0.12	0.03	0.23
debt12	0.09	1.00	0.19	0.25	0.12	1.00	0.12	0.09
debt134	0.03	0.19	1.00	-0.01	0.03	0.12	1.00	-0.02
debt02	0.17	0.25	-0.01	1.00	0.23	0.09	-0.02	1.00

3.3 잡음 추가 기법 적용 결과



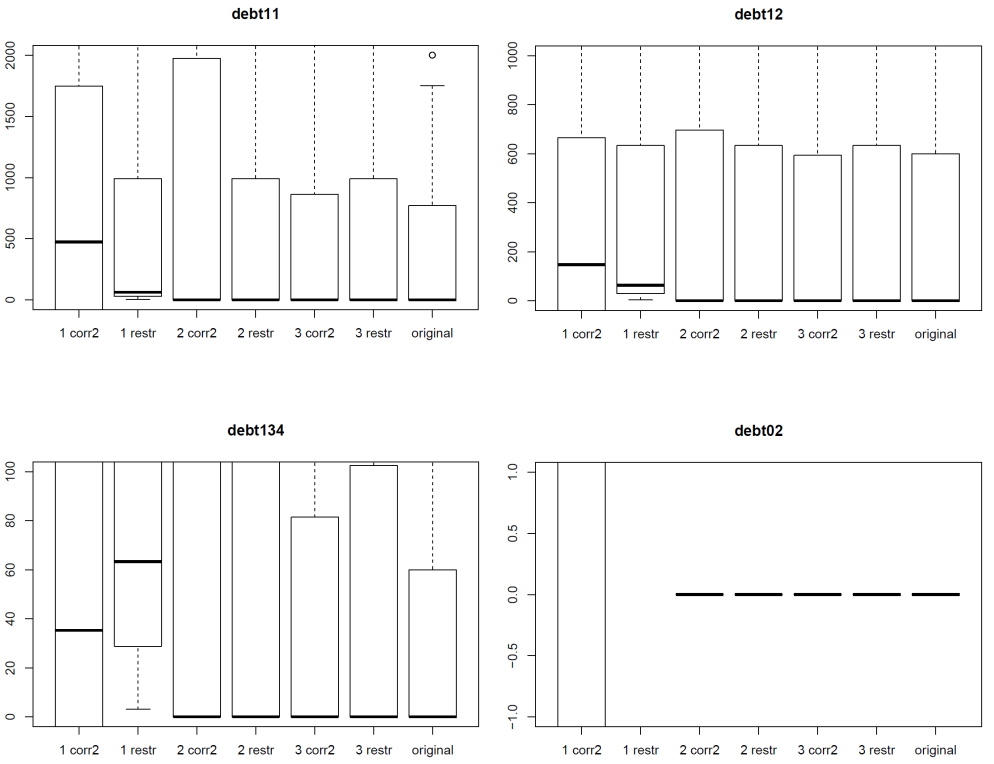
[부그림 6-5] 잡음 추가 시나리오별 자산 변수들의 상자그림



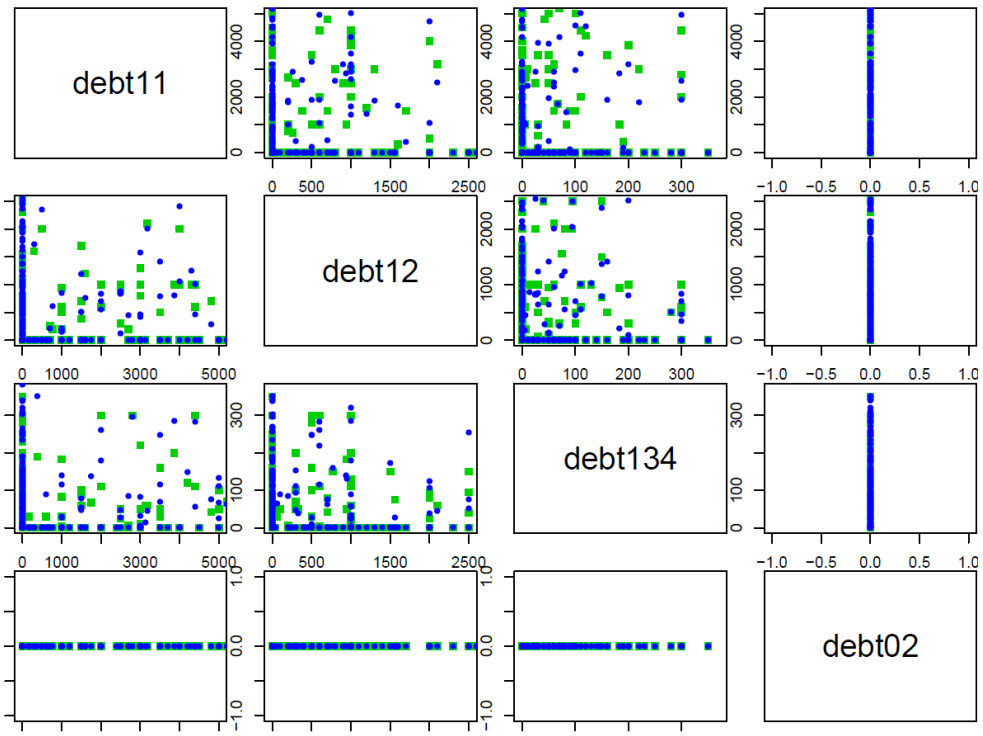
[부그림 6-6] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 자산 변수들의 산점도

〈부표 6-5〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 자산 변수들의 상관계수

	original				잡음 추가3(correlated2)			
	asset11	asset12	asset21	asset22	asset11	asset12	asset21	asset22
asset11	1.00	0.03	0.66	0.36	1.00	0.02	0.66	0.36
asset12	0.03	1.00	-0.02	0.05	0.02	1.00	-0.03	0.06
asset21	0.66	-0.02	1.00	0.43	0.66	-0.03	1.00	0.43
asset22	0.36	0.05	0.43	1.00	0.36	0.06	0.43	1.00



[부그림 6-7] 잡음 추가 시나리오별 부채 변수들의 상자그림

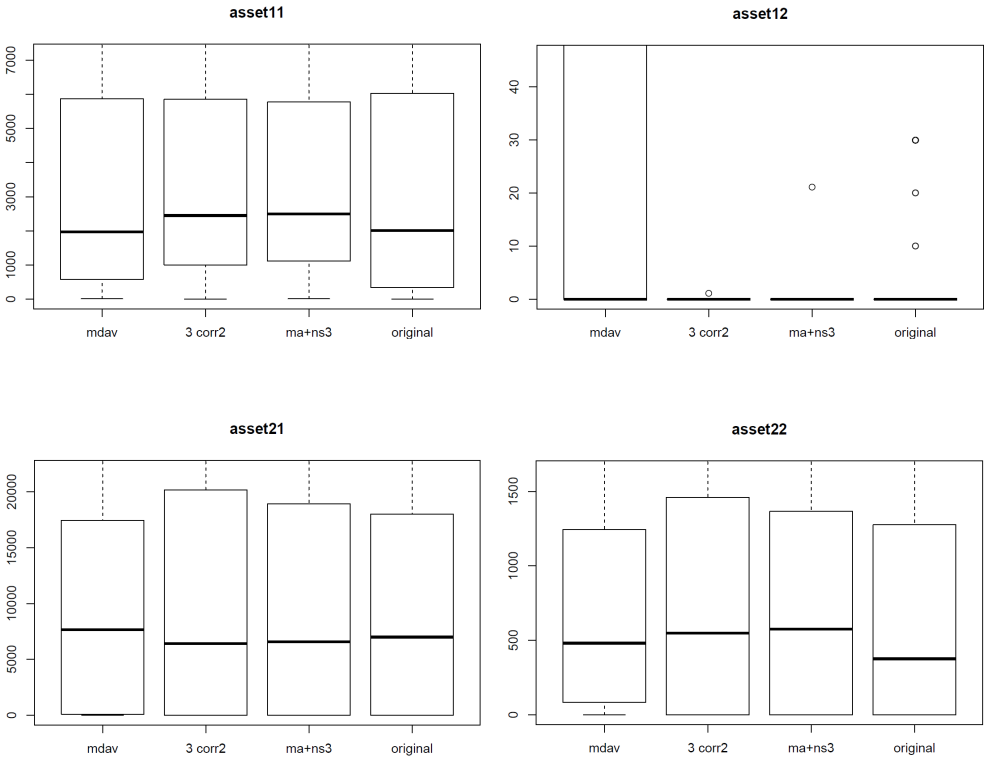


[부그림 6-8] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 부채 변수들의 산점도

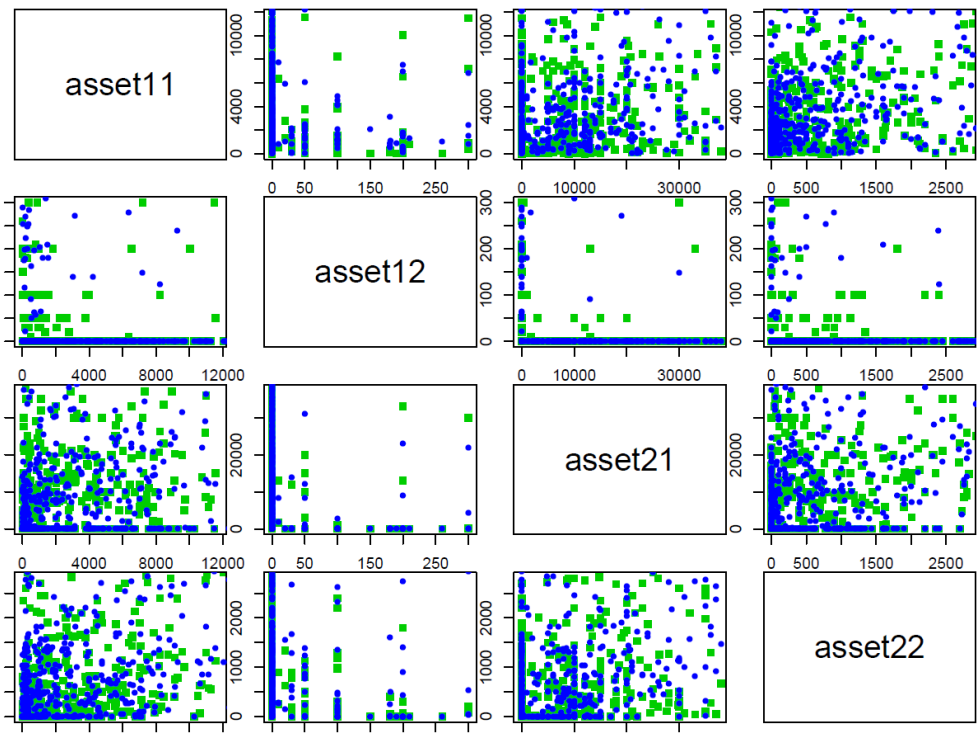
〈부표 6-6〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 부채 변수들의 상관계수

	original				잡음 추가3(correlated2)			
	debt11	debt12	debt134	debt02	debt11	debt12	debt134	debt02
debt11	1.00	0.09	0.03	0.17	1.00	0.09	0.03	0.18
debt12	0.09	1.00	0.19	0.25	0.09	1.00	0.19	0.26
debt134	0.03	0.19	1.00	-0.01	0.03	0.19	1.00	-0.01
debt02	0.17	0.25	-0.01	1.00	0.18	0.26	-0.01	1.00

3.4 결합안 적용 결과 및 비교



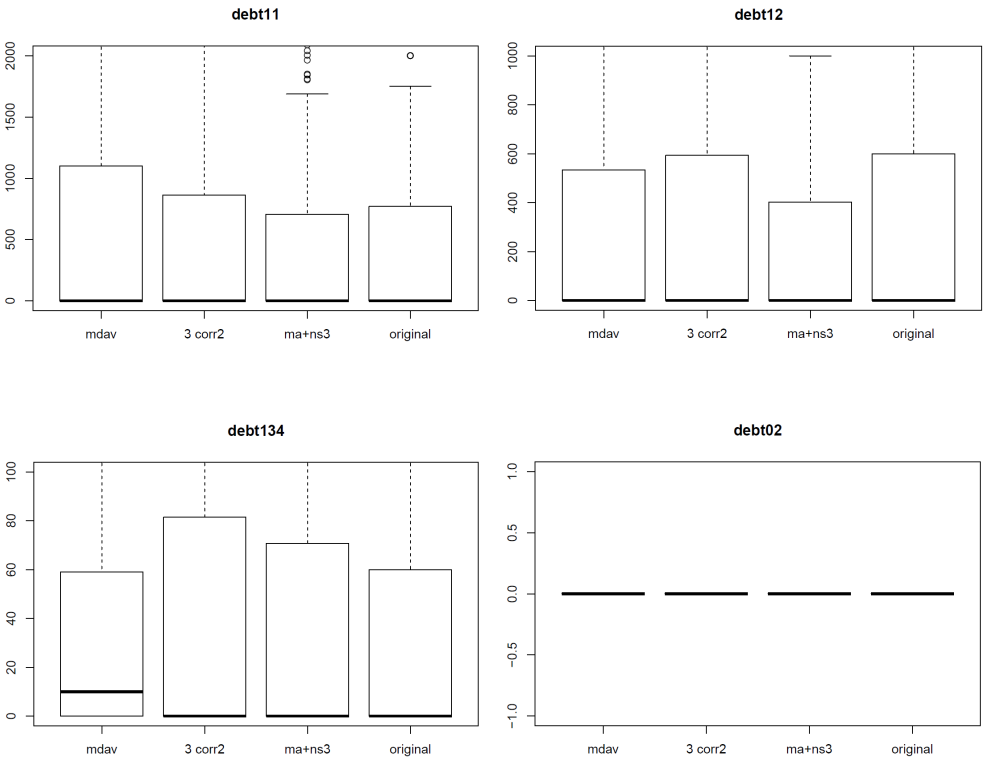
[부그림 6-9] 국소통합, 잡음 추가3 및 결합안3에 대한 자산 변수들의 상자그림



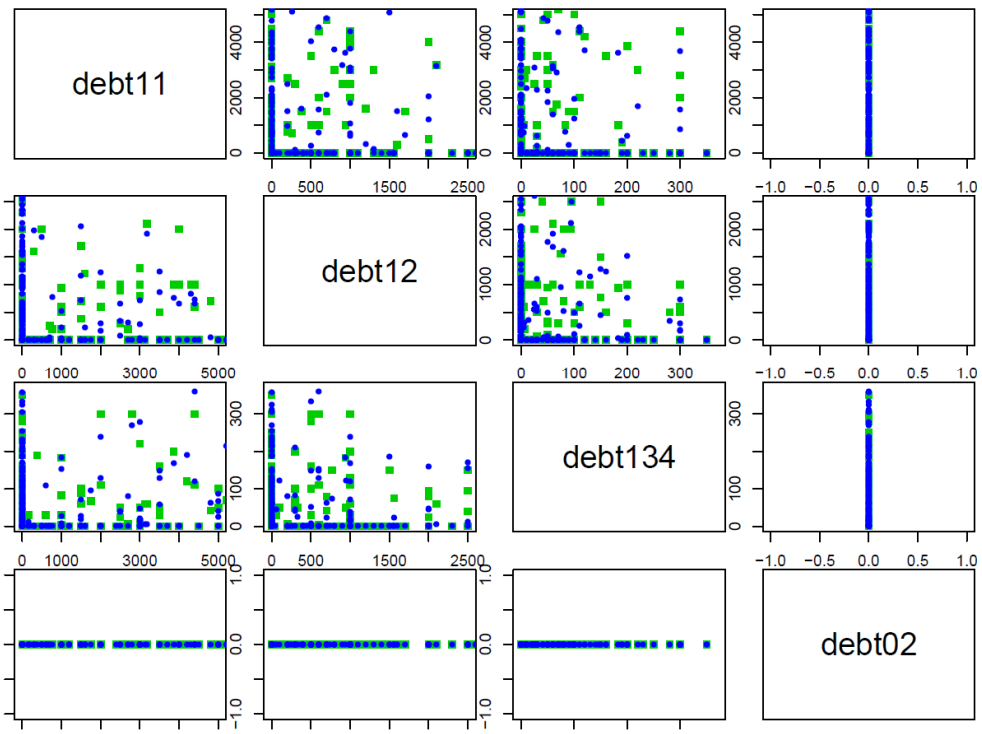
[부그림 6-10] 결합안3 적용 전후 자산 변수들의 산점도

〈부표 6-7〉 결합안3 적용 전후 자산 변수들의 상관계수

	original				결합안3			
	asset11	asset12	asset21	asset22	asset11	asset12	asset21	asset22
asset11	1.00	0.03	0.66	0.36	1.00	0.02	0.62	0.39
asset12	0.03	1.00	-0.02	0.05	0.02	1.00	-0.03	0.00
asset21	0.66	-0.02	1.00	0.43	0.62	-0.03	1.00	0.43
asset22	0.36	0.05	0.43	1.00	0.39	0.00	0.43	1.00



[부그림 6-11] 국소통합, 잡음 추가3 및 결합안3에 대한 부채 변수들의 상자그림



[부그림 6-12] 결합안3 적용 전후 부채 변수들의 산점도

〈부표 6-8〉 결합안3 적용 전후 부채 변수들의 상관계수

	original				결합안3			
	debt11	debt12	debt134	debt02	debt11	debt12	debt134	debt02
debt11	1.00	0.09	0.03	0.17	1.00	0.12	0.03	0.23
debt12	0.09	1.00	0.19	0.25	0.12	1.00	0.12	0.09
debt134	0.03	0.19	1.00	-0.01	0.03	0.12	1.00	-0.02
debt02	0.17	0.25	-0.01	1.00	0.23	0.09	-0.02	1.00

제7장

가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가

박민정

제1절 서 론

마이크로데이터를 공표할 때 개인정보의 노출을 제한하기 위한 통계기관의 노력은 다음 <표 7-1>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 7-1〉 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)]

✍ 이 연구는 2014년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

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 7-2>와 같다.

〈표 7-2〉 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

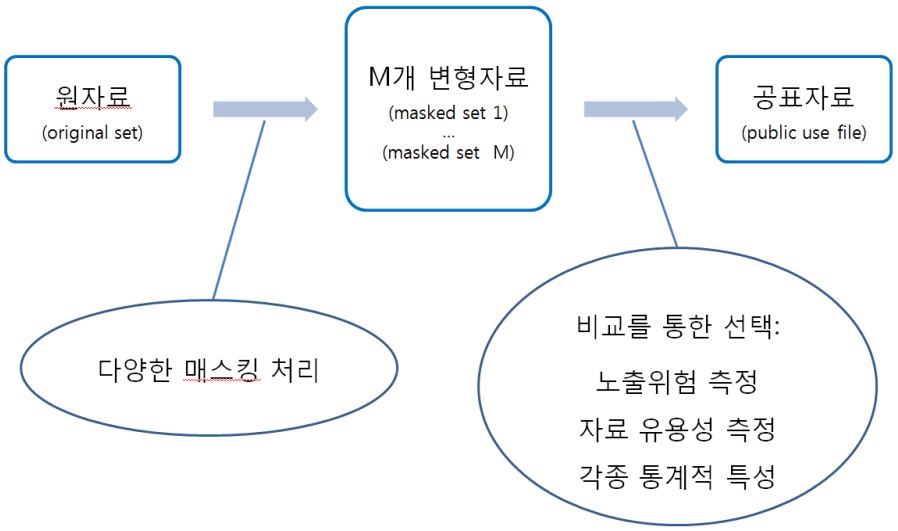
먼저 2006~8년에 인총 및 가계조사 자료를 대상으로 그룹화, 재코딩, 잡음추가 등의 몇몇 매스킹(masking) 기법들을 연구하였다. 이 연구들은 비밀보호의 중요성, 개념, 세부 기법을 소개한 의의를 가지기도 한다. 그 외에 표에 기록되지는 않았지만 각국 통계작성 기관의 비밀보호 현황에 관한 보고서들도 여러 편 작성되었고, 주요 물리적 노출 제한에 관한 각국의 현황도 정리되어 있다.

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 7-1] 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 네덜란드 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,

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

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 $\overrightarrow{y_j} = (y_{j1}, \ldots, y_{jp})$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

것으로 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.

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

이 때 $Z_{-j}^C=(\vec{z}_1^C,...,\vec{z}_{j-1}^C,\vec{z}_{j+1}^C,...,\vec{z}_n^C)$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. $\Pr(Z_{-j}^C|\vec{z}_j^C,J=j,\vec{t},Z^{Ad})$

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 $\overrightarrow{y_j^U}$의 분포를 예측하는 것은 회귀분석이나 시뮬레이션을 통해 이루어지며, $\overrightarrow{y_j^U}$까지 주어져 있을 때 $\overrightarrow{z_j^U}$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. $\Pr(Z_{-j}^C \mid \overrightarrow{z_j^C}, J=j, \vec{t}, Z^{Ad})$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.

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

이 성분은 $Z_{-j}^C = (\overrightarrow{z_1^C}, ..., \overrightarrow{z_{j-1}^C}, \overrightarrow{z_{j+1}^C}, ..., \overrightarrow{z_n^C})$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 7-3>은 지금까지 언급된 노출위험 측도들의 특징을 정리한 것이다.

〈표 7-3〉 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)	$\Pr(J=j) = \frac{\Pr(Z^C \mid J=j, \vec{t}, Z^{Ad})\Pr(J=j \mid \vec{t}, Z^{Ad})}{\sum_{j=1}^{n+1}\Pr(Z^C \mid J=j, \vec{t}, Z^{Ad})\Pr(J=j \mid \vec{t}, Z^{Ad})}$
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,

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

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 측도값을 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 7-4>는 지금까지 다루어진 자료유용성 측도들을 정리한 것이다.

〈표 7-4〉 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. $J = \frac{1}{p}\sum_{k=1}^{p} J_k,$ $J_k = \frac{1}{2}\left[\frac{U_k^{over} - L_k^{over}}{U_k^Y - L_k^Y} + \frac{U_k^{over} - L_k^{over}}{U_k^Z - L_k^Z}\right]$ 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 $U_c = \frac{1}{G}\sum_{j=1}^{G} w_j\left[\frac{n_{Yj}}{n_j} - c\right]^2, \quad c = \frac{N_Y}{N_T}$

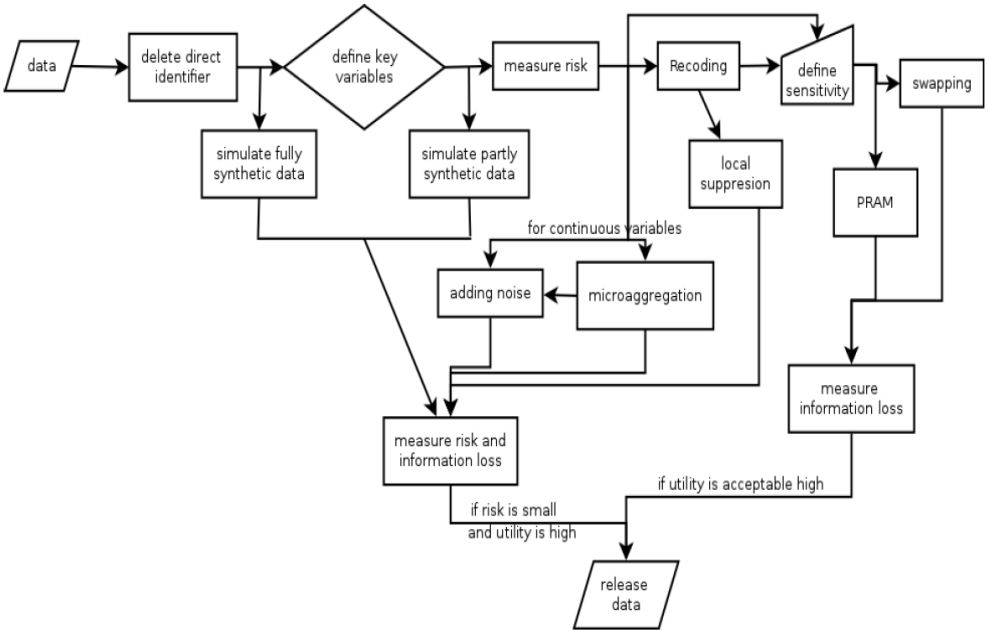
제4절 가계금융 · 복지조사 마이크로데이터 매스킹 방안 평가

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 7-2]와 같이 나타낼 수 있다(Meindl 등, 2013). 이 작업 흐름도를 살펴보면 마이크로데이터



[그림 7-2] 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 7-5>와 같다. 괄호 안은 변수별로 범주의 개수가 표시되어 있고, *은 재코딩 되었음을 의미한다. 키변수에 대한 매스킹 기법은 재코딩만을 고려하였다.

〈표 7-5〉 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 7-6>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 7-6〉 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 7-7>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 7-7〉 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 7-8>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 7-8〉 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.

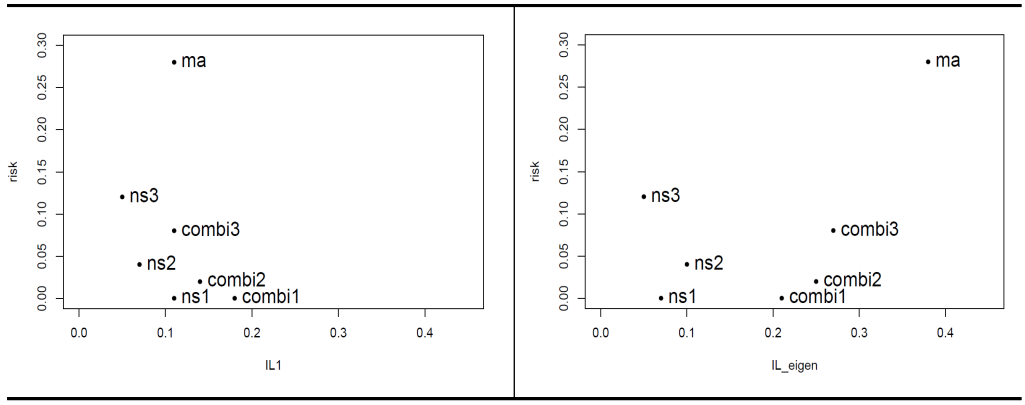
기존 연구의 결합안(combi)에서는 국소통합의 결과 값에 후 잡음추가 알고리즘에서 발생시킨 잡음을 더하였다[17]. 이때 잡음추가 후 민감변수에 대한 음의 값 처리는 위의 세 시나리오를 따라 수행하였다. 이들을 차례로 ma, ns1, ns2, ns3, combi1, combi2, combi3으로 각각 표기하면 상대적인 노출위험-정보손실은 다음 [그림 7-3]과 같이 나타난다.

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 7-3] 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]

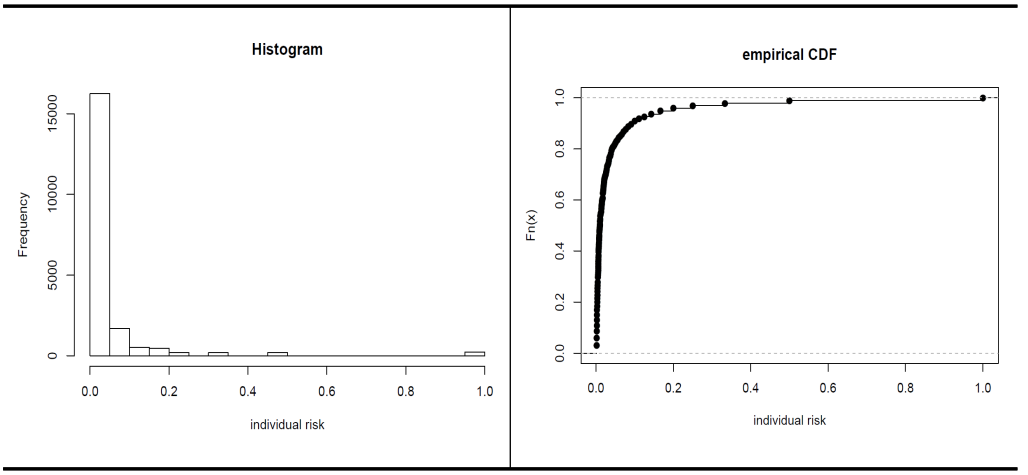
정보손실(information loss) 측도 IL1이나 IL_eigen 역시 거리 기반 측도이다. 매스킹 전후 레코드 사이의 거리를 계산하고 이를 표준 편차로 나눈 것이 IL1이며, 자료의 구조가 얼마나 변화되었는지 파악하고자 고유값들을 비교한 것이 IL_eigen에 해당한다. 매스킹의 목적은 노출위험과 정보손실 모두를 작게 만드는 것이나, 매스킹된 자료가 원자료와 가까울수록 노출위험(risk)은 커지고 정보손실(IL1)은 작아지게 되며 어느 하나를 줄이려면 다른 하나가 늘어나는 상충관계(trade-off)가 존재한다. 따라서 위 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 7-4]와 같다.



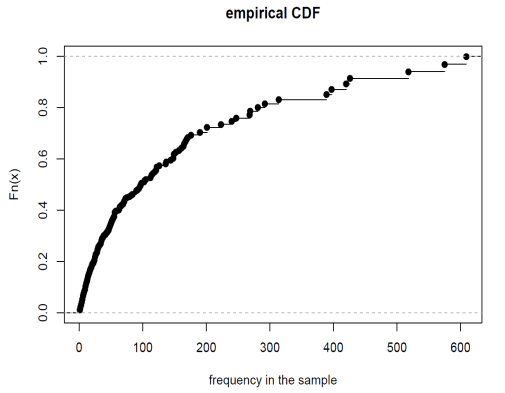
[그림 7-4] 0응답 여부에 따른 개별 노출위험 측도 값들의 분포

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 7-5]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 7-5] 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

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

(4-1)

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.

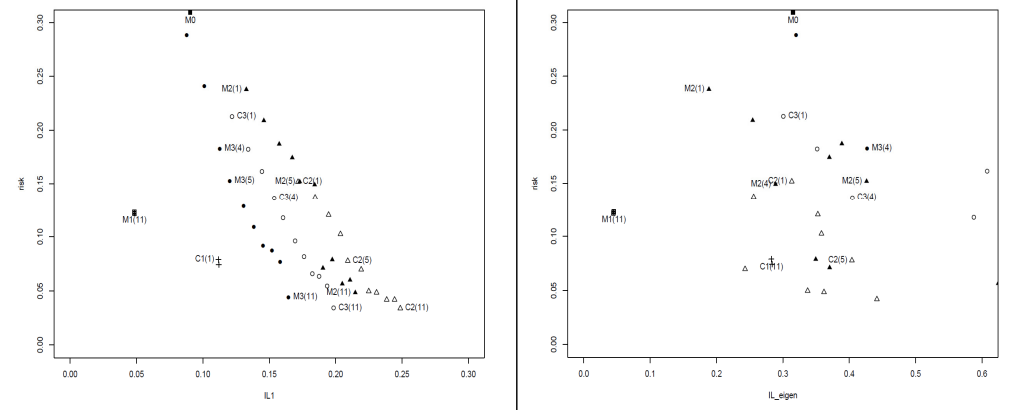
M2는 앞 절의 식 (4-1)을 전체 자료에 적용하며 매스킹 과정에서 음의 값이 발생하지 않는 잡음추가 기법이다. 이때 원래 0인 자료에 대해서는 c를 0으로 주어 2개 구역 차별처리(two-zone masking)를 통해 0응답 정보를 보존하였다. 다만 변수의 개수가 많아 원자료의 공분산 행렬이 양정치(positive definite) 행렬이 아니며, 이를 양정치 행렬로 변환시키기 위한 과정에서 많은 정보손실이 발생하는 문제가 있다. 따라서 식 (4-1)을 위해 잡음을 발생시킬 경우, 이론과 달리 매스킹 전후 상관계수 행렬에서 상대적으로 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 7-6] 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 결과값을 정리하면 다음 <표 7-9>와 같다. 전체 방안에 대한 노출위험 및 정보손실 측도에

대한 표는 <부록 3>에 첨부하였다.

〈표 7-9〉 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**

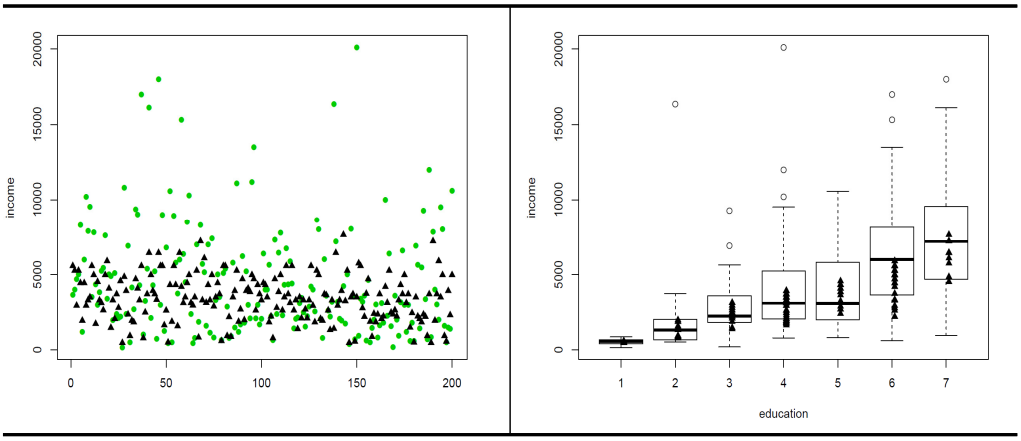
주: S는 상관계수 차이 절대값의 합

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 7-7]과 같다.



[그림 7-7] 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.

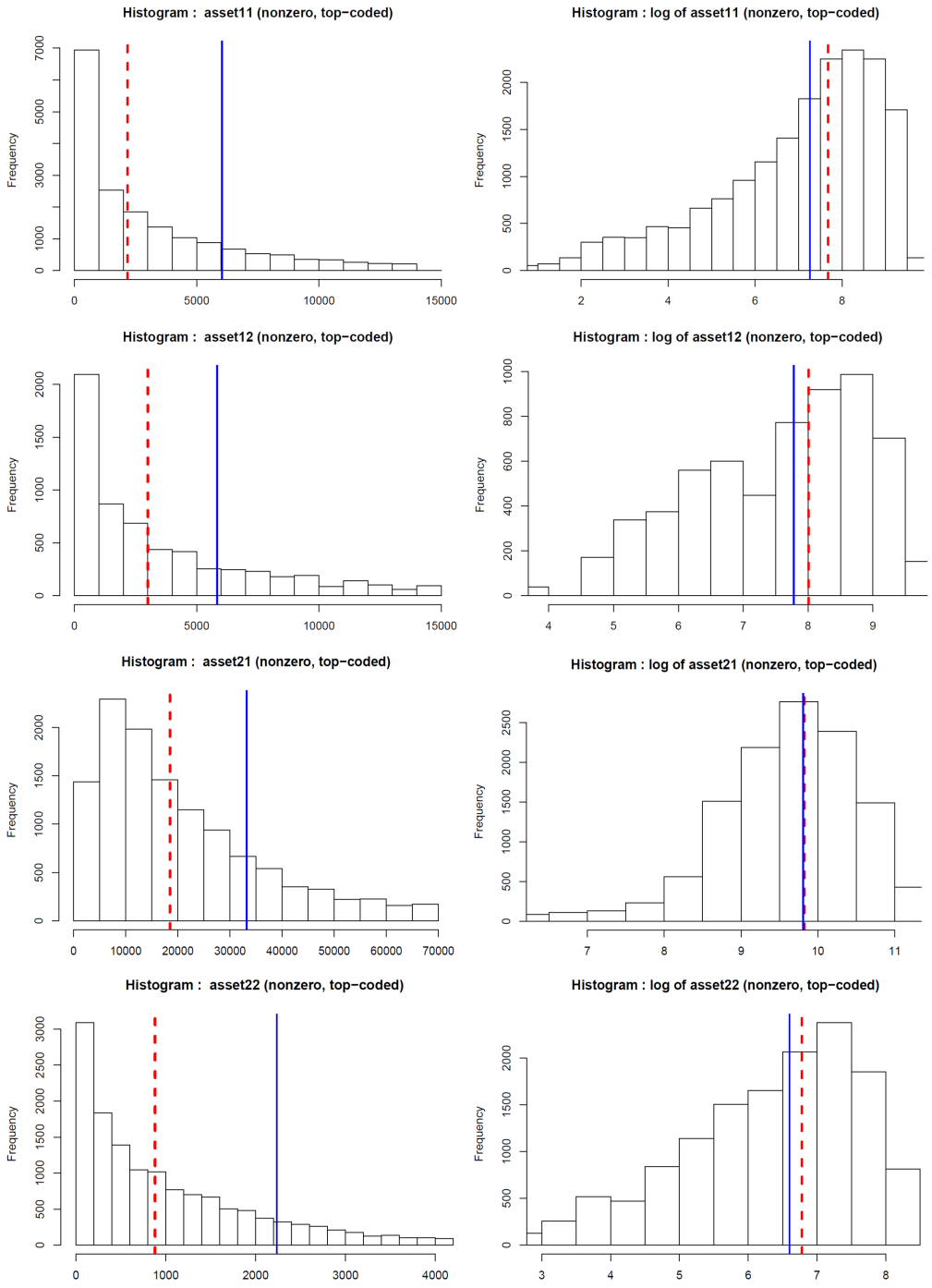
외부인이 표적이 표본에 있는 것을 아는 경우

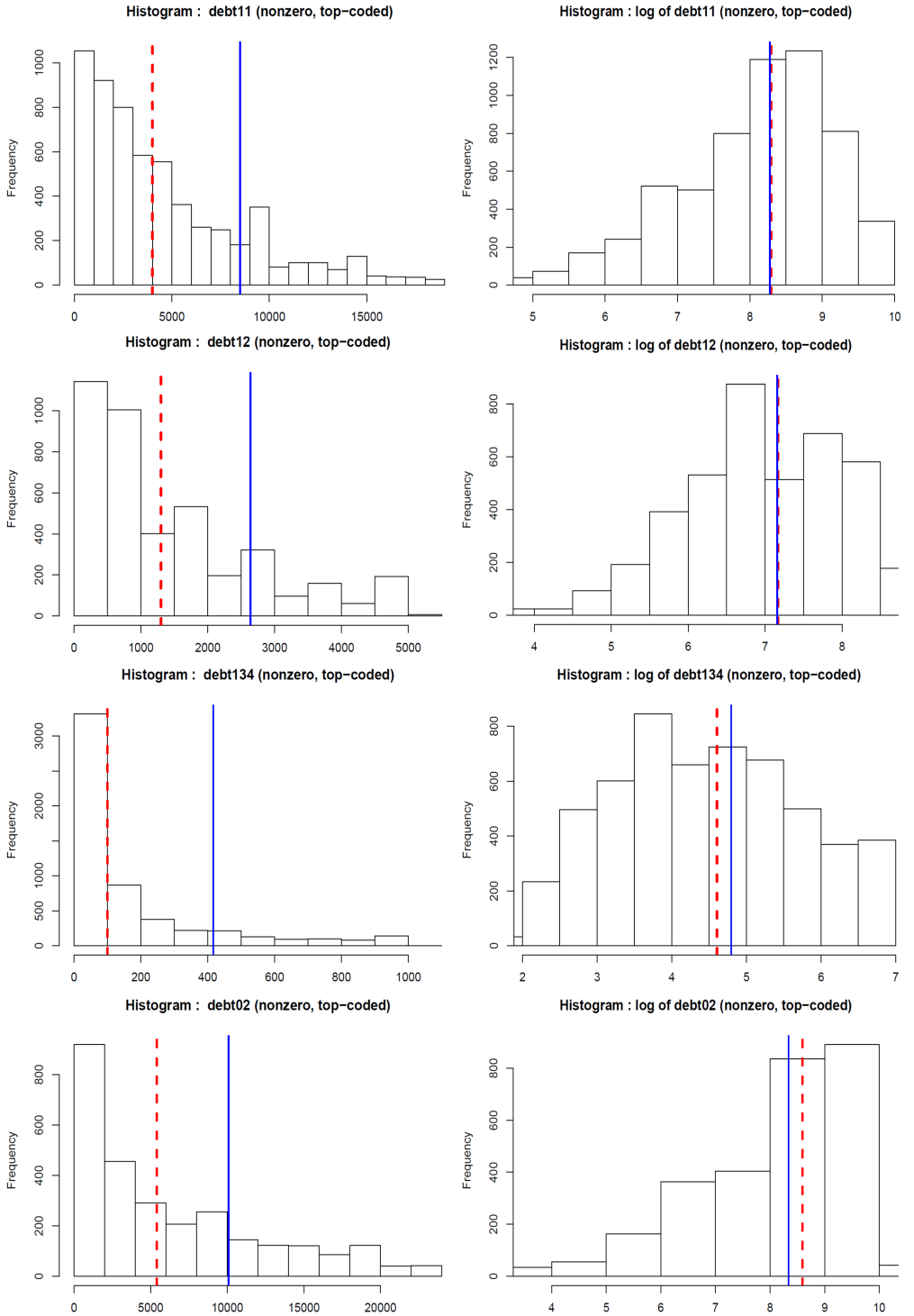
매스킹 처리	-	G	R,M	G,R,M
표적 유형 1 (Everyman)	.0022 (455)	.00045 (2230)	.0023 (410)	.00045 (2026)
표적 유형 2 (Unique)	1 (1)	1 (1)	.022 (10)	.0047 (49)
표적 유형 3 (Big I)	.0029 (345)	.00067 (1498)	.0031 (300)	.00069 (1355)
표적 유형 4 (Big P)	.0060 (166)	.0013 (776)	.0046 (169)	.00097 (836)

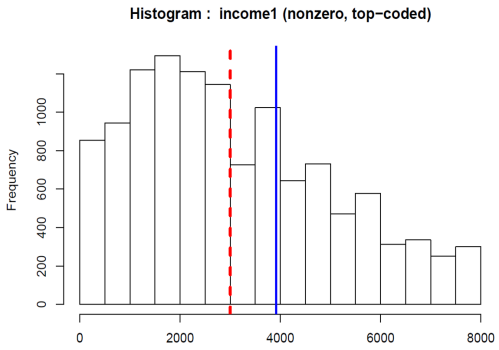
외부인이 표적이 표본에 있는지 모르는 경우

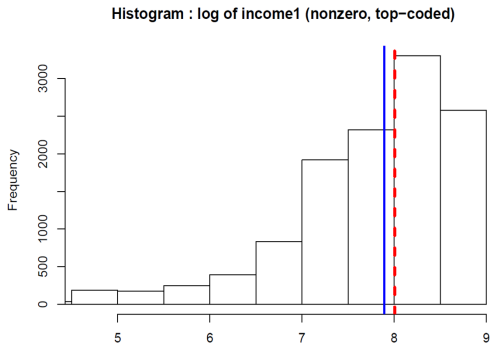
매스킹 처리	-	G	R,M	G,R,M
표적 유형 1 (Everyman)	.000001	.0000002	.000001	.0000002
표적 유형 2 (Unique)	.00032	.00032	.000001	.0000003
표적 유형 3 (Big I)	.000002	.0000003	.000002	.0000003
표적 유형 4 (Big P)	.000003	.0000006	.000002	.0000005

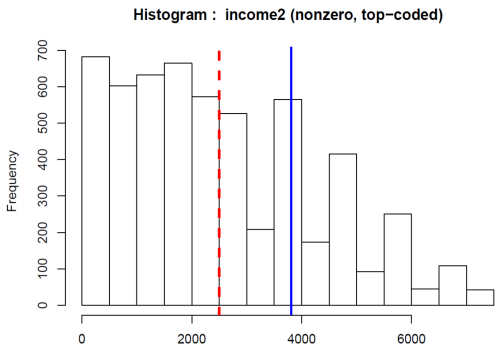
<부록 2>

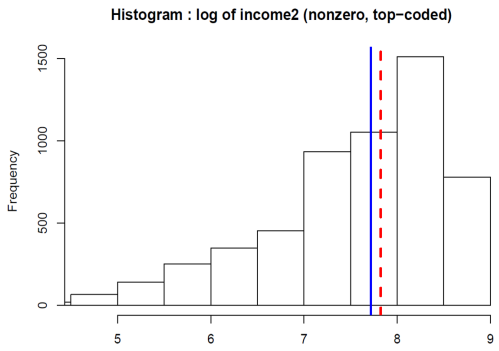

Histogram : asset11 (nonzero, top-coded)
Frequency
0
1000
3000
5000
7000
0
5000
10000
15000
Histogram : log of asset11 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
2
4
6
8
Histogram : asset12 (nonzero, top-coded)
Frequency
0
500
1000
1500
2000
0
5000
10000
15000
Histogram : log of asset12 (nonzero, 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

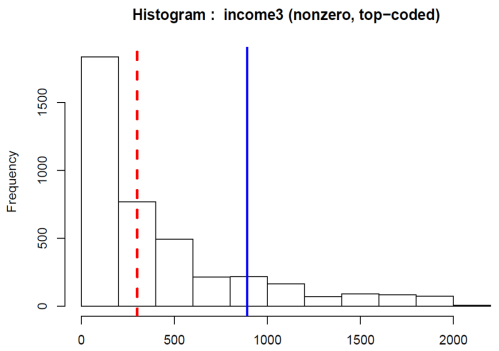

Histogram : debt11 (nonzero, top-coded)
Frequency
0
200
400
600
800
1000
0
5000
10000
15000
Histogram : log of debt11 (nonzero, top-coded)
Frequency
0
200
400
600
800
1000
1200
5
6
7
8
9
10
Histogram : debt12 (nonzero, top-coded)
Frequency
0
200
400
600
800
1000
0
1000
2000
3000
4000
5000
Histogram : log of debt12 (nonzero, top-coded)
Frequency
0
200
400
600
800
4
5
6
7
8
Histogram : debt134 (nonzero, top-coded)
Frequency
0
500
1000
2000
3000
0
200
400
600
800
1000
Histogram : log of debt134 (nonzero, top-coded)
Frequency
0
200
400
600
800
2
3
4
5
6
7
Histogram : debt02 (nonzero, top-coded)
Frequency
0
200
400
600
800
0
5000
10000
15000
20000
Histogram : log of debt02 (nonzero, top-coded)
Frequency
0
200
400
600
800
4
5
6
7
8
9
10

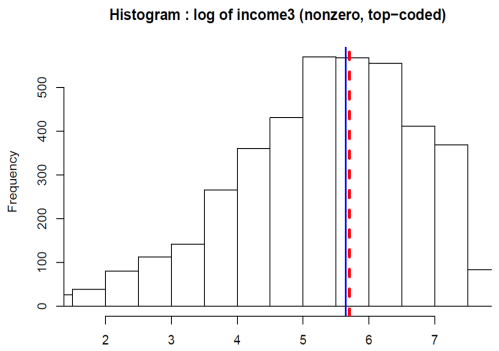

Histogram : income1 (nonzero, top-coded)
Frequency
0
200
400
600
800
1000
0
2000
4000
6000
8000

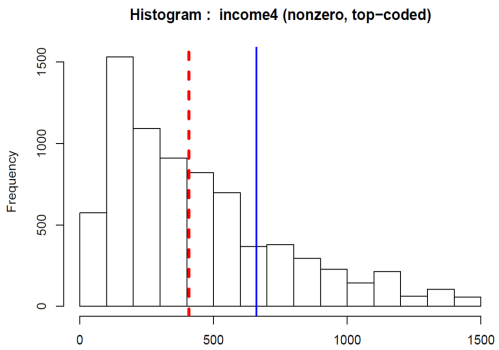

Histogram : log of income1 (nonzero, top-coded)
Frequency
0
500
1000
2000
3000
5
6
7
8
9

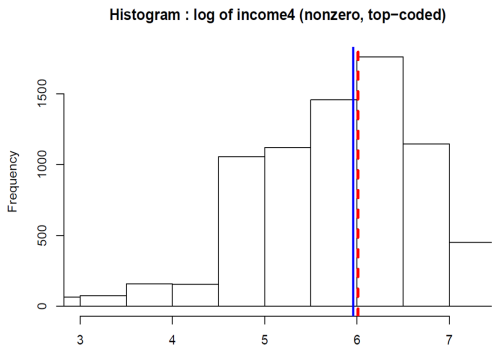

Histogram : income2 (nonzero, top-coded)
Frequency
0
100
200
300
400
500
600
700
0
2000
4000
6000


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

3부

매스킹 기법의 실무활용

8. 마이크로데이터 매스킹 기법 실무 활용 안내서

박민정

9. 통계작성기관 마이크로데이터 제공범위 설정 가이드라인 작성

박민정 · 성병찬 · 최유성

제8장

마이크로데이터 매스킹 기법 실무 활용 안내서

박민정

제1절 서론

요즘 우리나라 언론 보도를 살펴보면 과거에 비하여 통계를 이용해 작성된 기사가 점점 많아지고 있는 것을 느낄 수 있다. 통계청의 보도 자료를 단순하게 이용한 기사들도 있으나, 여러 자료를 모아 심층 분석하여 사회 현상을 해석하고 보도하는 사례들도 예전보다 빈번하게 찾아 볼 수 있다. 사회가 발전해 갈수록 국가 통계는 우리가 매일 접하는 뉴스부터 정부의 정책결정이나 학계의 학술 논문까지 뒷받침하는 공공재의 역할을 더욱 많이 수행해 나가며, 그 활용도가 점점 증가하게 됨을 부인하기 어렵다.

우리나라 통계청 등 각 통계작성 기관에서 생산하고 있는 각종 통계들은 대부분 매크로데이터라고도 불리는 집계 자료의 형태로 공표되고 있다. 매크로데이터 역시 매우 중요하며, 일반적으로 한 눈에 정보를 파악할 수 있도록 정리되어 발표되므로 그 활용도가 매우 높다고 할 수 있다. 매크로데이터는 개별 응답 자료들로 구성된 마이크로데이터를 사용해서 만들어지는데, 기관에서 정리해주지 않은 좀 더 많은 정보를 원하는 이용자들의 마이크로데이터에 대한 수요도 나날이 증가하고 있다. 또한, 정부3.0이라는, 공공정보를 국민에게 개방·공유하고 각종 서비스를 맞춤형으로 제공하고자 하며, 공공데이터 활용을 통한 창조경제의 활성화를 추구하는 현 시대에는 마이크로데이터의 공표가 더욱 중요한 문제이지 않을 수 없다. 즉, 효율적인 마이크로데이터 공표를 통한 통계 이용 활성화에 대한 시대적 요구가 커지고 있다고 할 수 있다.

✍ 이 연구는 2015년 상반기 통계개발원 연구보고서에 수록된 자료입니다.

이처럼 실생활에서 피부로 느낄 수 있는 언론 보도부터 각종 학술 연구와 정부의 정책 개발에 이르기까지 널리 활용 가능한 공공재로써 마이크로데이터는 매우 중요한 가치를 가지고 있으나, 이를 조사 내용 그대로 모두 공표하는 것은 위험할 수 있다. 조사에 응답한 개별 응답자들이 식별되어 원하지 않는 개인정보 유출의 위험이 조금이라도 있을 경우 문제가 될 수 있기 때문이다. 가상의 예를 들어, 제주도에 사는 50대 여성 군인이 지역·연령·성별·직업이라는 응답자 특성에서 전국에서 유일하면서 자산·부채·소득에 대하여 조사에 성실히 응답했는데, 그 이웃이 공표된 마이크로데이터에서 이 응답자를 찾아낼 수 있다면 개인정보 유출이 일어날 수도 있기 때문이다. 개인의 사생활이 중시되고 국가통계 작성을 위한 각종 조사의 응답률이 나날이 떨어지고 있는 환경 속에서, 마이크로데이터에 적절한 비밀보호 조치를 취하는 것은 개인 사생활의 보호일 뿐만 아니라 조사에 참여하는 응답자의 응답을 장려하는 가장 기본적인 작업이기도 하다.

마이크로데이터의 비밀보호를 위해 세계 각 통계작성 기관 및 학계의 연구자들은 활발한 연구 활동을 지속해 왔다. 비밀보호(confidentiality)는 통계적 노출제어(statistical disclosure control/limitation)라고도 불리며, 노출제어는 학술적으로 크게 매스킹, 인위자료 및 분석서버 기반 대안이라는 세 분야로 나눌 수 있다. 이 중 실무적인 차원에서 세계적으로 주로 활용되고 있는 것은 매스킹 기법들이라 할 수 있으며, 마이크로데이터 제공시 보다 완전한 통계적 노출제어를 위한 인위(synthetic)자료나 자료 제공·분석용 서버와 관련된 노출제어 연구는 계속해서 진행되고 있는 상황이다.

통계개발원에서는 최근 2년간 마이크로데이터의 비밀보호를 집중적으로 연구하였고, 해외 다른 기관들과 마찬가지로 마이크로데이터의 보다 완전한 통계적 노출제어를 위한 연구를 지속해나갈 계획이다. 다만, 이러한 흐름 가운데 현재 실무적인 차원에서 사용할 수 있는 매스킹 기법들의 적용 및 노출위험과 자료유용성 측정을 통한 매스킹 방안 평가의 노하우를 관련자들이 최대한 쉽게 습득할 수 있도록 관련 연구 결과를 안내서 형태로 정리하고자 한다. 본 보고서는 지난 2년('13~'14년)간의 통계개발원 연구보고서들을 바탕으로 재구성된 측면도 있어, 일부 내용은 기존 연구보고서들에서 요약 · 발췌(재인용)되기도 하였음을 밝혀 둔다.

본 보고서의 구성은 다음과 같다. 제2절에서는 통계적 노출제어의 종류를 소개하고, 몇몇 매스킹 기법들을 프로그램 설명과 함께 살펴본다. 이어 현재 연구되어 있는 노출위험 측정 방법론 및 자료유용성 측정 방법론을 제3절에서 소개한다. 최신 노출위험 측도를 이해하는 것이 어려울 수 있으나 이러한 개념을 어느 정도 습득하고 있는 것은 마이크로데이터 비밀보호 처리 담당자에게 필요할 것으로 생각된다. 관련 이론에 익숙한 독자는 이 부분들을 생략할 수 있다. 마지막으로 제4절의 요약 및 제언을 끝으로 보고서를 마치도록 한다. 이 보고서 전체에서 사용하는 프로그램은 R패키지인 sdcMicro이며, 사전에 알아야 할 관련 사용법은 <부록> 2.에 첨부하니 본문을 읽기에 앞서 참고하기 바란다.

제2절 통계적 노출제어와 매스킹

1. 노출제어의 종류

개인 식별 및 민감한 정보에 대한 노출제어는 각 통계기관의 자료 배포 과정에서 요구된다. 자료의 활발한 이용을 위해 마이크로데이터가 배포될 때 추구되는 원칙은 크게 다음 세 가지를 생각해 볼 수 있다(Reiter, 2004).

- 응답자의 식별 정보나 민감한 정보를 알려는 외부 시도에 대한 **안전성**
- 광범위한 통계적 분석을 뒷받침할 수 있는 **정보의 충분성**
- 보편적인 통계적 방법론들을 사용하는 이용자의 **접근 편의성**

통계기관이 이러한 원칙을 모두 만족시키면서 자료를 배포한다는 것은 쉬운 문제가 아니다. 조사된 상태 그대로의 정보를 모두 담아 마이크로데이터를 배포하면 개별 자료의 노출위험이 너무 커지고, 노출위험을 줄이기 위해 공표 범위를 줄이거나 자료를 변형하면 이용자가 원하는 수준에 비하여 정보 손실이 커지기 때문이다. 또한, 정보손실과 노출위험 모두 최소화하며 자료를 배포하기 위해서는 자료 이용자를 엄격히 통제하는 절차를 만들게 되므로 이용자의 접근 편의성이 떨어지게 된다. 위의 세 원칙을 모두 만족시키는 해결책을 찾은 사례는 현재까지는 없다고 보인다. 다만 지금까지 해외 통계기관의 비밀보호 처리 현황을 조사한 결과들을 살펴보면(김경미 외, 2007), 통계기관들은 다음 <표 8-1>과 같이 이용자의 유형에 따라 위의 세 목표 중 일부라도 달성할 수 있도록 자료를 배포하는 전략을 취해 해왔다고 할 수 있다.

〈표 8-1〉 노출제어의 종류와 배포되는 자료의 형식

종류	물리적 노출제어		통계적 노출제어(SDC[1])	
이용자	심층 이용자		불특정 다수	
전략/기법	접근제어	결과통제	매스킹처리	인위자료
배포형식	인가파일 데이터 실험실	가공 통계표 원격접속	공공이용파일	

주: 1) SDC/SDL: Statistical Disclosure Control/Limitation
출처: 가계금융·복지조사 마이크로데이터 제공을 위한 매스킹 방안(2013), 통계개발원

우선 마이크로데이터 이용자를 통계 소프트웨어를 이용하여 심층적으로 자료 분석을 하기 원하며 전문적 지식을 가진 이용자와 간단한 분석을 하기 원하는 불특정 다수의

이용자로 구분해 볼 수 있다. 심층 이용자는 조사된 마이크로데이터 원본을 원하는 경향이 있으며, 불특정 다수는 적절한 수준의 마이크로데이터라면 접근 편의성을 선호한다고 볼 수 있다. 이용자별로 노출제어 접근법을 물리적 노출제어와 통계적 노출제어로 구분해 볼 수 있는데 이에 대하여 하나씩 살펴보도록 하자.

가. 물리적 노출제어

먼저 자료의 심층 이용자들이 마이크로데이터를 통계적 목적으로만 이용하였음을 점검할 수 있고, 또한 생산된 결과물에 응답자의 노출위험이 없다면 마이크로데이터는 노출에서 안전하다고 할 수 있다. 물리적으로 심층 이용자의 접근을 통제하거나 결과물의 검열을 실시하는 것을 통해 그 목적을 이룰 수 있으며, 이러한 전략을 취할 때는 인가파일, 데이터 실험실, 가공 통계표 및 원격접속을 통해서 자료 배포가 이루어진다고 할 수 있다.

이 중 가공 통계표를 제공한다는 것은 이용자의 요구에 따라 통계기관이 마이크로데이터로부터 이용자가 원하는 결과물을 직접 작성하여 서비스하는 것을 말하며, 원격접속은 이용자의 접속을 원격으로 제어하며 생성된 결과물을 통계기관이 검열 후 제공하는 자료 배포 방식을 일컫는다. 이러한 결과물 통제 전략을 이용할 경우, 자료를 제공하는 통계기관의 부담이 크다고 할 수 있다.

다음으로 인가파일을 통한 자료 배포의 해외 사례를 살펴보면 다음과 같다. 미국 노동통계국은 자료 이용자의 소속 기관과 기관 대 기관 계약을 체결한 후 자료를 제공한다. 계약 후 자료 이용자 소속기관 책임자의 데이터 반환 및 소거 확인 절차, 자료 제공자인 노동통계국에 의한 성과물 검열 절차, 이용 기간 동안의 노동통계국 직원에 의한 현장 사찰 등의 과정을 거친다. 이러한 법적 절차와 물리적 검열 과정과 함께 제공하는 파일을 인가파일(licensed file)이라고 한다.

한편, 미국 센서스국에서는 총 9개의 데이터 실험실을 운영하고 있는데, 자료의 심층 이용자는 가까운 데이터 실험실을 찾아 센터의 서버에 접속하여 모든 필요한 작업을 수행할 수 있다. 얻어진 최종 성과물에 대해서는 심사위원회의 승인 후 학술지 투고가 허용되며, 전체적으로 자료 이용 제안서 제출부터 데이터 이용까지 총 6~12개월이 소요될 것을 예상하도록 권고하고 있다. 인가파일이나 데이터 실험실을 통한 자료 배포는 정보의 충분성 및 비밀보호 측면에서는 효율적일 수 있으나, 자료 이용자와 배포 기관 모두에게 복잡한 절차와 경제적 비용을 요구하게 된다.

이에 각국 통계기관은 불특정 다수가 용이하게 접근하며 통계적으로 자료를 이용할 수 있으나, 응답자의 정보가 식별되지 않는 공공이용파일(Public Use Microdata File)을 제공하여 마이크로데이터에 대한 수요를 좀 더 많이 충족시키고자 노력하고 있다. 매우 심층적인 마이크로데이터 분석이 필요한 이용자들은 복잡하고 비용이 발생하더라도

물리적 제어에 응할 것이나, 그러한 수요를 최대한 줄이고 통계적 노출제어 처리된 공공이용파일 배포 비중을 높인다면 기존 이용자의 접근성을 높일 뿐만 아니라 추가적인 자료 이용 활성화에도 도움이 될 것이기 때문이다.

나. 통계적 노출제어

통계적 노출제어 처리가 된 공공이용파일은 현재까지는 크게 매스킹 기법들 및 인위자료 생성 기법을 통해 생산되어 왔다고 할 수 있다. 매스킹(masking)은 전통적인 노출제어 기법으로 활용되어 왔으며, 자료에 적절한 변형을 가하여 변형된 자료와 변형 방법을 함께 공표하는 것을 말한다. 인위(synthetic)자료 생성 방법은 비교적 최근에 만들어진 것으로, 베이지안 기법을 활용하여 원자료의 분포를 따르지만 원래 값과는 다른 파일을 생성하여 배포하고자 하는 목적으로 연구되어 왔다. 먼저 인위자료에 대해 간략히 살펴보면 다음과 같다.

인위자료 생성의 기본 아이디어는 표본 조사를 통해 모집단을 대표하는 다수의 표본 추출이 가능한 것처럼, 조사된 표본과 구조가 유사한 인위자료 세트들을 통계적으로 생성하자는 것이며, 생성 방법은 베이지안 사고를 기반으로 한다. 목표로는 ①공표 자료 내 어떤 응답 단위도 원래 자료 내 민감변수의 값을 가지지 않고, ②원래 자료에서 가능한 통계 분석이나 추론이 공표 자료에서도 가능할 것을 추구한다(Rubin, 1993). 이것은 노출위험은 없이 자료 구조를 보존하여 마이크로데이터를 공표하고자 하는 것을 의미한다. 인위자료 생성 과정은 ①모집단에서 개체(레코드[1])를 추출해 인위자료 세트(들)을 구성하고 ②원래의 조사 자료를 기반으로 만든 모형을 가지고 다중대체(multiple imputation) 기법(Rubin, 1993)을 사용해 위의 인위자료 세트를 채우며 ③각 인위자료 세트들에서 얻은 통계량을 결합하는 방법(combining rule)을 제시하는 단계를 가진다. 관련 연구 결과로는 2001년 프랑스의 종단 연계 자료(Abowd와 Woodcock, 2001), 독일 IAB 기관 패널 조사(Drechsler와 Reiter, 2009) 및 미국 사업체 종단자료(Kinney 외, 2011)의 인위자료 생성 사례가 있다. 독일과 미국 사례에서는 전체 자료를 인위자료로 교체하지 않고 부분적으로 인위자료를 교체하는 방법을 사용하였다.

인위자료의 이용은 기존의 매스킹 기법만으로는 어려웠던 노출제어와 정보의 충분한 보존이라는 서로 상충되는 두 목적을 성취할 수 있는 길을 열어주는 획기적인 아이디어로 간주되었다. 장점은 생성되는 인위자료의 정확성 조정을 통해 노출위험 수준을 제어하고, 조사된 자료의 분포와 통계적으로 유사한 분포에서 인위자료를 생성하여 다양한 추론에 대한 통계적 타당성을 제공하며, 또한 소지역을 포함한 지리적 정보 공표의 가능성도 가진다는 것이다. 나아가 비밀보호 기법과 관련된 새로운 통계 이론이나 프로그램 습득이 불필요하고 원래 자료의 표본추출법에 관한 고려 없이 단순 추출을 가정하고 통계 분석이

1) 레코드란 한 응답자가 응답한 모든 값들을 일컫는 것으로 행렬 형식의 마이크로데이터에서 행에 해당한다고 이해할 수 있다. 레코드라는 말 대신에 응답자를 일컫는 개체라는 표현도 사용하도록 하겠다.

가능하여 이용자의 편의성을 크게 높이기도 한다. 반면 단점으로는 인위자료를 만드는 모형 및 가정에 지나치게 의존적이어서 추정량의 편의가 크게 된다는 측면이 있다. 공표되는 인위자료가 조사 자료와 완전히 다르기 때문에 베이지안 기법에 익숙하지 않은 자료 이용자들에게는 심리적 거부감을 줄 수 있다는 문제점도 가진다. 때문에 어느 통계기관도 인위자료 생성 방법을 전적으로 받아들이지는 않았으며, 인위자료 관련 연구는 준모수적 및 비모수적 모형 탐구, 다양한 구조를 가지는 모집단에 대한 추론 타당성 연구 및 부분적 인위자료 활용 연구로 발전하는 모습을 보이고 있다.

인위자료는 최근에 와서 집중적으로 연구되었고, 실제 자료에 적용하는 것은 실험적으로 이루어져 왔다고 할 수 있다. 참고로, 인위자료에 이어서 자료 배포에서 차등 정보보호 개념(Dwork, 2006; 이용희, 2013)이 적용된 자료 제공·분석용 서버 활용의 필요성이 제기되면서 이에 관한 연구 역시 최근 활발하게 진행되고 있다. 이상의 인위자료 및 차등 정보보호 개념 활용을 통한 서버 기반 대안 관련 내용은 간단히 소개만 하며, 본 보고서의 범위를 넘는 것으로 간주한다. 관심 있는 독자는 개인정보보호의 법률적 측면부터 차등 정보보호까지 소개한 문헌으로 오미애 외(2014)의 6장을 참고하기 바란다.

인위자료 등 최근에 통계적 노출제어를 위한 새로운 방법들이 발전하고 있지만, 실무에서 노출제어 처리된 마이크로데이터의 공표를 위해서는 전통적으로 매스킹 기법들이 주로 사용되어 왔다고 할 수 있다. 이러한 매스킹 기법들은 종류가 많고 각 기법별로도 세부 알고리즘이 다수 존재한다. 각각의 기법 및 알고리즘들은 오랜 시간에 걸쳐 연구·발표되어 왔고, 네덜란드 통계청이나 오스트리아 통계청의 후원으로 관련 프로그램들도 만들어져 배포되었다. 이에 대한 내용은 다음 2.2절에서 다루도록 한다.

본 보고서에서는 자료를 배포하고자 하는 실무자가 매스킹 기법을 적용하고자 할 때 어떻게 프로그램을 활용할 수 있는지를 기술하고자 하였다. 본 보고서를 따라 매스킹 기법들을 사용할 수 있지만 매스킹 작업이 노출위험 제어 및 정보손실 최소화에 가장 적합한 방법이 아닌 차선책이라는 한계를 가지고 있음을 강조하며, 인위자료 및 차등 정보보호 이론 등을 이용한 궁극적인 해결 방안 모색을 위한 연구가 현재 여러 나라의 통계기관 및 학계에서 이루어지고 있음을 다시 한번 언급하고자 한다.

2. 매스킹 기법의 종류

매스킹(masking)이란 원자료에 적절한 변환을 통해 간접적인 식별 정보를 가리는 것이며, 매스킹된 자료란 변환된 자료와 변환에 관한 정보 모두를 의미한다. 변환에 관한 정보는 이용자의 분석이 올바르게 진행되는데 필요하기 때문에 함께 제공되어야 한다. 여러 변수의 조합으로 인해 간접적인 식별 정보가 생길 때 자료들을 약간씩 변형하여 그러한 식별이 불가능

하도록 해주는 기법들을 매스킹 기법이라 하지만, 자료를 이용할 때 어떠한 변환이 행해졌는지에 대한 정보가 주어지는 것은 이용자의 적절한 자료 분석에 필수적이라 하겠다.

매스킹 기법들은 크게 네 가지 범주로 구분할 수 있다. 여기에는 ①부분적으로 자료 값을 제공하거나(sampling), ②자료 값을 가리거나(suppression), ③구분할 수 없는 그룹으로 값을 섞어 제공하거나(aggregation), 혹은 ④자료 값을 변조(perturbative data)하는 것들이 해당된다. 또한, 매스킹 기법들을 연속형이나 범주형 등 변수의 형식에 따라 적용할 수 있는 기법들로 분류하기도 한다. 이 절에서는 비교적 널리 사용되고 있는 매스킹 기법들을 하나씩 간단히 살펴보도록 한다. 각각의 기법들에는 세부 종류가 많고, 전문적으로 방대하게 연구되어 자세한 내용을 여기에서 모두 다루기는 불가능하다. 자세한 내용에 관심이 있는 독자들은 필요에 따라 기법별 관련 논문들을 참고하기를 권한다. 일반적으로 매스킹 기법들을 자세히 공부하고자 한다면 Duncan 외(2011)의 5장 및 관련 참고문헌들이나 통계자료 비밀보호론(충남대)을 권하는 바이다.

가. 국소 감추기(local suppression)

마이크로데이터를 행은 개체(레코드)들, 열은 속성 변수들로 구성된 행렬처럼 생각할 수 있다. 감추기(suppression)는 보통 레코드나 속성 차원에서 이루어진다. 예를 들면, 특정 금액 이상의 고소득자의 자료를 모두 감추거나(레코드 감추기), 레코드의 식별을 간접적으로 도울 수 있는 지역 변수를 모든 레코드에 대하여 감추는 것(속성 감추기) 등이다. 즉, 행렬에서 한 행 혹은 한 열을 통째로 감추는 것이다. 이러한 감추기는 비밀보호는 확실하나 너무 많은 정보의 손실을 야기한다.

국소 감추기(local suppression)는 노출위험을 높이는 속성변수 조합의 몇 개의 값만을 감추는 것을 말한다. 이 경우 마이크로데이터 행렬에서 한 셀을 감추는 것이 된다. 예를 들면, 변수 조합이 지역·연령·성별·직업이면서 노출위험이 높은 어떤 레코드의 해당 변수들의 값이 A지역·50대·여성·군인일 경우, 성별인 여성이라는 값이나 직업인 군인이라는 값을 감추는 것을 생각할 수 있다. 만약 자료 이용자가 직업에 관한 정보를 더 필요로 할 경우에는 직업 대신에 성별을 감추는 것이 유익할 것이다.

국소 감추기를 사용할 경우에는 레코드별 노출위험을 측정하여 노출위험이 높은 레코드를 선별하는 과정 및 해당 레코드의 값들 중 어느 변수의 값을 감출지 선택하는 과정이 필요하다. 먼저 노출위험은 공표하는 변수 조합들에 대하여 유일한 레코드가 발생하지 않도록 익명성[2]을 2 이상 확보하도록 할 수 있다. 다음으로 되도록 감추어지지 않기를 원하는 변수를 선택하는 알고리즘을 사용할 수 있다. 국소 감추기는 수행할 때마다 정보 손실을 초래하기 때문에 통계

2) 익명성(k-anonymity)이란, 특정 변수 조합에 대해 같은 값을 가지는 레코드의 수를 말한다.

개발원에서 수행하였던 기존 연구에서는 마이크로데이터에 국소 감추기 외의 매스킹 기법들을 적용한 후에, 마지막으로 필수적인 수준의 익명성 획득에 활용하고자 하였다.

프로그램에서 국소 감추기 등 각 매스킹 명령어들을 사용할 때는 입력변수는 matrix, data.frame 및 sdc 객체(object)일 수 있다. 본 보고서는 sdc 객체를 사용하는 경우를 중심으로 작성되었다. sdc 객체에 대하여는 <부록> 1.을 먼저 읽어보기 바란다. 이 절의 매스킹 명령어를 설명하는 예제에서는 sdcMicro 패키지에 주어진 마이크로데이터인 francdat를 이용한다. 데이터 francdat는 8개 변수와 8개 레코드를 가지며 data.frame 형식으로 되어 있다. 이 마이크로데이터를 실행하기 위한 코드는 다음 <표 8-2>와 같다. 앞으로 2절의 매스킹 명령어들을 사용하는 예에서는 francdat를 aa로 표현하여 사용하도록 한다.

〈표 8-2〉 sdcMicro의 예제 데이터 francdat 설명

```
library(sdcMicro)      # sdcMicro 패키지 불러오기
data(francdat)         # 마이크로데이터 예제로 francdat 불러오기
aa <- francdat         # 아래 표는 aa의 값들을 보여준다. 참고로 열 번호를 표시하였다.
```

aa	1	2	3	4	5	6	7	8
	Num1	Key1	Num2	Key2	Key3	Key4	Num3	w
1	0.3	1	0.4	2	5	1	4	18
2	0.12	1	0.22	2	1	1	22	45.5
3	0.18	1	0.8	2	1	1	8	39
4	1.9	3	9	3	1	5	91	17
5	1	4	1.3	3	1	4	13	541
6	1	4	1.4	3	1	1	14	8
7	0.1	6	0.01	2	1	5	1	5
8	0.15	1	0.5	2	5	1	5	92

R패키지 sdcMicro를 사용하여 국소 감추기를 시행할 경우, 코드, 예제 및 설명은 다음 <표 8-3>과 같다.[3] 표는 차례로 일반적인 명령어 사용 형식, 명령어 입력변수 설명, 출력내용 설명 및 실행 예제로 구성되어 있다. sdc 객체를 이용할 경우[4] localSuppression과 같은 매스킹 명령어의 입력변수 및 출력 변수가 모두 동일한 sdc object이지만, <부록>에 설명된 대로 @risk 등, 주어진 sdc

3) 통계패키지 R 및 sdcMicro에 대한 소개는 <부록>을 참고하길 바란다. <부록>에 설명된 바와 같이 마이크로데이터 aa에 대하여 aa.sdc를 마이크로데이터 aa를 이용해 만든 sdc 객체라고 하자.

4) 입력변수로 matrix 혹은 data.frame도 사용가능하나 출력 형식이 다르다. 본 보고서는 sdc 객체를 사용하는 경우를 중심으로 작성되었다.

object의 내부 값들이 모두 변하게 된다.[5] 옵션 중에서 k=2 대신에 k=3 등을 입력하여 익명성을 강화하여 정보손실이 크더라도 대신에 좀 더 강도 높은 비밀보호를 할 수 있다.

한편, 키변수가 모두 p개 있다고 할 때, 감추지 않기를 원하는 키변수부터 나타나도록 1부터 p의 숫자로 구성된 벡터를 옵션 중 importance에 입력하여 국소 감추기를 시행할 수 있으나, 실제 프로그램에서는 잘 작동하지 않는 경우가 많았다.

〈표 8-3〉 국소 감추기 명령어 형식 및 사용 예제

localSuppression(obj, keyVars= , k=, importance=NULL)

obj: 입력 객체 matrix, data frame 또는 sdc object
keyVars: 국소 감추기를 하려는 키변수 열을 나타내는 벡터 혹은 키변수 이름 벡터
k: 확보할 익명성 개수, default는 2

aa.sdc@localSuppression: 키변수별로 국소 감추기가 이루어진 자료 개수
aa.sdc@origData 및 aa.sdc@manipKeyVars: 처리 전후 마이크로데이터

```
aa.sdc <- createSdcObj(aa, keyVar=c(4,5,6))      # 4,5,6열을 키변수로 sdc 객체 생성
aa.sdc
```

```
Number of observations violating
 - 2-anonymity:  4     # 익명성 2를 확보하지 못한 레코드는 총 4개
 - 3-anonymity:  8     # 익명성 3을 확보하지 못한 레코드는 총 8개
```

```
aa.sdc <- localSuppression(aa.sdc, k=2)      # 익명성 2를 확보하는 국소 감추기
aa.sdc
```

```
Number of observations violating
 - 2-anonymity:  0 (orig:  4 )     # 익명성 2를 확보하지 못한 레코드는 총 0개
 - 3-anonymity:  6 (orig:  8 )     # 익명성 3을 확보하지 못한 레코드는 총 6개
```

```
aa.sdc@manipKeyVars      # 매스킹 이후 변화된 키변수 값들
                         # 4번과 7번 레코드의 변수 Key4의 값들이 감추어졌다.
```

	Key2	Key3	Key4	idvarextraforsls
1	2	5	1	NA
2	2	1	1	NA
3	2	1	1	3
4	3	1	NA	NA
5	3	1	4	5
6	3	1	1	6
7	2	1	NA	7
8	2	5	1	8

5) 매스킹 처리로 변환된 마이크로데이터는 aa.sdc@manipKeyVars에 나타나며, 매스킹 이전의 마이크로데이터는 aa.sdc@origData에 저장되어 있다. 기타 달라지는 노출위험 등 역시 aa.sdc@ 형태로 저장되어 출력해 볼 수 있다.

나. 전반적 재코딩(global recoding)

전반적 재코딩이란 특정 속성이 가지는 범주를 더 상위 범주(less-specific categories)로 묶는 것을 말한다. 예를 들면 직업이 의사와 간호사일 경우 모두 의료계 종사자로 묶는 것을 생각할 수 있다. 연령 변수도 마찬가지로 각 세 연령을 5세나 10세 단위로 묶어 전체 마이크로데이터의 익명성을 확보하는데 도움을 줄 수 있다. 참고로 전반적 재코딩 중에서 특정 값 이상/이하를 묶는 것을 top/bottom 코딩이라고 하며, 그 그룹에 속한 값들을 해당 그룹의 평균이나 중앙값으로 대체하여 표시하곤 한다. 보통은 각 키변수에 대하여 빈도수가 너무 적은 그룹이 생기지 않도록 재코딩 작업을 하는 것이 바람직하다고 할 수 있다. R패키지 sdcMicro를 사용하여 전반적 재코딩을 시행할 경우, 코드, 예제 및 설명은 다음 <표 8-4>와 같다.

〈표 8-4〉 전반적 재코딩 명령어 형식 및 사용 예제

globalRecode(obj, column= , breaks= , labels=)

obj: 입력 객체 matrix, data frame 또는 sdc object
column: 전반적 재코딩을 시행할 열 번호 벡터
breaks: 재코딩 경계를 나타내는 벡터
labels: 재코딩으로 인해 변환될 값을 나타내는 벡터

aa.sdc@origData 및 aa.sdc@manipKeyVars: 처리 전후 마이크로데이터

```
aa.sdc <- createSdcObj(aa, keyVar=c(2,4,5,6))
aa.sdc <- globalRecode(aa.sdc, column=1, breaks=c(0,3,7), labels=c("a", "b"))
# 키변수로 지정된 변수 중 첫 열에 대하여 전반적 재코딩 시행
aa.sdc@origData[, aa.sdc@keyVars]
aa.sdc@manipKeyVars      # 첫 키변수에 대한 재코딩 결과, 1,3 → a, 4,6 → b로 변환
```

aa.sdc@origData[,aa.sdc@keyVars]

	Key1	Key2	Key3	Key4
1	1	2	5	1
2	1	2	1	1
3	1	2	1	1
4	3	3	1	5
5	4	3	1	4
6	4	3	1	1
7	6	2	1	5
8	1	2	5	1

aa.sdc@manipKeyVars

	Key1	Key2	Key3	Key4
1	a	2	5	1
2	a	2	1	1
3	a	2	1	1
4	a	3	1	5
5	b	3	1	4
6	b	3	1	1
7	b	2	1	5
8	a	2	5	1

다. 국소통합(microaggregation)

국소통합이란 자료를 최소 k개(보통 3개) 이상의 레코드를 가지는 그룹으로 묶고, 각 그룹의 자료 값들을 그 그룹의 평균이나 중앙값 등의 동일한 한 값으로 대체하는 기법을 말한다. 레코드를 묶어 그룹을 만들 때, 단일 변수를 기준으로 할 수도, 여러 변수를 모두 고려하여 그룹을 정할 수도 있다. 여러 변수를 한꺼번에 고려하기 위해서는 주성분을 이용하거나, 네델란드 통계청에서 제안했던 자료 간 거리가 먼 것부터 묶어가는 MDAV[6] 알고리즘을 이용할 수도 있다(Statistics Netherlands, 2007). 그 외에도 다양한 종류의 그룹화 기법이나 그룹 내 자료의 수 k의 결정에 관해 비교한 연구들도 있으니 관심 있는 독자는 Mateo-Sanz와 Domingo-Ferrer(1998)를 참고하면 된다.

R패키지 sdcMicro를 사용하여 국소통합을 시행할 경우, 코드, 예제 및 설명은 다음 <표 8-5>와 같다. 그룹화 방법을 "onedims"로 선택한 경우, 각 변수에 대하여 크기 순서대로 3개씩 묶어 평균으로 대체하여 변수별로 값이 달라진 것을 확인할 수 있고, "mdav"로 선택한 경우, 묶인 그룹에 대하여 세 변수 모두 같은 값으로 대체된 것을 확인할 수 있다. 단일 변수에 대해 국소통합을 실시하는 알고리즘의 명령어로는 single, simple, onedims가 있으며, 여러 변수를 함께 고려하여 국소통합을 실시하는 명령어로는 mdav, pca, influence, clustpca 등이 있다. 그 외에 로버스트한 기법이나 군집 방법을 응용한 알고리즘들인 rmd, pppca, clustpppca, clustmcdpca, mcdpca 등도 있으나 실제로 자료 분석에 사용하였을 때, 잘 작동하지 않는 경우가 많았다. R패키지 sdcMicro의 영문 매뉴얼에서는 mdav를 사용할 것을 추천하였고, 통계개발원 기존 연구인 박민정 외(2013)에서는 적용 가능한 각 알고리즘들을 실제 자료에 모두 적용하고 노출위험과 정보손실 측도 결과를 이용하여 각 알고리즘을 평가하였다.

〈표 8-5〉 국소통합 명령어 형식 및 사용 예제

microaggregation(obj, variables= , method= , aggr=3)
obj: 입력 객체 matrix, data frame 또는 sdc object variables: 국소통합을 시행할 변수를 나타내는 열 번호 벡터, sdc 객체는 numVars에 실행 method: 국소통합의 세부 알고리즘을 나타내는 명령어로 pca, rmd, onedims, single, simple, clustpca, pppca, clustpppca, mdav, clustmcdpca, influence, mcdpca가 있음 aggr: 국소통합으로 묶이는 각 그룹당 자료의 개수, default는 3

6) Multivariate Microaggregation based on Maximum Distance to Average Vector

aa.sdc@origData 및 aa.sdc@manipNumVars: 처리 전후 마이크로데이터

```
aa <- francdat
aa.sdc <- createSdcObj(aa, keyVar=c(2,4,5,6), numVar=c(1,3,7))
aa.sdc <- microaggregation(aa.sdc, variables=c('Num1', 'Num2', 'Num3'),
                           method="onedims", aggr=3)
# 지정된 변수명을 가진 변수들에 대해 국소통합 시행
```

원래 변수값들	각 단일 변수 기준 국소통합	세 변수 동시 고려 국소통합
aa.sdc@origData[,aa.sdc@numVars]	method="onedims"일 경우 round(aa.sdc@manipNumVars, 2)	method="mdav"일 경우 round(aa.sdc@manipNumVars, 2)

	Num1	Num2	Num3
1	0.30	0.40	4
2	0.12	0.22	22
3	0.18	0.80	8
4	1.90	9.00	91
5	1.00	1.30	13
6	1.00	1.40	14
7	0.10	0.01	1
8	0.15	0.50	5

	Num1	Num2	Num3
1	0.88	0.21	3.33
2	0.12	0.21	29.60
3	0.88	2.60	29.60
4	0.88	2.60	29.60
5	0.88	2.60	29.60
6	0.88	2.60	29.60
7	0.12	0.21	3.33
8	0.12	2.60	3.33

	Num1	Num2	Num3
1	0.18	0.30	3.33
2	0.15	0.51	15.00
3	0.15	0.51	15.00
4	1.30	3.90	39.33
5	1.30	3.90	39.33
6	1.30	3.90	39.33
7	0.18	0.30	3.33
8	0.18	0.30	3.33

라. 잡음 추가(noise addition)

잡음 추가 기법은 연속형 변수의 자료에 주로 적용하며, 매스킹 처리로 인해 자료의 값이 변하게 되는 자료 변조적(perturbative) 기법 중 하나이다. 잡음 추가 기법은 통계적인 잡음을 원래 자료에 더하여 특정 변수의 정보가 공유된 다른 마이크로데이터와의 정확 매칭(exact matching)을 피하기 위해 사용된다.

세부 알고리즘으로는 먼저 각 변수별로 원래 자료의 평균과 분산을 가지는 정규 분포를 찾아 그 정규 분포에서 잡음을 발생시켜 더하는 것이 있으며, 이를 비상관(uncorrelated) 잡음 추가라고 한다. 다음으로 원래 자료에서 매스킹 처리를 하고자 하는 모든 변수들의 공분산 행렬에 비례하는 공분산을 가지는 분포에서 잡음을 발생시켜 원래 자료에 더할 수 있는데 이를 상관(correlated) 잡음 기법이라 한다(Brand, 2004). 비상관 잡음 추가보다는 상관 잡음 추가가 원래 자료의 구조를 더욱 잘 보존하는 것이 당연하나, 상관 잡음 추가 기법을 사용하여도 원래 자료에서 정규성 가정이 성립하지 않는다면 매스킹 후에 자료의 구조가 심각하게 왜곡될 수도 있다. 이에 따라 상관 잡음 기법의 로버스트한 알고리즘이 제시되었고(Templ과 Meindl, 2008), 이 알고리즘 역시 sdcMicro 패키지에 구현

되어 있다. 그 외에 자료의 크기를 고려하거나 자료 중에서 이상값들을 골라서 잡음을 더하는 알고리즘들도 존재하지만, 실제 자료 분석에서 이용할 경우 잘 작동하지 않는 경우가 많았다. 잡음 추가를 통한 매스킹에 대하여는 많은 연구가 이루어져 왔으며(Duncan과 Mukherjee, 2000; Yancy, 2002; Kim과 Winkler, 2003), 세부 알고리즘들 중에서 현재 sdcMicro에서 오류가 발생하지만 향후 활용 가치가 높은 알고리즘에 대해서는 추가 검토 및 구현 작업이 이루어지는 것이 바람직하다 하겠다.

R패키지 sdcMicro를 사용하여 잡음 추가를 시행할 경우, 코드, 예제 및 설명은 다음 <표 8-6>과 같다. 두 변수, 'Num1' 및 'Num2'를 동시에 고려하여 잡음 추가를 시행할 경우 잡음 추가 전후의 값이 달라진 것을 확인할 수 있다.

〈표 8-6〉 잡음 추가 명령어 형식 및 사용 예제

addNoise(obj, variables= , noise= , method=)

obj: 입력 객체 matrix, data frame 또는 sdc object
variables: 잡음을 추가할 변수를 나타내는 열 번호 벡터, sdc 객체는 numVars에 실행
noise: 추가하는 잡음의 정도(%)
method: 잡음을 추가하는 세부 알고리즘을 나타내는 명령어로
additive, correlated, correlated2, restr, ROMM, outdect가 있음

aa.sdc@origData 및 aa.sdc@manipNumVars: 처리 전후 마이크로데이터

```
aa.sdc <- createSdcObj(aa, keyVar=c(2,4,5,6), numVar=c(1,3,7))
aa.sdc <- addNoise(aa.sdc, variables=c('Num1', 'Num2'), noise=5, method='correlated2')
# 지정된 변수명을 가진 변수들에 대해 잡음 추가를 시행
```

aa.sdc@origData

	Num1	Key1	Num2	Key2	Key3	Key4	Num3	w
1	0.30	1	0.40	2	5	1	4	18.0
2	0.12	1	0.22	2	1	1	22	45.5
3	0.18	1	0.80	2	1	1	8	39.0
4	1.90	3	9.00	3	1	5	91	17.0
5	1.00	4	1.30	3	1	4	13	541.0
6	1.00	4	1.40	3	1	1	14	8.0
7	0.10	6	0.01	2	1	5	1	5.0
8	0.15	1	0.50	2	5	1	5	92.0

aa.sdc@manipNumVars

	Num1	Num2	Num3
1	0.4179395	0.5852167	4
2	0.1769261	0.3485025	22
3	0.1702867	0.3802250	8
4	1.8466465	8.9091616	91
5	1.1108758	0.9698924	13
6	1.0464650	2.0410695	14
7	0.1625727	0.2736963	1
8	0.1058896	0.8934963	5

마. 기타 매스킹 기법들

자료순위 교환(rank swapping) 기법은 범주형 변수들에 주로 사용되는 자료교환(swapping)[7] 기법의 확장으로 생각할 수 있으며, 연속형 변수에서 자료를 서로 교환하는 것으로 간주할 수 있다. 즉, 자료순위 교환이란 자료를 정렬하고 제한된 범위 내에서 자료를 서로 교환하는 것으로 자세한 내용은 Moore(1996)을 참고하면 된다. 자료순위 교환을 위한 프로그램이 안정적이라고 하기는 어려워, 예제는 생략하고 명령어만을 <표 8-7>에 정리한다.

〈표 8-7〉 자료순위 교환 명령어 형식

rankSwap(obj, variables=, TopPercent= , BottomPercent=)
obj: 입력 객체 matrix, data frame 또는 sdc object variables: 순위 교환할 변수를 나타내는 열 번호 벡터, sdc 객체는 연속형 keyVars에 실행 TopPercent 및 BottomPercent: 자료순위 교환할 범위(%) 지정

자료섞기(shuffling)[8]는 연속형 자료를 교환하는 방법을 확장시킨 것으로 매스킹된 자료를 생성하기 위해 조건부 분포를 이용한다(Muralidhar과 Sarathy, 2006). 조건부 분포를 이용한다는 것은 표본으로부터 모집단의 분포를 모수적으로 추정하여 새로운 자료 세트를 생성하고자 한다는 것을 말한다. 기본적으로 자료가 정규분포를 따른다고 가정하여 자료섞기 작업을 수행할 수 있으나, 가정을 확장하여 자료가 치우친 t분포를 따를 경우에 대한 자료섞기 알고리즘도 연구되어 있다(Lee 외, 2010). 세부 알고리즘으로는 일반적 가법 자료변조(GADP[9])를 이용하는 경우 및 정규 코풀라(normal copula)를 이용할 경우를 고려할 수 있다. 다만, 기존 통계개발원 연구에서 실제 자료에 적용하여 본 결과 자료섞기 기법이 노출위험 및 정보손실 측도에서 다른 방법에 비해 효율적이라 하기 어렵고, 이론적 난이도가 높아 실무에서 활용하기에 용이하지 않다고 판단된다. 또한 프로그램도 안정적이라고 하기 어려워 관련 설명은 생략하도록 한다.

한편, 자료섞기 기법은 그 특성상 다른 매스킹 기법들과 인위자료 생성 사이에 위치한다고도 볼 수 있으므로 좀 더 엄밀한 검토를 하는 것도 가치가 있겠으나, 이는 향후 과제로 남기도록 한다.

---

7) sdcMicro 패키지의 자료교환(swapping) 명령어를 이용할 경우에는 얻어진 결과에 오류가 있는 것으로 판단된다.

8) 잡음 추가나 자료교환 기법은 자료 변조적(perturbative) 기법으로 주로 언급되나, 자료섞기로 번역한 shuffling은 자료의 분포를 이용하여 자료를 완전히 바꾸므로 자료 변조(perturbation) 기법으로 구분되기도 한다.

9) General Additive Data Perturbation

바. 실제 자료 분석 프로그램 설명

다음 <표 8-8>에서는 실제 마이크로데이터에서 매스킹 기법들을 적용(박민정 외, 2013)하였던 프로그램의 일부를 활용하여 실무자들의 이해를 돕고자 한다.

〈표 8-8〉 가계금융 · 복지조사 마이크로데이터 매스킹을 위한 프로그램 일부

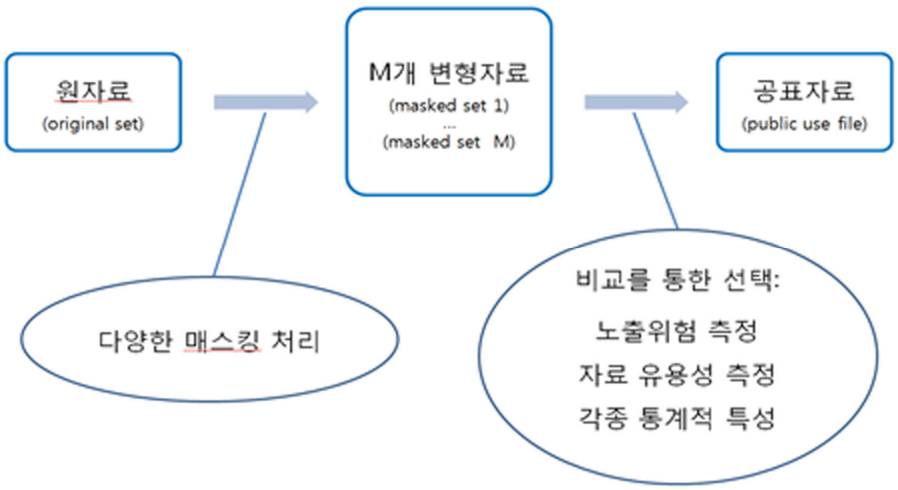
```
library(sdcMicro)
setwd("D:/2013/Data anal")      # 마이크로데이터가 들어있는 컴퓨터의 경로 지정
aa = read.csv(file="data.csv")    # 지정된 폴더 안의 자료 파일명 입력, csv 형식 사용 권장

keyVars.nm = c("region", "sex", "age", "edu1", "jstatus", "together", "marriage", "job",
                "fnum1", "htype", "area", "hposs")    # 파일의 첫 행대로 키변수명 지정
numVars.nm = c("asset1", "asset2", "asset3", "asset4", "debt1", "debt2", "debt3", "debt4",
                "income1", "income2", "income3", "income4")    # 연속형 민감변수명 지정
wVar.nm = c("weight")    # 가중값 변수명 지정

### 전반적 재코딩 예: age(각 세 → 5세), together(동거여부 범주 1,2,3,4 → 1,2,3)
lb = seq(30,80,by=5); lb = paste(rep("~", length(lb)), lb); lb = c(lb, "80+")
aa[,"age"] = globalRecode(aa[,"age"], breaks=c(0,seq(30,80,by=5),120), labels=lb)
aa[,"together"] = globalRecode(aa[,"together"], breaks=c(0,1,2,5), labels=1:3)

# 빈도표와 총 빈도수를 이용한 재코딩 결과 점검
table(aa[,"age"]); sum(table(aa[,"age"]))
table(aa[,"together"]); sum(table(aa[,"together"]))

### 국소통합: method에 저장된 세부 알고리즘별로 시행, 시행 후 변환된 민감변수 값 및
    노출위험과 정보손실 측도값을 각각 rst.num.ma 및 rst.rkutil.ma에 저장
method = c("single", "simple", "onedims", "mdav", "influence", "pca", "clustpca")
rst.num.ma = list(NULL); rst.rkutil.ma = NULL
for(case in 1:length(method)){
  aa.sdc = createSdcObj(aa, keyVars=keyVars.nm, numVars=numVars.nm, weightVar=wVar.nm)
  aa.sdc = microaggregation(aa.sdc, variables=numVars.nm, method=method[case])
  rst.num.ma[[case]] = sdc@manipNumVars    # 매스킹 후 자료값을 list에 저장
  rst.rkutil.ma = rbind(rst.rkutil.ma, c(sdc@risk$numeric, sdc@utility))    # 추후 설명
  rownames(rst.rkutil.ma)[case] = method[case]
}
colnames(rst.rkutil.ma) = c("risk", "il1", "il.eigen")
```

```
### (참고) 각 연속형 민감변수에 대해서 자료값이 0인 레코드수 점검
numVars = match(numVars.nm, colnames(aa))    # 민감변수명과 같은 열 번호 지정
rst = NULL; for(ii in 1:length(numVars)) rst = c(rst, sum(aa[, numVars[ii]]==0))
names(rst) = colnames(aa)[numVars]; rst    # 결과를 나타내는 rst에 변수별 0의 개수 저장

### (참고) 0을 제외한 변수들에 대하여 정규성 점검
index = list(NULL)    # 각 변수별로 0이 아닌 값들에 대한 행 번호를 저장하기 위한 list
for(ii in 1:length(numVars)) index[[ii]] = c(1:nn)[aa[, numVars[ii]]!=0]
ii = 0; ii = ii + 1
qqnorm(aa[index[[ii]], numVars[ii]], xlab="normal", ylab=colnames(aa)[numVars[ii]])
qqline(aa[index[[ii]], numVars[ii]])

qqnorm(log(aa[index[[ii]], numVars[ii]]), xlab="normal", ylab=colnames(aa)[numVars[ii]])
qqline(log(aa[index[[ii]], numVars[ii]]))    # log 변환 혹은 sqrt 변환 후 정규성 점검

### 잡음 추가, 0인 값들에 대해서도 잡음 추가해 income값이 0에서 음수가 될 수도 있음
method = c("additive", "correlated", "correlated2", "restr")  # 세부 알고리즘
rst.num.ns = list(NULL); rst.rkutil.ns = NULL
for(case in 1:length(method)){
  aa.sdc = createSdcObj(aa, keyVars=keyVars.nm, numVars=numVars.nm, weightVar=wVar.nm)
  aa.sdc = addNoise(aa.sdc, noise=5, method=method[case])

  rst.num.ns[[case]] = sdc@manipNumVars    # 매스킹 후 자료값을 list에 저장
  rst.rkutil.ns = rbind(rst.rkutil.ns, c(sdc@risk$numeric, sdc@utility))    # 추후 설명
}
colnames(rst.rkutil.ns) = c("risk", "il1", "il.eigen")

### 자료순위 교환, 0.1% 범위에서 교환
aa.sdc = createSdcObj(aa, keyVars=keyVars.nm, numVars=numVars.nm, weightVar=wVar.nm)
aa.sdc = rankSwap(aa.sdc, TopPercent=.1, BottomPercent=.1)
rst.num.rs = sdc@manipNumVars    # 매스킹 후 자료값을 list에 저장
rst.rkutil.rs = c(sdc@risk$numeric, sdc@utility)    # 추후 설명
```

제3절 노출위험 및 정보손실 측정

마이크로데이터를 매스킹 하는 방안은 여러 가지가 있을 수 있다. 예를 들면, 잡음을 추가할 수도 있고, 국소통합을 할 수도, 또 국소통합 처리된 자료에 잡음을 추가할 수도 있다. 여기에 세부 알고리즘까지 따진다면 수십 가지의 매스킹된 파일을 생성하는 것이 가능하게 된다. 매스킹 방안이 많을 수 있기 때문에 어쩌면 매스킹 기법 적용 자체보다도 매스킹 처리된 여러 마이크로데이터 중에서 어느 파일을 선택하여 공표하는 것이 타당한가를 결정하는 것이 더욱 어려운 문제일지도 모른다. 더 나아가서 마이크로데이터에서 노출위험이 높은 레코드만을 선별하여 매스킹 처리를 할 수 있다면, 더욱 효율적인 매스킹 방안을 모색할 수 있을 뿐만 아니라 매스킹이 아닌 다른 통계적 노출제어 기법에서도 그러한 개체(응답 단위)별 노출위험 측도가 활용될 수도 있다.

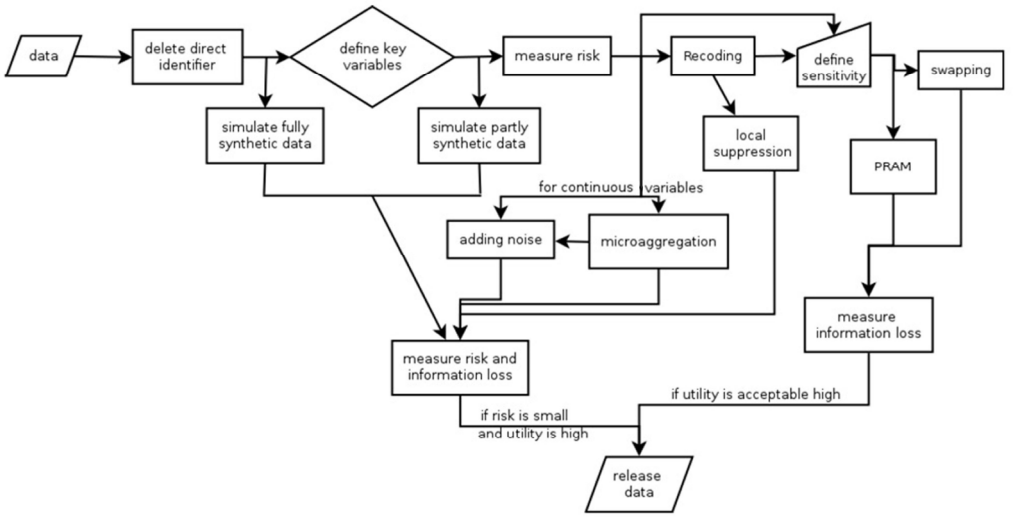
다음 [그림 8-1]은 이해를 돕기 위해 매스킹 절차를 단순하게 표현한 것이다. 앞의 2절의 내용이 다양한 매스킹 처리 부분을 위한 것이라면 이번 3절에서는 노출위험, 정보손실 및 각종 통계적 특성들을 파악하기 위해 sdcMicro 패키지를 어떻게 활용할 수 있을 것인가를 다루고 있다고 할 수 있다.



[그림 8-1] 마이크로데이터 매스킹 절차(박민정, 2014)

이해를 돕기 위해 단순하게 만들어진 [그림 8-1]보다는 복잡해 보일 수도 있으나 좀 더 일반적으로 매스킹 전과정을 표현한 순서도가 있어 다음 [그림 8-2]에서 소개하고자 한다. [그림 8-2]에서는 인위자료 활용을 포함하여 키변수, 민감변수 및 연속형 변수를 구분하여 다양한 기법을 적용하면서 매스킹 처리를 진행하는 과정을 보여주고 있으며, 순서도에서

노출위험(disclosure risk) 및 정보손실(information loss) 측도를 마이크로데이터의 매스킹 방안들을 평가하고 선택하는 기준으로 활용하여 공표할 파일을 선택하는 단계를 제시하고 있다. 노출위험과 정보손실 측도들에 관한 이론적인 내용은 박민정(2014)을 참고하기 바란다. 본 보고서에서는 반드시 알아야 한다고 생각되는 개념들과 현재 구현되어 활용 가능한 노출위험 및 정보손실 측도들을 간략히 설명하고, 실제 프로그램을 어떻게 사용할 수 있는지에 관해 설명하고자 한다. 이론적으로 연구되어 있는 다양한 노출위험 및 정보손실 측도들에 관한 내용 일부는 <부록> 2., 3.에 기술하여 관심 있는 독자는 참고하도록 하였다.



[그림 8-2] 일반적인 마이크로데이터의 매스킹 절차(Meindl 외, 2013)

1. 유일성 기반 노출위험 측정

노출위험에 관한 이론적인 내용을 본 보고서에서 모두 다룰 수는 없지만, 프로그램을 사용하기에 앞서 매스킹 담당자라면 반드시 알아야할 이론적 내용들은 여기서 간략히 설명하고자 한다. 노출위험 측정은 ①특정 레코드의 유일성[10]에 근거하여 이루어질 수도 있으며, ②마이크로데이터에서 개인정보를 얻고자 하는 외부인에 대하여 통계적 의사결정론을 적용하여 측정할 수도 있다. 후자인 침입자 의사결정론에 근거한 노출위험 이론은 유일성 기반 노출위험에 비하여 최근에 발표되었으며 많은 장점을 가지고 있으나, 프로그램으로 구현되어 발표되어 있지는 않고 이론적으로 어려운 내용을 포함하고 있다. 따라서 이에 관한 내용은

10) 서론에서 가상의 사례로 제시하였던 제주도·50대·여성·군인이 지역·연령·성별·직업이라는 응답자 특성에서 전국에서 유일한 경우임을 참고하기 바란다.

본 보고서에서는 다루지 않기로 한다. 본 보고서에서는 먼저 매스킹 담당자라면 반드시 알아야 하는 유일성에 근거한 노출위험 개념을 간략히 설명하고자 한다.

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 다음 3가지 조건을 모두 만족하는 경우에 발생한다고 한다.

- 어떤 사람이 특정 변수(조합)에 대해 모집단에서 유일하다.
- 그 사람은 어떤 공표되는 마이크로데이터에 포함되어 있다.
- 그 사람은 외부인이 가지고 있는 또 다른 마이크로데이터에도 포함되어 있다.

이때 노출위험은 확률적 모형으로 표현될 수 있고, 어떤 사람 A가 모집단에서도 유일하고 표본(공표되는 마이크로데이터)에도 있을 확률은 $\Pr[(A\in U_s)\cap(A\in U_p)]$이며, 이를 위해 사용된 기호들의 정의는 다음과 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 응답 단위들의 모임
U_s : 표본에서 유일한 응답 단위들의 모임

보수적으로 외부인이 A에 대하여 안다고($A\in S_2$) 가정하면 구하고자 하는 노출위험은 $\Pr(A\in S_1)\Pr(A\in U_p)$로 표현할 수 있으며, 이는 A가 표본에 포함되는 확률 $\Pr(A\in S_1)$과 모집단에서 유일한 확률 $\Pr(A\in U_p)$을 이용한 것이다. 표본으로 추출되는 확률 $\Pr(A\in S_1)$는 조사 단계에서 정해져 있으므로, 노출위험을 측정하는 문제는 $\Pr(A\in U_p)$을 어떻게 결정할 것인가 하는 문제와 직결된다고 할 수 있다. 즉 위의 주어진 세 조건 아래에서 노출위험을 측정하는 것은 표본을 기반으로 모집단에서 어떤 레코드가 유일할 확률을 추정하는 문제를 다루는 것으로 전환하여 생각해야 한다.

이제 조사된 표본을 기반으로 모집단에서 어떤 레코드가 유일할 확률을 추정하는 문제를 다루기 위해 먼저 자료의 형태를 바꾸어 이해할 필요가 있다. 여기서는 Bethlehem 외 (1990)에서 활용된 예제를 사용해 자료의 형태와 앞으로 다룰 모집단 유일성 추정 관련 모형에 대한 이해를 도모하고자 한다. 어떤 마이크로데이터가 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, 12, 2, 2개의 범주를 가지고 있다고 하자. 이를 일반적인 마이크로데이터 형식으로 표현하면 다음 <표 8-9>와 같은 행렬 형식으로 표현할 수 있다.

〈표 8-9〉 행렬 형식의 마이크로데이터 예제

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 및 S라는 2개의 키변수에 대해서는 n개의 자료를 2차원 표로 나타낼 수 있고, 이때 어떤 셀의 빈도수인 n_{ij}의 값이 1이면 마이크로데이터는 해당 셀이 가지는 조건에 대해 유일한 레코드를 가지는 것이 된다. M과 S는 각각 두 개의 범주를 가지므로 이 표에서 셀의 개수는 4개가 된다. 아래 <표 8-10>은 M 및 S에 대한 분할표를 보여주고 있다.

〈표 8-10〉 분할표 형식으로 정리된 마이크로데이터, 변수가 2개일 경우

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

한편, 셀의 개수는 해당 키변수 조합에 대한 레코드수를 나타내는 것이기도 하므로, 아래 <표 8-11>과 같이 변수 조합별 레코드수 형식으로[11] 표현할 수도 있다.

〈표 8-11〉 변수 조합별 레코드수 형식으로 정리된 마이크로데이터, 변수가 2개일 경우

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

이제 <표 8-10>을 확장시켜 변수의 개수가 많거나 변수별로 범주수가 많은 경우를 이해할 수 있다. 키변수로 H, A, M, S라는 4개 변수를 모두 사용한다면 다음의 <표 8-12>와

11) 앞에서 사용하였던 제주도·50대·여성·군인인 응답자가 전국 모집단에서 유일하다는 가상 사례의 경우, 변수 (M, S) 대신에, 변수가 (지역, 연령, 성별, 직업)이면서 변수 조합의 값이 (제주도, 50대, 여성, 군인)인 경우를 나타내는 셀의 빈도수가 1인 것으로 모집단 유일성을 이해할 수 있다.

같이 변수 조합별 레코드수를 정리할 수 있게 된다. 이 예제에서 키변수들의 값이 (1,1,1,1)부터 (24,12,2,2)까지인 경우를 나타내는 셀의 개수는 모두 K=1,108개가 되며, 이 중 어느 한 셀의 값이 1이면 해당 키변수 조합을 가지는 레코드는 유일한 레코드가 된다. 즉, K=1,108개의 셀에 해당하는 조건들에서 n=8,399개의 사건(레코드)들이 나누어 발생하는 문제로 이해할 수 있다. 이때 키변수의 개수가 늘어나거나 범주의 개수가 늘어나면 셀의 개수도 급격히 증가하여 분할표는 성긴 분할표(sparse table)가 되며, 어느 셀에 해당하는 레코드가 하나만 존재하는 일이 더욱 쉽게 일어날 수 있다. 따라서 변수의 개수나 범주가 늘어나면 해당 마이크로데이터 파일에서 유일성을 가지는 레코드들이 증가할 확률이 높다는 것을 이러한 형식의 분할표를 통해 명확히 이해할 수 있게 된다.

〈표 8-12〉 변수 조합별 레코드수 형식으로 정리된 마이크로데이터, 변수가 4개일 경우

(H,A,M,S)	(1,1,1,1)	(1,1,1,2)	...	(24,12,2,2)
n	$n_{1,1,1,1}$	$n_{1,1,1,2}$	...	$n_{24,12,2,2}$

이렇게 분할표 형식으로 마이크로데이터를 다루면 노출위험 측정은 크게 두 가지 방향에서 다루어질 수 있다. 첫 번째는 파일 단위의 노출위험 측정으로, 특정 마이크로데이터 파일에 대하여 해당 모집단에서 유일한 레코드($N_k = 1$) 수가 많다고 추정된다면 그 파일은 노출위험이 높다고 할 수 있다. 두 번째는 개체(응답 단위)별 노출위험 측정으로, 이를 통하여 어떤 레코드가 모집단에서 유일한 레코드일 확률을 추정하여 노출위험이 높은 레코드를 적게 만들어내는 매스킹 방안을 노출위험이 낮은 방안으로 판단할 수 있다. 각각에 관하여 하나씩 살펴보도록 하자.

가. 파일 단위의 노출위험 측정

매스킹 처리를 하여 변형된 파일들이 총 M개 있다고 하자. 이제 M개의 파일 중에서 노출위험이 가장 낮은 파일이 무엇인지 간단하게 파악하기 위해서는 파일 단위의 노출위험 측정이 필요하다고 할 수 있다. 앞에서 설명한 대로 표본 조사된 마이크로데이터에서 유일한 레코드를 가지는 변수 조합을 이용하여, 모집단에서 레코드 수가 1인 변수 조합 ($F_k = 1$)[12]의 수, 즉 유일한 레코드들의 수를 추정하는 문제가 파일 단위의 노출위험을 측정하는 문제가 된다.

이를 위하여 도입하는 통계적 모형으로 우선 포아송-감마 모형을 살펴볼 수 있다. 키변수 조합에 의한 셀들로 자료를 바꾸어 다루고 있으므로 통계학에서 범주형 자료를 다룰 때 흔히 사용하는 포아송 분포를 가정하고 모형을 만들어 갈 수 있으며, 각 셀에

12) k번째 키변수 조합에서 나타나는 모집단의 빈도수를 F_k라고 표현하자.

대한 기댓값을 추정할 때 모수에 대하여 감마분포를 가정하는 것이 포아송-감마 모형이라 설명할 수 있겠다. 노출위험 연구 초기의 포아송-감마 모형에 대한 좀 더 자세한 내용은 <부록> 2. 혹은 박민정(2014)을 참고하면 된다. 한편, 포아송-감마 모형은 파일 단위의 노출위험을 통해 매스킹 방안들을 비교하기보다는 주어진 유일성 비율을 만족하는데 얼마나 많은 표본이 필요한지를 파악하기 위해 주로 연구된 것으로 보인다.

이제 모집단에서 유일한 레코드들의 수를 추정하여 파일 단위의 노출위험을 측정하는 것이 포아송-감마 모형을 통해 이루어질 수 있음을 간략히 소개하였다. 이러한 모집단 유일성에 근거한 노출위험 추정을 위한 모형들은 포아송-감마 모형 이후에도 계속해서 발전하여 나갔으며 그러한 모형들을 초모집단(superpopulation) 모형들[13]로 묶어 설명할 수 있다. 여기에는 디리슈렛-다항정규(Dirichlet-multinomial) 모형[14], 음이항 모형[15], 로그-선형 모형[16], 다항정규(multinomial) 모형[17] 및 포아송-역 가우시안(Poisson-inverse Gaussian) 모형[18] 등이 있다고 알려져 있다. 그러나 각 모형들은 전반적으로 유일한 레코드의 개수를 과소 추정하는 문제가 있는 것으로 알려져 있다(포아송-감마 모형의 과소 추정 결과는 <부록> 2.의 표를 참고하라).

공표되는 키변수들에 대해서 모집단에서 유일한 레코드 수를 직접 추정하는 문제를 해결하기 위해 지금까지 소개한 대로 많은 모형들이 제시되어 있으나 대부분 과소 추정의 문제를 가지며, 과소 추정의 문제가 적게 발생하도록 연구된 최근의 고급 모형들은 실무에서 도입하기에는 이론적으로 어렵고 복잡한 측면의 어려움을 가지고 있다고 파악된다. 모집단에서 유일한 레코드 수를 직접 추정하여 파일 단위의 노출위험을 측정하기 위해서, 지금까지 소개한 여러 모형들 중에서 sdcMicro 패키지에는 로그 - 선형 모형[19]이 구현되어

---

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

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

15) 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.

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

17) 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.

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

19) 이 모형은 포아송-감마 모형과 유사하나 모수를 추정할 때, 감마 분포를 가정하는 대신에 로그 선형 모형을 이용하여 모수를 추정한다는 점에서 포아송-감마 모형과 차이가 있다.
Skinner, C. J. and Holmes, D. J. (1998), Estimating the re-identification risk per record in microdata. Journal of Official Statistics, 14, 361-371.

있으며 해당 명령어는 LLmodGlobalRisk이다.

한편 sdcMicro 패키지에는 모집단에서 유일한 레코드 수를 직접 추정하여 노출위험을 측정하는 것 외에도, 개별 레코드의 노출위험을 구하고 나서 그 값들의 합이나 평균을 파일 단위의 노출위험으로 활용하고 있기도 하다. 이제 개별 레코드 단위의 노출위험 추정에 관해 좀 더 살펴보도록 하자.

나. 개체(레코드) 단위의 노출위험 측정

먼저 레코드 단위의 노출위험 추정을 위한 로그 선형 모형을 소개하고자 한다. 여기서 소개하는 로그 선형 모형은 앞에서 소개한 파일 단위의 노출위험 측정을 위한 모형이 아니며, 레코드 단위의 노출위험 측정을 위한 최신 로그 선형 모형에 해당한다(Skinner와 Shlomo, 2008). 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 직접 추정하여 노출위험 측도로 삼았었던 반면, 로그 선형 모형에서는 표본에서 유일한 개체가 모집단에서 유일할 확률이나(r_{1k}), 모집단 개체수 역수의 기댓값을(r_{2k}) 이용해 노출위험 측도를 다음과 같이 정의한다. 두 노출위험 측도 r_{1k} 및 r_{2k}의 추정 과정에 관한 내용은 <부록> 3.을 참고하면 된다.

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

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

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.

이러한 개체 단위의 노출위험 측정을 위한 로그 선형 모형은 과소 추정 문제를 더욱 완화하기 위해 계속 연구되어, 복잡한 로그 선형 모형이나(Siknner와 Shlomo, 2008) 베이지안 GoM(Bayesian version of grade of membership) 모형(Manrique-Vallier와 Reiter, 2012)[20]으로 발전해나가는 모습을 보이고 있다. 두 모형은 모두 미국 센서스 자료를 이용하여 제안한 노출위험 측도를 적용하고 평가하였다.

한편, sdcMicro 패키지에 구현된 개별 레코드의 노출위험은 네델란드 통계청에서 만든 마이크로데이터 매스킹 프로그램인 μ-Argus에 구현된 노출위험을 동일하게 따라 만든

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

것으로 음이항 모형을 이용하고 있다.[21] 개별 레코드의 노출위험을 구하고 나서 그 값들의 합이나 평균을 파일 단위의 노출위험으로 활용하며, 이 개별 노출위험의 합을 노출이 기대되는 레코드 수로 해석하기도 한다.

다. sdcMicro에 구현되어 있는 노출위험 측도의 활용

먼저 모집단에서 유일한 레코드 수를 추정하여 파일 단위의 노출위험을 측정하는 로그 선형 모형에 의한 노출위험 결과를 제공하는 LLmodGlobalRisk 명령어에 대한 코드, 예제 및 설명은 다음 <표 8-13>과 같다. sdc 객체를 입력 변수로 사용하면 @risk 부분에 노출위험 결과를 제공한다. 로그 선형 모형을 적용할 때 사용하는 변수에 따라 파일 단위의 노출위험(global risk)의 값들이 달라지며, 결과로 제공되는 gr1 및 gr2에 대한 설명은 sdcMicro 패키지와 관련된 문헌들에서 찾기가 어려웠다. 따라서 sdcMicro를 사용할 때, 모집단에서 유일한 레코드 수 추정에 근거한 파일 단위 노출위험 측도는 참고로 활용하며 개체별 노출위험 측도의 결과를 활용하여 파일 단위 노출위험으로 사용하기를 권하는 바이다.

〈표 8-13〉 유일한 레코드 수에 근거한 파일 단위의 노출위험 측정

LLmodGlobalRisk(obj, form=~)

obj: 입력 객체 matrix, data frame 또는 sdc object
form: 모형을 나타내는 식

aa.sdc@risk$model: 노출위험 측도 값

```
aa <- francdat
aa.sdc <- createSdcObj(aa, keyVar=c(2,4,5), numVar=c(1,3,7), w='w')
aa.sdc <- LLmodGlobalRisk(aa.sdc, form = ~ Key1 + Key2 + Key3)
#(참고) 입력 객체가 matrix일 경우
rst <- LLmodGlobalRisk(aa[ , c("Key1", "Key2", "Key3")],
                       form = ~ Key1 + Key2 + Key3, inclProb=1/aa[,"w"])
```

aa.sdc@risk$model	rst
> aa.sdc@risk$model $model$gr1 [1] 0.1818401 $model$gr2 [1] 0.2140609	> rst $gr1 [1] 0.1818401 $gr2 [1] 0.2140609

21) 본문에서 설명하고 있는 로그 선형 모형은 sdcMicro에 구현되어 있는 음이항 모형보다 나중에 발표된 것이기는 하지만 개체 단위의 노출위험을 전반적으로 이해하기에 좀 더 적절하다고 생각하여 설명하였다.

다음으로 sdcMicro에서 키변수 조합을 기준으로 모집단 유일성에 근거한 개체별 노출위험 측도 결과를 얻기 위한 과정을 살펴보도록 하자. 이는 sdc 객체를 생성하는 과정에서 자동으로 계산되어 출력된다. 관련 코드, 예제 및 설명은 다음 <표 8-14>와 같다. 결과에 나타나는 @risk$individual에는 세 개의 열이 있으며, 두 번째 열의 fk는 지정된 키변수에 대하여 주어진 마이크로데이터에서 동일한 키변수 조합을 가지는 레코드 수를, 세 번째 열의 Fk는 가중값을 고려하여 추정된 모집단에서 동일한 키변수 조합을 가지는 레코드 수를 나타내고, 마지막으로 첫 번째 열의 risk는 네델란드 통계청의 마이크로데이터 매스킹 프로그램인 μ-Argus에 구현된 것과 동일한 음이항 모형을 이용해 추정된 모집단 유일성에 근거한 레코드별 노출위험을 말한다. 한편, 레코드별 노출위험의 평균으로 파일 단위의 노출위험을 측정하기도 하는데 @risk$global에 속해 있는 $global$risk에 그 결과가 출력된다.

〈표 8-14〉 유일성에 근거한 레코드별 노출위험 측도

createSdcObj(obj, keyVar=, numVar=, w=)

data: 행렬이나 data frame의 형태로 매스킹할 대상이 되는 마이크로데이터
keyVars: data의 열에서 키변수에 해당하는 열 번호 벡터나 변수명
numVars: data의 열에서 연속형 변수에 해당하는 열 번호 벡터나 변수명
w: 표본 가중값 벡터에 해당하는 열 번호 벡터나 변수명

aa.sdc@risk$individual: 레코드별 노출위험 측도 값
aa.sdc@risk$global: 파일 전체를 대상으로 레코드별 노출위험의 평균

```
aa <- francdat
aa.sdc <- createSdcObj(aa, keyVar=c(2,4,5), numVar=c(1,3,7), w='w')
```

aa.sdc@risk$individual	aa.sdc@risk$global
> aa.sdc@risk$individual risk fk Fk [1,] 0.017144262 2 110.0 [2,] 0.022042326 2 84.5 [3,] 0.022042326 2 84.5 [4,] 0.177075834 1 17.0 [5,] 0.003581243 2 549.0 [6,] 0.003581243 2 549.0 [7,] 0.402359478 1 5.0 [8,] 0.017144262 2 110.0	> aa.sdc@risk$global $global$risk [1] 0.08312137 $global$risk_ER [1] 0.664971 $global$risk_pct [1] 8.312137 $global$threshold [1] 0.003581243 $global$max_risk [1] 0.01

그 외에도 sdcMicro는 키변수 조합에 따라 발생하는 유일성에 근거한 노출위험 측도로 k-익명성(k-anonymity), l-다양성(l-diversity) 및 SUDA(Speicial Uniques Detection Algorithm) 측도 등을 제공하고 있다. 이러한 노출위험 측도들은 유일한 레코드를 만들지 않기 위해 국소 감추기 매스킹 기법을 적용하고자 할 때 활용하기에 적당하다고 할 수 있다.

지금까지 sdcMicro에 구현되어 있는 유일성에 근거한 노출위험 측도들을 소개하였다. 이러한 노출위험은 키변수 조합에 의한 유일성을 이용하므로 연속형 변수들에 매스킹 기법을 적용할 경우 매스킹 전후의 노출위험 변화를 탐지할 수 없다는 단점을 가지고 있다. 앞에서 언급하였던 침입자 의사결정론 기반 노출위험 측도들은 연속형 변수들에 잡음 추가 등의 매스킹 처리를 할 경우 노출위험 변화를 나타낼 수 있으나, 구현되어 있는 프로그램이 없고 이론적으로 실무에 적용하기에는 쉽지 않음을 설명하였다. 다행히 sdcMicro는 연속형 변수들에 대한 매스킹 처리 전후의 노출위험 변화를 탐지하기 위해 거리 기반 노출위험을 구현하여 제시한다. 이는 매스킹 이후의 레코드에 어떤 구간을 만들고 매스킹 이전의 레코드가 그 구간 내에 있는지 없는지를 통하여 노출위험을 측정하는 방식이다. 엄밀하다고 하기는 어렵지만 여러 매스킹 방안들을 비교하는데 사용할 수 있다. 이러한 거리 기반 노출위험 역시 sdc 객체를 생성하는 과정에서 자동으로 계산되어 @risk에 $numeric으로 출력된다. 관련 코드, 예제 및 설명은 다음 <표 8-15>와 같다.

〈표 8-15〉 거리 기반 노출위험 측도

createSdcObj(obj, keyVar=, numVar=, w=)

data: 행렬이나 data frame의 형태로 매스킹할 대상이 되는 마이크로데이터
keyVars: data의 열에서 키변수에 해당하는 열 번호 벡터나 변수명
numVars: data의 열에서 연속형 변수에 해당하는 열 번호 벡터나 변수명
w: 표본 가중값 벡터에 해당하는 열 번호 벡터나 변수명

aa.sdc@risk$numeric: 거리 기반 노출위험 측도 값

```
aa <- francdat
aa.sdc = createSdcObj(aa, keyVar=c(2,4,5,6), numVar=c(1,3,7))
aa.sdc = addNoise(aa.sdc, variables = c('Num1', 'Num2'), noise=5,
                  method="correlated2")
```

```
> aa.sdc@risk
$numeric
[1] 0.75
```

2. 거리 기반 정보손실 측정

노출위험만을 고려하여 매스킹 처리를 한다면 원래의 자료와 거리가 멀도록 자료를 많이 변환하면 된다. 그러나 이는 많은 정보손실을 초래하여 마이크로데이터를 쓸모없게 만들 수 있다. 반대로 정보손실을 적게 하려면 자료를 최대한 보전해야 하므로 노출위험이 높아지게 된다. 즉 노출위험과 정보손실 사이에는 어느 하나를 얻기 위해서는 다른 하나를 포기해야하는 일종의 상충 관계(trade-off)가 존재한다고 할 수 있다. 이처럼 매스킹 과정에는 상충 관계에 있는 노출위험과 정보손실을 동시에 최소화해야 하는 어려움이 있으며, 두 측도를 함께 최소화하는 효율적인 방안을 찾는 것은 쉬운 일은 아니다.

Duncan 외(2001)에서는 노출위험과 정보손실을 함께 나타내는 위험-유용성[22] 지도를 제안하여 매스킹 방안 선택의 기준을 제시한 바 있다. 보통 세로축에 노출위험을, 가로축에 정보손실 측도를 나타내며, 각 매스킹 방안별로 좌표가 원점에 가까울수록 노출위험과 정보손실 모두가 최소화된 것으로 간주하여 좀 더 효율적인 매스킹 방안으로 간주할 수 있다. 그러나 만약 원점에 대해 볼록한 어떤 무차별 곡선이 존재하고, 두 매스킹 방안의 노출위험과 정보손실 측도가 그 곡선 위에 있다면 두 방안 사이에서 우위를 논하기 어렵다는 문제가 있기도 하다. 따라서 이 절에서 설명할 거리 기반 정보손실 측도를 이용해 노출위험-정보손실 지도에서 각 방안들을 표현하여 상대적으로 효율적인 안들을 선택하고, 다음 절에서 알아볼 특정 통계량 기반 정보손실 측도를 참고하여 종합적으로 최종 매스킹 방안을 선택하는 것을 추천하고자 한다. 이제 거리 기반 정보손실 측도들에는 어떠한 것들이 있는지 살펴보고 이를 측정하기 위해서 sdcMicro를 어떻게 이용할 수 있는지 알아보도록 하겠다.

노출위험에 비하여 정보손실 측도는 상대적으로 단순하다고 할 수 있다. 먼저 Duncan 외(2001)에서 제안된 노출위험은 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.

한편 평균을 넘어서 좀 더 다양한 기초 통계량들의 변화를 매스킹 처리된 자료가 원래 자료와 얼마나 다른지를 나타내는 정보손실 측도로 사용할 수 있다. 변수 p개를 가지는 연속형 자료에 대해서 기초 통계량들로 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

22) 자료 유용성(data utility)과 정보손실(information loss)은 서로 역의 관계에 있으나 동일한 의미를 가지는 측도를 나타내는 명칭으로 간주하고 참고문헌에 따라 의미 구분 없이 사용하기로 한다.

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

한편, 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 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 외, 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 외, 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.

이처럼 문헌들에 연구되어 있는 여러 정보손실 측도들이 있으나 sdcMicro에서 사용하는 것은 자료값이나 고유값을 이용한 거리 기반 정보손실 측도들이다.

가. sdcMicro에 구현되어 있는 정보손실 측도의 활용

여러 정보손실 측도들 중에서 sdcMicro에는 연속형 변수에 대한 정보손실 측도들이 구현되어 있다. 거리 기반 정보손실 역시 sdc 객체를 생성하는 과정에서 자동으로 계산되어 @utility에 $il1 및 $eigen으로 출력된다. 각 측도의 값들은 매스킹 전후 통계량들의 거리를 표준 편차로 나누어 표준화된 거리를 나타내며, il1은 자료값들을, eigen은 고유값들을 활용하여 계산된다. 한편, 로버스트한 고유값(robust eigenvalues)들을 이용하여 표준화된 거리를 계산하여 제공하기도 하는데 이를 위해서는 행렬 형태의 자료 입력과 dUtility라는 함수의 사용이 필요하다. 관련 코드, 예제 및 설명은 다음 <표 8-16>과 같다.

〈표 8-16〉 거리 기반 정보손실 측도

createSdcObj(obj, keyVar=, numVar=, w=)

data: 행렬이나 data frame의 형태로 매스킹할 대상이 되는 마이크로데이터
keyVars: data의 열에서 키변수에 해당하는 열 번호 벡터나 변수명
numVars: data의 열에서 연속형 변수에 해당하는 열 번호 벡터나 변수명
w: 표본 가중값 벡터에 해당하는 열 번호 벡터나 변수명

aa.sdc@utility: 거리 기반 정보손실 측도 값

```
aa <- francdat
aa.sdc = createSdcObj(aa, keyVar=c(2,4,5,6), numVar=c(1,3,7))
aa.sdc = addNoise(aa.sdc, variables = c('Num1', 'Num2'), noise=5,
                  method="correlated2")
rst.il1 = dUtility(aa.sdc@origData[, aa.sdc@numVars, xm=aa.sdc@manipNumVars)
rst.eigen = dUtility(aa.sdc@origData[, aa.sdc@numVars,
                   xm=aa.sdc@manipNumVars, method="eigen")
rst.robeigen = dUtility(aa.sdc@origData[, aa.sdc@numVars,
                     xm=aa.sdc@manipNumVars, method="robeigen")
```

sdc 객체 활용	dUtility 함수 활용
> aa.sdc@utility $il1 [1] 0.0740877 $eigen [1] 0.2799216	> rst.il1 [1] 0.0740877 > rst.eigen [1] 0.2799216 > rst.robeigen [1] 78.34436

3. 특정 통계량 기반 정보손실 측정

매스킹 전후 사용할 수 있는 정보손실을 측정하기 위해서 각종 통계분석을 시행한 결과가 얼마나 충실하게 유지되었는지를 살펴보는 것을 고려할 수 있다. 자료 이용자가 임의의 통계분석을 시행하였을 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 통계분석 결과를 만들어내야 한다는 관점에서 관심의 대상이 되는 주요 통계량들을 매스킹 전후의 각 자료에서 얻고, 이를 비교하는 것은 의미가 있다고 할 수 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 IO를 p개 회귀계수에 대해 다음과 같이 정의할 수 있다(Karr 외, 2006).

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

여기서 I_k의 첫째 항은 매스킹 후 자료 Z로부터 얻어진 β_k^Z의 사후분포에 매스킹 전 자료 Y에서 얻은 β_k^Y에 대한 신뢰구간 $(L_k^Y,\ U_k^Y)$를 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 들면 매스킹 전 자료에서 얻은 신뢰구간이 $(L_k^Y,\ U_k^Y) = (8, 10)$일 때, 각각 서로 다르게 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_k 역시 동일한 판단을 하도록 한다. 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.

측도 I를 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 외, 2006).

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

측도 J는 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.

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

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

한편, 여러 회귀계수를 동시에 고려하여 신뢰구간의 중복을 측정하는 것도 생각해 볼 수 있다. 이를 타원형 신뢰구간 중복(Ellipsoid Overlap, EO)이라고 하며 I_k와 비슷한 개념으로 정의되나 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, Y 및 $\hat{\beta}_Y$, $\hat{\sigma}_Y^2$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 EO를 계산하기 위해 단계별로 몬테카를로 모의실험이 필요하다. 기타 자료 유용성 측도는 <부록> 4.에 소개하였으니 참고하기 바란다.

가. 실제 자료 분석 프로그램 설명

다음 <표 8-17>에서는 실제 마이크로데이터에서 매스킹 전후 노출위험 및 정보손실 측도를 살펴보았던(박민정, 2014) 프로그램의 일부를 활용하여 실무자들의 이해를 돕고자 한다. 앞의 2절에서 소개하였던 매스킹 프로그램 중 국소통합 부분에 대하여 노출위험 및 정보손실 측도들을 계산하고 이를 그림으로 나타내는 과정을 보여주고 있다.

〈표 8-17〉 가계금융 · 복지조사 마이크로데이터 매스킹 방안의 노출위험 –정보손실 지도 작성을 위한 프로그램 일부

```
### 국소통합: method에 저장된 세부 알고리즘별로 시행, 시행 후 변환된 노출위험과 정
    보손실 측도값을 각각 rst.num.ma 및 rst.rkutil.ma에 저장
method = c("single", "simple", "onedims", "mdav", "influence", "pca", "clustpca")
rst.rkutil.ma = NULL
for(case in 1:length(method)){
  aa.sdc = createSdcObj(aa, keyVars=keyVars.nm, numVars=numVars.nm, weightVar=wVar.nm)
  aa.sdc = microaggregation(aa.sdc, variables=numVars.nm, method=method[case])
  rst.rkutil.ma = rbind(rst.rkutil.ma, c(sdc@risk$numeric, sdc@utility))
  rownames(rst.rkutil.ma)[case] = method[case]
}
colnames(rst.rkutil.ma) = c("risk", "il1", "il.eigen")

### risk information loss map
rst = rst.rkutil.ma[,c(2,1)] # 정보손실 측도 il1 이용
rst = rst.rkutil.ma[,c(3,1)] # 정보손실 측도 eigen 이용

plot(rst, pch=20, xlim=range(rst[,1])*1.1)
text(rst, labels=method, pos=4)
```

다음으로 특정 통계량 기반 정보손실 측정을 위해서는 앞에서 소개한 구간 중복 등을 대신하여 매스킹 전후 각 통계량들 사이의 절대적 차이에 대한 상대적인 비율(ARB)을 이용하였다. 표본 추출 과정의 층 등, h개의 영역에 대해 통계량을 θ_i로 표현하였다.

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

사용한 특정 통계량[23]은 평균, 상관계수, 지니 계수(Gini coefficient) 및 성별에 따른 소득 차이(Gender Pay Gap, GPG) 등이며, 각 통계량에 대해 전체적, 층별, 시도별 및 아래 모형에 대하여 ARB를 비교하였다(박민정, 2014).

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

사용한 프로그램의 일부는 다음 <표 8-18>과 같다. 프로그램 사용 결과는 통계개발원 보고서들을 참고하기 바란다.

〈표 8-18〉 가계금융 · 복지조사 마이크로데이터 매스킹 방안의 특정 통계량 기반 –정보손실 측정 프로그램 일부

```
### sdc 객체 생성 및 매스킹 처리
aa.sdc = createSdcObj(aa, keyVars=keyVars.nm, numVars=numVars.nm, weightVar=wVar.nm)
aa.sdc = microaggregation(aa.sdc, variables=numVars.nm, method="mdav")

### 모형에 사용할 키변수 지정
keyVars.Mtx = aa.sdc@origData[ ,keyVars]
age = as.numeric(keyVars.Mtx[,"age"])
gender = as.factor(keyVars.Mtx[,"sex"])
edu = as.factor(keyVars.Mtx[,"edu1"])

### 매스킹 이전 자료에 대한 모형 적합값
snsVars.Mtx.o = aa.sdc@origData[ ,numVars]
income.o = rowSums(snsVars.Mtx.o[, 9:12])
income.o[income.o==0] = rep(1, sum(income.o==0)
mdl.o = lm(log(income.o)~age+gender+edu, weights=weight.Vec)
hat.income.o = exp(mdl.o$fitted.values)
```

23) 정보손실 측정을 위해 사용하는 통계량을 문헌들에서는 벤치마킹 통계량(benchmarking indicators)으로 언급되고 있다.

```
### 매스킹 이후 자료에 대한 모형 적합값
snsVars.Mtx.m = aa.sdc@manipNumVars
income.m = rowSums(snsVars.Mtx.m[, 9:12])
income.m[income.m==0] = rep(1, sum(income.m==0)
mdl.m = lm(log(income.m)~age+gender+edu, weights=weight.Vec)
hat.income.m = exp(mdl.m$fitted.values)

indx.M = c(1:nn)[gender==1]
indx.F = c(1:nn)[gender==2]

### 전체 자료에 대하여 GPG 및 ARB 계산
GPG.data.o = (sum(income.o[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M])
              - sum(income.o[indx.F]*weight.Vec[indx.F])/sum(weight.Vec[indx.F]) )/
              (sum(income.o[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M]))
GPG.data.m = (sum(income.m[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M])
              - sum(income.m[indx.F]*weight.Vec[indx.F])/sum(weight.Vec[indx.F]) )/
              (sum(income.m[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M]))
ARB.data = abs((GPG.data.o - GPG.data.m)/GPG.data.o)

### 모형 적합값에 대하여 GPG 및 ARB 계산
GPG.model.o=(sum(hat.income.o[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M])
             -sum(hat.income.o[indx.F]*weight.Vec[indx.F])/sum(weight.Vec[indx.F]))/
            (sum(hat.income.o[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M]))
GPG.model.m=(sum(hat.income.m[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M])
             -sum(hat.income.m[indx.F]*weight.Vec[indx.F])/sum(weight.Vec[indx.F]))/
            (sum(hat.income.m[indx.M]*weight.Vec[indx.M])/sum(weight.Vec[indx.M]))
ARB.model = abs((GPG.model.o - GPG.model.m)/GPG.model.o)
```

제4절 요약 및 제언

통계개발원에서는 마이크로데이터를 배포하는 과정에서 필요한 통계적 노출제어 방안들 중, 전통적으로 꾸준히 사용되어 온 매스킹 기법의 전반적인 활용을 위해 각 기법들의 세부 알고리즘 파악부터 노출위험과 정보손실 측도들을 활용하여 특정 매스킹 방안을 평가하는 문제에 이르기까지 매스킹 전체 과정에 대하여 지난 2년간 연구하였다. 이는 단편적으로 각 매스킹 기법들을 파악하였던 기존 연구에서 진일보하여 매스킹 기법들을 활용하는 통계 선진국들의 최근 현황을 파악하고 매스킹 전체 과정을 우리나라 마이크로데이터에 적용하였다는 의의를 가진다고 할 수 있다. 본 보고서는 마이크로데이터 배포와 관련된 실무자들이 최신 매스킹 기법들을 어떻게 사용할 수 있을지 알 수 있도록 현재 안정적으로 이용이 가능한 범위 안에서 마이크로데이터 매스킹 전체 과정을 다룬 프로그램 활용 사례를 더하여 일종의 안내서로 작성되었다.

마이크로데이터의 통계적 노출제어 문제 해결은 전통적으로 매스킹 기법들에 의존해 왔으며 또한 매스킹 기법들이 실무적으로 가장 널리 활용되고도 있지만, 서론과 제2절에서 언급하였던 것처럼 매스킹 기법은 노출위험과 정보손실을 동시에 축소하기에는 한계를 가진다. 이는 자료 공급자 및 자료 이용자를 모두 만족시키기 어려운 한계를 가진다는 의미이기도 하다. 때문에 미국, 독일 등에서는 인위자료(synthetic data)를 활용하여 공공이용파일을 작성하는 것을 최근 꾸준히 연구해왔으나, 인위자료의 활용 역시 많은 한계를 가지고 있다. 이제 여러 통계 기관들 및 학계의 관심은 인위자료 활용보다는 차등 정보보호 이론을 적용한 분석서버를 통한 자료 배포로 옮겨가고 있다고도 보인다.

이렇게 많은 변수들을 가지는 마이크로데이터를 배포할 때 발생하는 비밀보호의 문제에 대하여 세계적으로 해결 방안을 찾기 위해 애쓰고 있으나, 실무에서 사용할 정도로 주목되는 성과를 찾기는 어려우며 궁극적인 해결 방안을 찾기 위해 다양한 노력이 이루어지고 있다고 판단된다. 우리 역시 유용한 매스킹 기법들의 안정적 사용, 개별 노출위험 측정의 효율적인 활용, 그리고 통계 선진국들과 어깨를 나란히 하여 마이크로데이터 노출제어의 궁극적인 해결 방안 모색을 위한 노력을 지속해 나가기를 기대한다.

<참고문헌>

김경미 · 이의규 · 정미옥(2007), 마이크로데이터 제공에 관한 해외사례연구, 통계개발원.

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

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

박민정(2014), 가계금융·복지조사 마이크로데이터 매스킹 방안 평가, 통계개발원.

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

이용희 (2013), 정보공개 환경에서 개인정보 보호와 노출 위험의 측정에 대한 통계적 방법, 한국데이터정보과학회지, 24(5), 1029-1041.

오미애 · 최현수 · 김용대 · 이용희 · 진재현(2014), 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안, 한국보건사회연구원.

충남대학교 통계학과 대학원. 통계자료 비밀보호론 [매스킹(masking)이론].

Abowd, J. M. and Woodcock, S. D. (2001), Disclosure limitation in longitudial 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, 215-277, Amsterdam: North-Holland.

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

Brand, R. (2004), Microdata protection through noise addition, In Priavcy in Statistical Database, Lecture Notes in Computer Science, Springer, 347-359.

Domingo-Ferrer, J. and Torra, V. (2001), Chapter 5. Disclosure control methods and information loss for microdata. Confidentiality, Disclosure and Data Access, North Holland.

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.

Duncan, G. T. and Mukherjee, S. (2000), Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise, Journal of the American Statistical Association, 95(451), 720-729.

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

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

Dwork, C. (2006), Differential privacy. In 33rd International colloquium on Automata, Languages and Programming, part II (ICALP 2006), springer, Venice, Italy, 1-12.

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

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.

Kim, J. J. and Winkler, W. E. (2003), Multiplicative noise for masking continuous data, Research Report Series Statistics #2003-01, U.S. Bureau of the Census.

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.

Lee, S., Genton, M. G. and Arellano-Valle, R. B. (2010), Perturbation of numerical confidential data via skew-t distributions, Management Science, 56(2), 318-333.

Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1998), A comparative study of microaggregation methods, Qiiestiio, 22: 511-526.

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

Moore, Jr. R. (1996), Controlled data-swapping techniques for masking public use microdata, U.S. Bureau of Census Statistical Research Division Report Series, RR 96-04.

Muralidhar, K. and Sarathy, R. (2006), Data shuffling - a new masking approach for numerical data. Management Science, 52(5), 658-670.

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

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

Skinner, C. and Shlomo, N. (2008), Assessing identification risk in survey micordata using log-linear models, Journal of the American Statistical Association, 103(483), 989-1001.

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

Templ, M. and Meindl, B. (2008), Robustification of microdata masking methods and the comparison with existing method. Privacy in Statistical Database. Lecture Notes in Computer Science, Springer, 5262: 177-189.

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.

Yancy, W. E. (2002), Working papers for mixture model additive noise for microdata masking, Research Report Series Statistics #2002-03, U.S. Bureau of the Census.

<부 록>

1. R패키지 sdcMicro 사용법

마이크로데이터 매스킹을 위한 종합적인 프로그램으로는 현재까지 μ-Argus라는 네델란드 통계청에서 만든 프로그램과 오스트리아 통계청의 후원으로 만들어진 R패키지 sdcMicro가 있다. 이 중 μ-Argus는 4.1.0 버전이 2002년 발표되었으며 관련 매뉴얼(4.2버전)도 배포되어 있다(Statistics Netherlands, 2007). 네델란드 통계청의 μ-Argus는 방대한 프로젝트로 오랫동안 만들어져 안정성이 있으며 여러 가지 매스킹 기법들을 구현하였다. 또한 명령어 형식뿐만 아니라 사용자 친화적인 GUI로도 프로그램이 구현되어 있다. 다만, μ-Argus가 독립된 프로그램이므로 이를 이용하려면 해당 프로그램을 구하여 설치하는 과정과 비밀보호 담당자가 새로운 프로그램에 친숙해지는 과정이 필요하며, μ-Argus가 비교적 오래 전에 만들어져 최신 이론의 반영이 부족한 측면이 있다는 단점이 있다.

이에 반해 sdcMicro는 통계 관련자들에게 널리 사용되고 있는 통계 분석 프로그램인 R[1]에서 사용할 수 있도록 만들어진 패키지로 R에 익숙한 통계 담당자들이라면 친숙하게 사용 가능하며 또한 확장성이 좋은 R의 장점을 이어 받는다. 또한 μ-Argus를 기반으로 하여 비교적 최근에 만들어졌기 때문에 최신 매스킹 알고리즘들 및 각종 측도들을 좀 더 많이 구현하였으며, μ-Argus와 마찬가지로 GUI도 구현되어 있다.[2] R패키지 sdcMicro의 단점은 소수 인원으로 비교적 짧은 기간 동안에 만들어져 프로그램이 불안정한 측면이 있어 주의가 요구된다는 것이다. 본 보고서에서는 안정적인 명령어들과 그렇지 않은 명령어들을 구분하여 설명하였으므로 참고하기 바란다. 좀 더 자세한 내용을 원하는 독자는 sdcMicro를 다운로드하여 생성되는 하위 폴더들 중 doc 폴더에서 찾을 수 있는 Guidelines(Meindl 외, 2013) 문헌을 참고하기 바란다.

1) R에 관한 내용은 관련 문헌들을 참고하기 바라며, 검색을 통해 다운로드 받을 수 있는 R-studio를 함께 사용한다면 개인적인 기호에 따라 다를 수는 있겠으나 일반적인 R의 이용에 좀 더 편리함을 밝혀 두고자 한다.

2) 다만, 실제로 사용하여 본 결과 매스킹 기법의 습득을 위한 단편적인 이용이 아닌 마이크로데이터 매스킹 전 과정을 위해 프로그램을 이용할 때는 직접적인 명령어 활용이 효율적이라 판단하여 본 보고서에서는 명령어 입력을 통한 프로그램 사용법을 설명하도록 한다.

가. sdc object 설명

우선 sdcMicro에서는 다른 패키지들과 다르게 새로운 객체인 sdc 객체(object)를 정의하여 사용하고 있다. R에서 주로 사용되는 자료 형식들은 행렬(matrix), 데이터 프레임(data frame), 리스트(list) 등이 있는데, sdc object는 $를 사용하는 list와 유사하게 @를 사용하여 여러 성분들을 저장할 수 있다. 다음은 sdc object를 만드는 명령어를 설명한 것이다.

---

createSdcObj(data, keyVars=, numVars=, weightVar=, ...)

---

data: 행렬이나 data frame의 형태로 매스킹할 대상이 되는 마이크로데이터
keyVars: data의 열에서 키변수에 해당하는 열 번호 벡터나 변수명
numVars: data의 열에서 연속형 변수에 해당하는 열 번호 벡터나 변수명
weightVar: 표본 가중값 벡터에 해당하는 열 번호 벡터나 변수명
... 그 외에 pramVars, hhId, strataVar, sensibleVars, options 등이 있음

예를 들어, 행렬 aa를 n개의 레코드와 성별(gender), 연령(age), 직업(job), 거주지(region), 소득(income), 자산(asset), 부채(debt) 및 가중값(w)으로 구성된 8개의 변수를 가지는 마이크로데이터라고 하자. 그러면 응답자 특성을 나타내는 키변수는 성별, 연령, 직업, 거주지가 되고, 연속형 민감변수는 소득, 자산, 부채라고 할 수 있다. 이제 sdc object는 다음과 같이 정의할 수 있다.

```
colnames(aa) <- c('gender', 'age', 'job', 'region', 'income', 'asset', 'debt', 'w')
aa.sdc <- createSdcObject(aa, keyVars=c(1,2,3,4), numVars=c(5,6,7), wegithVar=c(8))
aa.sdc <- createSdcObject(aa, keyVars=c('gender', 'age', 'job', 'region'),
                          numVars=c('income', 'asset', 'debt'), wegithVar=c('w'))
```

참고로, 보통은 굉장히 많은 변수를 가지는 마이크로데이터 파일에서 특정 변수들을 선택하여 매스킹 처리를 해보는 것이 실무자들에게 흔히 일어나는 일이므로 열 번호를 입력하는 것보다는 다음과 같이 처리하면 편리하다. 기존 데이터 파일에서 변수명을 바꾸지만 않으면 데이터 추가 및 축소에 의해 열의 위치가 바뀌는 것과 같은 변화에 손쉽게 대처할 수 있기 때문이다. 다음 예제의 data.csv에서는 위에 소개한 8개 변수를 포함하여 100개의 변수가 들어 있다고 가정하자.

```
aa <- read.csv(file="data.csv")
# 자료를 입력받음, 마이크로데이터가 csv파일 형태로 저장되어 있으며
# 첫 행은 변수명이라고 가정함
keyVars.name <- c('gender', 'age', 'job', 'region')
# 마이크로데이터의 변수명 중 네 개의 키변수 이름을 지정
numVars.name <- c('income', 'asset', 'debt')
# 마이크로데이터의 변수명 중 세 개의 연속형 변수 이름을 지정

aa.sdc <- createSdcObj(aa, keyVars=keyVars.name, numVars=numVars.name,
                       weightVar='w')
# 지정된 키변수, 연속형 변수 및 가중값 변수를 가지는 sdc object를 생성함
```

이제 aa.sdc는 sdc object이며 list object와 유사하게 여러 성분들을 가지게 된다. 다음과 같이 sdc object에 어떤 매스킹 처리를 수행하면 R의 명령어 입력창에서는 아무 일도 일어나지 않지만, sdc object 내부의 성분들이 변하여 매스킹 전후의 값을 저장하게 된다.

```
aa.sdc <- masking(aa.sdc, options)
```

masking: globalRecode, addNoise 등과 같은 매스킹 명령어
options: 주어진 매스킹 명령어에 수반되는 명령 옵션들

@origData: 입력받은 매스킹 이전의 마이크로데이터
@keyVars: 매스킹 이전의 마이크로데이터에서 키변수로 지정된 열 번호
@numVars: 매스킹 이전의 마이크로데이터에서 연속형 변수로 지정된 열 번호
@weightVar: 매스킹 이전의 가중값 변수로 지정된 열 번호
@originalRisk: 매스킹 이전의 지정된 키변수 조합의 유일성에 근거한 노출위험
... 그 외에 @pramVars, @hhId, @strataVar, @sensibleVar 등이 있음

@manipKeyVars: 매스킹 이후 변화된 키변수 data frame
@manipNumVars: 매스킹 이후 변화된 연속형 변수 data frame
@risk: 매스킹 이후 노출위험
@utility: 매스킹 이후 거리기반 자료유용성 측도
@set: 매스킹 처리에 대한 정보 모음
... 그 외에 @manipPramVars, @manipStrataVar, @pram, @options 등이 있음

이러한 sdc object를 사용하게 되면 행렬들을 이용해 매스킹 처리를 할 경우와 다르게 각 변수들 및 매스킹 전후 생겨나는 변화를 일일이 저장할 필요가 없어 매우 편리하다. 참고로 행렬을 사용할 경우의 예를 살펴보면 다음과 같다. 다음 예제는 연속형 변수들에 mdav 알고리즘을 사용하여 microaggregation 매스킹 기법을 적용하면서, 행렬로 마이크로데이터를 다룰 경우이다. 앞의 예제와 같이 마이크로데이터 aa, 키변수, 연속형 변수 및 가중값 변수 위치 등이 지정되었다고 하자.

```
numVars.mtx <- aa[ , numVars.name]
# 연속형 변수들의 행렬을 따로 지정한다.
numVars.mtx.masked <- microaggregation(numVars.mtx, method="mdav")
# 매스킹 이후 행렬을 얻는다.
risk.after.masking <- dRisk(numVars.mtx, numVars.mtx.masked, k=0.05)
utility.after.masking <- dUtility(numVars.mtx, numVars.mtx.masked)
# 매스킹 이후 노출위험 및 자료유용성 측도를 얻는다.
```

이처럼 행렬 그대로 매스킹 처리를 하게 되면 매스킹으로 인해 변화된 행렬, 변화된 노출위험 및 자료유용성(정보손실) 측도 값 등, 다루어야 할 요소들이 많아지지만, 다음과 같이 sdc object를 사용하면 한 줄로 간단하게 처리할 수 있으며, 변화된 값들이 자동으로 aa.sdc@manipNumVars, aa.sdc@risk, aa.sdc@utility 등에 저장된다.

```
aa.sdc <- microaggregation(aa.sdc, variables=numVars.name)
```

행렬을 사용할 경우와 sdc object를 사용할 경우 아래와 같이 짝지어진 변수들에서 동일한 결과를 얻을 수 있다. 즉, 처리하는 명령어는 한 줄로 줄어들지만 얻어지는 결과는 동일하여 sdc object를 사용하여 매스킹 전 과정을 수행하는 것이 효율적이라 하겠다.

행렬을 사용할 경우	sdc object를 사용할 경우
numVars.mtx	aa.sdc@@origData[,aa.sdc@numVars]
numVars.mtx.masked	aa.sdc@@origData[,aa.sdc@manipNumVars]
risk.after.masking	aa.sdc@risk
utility.after.masking	aa.sdc@utility

2. 파일 단위 노출위험 측정을 위한 포아송-감마 모형

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[3]을 가정한다. 그러면 모집단에서 유일한 레코드들의 수 U_p는

$$U_p = \sum_{i=1}^{K} \mu_i e^{-\mu_i}$$

이며, 이를 추정하기 위해서 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,

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

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 최대우도추정량(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 $NB(N_j, \alpha_j, \beta_j)$을 가정할 수 있다.[4] 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.

---

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

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

이러한 U_p의 추정을 노출위험을 일정 수준 이하로 제어하는데 이용할 수 있다. 첫째 표본의 유일한 레코드의 수를 일정한 값 이하[5]로 유지하거나, 둘째 표본과 모집단에서 유일한 레코드의 비율[6]을 일정한 값 이하[7]로 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 지금까지 다루어 온 Bethlehem 외(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. 아래 표를 살펴보면 포아송-감마 모형에 의한 추정은 실제 결과를 과소 추정하는 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 모집단의 유일성 비율 0.1% 기준에서 4개 변수를 안전하게 공표하기 위해서는 46,228보다 작은 인구수를 가지는 지역단위를 포함해서는 안 된다는 결론을 얻을 수 있다. 더 자세한 내용은 Bethlehem 외(1990)를 참고하라.

키변수	키변수 조합 수 (셀의 수)	표본의 유일성 개수		모집단 최소 크기
		관측 결과	추정 결과	
H	24	0	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) $\widehat{U_{ps}} < C_a$

6) $U_{ps}/n = U_p/N$

7) $\widehat{U_p}/N < C_r$

3. 개체(레코드)별 노출위험 측정을 위한 로그 선형 모형

본문에서 표본에서 유일한 개체가 모집단에서 유일할 확률이나(r_{1k}), 모집단 개체수 역수의 기댓값을(r_{2k}) 이용해 노출위험 측도를 다음과 같이 정의하며

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

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

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 한다고 하였다.

이러한 로그 선형 모형에서 각 노출위험을 추정하기 위해서는 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 이용하면 위의 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_k를 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 $\hat{\lambda}_k = exp(x_k' \hat{\beta})$을 얻어 추정된 모수를 이용하여 각 노출위험 수치를 구할 수 있다.

로그 선형 모형을 이용해 노출위험 측도를 추정하는 연구는 계속해서 이루어져 왔으나, Skinner와 Shlomo(2008)에서는 노출위험 측도에 대한 추론이 가능하도록 검정 통계량들을 제시

하였다. 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.

4. 기타 자료 유용성 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 여기에서는 이들을 간략히 살펴보자(Woo 외, 2009).

본문먼저, 성향점수(propensity score)를 이용한 유용성 측도는 다음과 같은 세 단계를 거쳐 얻어진다. ① 매스킹 전후 자료를 합친 $T=(Y,Z)$을 만들고, ② 로지스틱 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.

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

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.

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

이러한 측도들은 본문에서 언급한 다른 측도들과 함께 자료유용성을 평가하는데 참고할 수 있다.

제9장

통계작성기관 마이크로데이터 제공범위 설정 가이드라인 작성

박민정 · 성병찬 · 최유성*

요 약

마이크로데이터 제공범위 설정 문제는 다양한 이슈를 포함한다. 본 가이드라인 연구 과제는 그러한 이슈들 중에서 “변수의 제공범위에 따라 달라지는 개인정보 노출위험의 측정”을 다룬다.

실무자가 직접 R패키지를 사용해서 노출위험을 측정하기 위한 가이드라인은 2015년에 이미 발간된 통계개발원 보고서에 포함되어 있다. 이번 연구 과제는 기존의 문서 형식의 가이드라인이 아닌, “프로그램 형식으로 가이드라인을 작성”했다는 점에서 차별성을 가지고 있다. 프로그램 형식의 가이드라인 작성을 위해 Eurostat의 사용자 친화적인 엑셀 인터페이스를 가진 가이드라인 프로그램을 벤치마킹했다.

한편 사용자 친화적인 프로그램 형식의 가이드라인을 작성하기 위해, 이용자에게 친숙한 엑셀 프로그램과 노출위험 분석을 담당하는 R패키지를 연결할 필요가 있다. 이에 엑셀과 R을 연동하기 위한 방법을 연구했으며, “엑셀과 R을 연동해서 프로그램 형식의 가이드라인을 작성하는 방안”을 제시했다.

따라서 본 보고서는 독립적인 두 부분으로 구성된다. 보고서 앞 부분은 “노출위험 측정 이론 및 가이드라인 프로그램 이용법”을, 후반부는 “엑셀과 R을 연동한 프로그램(가이드라인) 작성 방안”을 다룬다.

주요 용어 : 변수의 제공범위, 노출위험, R, 엑셀, 가이드라인 프로그램

✍ 이 연구는 2018년 하반기 통계개발원 연구보고서에 수록된 자료입니다.

* 자료 협조

제1절 서 론

마이크로데이터[1]를 제공할 때 개인정보보호 여부는 반드시 검토되어야 하는 문제이다. 물론 통계작성기관은 개인정보보호를 위해 이름, 주소, 전화번호와 같은 직접식별자를 모두 제거하고 자료를 제공한다. 하지만 자료 수집 목적에 부합하는 자료분석을 위해서는 각 응답 개체의 특성을 나타내는 변수들이 제공되어야 한다. 이러한 변수들의 정보가 모이면 간접 식별자로서 기능할 수 있다. 즉, 자료 분석에 필수적인 응답자 특성을 나타내는 변수들의 조합으로 각 개체가 식별될 수도 있다는 문제가 발생한다. 예를 들어, 지역, 연령, 성, 직업군별 소득 분포를 파악하기 위해 마이크로데이터가 제공되고, 데이터에 포함된 '제주도 거주 57세 여성 군인'이 전국에서 단 한 명이라고 하자. 그러면 배포된 마이크로데이터를 통해 이 개인이 식별되고, 성실하게 응답한 개인의 소득에 관한 정보가 노출될 수 있다. 따라서 안전하게 마이크로데이터를 제공하기 위해서는, 직접 식별자를 제거한 이후에도 자료에 대한 노출위험 검토 과정이 반드시 필요하다.

마이크로데이터를 안전하게 만들기 위한 여러 기법들을 전통적으로 매스킹(masking) 기법이라고 불러왔다. 매스킹 처리란 원래 자료에 적절한 변환을 가해서 간접적인 식별정보를 가리는 과정이라고 이해할 수 있다. 예를 들면, 민감하다고 생각되는 변수나 응답 내용을 가리거나(감추기, suppression), 각 세 연령을 5세 단위로 바꾸거나(전체재범주화, global recoding), 잡음을 추가하거나(잡음 첨가, noise addition), 어떤 변수에 대해 유사한 조건을 가지는 응답 개체들의 값을 서로 바꾸거나(교환, swapping) 혹은 몇몇 응답 개체들을 묶어 동일한 값을 할당하는(군집화, microaggregation) 등의 처리가 가능하다. 참고로, 매스킹 처리된 자료를 제공할 때는 이용자가 올바른 분석을 할 수 있도록 변환된 자료와 함께 변환에 관한 정보를 모두 제공해야 한다. 마이크로데이터 매스킹 처리에 관해 좀 더 자세히 알고 싶은 독자는 박민정 외(2013) 및 박민정(2015)을 참고할 수 있다.

마이크로데이터를 매스킹 처리하면 각 개체에 대한 식별 가능성은 줄어들지만, 자료에 정보손실이 발생한다. 이는 매스킹 처리 이전 자료에 비해 자료 유용성이 감소한다는 것을 의미하고, 더 부정확한 자료를 받게 되는 것이므로 자료 이용자 입장에서는 바람직

1) 마이크로데이터(microdata)란 각 응답 개체의 정보를 행렬 형식으로 쌓은 데이터를 말한다. 보통 각 열은 변수를 나타내고, 각 행은 변수에 대한 응답 개체들의 정보로 구성된다. 통계작성기관은 조사 등의 자료수집 과정을 통해 마이크로데이터를 구축하고, 이 마이크로데이터에서 각종 통계를 생산해 공표한다. 자료 이용자가 기관에 의해 공표된 각종 통계 범위를 벗어나서 심층적인 자료 분석을 원할 때, 마이크로데이터 사용을 요청하게 된다.

하지 않다. 가상의 '제주도 57세 여성 군인' 사례의 경우, 개인정보보호를 위해 군인이라는 정보를 가리고 좀 더 안전하게 만들 수 있다. 하지만 여성 군인의 소득에 관심이 있는 자료 이용자라면 이러한 자료를 수용하기 어렵다. 이처럼 개인정보보호와 자료 유용성 확보는 언제나 충돌하는 관계에 있다.

자료 제공 시 통계적 비밀보호[2]를 위해 위에 언급한 다양한 방법들을 사용할 수 있지만, 대부분의 경우 통계작성기관은 비밀보호 기법을 적용한 여러 파일들 중 하나를 선택해 제공해야 한다. 즉, 비밀보호 처리된 각 파일들에 대해서 노출위험과 정보손실을 측정하고 종합적으로 검토해서 이용자에게 제공할 하나의 결과물을 선정하는 과정이 필요하다. 그러나 이러한 과정은 복잡하고, 특히 매스킹 기법만을 사용하는 경우에는 적절한 결론을 내리는데 더 많은 어려움이 있다(박민정, 2014). 노출위험과 정보손실을 종합적으로 고려한 자료제공 방안을 도출하는 문제는 학계, 통계연구기관, IT 산업 등 다양한 분야에서 연구주제로 다루어지고 있다. 하지만 안타깝게도 아직은 실무적으로 바로 사용할 수 있는 대안이 제시되어 있다고 하기는 어렵다. 그럼에도 불구하고 통계작성기관에서 자료 제공 시 노출위험을 과학적으로 검토하는 것은 당장 해결해야 하는 시급한 과제이다.[3]

현재 노출위험 검토에는 두 개의 걸림돌이 있다. 첫 번째는 노출위험을 측정하는 통계적 모형이 완전하지 않다는 것이다. 두 번째는 통계작성기관의 업무 담당자가 그런 복잡한 이론을 깊이 있게 이해하고 R, SAS 등의 통계 프로그램을 이용해 노출위험을 측정한 후, 자료 제공범위를 결정한다는 것이 현실적으로 어렵다는 것이다. 본 연구과제는 이러한 현실적인 한계 속에서도 통계작성기관의 업무 담당자가 기본적인 수준의 마이크로데이터의 노출위험 검토를 할 수 있도록 가이드라인을 마련하기 위해 진행되었다. 노출위험 검토 과정이 전혀 없는 통계작성기관이 많은 상황에서, 제공범위에 따른 노출위험을 수치적으로, 그러나 손쉽게 검토하는 과정을 시작하는 것이 본 과제의 의의라고 할 수 있겠다.

2) 자료 제공 시 개인정보보호와 정보손실 문제를 다루는 분야를 매스킹(masking), 통계적 노출제어/제한(statistical disclosure control/limitation), 통계적 비밀보호(statistical data confidentiality) 등으로 지칭하며, 최근 데이터 프라이버시(data privacy)라는 용어도 사용되고 있다. 이러한 통계적 비밀보호 분야의 연구자들은 노출위험과 정보손실을 동시에 낮추기 위한 방법을 연구하고 있다. 전통적인 매스킹 기법의 한계를 뛰어넘기 위해, 최근 국내에서도 재현자료(synthetic data), 차등정보보호(differential privacy) 등에 관한 연구가 시작되었다(박민정과 김항준, 2016; 박민정, 2016; 박민정과 김정연, 2017; 박민정 외, 2018).

3) 4차 산업혁명 시대의 주요 자원인 데이터의 제공 및 유통을 위해 개인정보보호와 자료 유용성 확보는 너무나 중요한 문제이기 때문이다.

보통 가이드라인이라 하면 문서를 생각하는 경우가 많다. 문서 형식의 가이드라인은 2015년 통계개발원이 발간했던 '마이크로데이터 매스킹 기법 실무 활용 안내서'라는 이름의 보고서로 제공되어 있다. 해당 보고서는 매스킹 처리 및 노출위험과 정보손실 측정을 전반적으로 다루고 있다. 반면 본 보고서의 제목에 기술된 '가이드라인'은 'R과 연동된 엑셀 프로그램'을 일컫는다. 아직 국내에 가이드라인을 프로그램으로 제시한 사례는 찾기 어려우며, 보통 가이드라인을 문서로만 생각할 수 있으므로 혼란의 방지를 위해 여기에서 부연 설명하는 바이다.[4)]

참고로 가이드라인 프로그램에서 사용하는 매스킹 기법은 전체 재범주화(global recoding)로 한정했다. 이는 가장 기본적인 수준의 노출위험만을 측정하면서, 정보손실을 검토하는 문제를 논외로 하기 위함이다. 한편 노출위험 측도는 네덜란드 통계청에서 제안한 노출위험 추정 모형(Statistics Netherlands, 2008)과 오스트리아 통계청의 후원으로 만들어진 R패키지 sdcMicro에서 제공하는 함수들(Meindl, 2013)을 사용했다. 기초적인 수준의 노출위험 측정 방법을 이용하는 본 가이드라인 프로그램은 향후 통계적 비밀보호 분야의 발전에 따라 업데이트 되는 것이 바람직하다.

보고서는 두 가지 목적을 가지고 작성되었다. 첫째는 노출위험 측정 모형 중에서 가장 기초적인 방법을 일반적인 업무 담당자가 쉽게 이해하고, 실무적으로 활용할 수 있도록 하는 것이다. 이를 위해 복잡한 주제를 설명하는 가이드라인 문서를 대신해, 가이드라인으로서 엑셀을 통해 구동되는 프로그램을 만들었다. 가이드라인의 노출위험을 측정하는 내용은 R로 구현하고 엑셀에 연동시켰다. 둘째 목적은 이러한 프로그램을 만드는 방법을 기술하는 것이다. 즉, 엑셀과 R을 연동하는 방법, 프로그램 형식의 가이드라인 작성 방안을 제시하고자 했다. 따라서 보고서는 독립적인 두 부분으로 구성되어 있다.

보고서 두 번째 부분의 주요 내용인 R-엑셀 연동 프로그램의 기본적인 아이디어는 다음과 같다. R을 이용해 복잡한 분석을 수행하는 경우, R과 엑셀을 연동해 원하는 분석은 R에서 수행하고, 사용자 인터페이스는 엑셀이 제공한다. 만약 연구자가 복잡한 분석을 R을 이용해 수행했고, 일반 이용자가 동일한 분석을 반복하기 원한다고 하자. 일반 이용자가 엑셀 인터페이스로 연구자의 R코드를 손쉽게 사용할 수 있다면, 문서로만은 습득하기 어려운 연구 성과물이 좀 더 활발히 공유될 수 있을 것이다. 본 보고서의 후반부에는 R-엑셀 연동을 구현하는 방법을 제시한다.

---

4) 또한 본 보고서는 마이크로데이터 변수 제공범위 설정과 관련된 일반적인 행정 사항까지 포괄하는 가이드라인 문서가 아니다. 과제 기획 단계에서부터 변수 제공범위를 다르게 할 때 노출위험 변화를 측정하기 위한 가이드라인용 프로그램을 만드는 것이 목적이었다.

보고서의 구성은 다음과 같다. 2절에서는 마이크로데이터 노출위험 측정 방법과 가이드라인 프로그램 사용법을 소개한다. 키변수와 민감변수의 구분, 노출위험 측정 모형 등을 간략히 소개한 후, 음이항 모형과 R패키지 sdcMicro의 기능을 설명한다. 이는 주요 개념을 이해하기 위함이다. 이어 모형을 이용해 계산된 노출위험을 해석하는 방법 및 가이드라인 프로그램 이용법을 다룬다. 향후 가이드라인 프로그램을 배포할 때, 프로그램 사용법 안내 용도로 2절의 내용을 사용할 수 있다. 실제로 통계작성기관 업무 담당자들은 2절과 프로그램을 사용하면 된다.

보고서의 두 번째 목적인 R과 엑셀을 활용해 프로그램을 구현하는 방법은 3절에 기술된다. 3절의 내용은 노출위험과 관련된 주요 이론적인 내용과 상관이 없다. 다만 가이드라인용 프로그램을 만드는 방법만을 다룬다. 먼저 엑셀 VBA를 이용한 가이드라인을 작성한 사례인 Eurostat의 환경경제계정 관련 매뉴얼, 엑셀 VBA를 이용한 통계분석 프로그램인 KESS, R과 엑셀 VBA를 이용한 REx를 소개한다. 또한 자료 익명성 확보를 위한 독립적인 프로그램인 ARX도 소개하고, R과 엑셀을 이용한 가이드라인 프로그램의 장점을 살펴본다. 4절에서는 결론으로 향후 발전 방향을 논의한다.

제2절 마이크로데이터의 노출위험 측정 및 가이드라인

1. 키변수와 민감변수

두 개의 데이터베이스를 연결할 때 사용할 수 있는 공통 변수를 IT 분야에서는 보통 키변수라고 부르며, 식별자(identifier)나 유사 식별자(quasi-identifier)들이 키변수로 사용되곤 한다. 통계적 비밀보호 분야에서도 키변수라는 용어가 사용되지만 그 정의는 문헌마다 다르다.[5] 키변수 정의에 관해, 본 연구과제에서는 네덜란드 통계청이 주도해 만든 매스킹 프로그램인 μ-Argus[6]의 매뉴얼에 제시된 기준을 따르도록 한다. 매뉴얼에 따르면 키변수는 각 개체의 특성을 나타내는 범주형 변수들 중에서 선택된다. 그러한 변수의 조합이 서로 다른 두 개의 데이터베이스를 연결하는 역할을 할 수 있기 때문이다. 예를 들어, 성, 연령, 지역 등의 변수가 키변수가 될 수 있다. 또한 노출을 일으키려는 외부인(intruder)이 오직

5) 특히 R패키지 sdcMicro 폴더 내부에 제공되는 가이드라인 문서에서는 범주형과 연속형 변수 모두 키변수가 될 수 있는 것으로 취급한다. 하지만 연속형 변수를 키변수로 사용하면 모든 개체가 유일해지는 문제가 발생할 수 있다.

6) μ-Argus 및 τ-Argus 프로그램은 각각 마이크로데이터 및 집계자료(매크로데이터)의 매스킹 처리 및 노출위험 측정을 위한 프로그램이다. 이 프로그램들은 통계적 비밀보호 체계 구축을 위해 진행되었던 유럽의 CASC(Computational Aspects of Statistical Confidentiality) 프로젝트의 결과물로 만들어졌다.

키변수만으로 개체들을 연결(match)하려 하고, 재식별(re-identification) 관점에서 노출위험을 평가(assessment)한다고 가정한다.

한편 민감변수가 무엇인지 결정하는 기준은 자의적일 수밖에 없다. 나아가 어느 변수가 민감한지에 대한 판단은 상황에 따라 다를 수 있다. 보통 연속형 변수들은 민감변수로 취급되겠지만, 그 연속형 변수 자체의 속성이 민감하지 않을 수도 있다. 반대로 키변수 역할을 하는 혼인 상태나 교육 수준과 같은 범주형 변수들에 대해, 그 자체로 노출에 민감하다고 판단해서 민감변수로 취급하고 매스킹 처리를 할 수도 있다.

따라서 데이터를 가장 잘 아는 통계작성기관이 사전에 키변수와 민감변수를 구분한다고 가정하도록 한다. 예를 들어 6개의 변수가 있는 마이크로데이터가 있다고 할 때, 다음 <표 9-1>과 같이 키변수와 민감변수에 대해 다양한 구분이 가능하다. 본 보고서에서는 키변수와 민감변수의 구분은 주어진 것으로 간주하고, 주어진 키변수들에 대해서 통계적으로 노출위험을 측정하는 문제를 다루도록 하겠다.

〈표 9-1〉 다양한 키변수와 민감변수의 구분

	성별	연령대	교육수준	혼인상태	소득	지출
구분1	←키변수→		민감변수			
구분2	←키변수→			민감변수		
구분3	←키변수→				민감변수	

비밀보호 처리 과정에는 노출위험과 정보손실을 모두 측정하는 것이 필요하다. 그러나 두 측면을 함께 고려하기에는 매스킹 기법 자체에 한계가 많으며, 이는 새로운 기법의 도입과 관련된 부분으로 본 연구과제의 범위를 넘어선다. 따라서 정보손실 측정을 배제하고 연구범위를 노출위험 측정에 한정하기 위해, 본 연구과제에서는 매스킹 기법들 중 키변수에 대한 전체 재범주화만을 고려한다. 이에 따라 본 연구과제는 재범주화된 키변수들 조합에 대해 노출위험을 측정하고, 변수들의 제공범위가 바뀔 때 노출위험이 상대적으로 어떻게 변하는지에 대해 사용자 친화적인 방식으로 알려주는 것을 목표로 한다.

2. 노출의 종류

마이크로데이터에서 발생하는 노출에는 식별, 속성노출, 추론에 의한 노출 등이 있다. 이에 관해 하나씩 살펴보자. 키변수를 기준으로 노출 여부를 판단하기 가장 손쉬운 방법은 빈도표를 작성해 보는 것이다. 우선 다음 <표 9-2>와 같은 마이크로데이터가 있다고 가정하자.

〈표 9-2〉 마이크로데이터 예제

ID	H	A	M	S
1	20	8	1	2
2	3	5	2	1
...	...	...	...	...
8000	17	2	1	2

먼저 이 마이크로데이터에서 M 및 S 두 개의 변수만을 제공하는 경우를 생각하자. 그러면 다음 <표 9-3>과 같이 빈도표가 만들어진다. 이 빈도표의 셀 값 중 하나가 1이라면 그 키변수 조합을 가지는 개체는 마이크로데이터에서 식별된다고 할 수 있다. 하지만 8000개 개체를 4개의 유형으로 나누면 어느 한 셀에서 1의 값이 나오기 쉽지 않다[7]. 따라서 제공범위를 변수 M 및 S로 한정하는 경우, 안전한 마이크로데이터를 제공하게 된다고 판단할 수 있다.

〈표 9-3〉 2개 변수를 제공할 때 이차원 빈도표

	M=1	M=2
S=1	n_{11}	n_{12}
S=2	n_{21}	n_{22}

한편 변수가 많은 경우에는 위와 같은 이차원 분할표(contingency table)를 작성하기 어렵다. 그래서 위의 빈도표를 다음 <표 9-4>와 같은 형식(flattened frequency table)으로 표현한다. 왼쪽 표가 두 개 변수에 대한 빈도표를 나타낸다. 변수가 많은 경우에 대해서는 <표 9-4>의 오른쪽과 같이 빈도표를 작성할 수 있다. 이차원 분할표와 마찬가지로 빈도가 1인 경우 해당 키변수 조합을 가지는 개체는 마이크로데이터 내에서 유일하기 때문에 식별된다고 볼 수 있다. 변수를 4개 제공한다면 변수별 범주가 각각 24, 12, 2, 2개이므로 키변수 조합은 총 1,152개의 경우가 생기며, 8,000개의 개체가 1,152개의 경우로 나뉘므로 <표 9-3>의 경우보다는 식별 가능성이 훨씬 높아진다. 즉 키변수를 2개 더 추가하여 8,000개의 개체를 분석한다면, 빈도가 1인 키변수 조합이 훨씬 많이 생길 수 있다. 이는 식별 가능성이 훨씬 높아지게 된다는 것을 의미한다.

7) 주어진 변수에 대해 자료가 지나치게 치우친 분포를 가지지 않는, 일반적인 경우를 가정한다.

〈표 9-4〉 2개 변수 및 4개 변수를 제공할 때 빈도표

M	S	빈도
1	1	n_{11}
1	2	n_{12}
2	1	n_{21}
2	2	n_{22}

H	A	M	S	빈도
1	1	1	1	$n_{1,1,1,1}$
1	1	1	2	$n_{1,1,1,2}$
…	…	…	…	…
24	12	2	2	$n_{24,10,2,2}$

다음으로 속성노출이란 식별된 개체에 대해 민감변수의 값이 드러나는 것을 말한다. 주어진 키변수 조합에서 유일한 어느 한 개체가 가지는 민감변수 속성은 쉽게 노출될 수 있다. 한편, 마이크로데이터에서 어떤 개체가 유일하지 않아도 속성이 노출될 수 있다. 특정 키변수 조합에 속한 모든 개체가 같은 민감변수 값을 가지는 경우가 여기에 해당하며, 이러한 경우 그룹 속성노출이 일어났다고 한다.

마지막으로 추론 혹은 확률 노출이 있다. 이는 모형을 통해 개인의 정보가 상당히 정확히 추정되는 경우를 말한다. 본 연구과제에서는 추론에 의한 노출은 연구범위를 넘는 것으로 간주한다. 식별과 매스킹 처리에 대한 좀 더 자세한 내용은 Hundepool 외 (2012)를 참고할 수 있다.

3. 프라이버시 모형

개인정보 보호기술이나 평가 측도를 프라이버시 모형(privacy model)이라고 부른다. 일반적으로 거론되는 프라이버시 모형으로 k-익명성(anonymity)[8], l-다양성(diversity)[9], t-근접성(closeness)[10] 등이 있다. 그 외에 δ-presence라는 것도 있으나 여기서는 설명을 생략한다. 프라이버시 모형은 노출위험 평가 측도로서 기능을 가지며, 기준으로 정한 측도가 일정 수준 이상이 되도록 자료를 변형시켜 안전성을 향상시키도록 유도하는 역할도 한다. 세 측도에 대해 간단히 살펴보자.

어떤 키변수 조합에 속한다는 것은 해당 개체들이 동일한 특성을 가진다는 것을 의미한다. 먼저 k-익명성을 확보했다는 것은 동일한 특성을 가지는 어떤 그룹이 최소

8) Sweeney, L.(2002), k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10, pp. 557-579.

9) Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M.(2006), l-Diversity: Privacy Beyond Kappa-Anonymity, in Proceedings of the 22nd International Conference on Data Engineering (ICDE '06).

10) Li, N., Li, T. and Venkatasubramanian, S.(2007), t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, 2007 IEEE 23rd International Conference on Data Engineering.

k개의 개체들로 구성되어 있다는 것을 말한다. 보통 k-익명성은 식별 위험성과 관련해 사용된다.

다음으로 l-다양성을 가진다는 것은 동일한 특성을 가진 그룹에 속한 개체들이 주어진 민감변수에 대해 l개 이상의 다양한 범주를 가진다는 것을 의미한다. 어떤 변수에 대해 충분히 큰 l-다양성을 가지면, 그 변수의 속성노출 위험은 그만큼 감소한다고 판단할 수 있다.

마지막으로 t-근접성은 l-다양성만으로 속성노출 위험을 제어하기 어렵다고 생각해 제안된 개념이다. 이는 주어진 민감변수에 대해 동일한 특성을 가지는 그룹에 속한 개체들의 분포가 전체 개체들의 민감변수 분포와 유사해야 함을 의미하며, 기준 절단값으로 t를 사용한다. t-근접성을 만족한다는 것은 해당 그룹이 전체 마이크로데이터의 분포와 주어진 기준값에 따른 유사성을 유지해야 한다는 것을 의미하므로 정보손실을 크게 만든다. 본 과제에서는 지나친 정보손실 발생 문제로 t-근접성은 논의하지 않도록 한다.

이제 예를 통해 k-익명성과 l-다양성을 이해해보자. 아래 <표 9-5>에 두 개의 키변수와 하나의 민감변수로 구성된 마이크로데이터가 있다. 처음 세 개체는 익명성 3을 확보했고, 4번 개체는 키변수 조합 (1, 2)가 유일하므로 익명성이 1이다. 5번과 6번 개체는 익명성 2를 확보한다. 따라서 익명성만을 기준으로 생각하면 5번과 6번 개체가 4번 개체보다 안전하다고 판단할 수도 있다.

하지만 5번과 6번 개체는 둘 다 민감변수 값이 62이므로 다양성 측면에서는 모두 1의 값만을 확보하고 있다. 따라서 다양성 측면에서는 4번, 5번, 6번 세 개체 모두 동일한 위험성을 가진다고 해석할 수 있다. 민감변수가 질병번호이고 62번이 에이즈 질병을 나타낸다면, 이 마이크로데이터가 제공될 때 5번과 6번 개체는 노출위험이 심각하다고 판단할 수도 있다. 같은 자료에 대해 <표 9-6> 또는 <표 9-7>과 같이 빈도표를 작성하면 k-익명성을 한 눈에 파악하기 용이하다.

〈표 9-5〉 마이크로데이터에서 k-익명성과 l-다양성

	키변수1	키변수2	민감변수	k-anonymity	l-diversity
1	1	1	50	3	2
2	1	1	50	3	2
3	1	1	42	3	2
4	1	2	42	1	1
5	2	2	62	2	1
6	2	2	62	2	1

〈표 9-6〉 이차원 분할표에서 k-익명성 파악

	키변수1 = 1	키변수1 = 2
키변수2 = 1	3	0
키변수2 = 2	1	2

〈표 9-7〉 빈도표(flattened frequency table)에서 k-익명성 파악

키변수1	키변수2	빈도수
1	1	3
1	2	1
2	1	0
2	2	2

지금까지 살펴본 k-익명성, l-다양성 및 t-근접성은 2016년 관계부처 합동으로 발간된 「개인정보 비식별화 조치 가이드라인」(관계부처 합동, 2016)에서 사용하고 있는 프라이버시 모형이기도 하다. 하지만 이러한 KLT 프라이버시 모델을 사용한다면 정보손실이 너무 커져서 자료가 쓸모없어지는 문제가 생긴다. 통신사 자료와 같이 매우 거대한 마이크로데이터에서는 이런 모형을 기준으로 노출위험을 평가하며 매스킹 처리를 하더라도 자료 유용성이 남아 있을 수 있다. 하지만 프라이버시 모형에서 검토하지 않은 민감변수를 다시 키변수로 사용해 자료를 공격적으로 분석한다면, 여전히 익명성과 다양성을 만족시키기는 불가능하다고 판단된다. 본 연구과제에서는 k-익명성과 l-다양성 측도는 참고로 사용하도록 한다.

4. 노출위험을 추정하는 통계적 모형

앞에서 설명한 프라이버시 모형들을 적용하는 경우, 적절한 기준 이하로 노출위험을 낮추기 위해 정보손실이 너무 크게 발생하는 문제가 있음을 언급했었다. 그러한 프라이버시 모형들에서는 정보손실 발생 이외에도 여러 단점들이 존재한다. 만약 k-익명성을 이용해 제공범위에 따른 노출위험을 측정한다면, 제공변수를 늘릴 때 주어진 k-익명성을 만족하지 못하는 개체의 숫자가 얼마나 늘어나는지만 제시할 수 있을 뿐이다. 또한 k-익명성을 만족하더라도 l-다양성을 만족하지 못하면 개별정보의 노출위험은 여전히 높다고 판단할 수도 있다. 나아가 표본 조사로 얻어진 마이크로데이터의 경우, k-익명성은 가중값의 영향을 전혀 반영하지 못하는 단점도 가진다.

이러한 k-익명성은 개체의 유일성에 근거한 노출위험에 속한다. 개체의 유일성을 기준으로 삼을 때는 어떤 개체가 모집단에서 유일하면 식별이 되고, 노출위험 값이 최대치인 1이라고 판단한다. k-익명성을 노출확률로 표현한다면 익명성 1인 경우에 노출확률은 1, 익명성 2인 경우에 0.5, 익명성 100인 경우는 0.01이라는 익명성 k의 역수로 노출확률을 표현할 수도 있다. k-익명성에 가중값을 반영하기 위해 단순히 익명성 k의 역수에 가중값의 역수를 곱해서 노출확률을 만들어 낼 수도 있다. 그러나 이러한 노출위험 측도로는 정확성을 담보하기 어렵다.

학계에서는 통계적 이론을 적용해 엄밀한 노출위험 측도를 제안하고 점검하는 작업이 오랫동안 이루어져 왔다. 개체의 유일성을 근거로 노출위험을 측정하기 위해서, 표본에서 유일한 개체가 모집단에서도 유일할 확률을 통계적 모형을 이용해 추정한다. 그런 모형들로는 SUDA[11], 포아송-감마 모형[12], Argus 모형[13], 로그 선형 모형[14][15], GoM 모형[16] 등이 있다. 이런 모형들을 사용해서 파일의 노출위험이나 각 개체별 노출위험을 계산할 수 있으며, 일부 모형들을 비교해 분석한 문헌[17]도 있다. 다만 각 모형들을 이용해 추정한 노출위험은 편의를 가지는 경우가 많다는 것을 주의해야 한다.

한편, 유일성에 근거한 노출위험 측도의 한계를 뛰어넘기 위해 자료를 공격하는 외부인에 대해 의사결정론(decision theory)을 적용[18]해 노출위험을 계산할 수도 있다. 이런 내용은 이론적으로 어렵고 프로그램을 범용적으로 구현하기가 거의 불가능하다. 다만 본 보고서에서 의사결정론 기반 노출위험 모형의 존재를 언급하는 이유는 다양한 노출위험 측도가 있다는 사실을 강조하기 위함이다. 각 모형에 대한 자세한 내용은 각주의 참고문헌과 박민정(2014)을 참고할 수 있다.

---

11) M. J. Elliot, A. Manning, K. Mayes, J. Gurd and M. Bane(2005), SUDA: A Program for Detecting Special Uniques, Using DIS to Modify the Classification of Special Uniques, Joint UNECE/Eurostat work session on statistical data confidentiality 2005.

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

13) μ-Argus manual 4.2

14) Skinner, C.J. and Holmes, D.J.(1998), Estimating the re-identification risk per record in microdata. Journal of Official Statistics, 14:361-372.

15) 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.

16) 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.

17) Rinott, Y.(2003), On Models for Statistical Disclosure Risk Estimation, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

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

정리하면, 노출위험을 측정할 때 k-익명성이나 l-다양성을 활용할 수도 있고, 어떤 개체가 모집단에서 유일할 확률을 통계적 모형을 이용해서 추정할 수도 있으며, 외부 공격자에게 의사결정이론을 적용하는 등의 더 엄밀한 방법을 활용할 수도 있다. 본 연구 과제에서는 k-익명성의 한계를 벗어나 노출위험을 측정하고자 한다. 그러나 실무적으로 손쉽게 노출위험을 측정할 수 있어야 한다. 따라서 복잡한 의사결정이론 등은 배제하고, 유일성에 근거한 노출위험 측정 모형을 활용하기로 결정했다. 따라서 가이드라인 프로그램에서는 네덜란드 통계청이 제안했던 모형을 이용해, 키변수 조합의 유일성에 근거해서 개체별로 노출위험을 측정한다. 마이크로데이터 파일 수준의 노출위험은 개체별 노출위험의 평균값을 활용한다.

개체의 유일성 개념에 근거해서, 즉 어떤 키변수 조합이 모집단에서 유일할 확률을 이용해 노출위험을 측정하는 모형은 프로그램으로 이미 구현되어 있다. 이 노출위험 측정 모형은 R패키지 sdcMicro에 구현되어 있으며, 가이드라인 프로그램은 R패키지 sdcMicro를 활용한다. 단순히 각 개체별로 k-익명성을 표현하는데 머무르기보다는, 가중값 정보도 활용하는 통계적 모형을 이용해 각 개체별로 노출위험을 추정하는 것 바람직하다. 주의할 것은 이러한 통계적 모형들을 이용한 추정량에 편의가 존재한다는 것이다. 따라서 통계적 모형을 이용해 제공범위에 따른 노출위험을 측정한 결과를 수치 그대로 신뢰하는 것은 바람직하지 않다. 제공하는 변수가 추가될 때, 노출위험이 얼마나 증가하는지 상대적으로 비교하는 방식이 바람직하다.

5. 노출위험 측정 및 해석

본 보고서에서 노출위험 측정을 위해 R패키지 sdcMicro를 이용한다. 이 패키지는 μ-Argus의 노출위험 추정 모형과 동일하게 개별 레코드의 유일성 기반 노출위험을 계산해 제공한다. 모형을 이해하기 위해 다음 <표 9-8>과 같이 마이크로데이터를 빈도표 형식으로 바꾸어 보자.

이제 i번째 키변수 조합에 대해 주어진 마이크로데이터에서 계산한 빈도수를 f_i라 하고, 알려지지 않은 모집단의 빈도수를 F_i라고 하자. 어떤 개체가 k번째 키변수 조합에 속한다고 하면, 이 i번째 개체의 노출위험을 다음과 같이 정의한다.

$$r_i = E\left(\frac{1}{F_k} \mid f_k\right).$$

따라서 k번째 키변수 조합에 속하는 모든 개체는 동일한 노출위험을 가진다. 이 확률을 노출위험으로 정의하고, 통계적 모형을 적용해 이 확률을 계산한다. 그러면 각 개체에 대해 노출위험을 구할 수 있다.

〈표 9-8〉 마이크로데이터와 빈도표

마이크로데이터

ID	V_1	⋯	V_P
1	1	⋯	1
2	1	⋯	2
⋯	⋯	⋯	⋯
⋯	⋯	⋯	⋯
i	⋯	⋯	⋯
⋯	⋯	⋯	⋯
⋯	⋯	⋯	⋯
n	⋯	⋯	⋯

빈도표

키조합	V_1	⋯	V_P	표본 빈도수	모집단 빈도수
1	1	⋯	1	f_1	F_1
2	1	⋯	2	f_2	F_2
⋯	⋯	⋯	⋯	⋯	⋯
k	⋯	⋯	⋯	f_k	F_k
⋯	⋯	⋯	⋯	⋯	⋯
K	C_1	⋯	C_P	f_K	F_K

한편 제공범위를 어느 변수들로 구성하느냐의 문제를 결정하기 위해서는 파일 단위에 대한 노출위험이 필요하다. 이때는 각 개체별 노출위험의 평균 혹은 합을 이용할 수 있다. 각 개체별 노출위험을 계산하는 R코드의 개념은 키변수와 가중값을 어떤 함수에 넣고 계산하면 노출위험 측정값이 산출된다는 것이다. 간단히 도식화하면 다음과 같다.

노출위험 측정값 ← 노출위험 계산 함수 (키변수, 가중값)

노출위험 함수는 다음 <표 9-9>와 같이 각 개체별로 노출위험 계산 결과를 반환한다. 노출확률은 통계적 모형을 이용해 계산하고, 표본 빈도수로는 주어진 마이크로데이터를 관측한 결과를 제시하며, 모집단 빈도수는 표본 빈도수에 가중값을 곱해서 간단히 계산한다. 이때 표본 빈도수가 1이면, 표본에서 해당 개체의 k-익명성은 1인 것이다.

참고로 통계적 모형을 이용해 계산한 노출위험 결과값은 모형이 가지는 문제점(편의 발생)을 그대로 반영하므로, 노출확률 숫자 자체에 절대적인 의미를 부여해서 해석하는 것은 바람직하지 않다는 점을 다시 한 번 강조한다.

이제 제공범위를 결정하기 위해, 제공범위를 바꿀 때 노출위험이 어떻게 달라지는지를 살펴볼 차례이다. 각 개체별 노출위험 변화를 일일이 관찰하는 것은 비효율적이므로 개체별 노출위험의 평균, 합 등을 활용한다.

〈표 9-9〉 개체별 노출위험 계산 결과

ID	V_1	…	V_P	노출확률 (추정값)	표본 빈도수 (관측값)	모집단 빈도수 (추정값)
1	1	…	1	r_1	f_1	F_1
2	1	…	2	r_2	f_2	F_2
…	…	…	…	…	…	…
i	…	…	…	r_i	f_i	F_i
…	…	…	…	…	…	…
n	…	…	…	r_n	f_n	F_n

일례로, 가계금융・복지조사 마이크로데이터를 이용해 노출위험을 계산해 보았다. 아래 <표 9-10>에서는 총 12개 변수 중에서 5개 변수는 반드시 제공하는 것으로 가정했다. 나머지 7개 변수들에 대해서는 제공여부를 0과 1로 표현해서 시나리오를 만들었다. 전체적으로 검토할 수 있는 시나리오는 $2^7 = 128$ 개가 된다. 이제 변수를 하나도 추가하지 않는 경우부터 7개 변수를 모두 추가하여 제공하는 것까지, 128 종류의 제공범위를 비교하고자 한다.

2만여 개의 각 개체들에 대해 개별노출위험을 계산해서 비교하기는 어려우므로, 그 대신에 개별노출위험의 평균을 'risk'값으로 기록했다. 또한 표본 빈도수가 1인 개수를 'U1', 2인 개수를 'U2'로 표기했다.

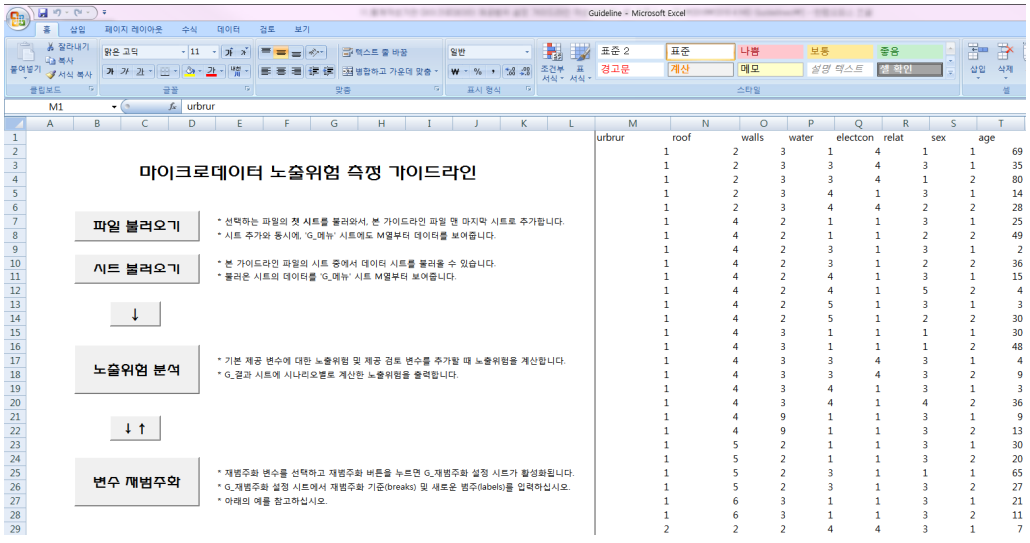
아래 <표 9-10>은 128개 시나리오 중 일부의 결과를 보여준다. 시나리오 S1은 5개 변수만을 제공하는 경우를 말하며, 노출위험의 각 열들은 익명성 1인 개체수, 익명성 2인 개체수 및 통계적 모형을 이용한 개체별 노출위험의 평균값을 나타낸다. 여기서는 S1을 기본 노출위험 시나리오라고 하겠다. 시나리오 S2, S3, S5는 시나리오 S1에 각각 V12, V11, V10을 추가로 제공하는 경우를 말한다. 노출위험은 기본 노출위험 0.07에서 0.14, 0.17, 0.13으로 약 두 배 가량 증가한다. 만약 어떤 변수를 추가할 때 노출위험이 특별히 많이 증가한다면, 자료제공자는 해당 변수를 다시 점검하고 전체 재범주화 등의 조치를 취할 필요가 있다.

〈표 9-10〉 노출위험 계산 결과 예제

시나리오	키변수 포함 여부												노출위험		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	U1	U2	risk
S1	1	1	1	1	1	0	0	0	0	0	0	0	1938	1740	0.07
S2	1	1	1	1	1	0	0	0	0	0	0	1	4854	2826	0.14
S3	1	1	1	1	1	0	0	0	0	0	1	0	5746	3178	0.17
S4	1	1	1	1	1	0	0	0	0	0	1	1	9001	3626	0.25
S5	1	1	1	1	1	0	0	0	0	1	0	0	4369	2748	0.13
S6	1	1	1	1	1	0	0	1	0	1	0	0	8245	3655	0.22
S7	1	1	1	1	1	0	0	1	0	1	0	1	11626	3166	0.30
S8	1	1	1	1	1	0	0	1	0	1	1	0	12440	3426	0.32
S9	1	1	1	1	1	0	0	1	0	1	1	1	14856	2623	0.37
S10	1	1	1	1	1	0	0	1	1	0	0	0	9630	3724	0.26

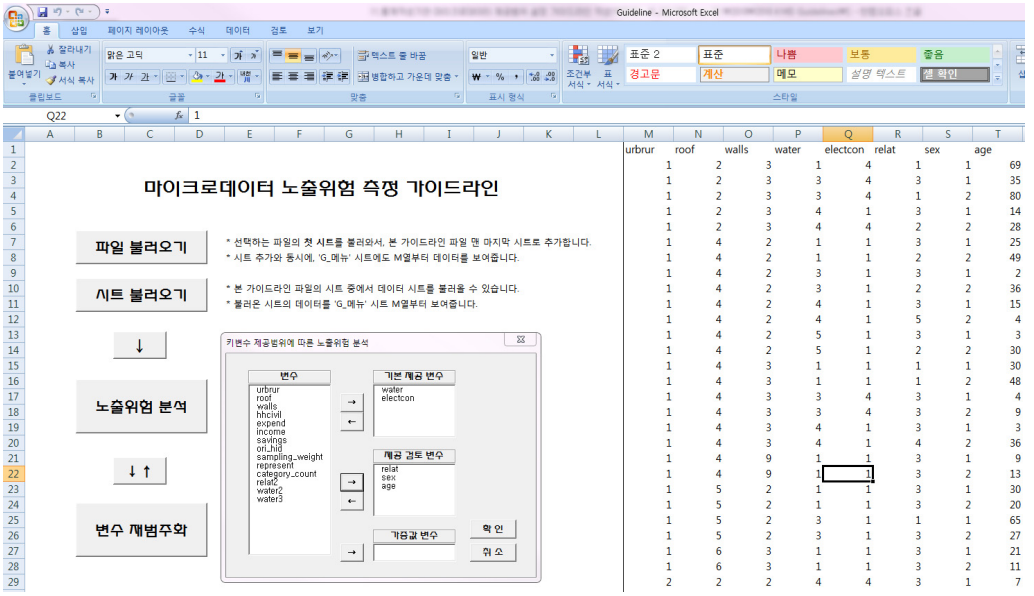
6. 가이드라인 프로그램 이용법

앞에서 설명한 시나리오별 노출위험 변화를 파악하기 위해 본 연구에서 제공하는 가이드라인 프로그램을 활용할 수 있다. 가이드라인 엑셀파일을 열면 다음[그림 9-1]과 같은 화면이 'G_메뉴' 시트에 나타난다. 시트 이름에서 'G'는 가이드라인을 의미한다.



[그림 9-1] 가이드라인 엑셀 프로그램

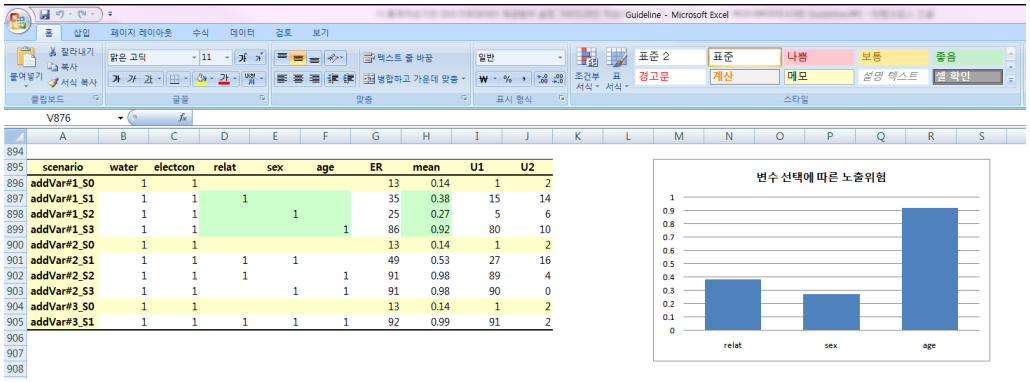
이용 순서는 화면에 나타난 것과 같다. 먼저 '파일 불러오기' 또는 '시트 불러오기' 버튼을 사용해서 데이터를 불러오면, 가이드라인 'G_메뉴' 시트의 오른쪽에 데이터가 표시된다. '파일 불러오기'를 실행하면 엑셀에서 파일 불러오기를 하는 것처럼 불러온 데이터가 독립된 시트로도 제공된다. 다음으로 '노출위험 분석' 버튼을 눌러서 반드시 제공하는 변수들을 '기본 제공 변수'로 선택하고, 시나리오별로 노출위험을 비교하고 싶은 변수들을 '제공 검토 변수'로 선택한다. 가중값 변수가 있으면 함께 선택해 준다. 다음 [그림 9-2]는 데이터를 불러온 후, 노출위험 분석 버튼을 누르면 나타나는 화면을 보여준다.



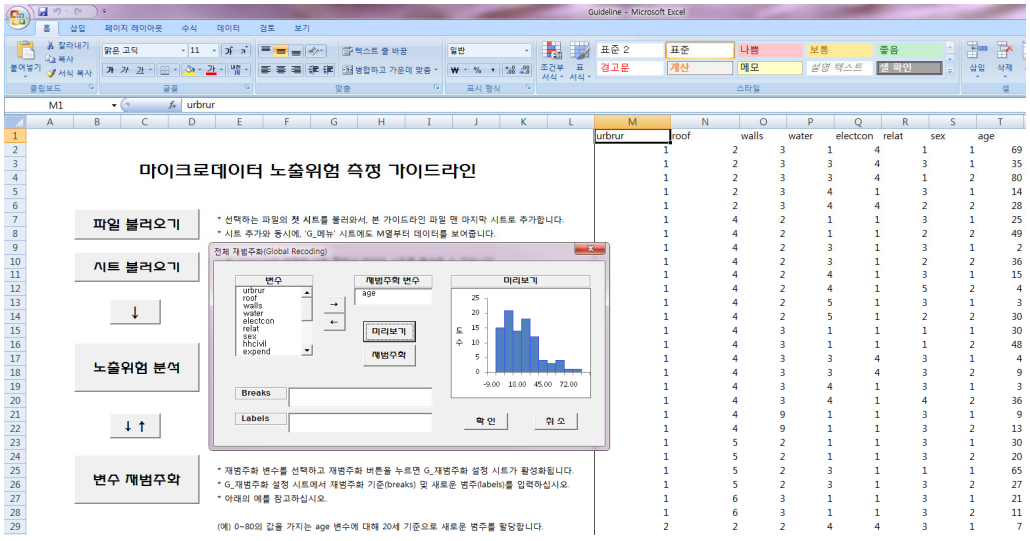
[그림 9-2] 'G_메뉴' 시트에서 노출위험 분석 수행

변수 선택 후, '확인' 버튼을 눌러 노출위험 분석을 수행하면 다음 [그림 9-3]과 같은 화면을 'G_결과' 시트에서 얻을 수 있다. addVar#1은 변수를 하나 추가한 시나리오를 의미한다. [그림 9-3]에서 addVar#1_S1부터 addVar#1_S3은 기본 제공 변수 2개에 변수를 하나 더했을 때 모집단에서 유일할 것으로 추정되는 개체들의 수(ER), 평균 노출위험(mean), 표본에서 유일한 개체들의 수(U1), 표본에서 익명성 2를 가지는 개체들의 수(U2)를 차례로 보여준다. 또한 오른쪽의 히스토그램은 각각의 변수를 하나 추가할 때 시나리오인 addVar#1들의 평균 노출위험(mean)을 보여준다. 이 예에서는 기본 제공 변수 2개(water, electcon)에 연령(age) 변수를 추가하면, 다른 변수(relat, sex)를 추가할 때에 비해 노출위험이 급격히 증가하는 것을 확인할 수 있다. 따라서 연령 변수를 재범주화할 필요가 있다는 판단을 할 수 있다.

참고로 각 변수의 범주 개수를 통해서 이러한 판단을 내릴 수도 있다. 연령의 범주가 각 세 단위이므로 다른 변수들에 비해 빈도표를 더 세분화하게 되고, 따라서 노출위험이 높아지는 것은 자명하다. 다른 두 변수는 2~3개의 범주로 이루어지므로 노출위험을 덜 높일 것이 예상된다. 이렇게 업무담당자가 각 변수의 범주 개수를 일일이 검토해서 노출위험을 높이는 변수들을 짐작해낼 수도 있다. 반면 가이드라인 프로그램을 사용한다면 그러한 결과를 숫자로 확인하거나, 그림을 통해 상대적으로 노출위험을 높이는 변수를 발견할 수 있다. 즉 짐작 대신에 정량적인 검토를 손쉽게 할 수 있다는 장점이 있다. 나아가 변수 범주의 개수는 적지만 자료의 분포가 특이해서 노출위험이 높아지는 경우를 걸러낼 수 있다.



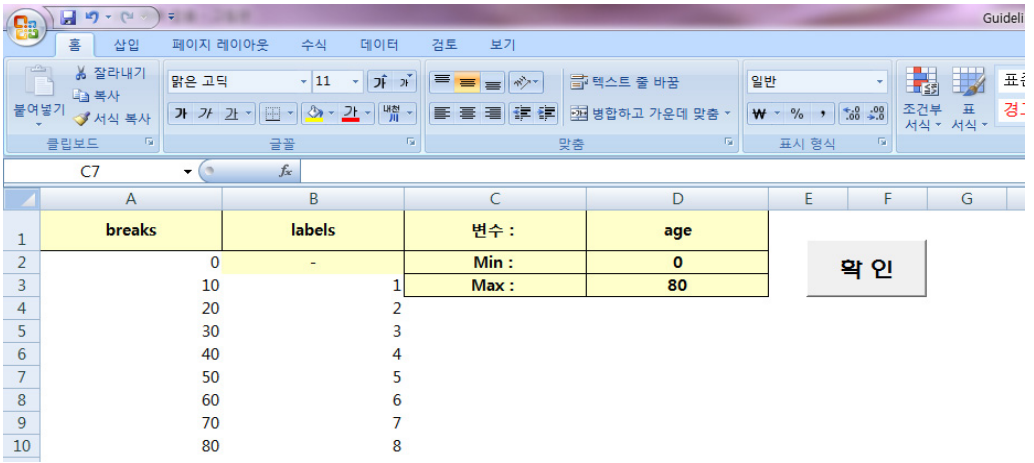
[그림 9-3] 'G_결과' 시트에 나타나는 노출위험 분석 결과



[그림 9-4] 'G_메뉴' 시트에서 변수 재범주화 수행

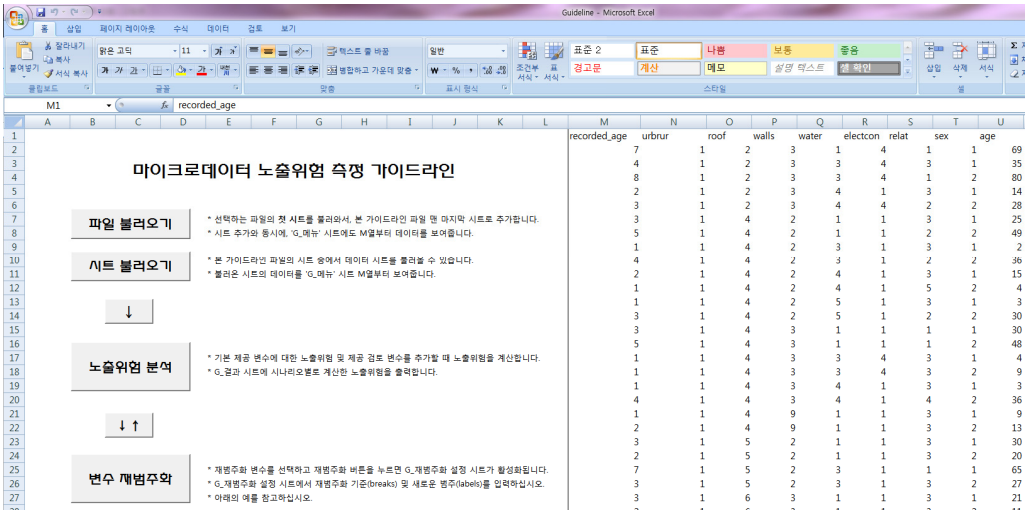
과도하게 노출위험을 증가시키는 변수가 발견되면 재범주화 조치를 취한다. 재범주화를 위해서는 [그림 9-4]와 같이 'G_메뉴' 시트에서 '변수 재범주화' 버튼을 누르고 재범주화할 변수를 선택하면 된다.

'미리보기' 버튼 아래 있는 '재범주화' 버튼을 누르면, 다름 [그림 9-5]에서처럼 'G_재범주화 설정' 시트의 D열에 선택한 변수의 최소값(Min)과 최대값(Max)이 나타난다. 선택한 변수에 대해 재범주화하고 싶은 기준값들을 'breaks'열에 적어주고, 새로 부여하는 범주들은 'labels'열에 입력한다. [그림 9-5]는 각 세 연령을 10세 단위로 바꾸는 예를 보여준다.



	A	B	C	D
1	breaks	labels	변수 :	age
2	0	-	Min :	0
3	10	1	Max :	80
4	20	2		
5	30	3		
6	40	4		
7	50	5		
8	60	6		
9	70	7		
10	80	8		

[그림 9-5] 'G_재범주화 설정' 시트에서 재범주화 기준(breaks) 및 새로운 범주(labels)의 입력



[그림 9-6] 'G_메뉴' 시트에 M행에 나타난 재범주화 처리된 변수

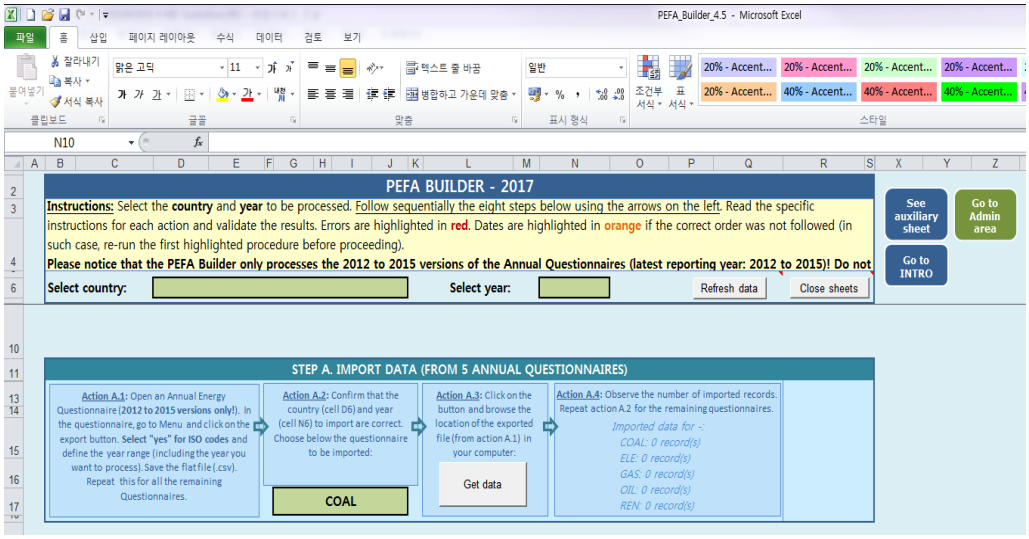
이제 '확인' 버튼을 누르면 R에서 처리한 재범주화 결과가 'G_메뉴' 시트 오른쪽에 나타나며, 재범주화된 변수가 'recorded_변수명'으로 나타난다. [그림 9-6]은 연령에 대해 재범주 처리한 결과를 보여준다.

제3절 엑셀과 R을 연동한 프로그램(가이드라인) 작성 방안

1. 가이드라인용 프로그램 개발을 위한 벤치마킹

본 보고서의 2절과 3절은 독립적인 내용으로 구성되어 있다. 3절에서는 2절에서 설명한 노출위험을 사용자 친화적으로 보여주기 위해 가이드라인 프로그램을 만드는 법을 논의한다. 이는 연구, 개발된 새로운 방법론을 일반적인 이용자가 편리하게 사용할 수 있도록 프로그램 형식으로 가이드라인을 작성하는 방안을 제시하는 것이다.

보통 가이드라인은 전문가가 아니더라도 손쉽게 따를 수 있도록 작성되는 것이 바람직하다. 하지만 통계 생산과 관련해 기술적인 방법을 전하려고 할 때는 가이드라인이 복잡해질 수밖에 없다. 게다가 조직에서 업무 담당자가 계속해서 바뀌는 상황이라면 가이드라인 습득을 위한 시간이 필요하므로 업무 수행이 지연되는 상황이 발생할 수 있다. 유럽연합 통계청(Eurostat)에서는 환경경제계정을 작성하도록 권고하면서 이런 문제를 파악하고, 프로그램으로 만들어진 매뉴얼을 제공했다. 아래 [그림 9-7]은 환경

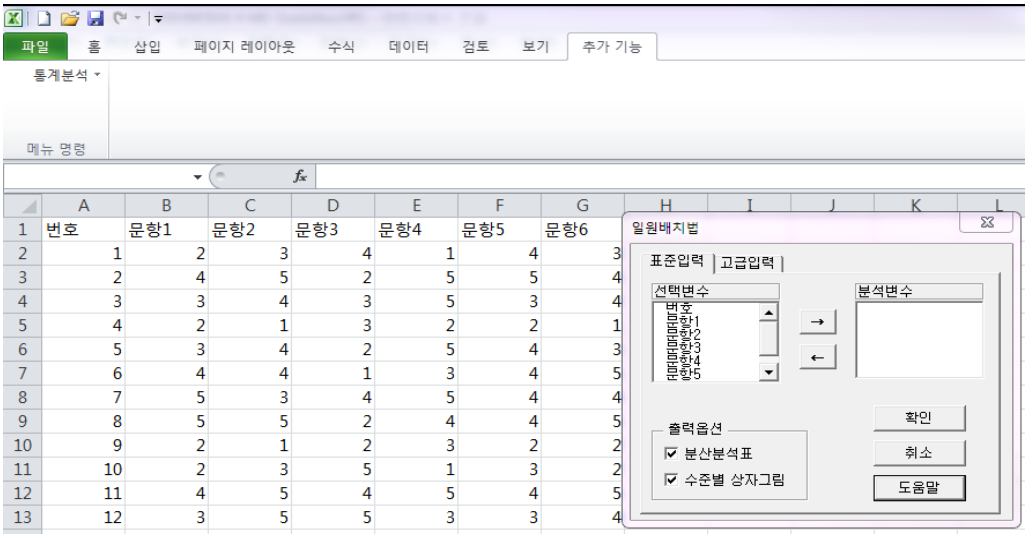


[그림 9-7] Eurostat의 PEFA tools 2017

경제흐름(flow)계정 중 에너지흐름계정 작성을 위한 가이드라인 프로그램인 PEFA (Eurostat Physical Energy Flow Accounts Manual) tools 2017 중 일부 화면을 보여준다.

이 프로그램은 엑셀 VBA(Visual Basics for Applications)만을 이용해 만들어졌다. 환경경제계정 작성 담당자가 복잡한 계정 작성 과정을 직접 수행하는 대신에, 프로그램이 제공하는 안내에 따라 각 버튼을 클릭하고 지시 사항을 수행하도록 구성되어 있다. 프로그램에서 요구하는 항목들을 모두 입력하면, 가이드라인이 지시하는 통계 결과가 작성되어 새로운 파일로 저장된다. 이런 방식의 프로그램으로 제공되는 매뉴얼 혹은 가이드라인의 장점은 업무 담당자가 해당 이론에 해박하지 않아도 큰 오류 없이 필요한 업무를 빠르게 수행할 수 있다는 것이다. 본 과제는 마이크로데이터 제공범위에 따른 노출위험을 측정하는 가이드라인을 제공하는 것이 주요 목적이므로 Eurostat의 매뉴얼을 참조했다.

한편, 국내에도 엑셀 VBA를 활용한 훌륭한 통계 분석 프로그램이 있다. KESS[19]는 교육용 통계 소프트웨어이며, 장기간의 메뉴 업데이트 과정과 점검을 거쳤다. 때문에 프로그램이 안정적이며, 무엇보다 인터페이스가 사용자 친화적으로 뛰어난 장점이 있다. 엑셀 메뉴로 구동되고, 대화창이 간결하며, 분석 결과물이 엑셀 시트위에 제공되므로 편집하기 쉽고, 그래프도 깔끔하다. 노출위험 측정 프로그램 구현을 위해 본 과제에서는 KESS의 사용자 친화적인 인터페이스를 따르고, 허락을 구하여 일부 소스 코드를 차용했다. 참고로 아래 [그림 9-8]은 KESS 화면 일부를 보여준다.



	A	B	C	D	E	F	G
1	번호	문항1	문항2	문항3	문항4	문항5	문항6
2	1	2	3	4	1	4	3
3	2	4	5	2	5	5	4
4	3	3	4	3	5	3	4
5	4	2	1	3	2	2	1
6	5	3	4	2	5	4	3
7	6	4	4	1	3	4	5
8	7	5	3	4	5	4	4
9	8	5	5	2	4	4	5
10	9	2	1	2	3	2	2
11	10	2	3	5	1	3	2
12	11	4	5	4	5	4	5
13	12	3	5	5	3	3	4

[그림 9-8] KESS의 사용자 친화적인 인터페이스

19) http://203.252.193.107/xe/

KESS의 단점은 분석 기능을 모두 VBA로 직접 구현해야 한다는 것이다. 따라서 새로운 기능을 추가하기 위해서는 그 기능과 관련된 모든 요소를 추가로 구현해야 한다. 만약 통계적 모형을 적용해 노출위험을 측정할 때 KESS 방식을 따른다면, 관련 함수를 모두 VBA로 새로 짜야 하는 어려움이 있다. 향후 다른 주제로 새로운 가이드라인용 프로그램을 작성하려 해도 연구자가 다른 언어로 작성한 코드를 VBA로 다시 코딩해야 한다. 반면 통계와 관련된 각종 프로그램들은 R로 이미 구현되어 있는 경우가 많다.

통계개발원의 내부 보고서인 박민정과 김경미(2013)에서는 원격접근서버 발전 방향을 논의하며 엑셀과 R의 연동을 제안한 바 있다. 이를 위해 당시에 R과 엑셀을 연결해주는 RExcel이라는 프로그램을 검토한 적이 있으나, 잘 작동하지 않아 후속 연구를 진행하지 못했다. 그러나 최근 국내에서 성공적으로 엑셀과 R을 연동해 통계 분석 프로그램을 구현한 REx[20]라는 이름의 통계 분석 프로그램이 2017년 12월부터 제공되고 있다.

REx의 성공을 인지하여 통계개발원에서는 기존의 '마이크로데이터 매스킹 기법 실무 활용 안내서(박민정, 2015)'를 대신할 가이드라인 프로그램을 작성하는 과제를 시작하였다. REx와 같은 방식으로 엑셀 VBA를 활용하면, 현재 존재하는 다양한 R프로그램들을 그대로 활용할 수 있는 장점이 있기 때문이다. 이제 본 보고서에서는 엑셀과 R을 연동하는 방법을 제시하고자 한다.[21] 이는 이번 가이드라인뿐만 아니라 다른 주제에 대해서도 유사한 시도를 할 수 있도록 지원하기 위함이다.

참고로 노출위험을 측정하고 매스킹 처리를 수행하는 ARX라는 프로그램을 소개한다. 이 프로그램은 익명화 처리를 위해 최근에 만들어졌다. k-익명성 등을 보여주고, 매스킹 처리에 따른 변화도 즉시 화면으로 나타낸다. 본 과제에서 사용하는 R패키지 sdcMicro의 거의 모든 기능을 가지고 있으며, 사용자 친화성이 강한 장점이 있다.

ARX의 단점은 다음과 같다. ARX는 매스킹 처리 및 노출위험 측정에 관한 새로운 프로그램이므로, 업무 담당자가 새로운 프로그램을 설치하고 적응해야 하는 단점이 있다. 또한 언제든 유료화가 될 수도 있다. 반면 엑셀은 사용자 친화성 측면에서 가장 강력한 프로그램이다. 독립적인 프로그램들은 나름의 장점이 있겠지만, 가능하다면 이미 사용자의 컴퓨터에 설치되어 있는 엑셀과 R을 활용할 수 있다면 좋을 것이다. 일례로 네덜란드 통계청이 만든 μ-Argus라는 프로그램이 있음에도, 사용자 친화성이 강한 R패키지인 sdcMicro가 만들어진 바 있다.

20) http://rexsoft.org/
21) 본 연구를 통해 알아낸 R과 엑셀을 연동하는 방식은 REx의 방식과 다르다.

본 연구과제에서는 지금까지 설명한 여러 프로그램들을 검토하여 노출위험 측정 가이드라인 프로그램의 구현 방향을 결정했다.

2. 엑셀과 R을 연동하는 프로그램의 구현

엑셀은 기본적인 자료 관리에서부터 보고서 작성 기능까지 갖추고 있어 다양한 용도로 활용되고 있다. 또한 대단히 폭넓은 사용자층을 가지고 있다. 따라서 엑셀에 R의 고급 자료 분석 기능까지 추가할 수 있다면, 보다 넓은 범위의 이용자들이 다양한 기능들을 편리하게 사용할 수 있을 것이다. 이를 위해 분석은 R에서 수행하고, 엑셀은 사용자 친화적인 인터페이스 역할을 한다는 아이디어를 제시한다. 즉, 엑셀은 자료와 사용자의 요구사항을 R로 전달하고, R의 분석결과를 받아서 사용자 친화적으로 보여주는 역할을 하도록 한다.

이 절에서는 R과 엑셀이 직접 통신하는 방법을 설명한다. 먼저 엑셀과 R이 연동되도록 경로를 지정하는 방법을 보여준다. 다음으로 엑셀에서 R로 분석 대상 자료 및 분석 요구 사항을 보내는 방법을 자세히 설명한다. 경로 지정 및 엑셀에서 R로 자료를 전송하는 방법이 프로그램 구축에 있어 핵심 요소가 된다.

R은 엑셀이 전송한 데이터와 요구사항에 따라 분석을 수행하고, 분석 결과는 클립보드에 저장한다. 엑셀은 클립보드의 내용을 결과시트[22]에 보여준다. 엑셀이 결과 시트를 보여주는 작업은 KESS와 같이 엑셀 VBA를 이용한다.

참고로 엑셀에서 R이 이해하는 형식으로 자료를 보내기 위해, R에서 사용하는 객체의 종류를 이해할 필요가 있다. 이를 위해 관련 내용을 <부록> 1.에서 간략히 소개했다. 한편 이 절에서 설명하는 엑셀과 R의 직접 연동 대신에, 엑셀과 R이 통신하기 위해 외부 프로그램을 사용할 수도 있다. 본 연구과제에서는 노출위험 측정 가이드라인을 구현하기 위해, 외부 프로그램을 사용하지 않고 엑셀과 R을 직접 통신시키는 방식을 채택했다. 외부 프로그램을 사용하여 엑셀과 R을 통신하도록 하는 방법은 관심 있는 독자들이 참고할 수 있도록 <부록> 2.에 정리했다.

엑셀에서 R과 직접 통신하기 위해서는 다음 사항들이 해결되어야 한다.

---

22) 보다 손쉬운 방법으로 분석 결과를 html 형식으로 저장해 보여줄 수도 있다. 이는 REx가 사용하는 방식이다. html 형식 대신에 분석 결과를 엑셀 시트위에 보여주기 위해서는 엑셀 VBA를 이용해 프로그램을 구현하는 과정이 추가로 필요하지만, 사용자는 엑셀에서 계속하여 작업할 수 있는 장점이 있다.

(i) R이 설치된 path를 시스템이 알고 있어야 한다.
(ii) 엑셀에서 R을 호출하는 VBA 코딩 방법을 알아야 한다.
(iii) R에서 엑셀이 보낸 인수를 받기 위한 방법을 알아야 한다.

가. path 구성

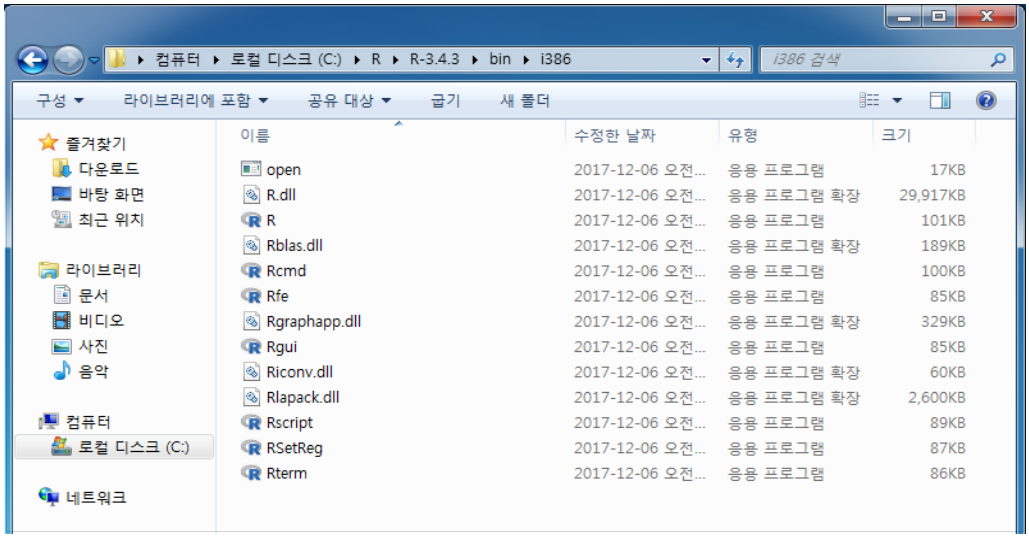
엑셀이 R과 통신하기 위해서 R이 설치된 path를 쉽게 알 수 있어야 한다. 이를 위한 여러 방법들이 있다. 기본적인 방법 중 하나를 단계별로 설명하면 다음과 같다.

1단계:

[그림 9-9]처럼 R이 설치된 폴더를 열어서 폴더 경로를 클립보드에 복사한다.

R은 보통 C나 D 드라이브에 있는 Program Files 하에 설치된다. 예를 들면, 다음과 같은 폴더이다.

C:\Program Files\R\R-3.5.1\bin\x64



[그림 9-9] R이 설치된 폴더 경로 복사

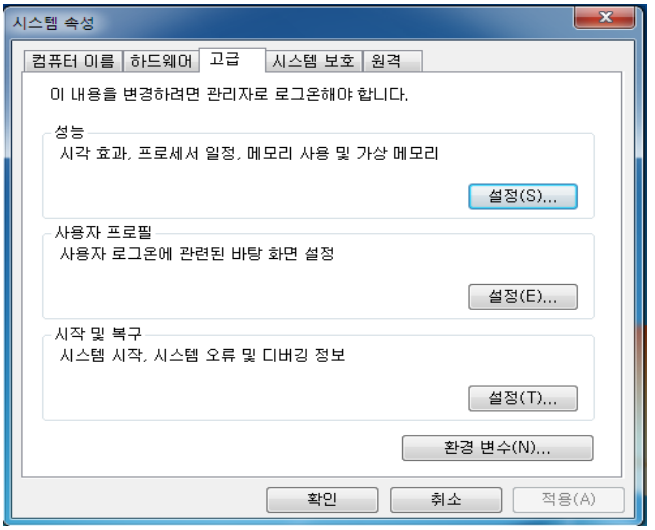
이 폴더 내에는 R.exe 및 Rscript.exe가 들어 있다. 엑셀이 R과 통신하기 위해서 필요한 실행 파일이 바로 Rscript.exe이다.

2단계:

이제 복사된 경로를 기존의 path에 추가하는 단계이다. 이를 위해서 윈도우 탐색기에 있는 [컴퓨터](혹은 [내pc])의 아이콘에 마우스를 대고 오른쪽 버튼을 클릭하면 나오는 메뉴 중, [속성]을 선택한다. 그러면 [그림 9-10]와 같은 화면을 볼 수 있다. 왼쪽 메뉴 중에서 [고급 시스템 설정]을 클릭하면 나타나는 [그림 9-11]와 같은 시스템 속성 화면에서 아래 쪽의 [환경 변수] 단추를 클릭한다.

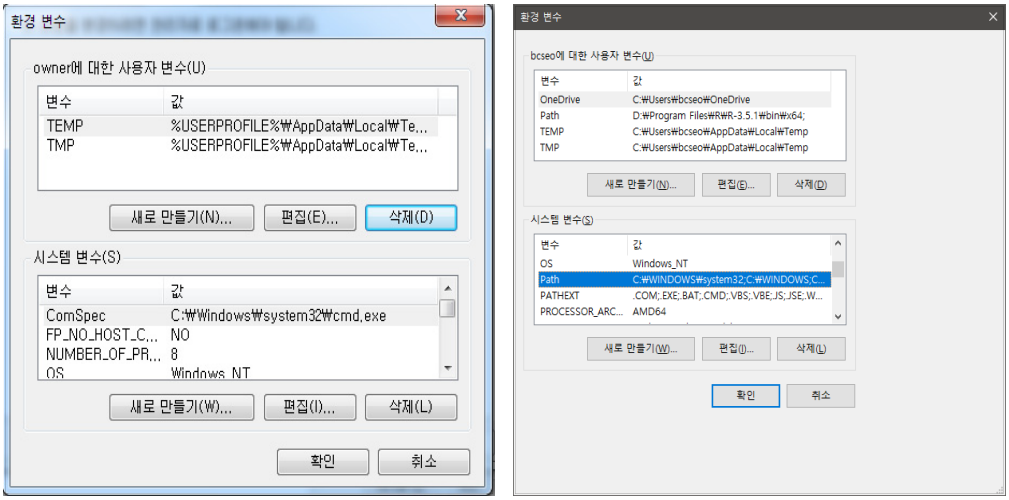


[그림 9-10] [컴퓨터]의 속성

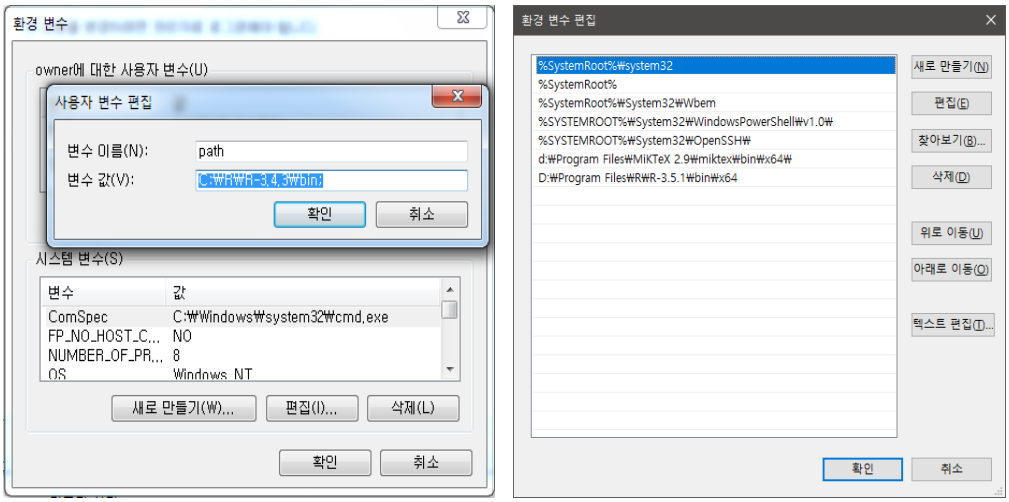


[그림 9-11] 시스템 속성

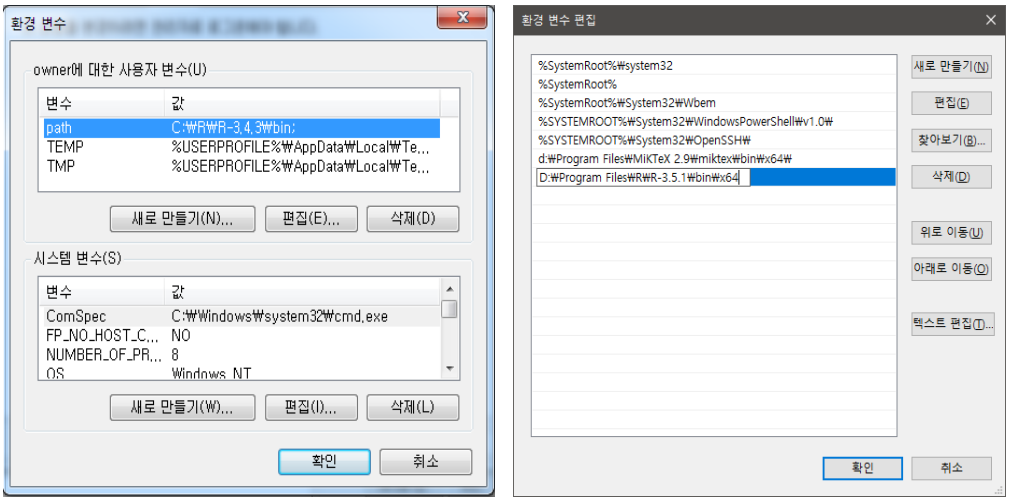
다음으로 [그림 9-12]와 같은 환경 변수 대화상자에서 [시스템 변수]에 나오는 변수들 중에서 [path]를 선택하고, [편집] 단추를 누른다. [그림 9-13]의 왼쪽 그림과 같은 상자에서 1단계에서 복사해둔 경로를 [변수 값]에 붙여 넣는다. 윈도우 버전에 따라 [그림 9-13]의 오른쪽과 같은 [환경 변수 편집] 대화상자에서 [새로 만들기] 단추를 누른 후 1단계에서 복사해둔 경로를 붙여 넣기도 한다. 결과로 출력되는 화면은 예를 들면, [그림 9-14]와 같이 보일 것이다. 이제 [확인] 단추를 눌러서 모든 설정화면을 빠져나간다.



[그림 9-12] 환경 변수 대화상자



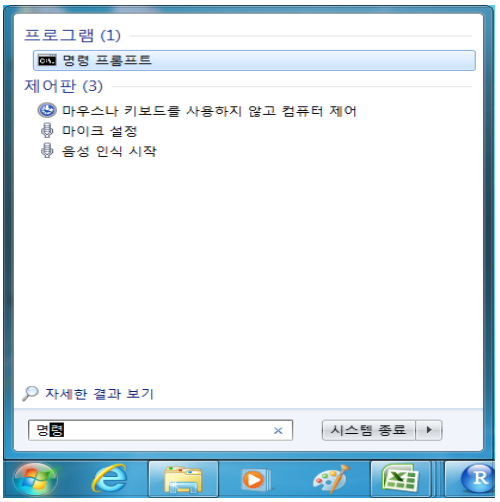
[그림 9-13] 환경 변수 편집 대화상자



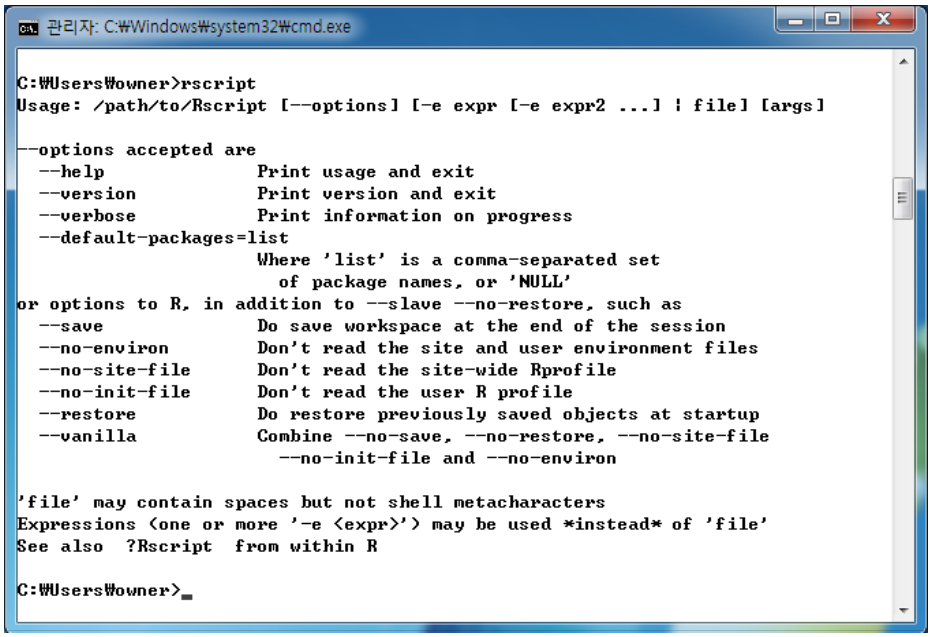
[그림 9-14] R의 path를 추가한 이후 화면

3단계:

R의 path가 잘 설정되었는지를 다음 과정을 통해 확인한다. [그림 9-15]와 같이 [명령 프롬프트]를 찾아 실행시킨다. [명령 프롬프트] 도스 화면에서 rscript를 타이핑하고 엔터키를 누른다. 이때 [그림 9-16]과 같은 화면이 나오면 1단계와 2단계를 잘 수행한 것이다. 참고로 보통 [그림 9-16]의 화면은 검은 색으로 나타나지만, 보고서 인쇄를 위해 흰 색으로 변경했다. 이제 엑셀과 R을 연동하기 위한 기본 설정을 마쳤다.



[그림 9-15] 명령 프롬프트의 찾기 및 실행



[그림 9-16] 명령 프롬프트 도스 화면에서 rscript를 실행한 화면

나. 엑셀 VBA 코드

엑셀에서 R을 통제하기 위해서는 엑셀 VBA를 사용한다. 아래의 <표 9-11>은 VBA 코드의 예를 보여주고 있다. 기본적인 방식은 엑셀에서 VBA.CreateObject("WScript.Shell")와 같은 shell 명령문을 통하여 path로 지정해 둔 경로에 있는 RScript가 특정 R 코드를 실행하도록 하는 것이다. 이를 수행하는 함수는 shell.Run이며, <표 9-11>에서 shell.Run의 3가지 인수는 다음과 같다.

path: 실행 파일명(RScript)과 실행 파일을 위한 인수(R 코딩된 파일명)이다.
style: 실행되는 과정을 명령 프롬프트 창으로 보여 줄 것인지 설정한다.
waitTillComplete: VBA가 실행이 끝날 때까지 기다릴 것인지 설정한다.

첫 번째 인수 'path'는 string 형식을 취하며, 여기서는 RScript, 공백, 분석을 코딩한 R파일의 경로로 구성된다. 두 번째 인수인 'style'은 보통 실행되는 과정이 나타나지 않도록 설정한다. 마지막 'waitTillComplete'는 실행이 끝날 때까지 기다리도록 설정할

필요가 있다. 그렇지 않으면 R이 실행되고 있는 잠시 동안 이용자가 엑셀 작업을 수행해서 오류를 일으킬 소지가 있다.

〈표 9-11〉 shell 명령어를 통한 RScript의 실행

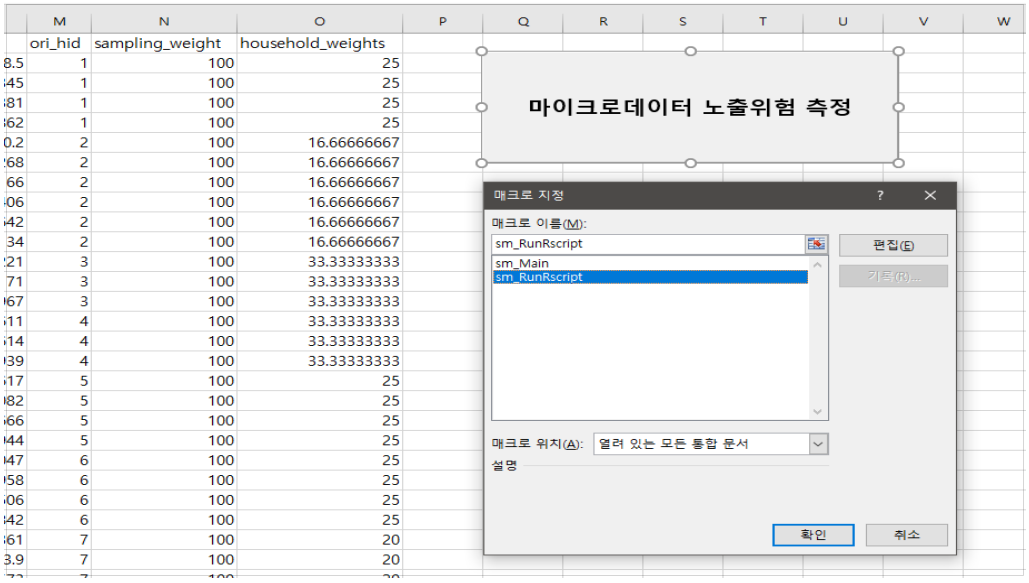
```
Sub sm_RunRscript()

    Dim shell As Object
    Set shell = VBA.CreateObject("WScript.Shell")
    Dim waitTillComplete As Boolean: waitTillComplete = True
    Dim style As Integer: style = 0
    Dim errorCode As Integer
    Dim path As String

    path = "RScript d:\test\analysis_code.R"
    errorCode = shell.Run(path, style, waitTillComplete)

End Sub
```

한편 위의 표에서 path 변수를 만들 때, 'RScript' 대신 RScript가 있는 폴더 경로까지 포함시킬 수도 있다. 예를 들면 아래와 같으며, 이렇게 처리하는 경우 앞 소절에서 설명한 path 구성을 생략할 수 있다.

```
path = "D:\Program Files\R\R-3.5.1\bin\x64\RScript d:\test\analysis_code.R"
```

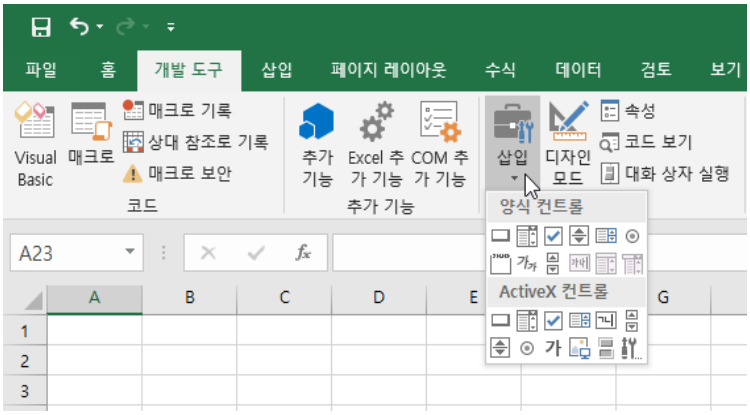
shell.Run의 수행이 성공적인 경우, errorCode 값은 0이 된다. RScript 다음에 오는 "analysis_code.R"은 연구자가 작성한 R 코드 파일로서 엑셀에서 송신된 정보를 받고 이를 연구자가 원하는 R 함수 및 코드로 계산하고 결과물로 만들어 낼 것이다. 본 가이드라인의 경우, 노출위험을 시나리오별로 계산한 결과물 또는 전체 재범주화를 수행해 변형된 변수 데이터 등을 반환하는 코드가 된다.

엑셀 VBA 코드를 편리하기 사용하기 위해서는 추가적인 작업들이 필요하다. 예를 들면, 위 표에 있는 sm_RunRscript() 매크로를 엑셀 시트와 연결해서 사용할 수 있다. 간단한 방법 중 하나로서 [그림 9-17]과 같이 명령단추를 이용하여 매크로와 시트를 연결할 수 있다.



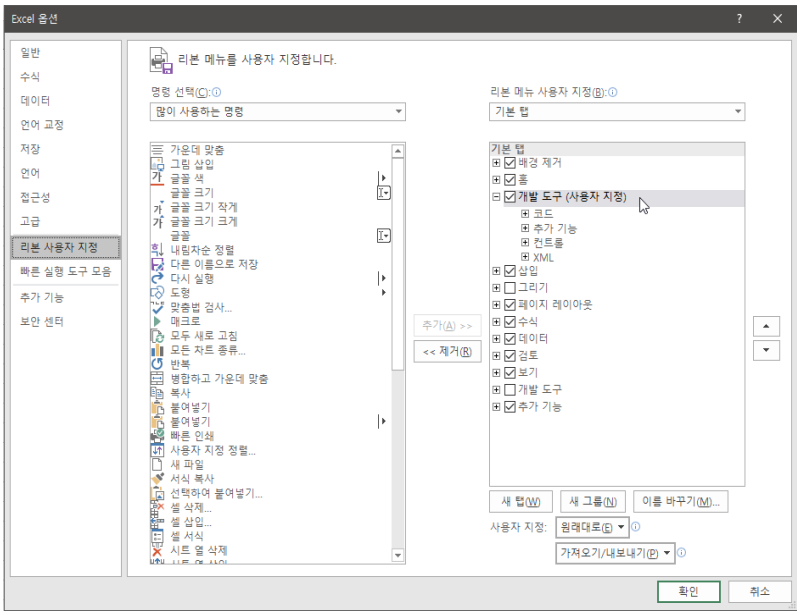
[그림 9-17] 명령단추와 매크로의 연결

참고로 명령단추는 [그림 9-18]과 같이 [개발 도구] 메뉴의 [삽입] 단추를 누르고, 첫 번째 아이콘( )을 클릭한 후 엑셀 시트 상에 드래그앤드롭(drag & drop)하여 만들 수 있다.



[그림 9-18] 명령단추 만들기

엑셀의 리본 메뉴에 [개발 도구]가 없을 경우, [파일]-[옵션]에 들어가서 [리본 사용자 지정]에서 [그림 9-19]와 같이 '개발 도구 (사용자 지정)'을 체크하여야 한다.



[그림 9-19] 개발 도구 메뉴 보이기

다. R 코드

엑셀에서 송신된 정보 또는 자료는 R에서 가공되어 엑셀로 보내질 것이다. 이를 위해 다음과 같은 2가지를 결정해야 한다.

(i) 엑셀 정보를 어떻게 가져올 것인가?
(ii) R 수행 결과를 어떻게 엑셀로 보낼 것인가?

실행 속도 및 편리성을 제고하기 위하여, 엑셀 정보는 클립보드를 통해서 가져오고 R 수행 결과는 html 형태로 출력할 수 있다. 텍스트 결과는 html에 직접 출력될 것이며, 그래프 같은 개체는 파일로 만든 이후에 출력될 것이다. 단, html 출력을 위해서 R 패키지 R2HTML을 이용하며, R의 data.frame이나 table 같은 경우 html에 그대로 출력하는 것은 비효율적이므로 csv파일 형태로 저장한 후 엑셀 시트로 열 수 있도록 한다.

[그림 9-20]은 엑셀과 통신하기 위한 R 코딩의 예이다. 1행부터 4행까지는 필요한 R 패키지를 불러온다. 특히, 1행과 2행에 관련 패키지가 설치되어 있지 않은 경우 설치를 수행하기 위한 R명령어가 있다. 3행과 4행에서는 설치되어 있는 패키지를 호출한다.

```
Source on Save
1  if (!require("sdcMicro")) install.packages("sdcMicro")
2  if (!require("R2HTML")) install.packages("R2HTML")
3  library(sdcMicro)
4  library(R2HTML)
5
6  aa=read.table("clipboard",sep="\t",header=TRUE)
7
8  setwd("g:/")
9  myfile <- "output.html"
10 HTML(as.title("분석 결과"),HR=1,file=myfile,append=FALSE)
11
12 HTML(as.title("요약 통계량"), HR=2,file=myfile,align="left")
13 res = summary(aa)
14 HTML(res, HR=2,file=myfile,align="left")
15
16 png('plot.png')
17 hist(aa$age)
18 dev.off()
19 HTMLInsertGraph(GraphFileName="plot.png",file=myfile)
20
21 browseURL(myfile)
22
```

[그림 9-20] 엑셀과 통신하기 위한 R 코딩 예

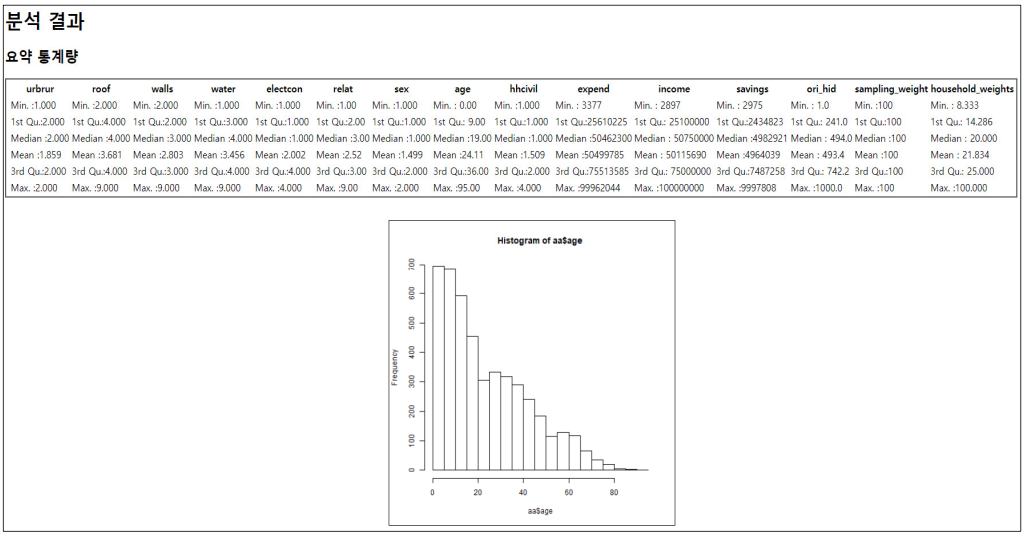
6행은 엑셀의 자료를 R로 가져오는 방식이다. 즉, 먼저 엑셀 시트의 데이터를 클립보드에 복사해 두었다고 가정할 때 R에서 이 클립보드 정보를 붙여넣기하는 것이다.

이와 같은 클립보드를 통한 통신 방식은 속도가 빠른 장점이 있다. 만약, 추가적으로 엑셀에서 다른 정보를 R에 보내야 한다면, RScript 명령문의 인수를 사용할 수 있다. 예를 들어, 추가 정보가 담긴 arg_array 변수를 인수로 해서 R로 보내야 할 경우 <표 9-11>의 path 변수는 다음과 같이 지정될 수 있다. 주의하여야 할 점은, 'analysis_code.R'과 큰따옴표(") 사이에 공백이 있다는 점이다.

```
path = "RScript d:\test\analysis_code.R " & arg_array
```

8행부터 14행은 R의 텍스트 결과를 html 파일에 출력하는 방법을 보여준다. 먼저, 8행과 9행은 html 파일을 만들 경로와 파일 이름을 설정하고, 10행 및 12행에서 HTML() 함수를 이용하여 제목을 출력한다. HR 인수는 출력될 내용의 레벨을 나타낸다.

16행에서 19행까지는 그래프를 html에 출력하는 방법을 보여준다. 16행에서 먼저 파일 이름을 설정하며, 17행에서 관련 그래프를 그리고 18행에서 파일을 만든 이후, 19행에서 HTMLInsertGraph() 명령어로 html에 그래프 파일을 붙여 넣는다. 21행에서는 만들어진 html 파일을 연다. 최종적으로 열리는 파일은 [그림 9-21]과 같다.



분석 결과

요약 통계량

urbrur	roof	walls	water	electcon	relat	sex	age	hhcivil	expend	income	savings	ori_hid	sampling_weight	household_weights
Min. :1.000	Min. :2.000	Min. :2.000	Min. :1.000	Min. :1.000	Min. :1.00	Min. :1.000	Min. : 0.00	Min. :1.000	Min. : 3377	Min. : 2897	Min. : 2975	Min. : 1.0	Min. :100	Min. : 8.333
1st Qu.:2.000	1st Qu.:4.000	1st Qu.:2.000	1st Qu.:3.000	1st Qu.:1.000	1st Qu.:2.00	1st Qu.:1.000	1st Qu.: 9.00	1st Qu.:1.000	1st Qu.:25610225	1st Qu.: 25100000	1st Qu.:2434823	1st Qu.: 241.0	1st Qu.:100	1st Qu.: 14.286
Median :2.000	Median :4.000	Median :3.000	Median :4.000	Median :1.000	Median :3.00	Median :1.000	Median :19.00	Median :1.000	Median :50462300	Median : 50750000	Median :4982921	Median : 494.0	Median :100	Median : 20.000
Mean :1.859	Mean :3.681	Mean :2.803	Mean :3.456	Mean :2.002	Mean :2.52	Mean :1.499	Mean :24.11	Mean :1.509	Mean :50499785	Mean : 50115690	Mean :4964039	Mean : 493.4	Mean :100	Mean : 21.834
3rd Qu.:2.000	3rd Qu.:4.000	3rd Qu.:3.000	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:3.00	3rd Qu.:2.000	3rd Qu.:36.00	3rd Qu.:2.000	3rd Qu.:75513585	3rd Qu.: 75000000	3rd Qu.:7487258	3rd Qu.: 742.2	3rd Qu.:100	3rd Qu.: 25.000
Max. :2.000	Max. :9.000	Max. :9.000	Max. :9.000	Max. :4.000	Max. :9.00	Max. :2.000	Max. :95.00	Max. :4.000	Max. :99962044	Max. :100000000	Max. :9997808	Max. :1000.0	Max. :100	Max. :100.000

[그림 9-21] R 수행 결과로 생성된 html 파일

본 가이드라인 프로그램에서는 HTML 방식을 사용하지 않고, R이 출력한 문서를 클립보드에 복사해서 엑셀에서 불러들인 다음 새로운 시트에 출력하는 방식을 취한다. 이 부분은 VBA로 직접 코딩했고 KESS 코드를 일부 이용했으며, 이에 관한 자세한 내용은 가이드라인 프로그램 내의 VBA 코드와 신봉섭(2017) 등을 참고하면 된다.

라. 가이드라인 프로그램에서 엑셀과 R의 연동 결과

본 연구에서 구축한 가이드라인 프로그램에서 엑셀과 R의 연동 결과를 살펴보자. 먼저 엑셀 VBA에서 R로 자료를 전송하는 부분은 다음과 같다. 엑셀 VBA 폼에서 입력받은 내용을 key1, key2, weight 라는 변수에 저장하고, R과 연동하는 함수인 RunRscript로 전달한다.

```
Sub RunRscript(key1, key2, weight)

    Dim shell As Object
    Set shell = VBA.CreateObject("WScript.Shell")
    Dim waitTillComplete As Boolean: waitTillComplete = True
    Dim style As Integer: style = 0
```

```
    Dim errorCode As Integer
    Dim arg_array As String
    Dim path As String

    On Error Resume Next

     ' 분석 시트에 있는 데이터 클립보드로 복사
    mainSht.Activate
    Range("A1").Cells(1, menuSpace + 1).Select
    Range(Selection, Selection.End(xlToRight)).Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.Copy

    ' R프로그램에서 \를 인식 못하기 때문에 / 로 바꿈
    folderpath = Replace(Application.ActiveWorkbook.path, "\", "/")

    arg_array = folderpath
    arg_array = arg_array & "#" & key1
    arg_array = arg_array & "#" & key2
    arg_array = arg_array & "#" & weight & "#"

    path = "RScript " & folderpath & "/DR_anal.R " & arg_array

    errorCode = shell.Run(path, style, waitTillComplete)

End Sub
```

R코드 일부는 다음과 같다. 필요한 패키지가 설치되어 있는지 확인하고, 없으면 설치한다. 클립보드로 전송받은 데이터를 'aa'에 저장하고, 엑셀 VBA에서 실행한 'shell.Run'을 통해 전달받은 'path'를 통해 R에서 분석할 때 사용할 정보를 추출한다. 가이드라인 프로그램에서는 각 정보를 구분하기 위해 #를 사용했다.

```
if (suppressMessages(suppressWarnings(!require("sdcMicro",
quietly=T,character.only=T,warn.conflicts=F))))
install.packages("sdcMicro", repos="https://cloud.r-project.org",
quiet=T)

aa = read.table("clipboard", sep="\t", header=TRUE)
arg = commandArgs(trailingOnly=T)
arg = strsplit(arg, "#")[[1]]

path = arg[1]
key1 = arg[2]; key1 = strsplit(key1, ",")[[1]]
key2 = arg[3]; key2 = strsplit(key2, ",")[[1]]
weight = arg[4]; weight = strsplit(weight, ",")[[1]]
```

R에서 필요한 분석을 마치고 분석결과를 아래와 같이 데이터 프레임 'rst'에 저장하고 클립보드를 이용해 엑셀로 전달한다.

```
write.table(rst, "clipboard", sep="\t", row.names=FALSE)
```

엑셀에서는 아래와 같이 간단하게 클립보드의 내용을 정해놓은 위치, 여기에서는 결과를 나타내는 시트 'rstSht'에 붙여 넣는다. 이제 결과시트에서 이용자에게 결과를 보여주기 위한 편집 작업을 VBA로 수행하면 된다.

```
ActiveSheet.Paste Destination:=rstSht.Cells(n1, 1)
```

지금까지 프로그램에 사용한 엑셀과 R을 연동하는 핵심 코드를 소개했다. 더 자세한 코드를 알기 위해서는 가이드라인 프로그램을 직접 참조할 수 있다.

제4절 결론

유럽연합의 일반개인정보보호법(GDPR)이 2018년 5월부터 시행되고 있다. 이는 법적인 측면에서도 자료제공 및 유통과 관련하여 개인정보보호에 힘써야 하는 상황이 도래했음을 의미한다. 그러나 기술적인 측면에서 비밀보호 방안은 여전히 연구[23] 중이다. 이러한 환경 속에서, 본 연구 과제는 지금 당장 시행할 수 있는 방안을 제시하기 위해 수행되었다. 다양한 마이크로데이터의 제공범위를 결정해야 하는 일반 업무담당자들을 대상으로, 제공하는 자료의 노출위험을 수치적으로 점검할 수 있도록 지원할 필요가 있기 때문이다.

다만 본 과제는 통계적 비밀보호 분야의 발전 상황과 맞물려 몇 가지 제약사항을 가진다. 먼저 본 가이드라인 프로그램에서 제시하는 노출위험은 범주형 변수 제공범위 변화에 따른 노출위험의 상대적인 비교를 목적으로만 사용할 수 있다. 프로그램 내부에서 사용하는 노출위험 측정 통계 모형 자체가 편의를 가지고 있기 때문에 노출위험 숫자를 절대적인 의미로 해석하는 것은 곤란하다. 다음으로 본 프로그램을 적용한 결과를 살펴보면 범주의 개수가 많은 변수가 노출위험이 높게 나오는 것을 확인할 수 있다. 이는 수치적으로 확인하지 않아도 짐작할 수 있는 결과이다. 다만 같은 범주 개수를 가지는 변수라도 자료의 분포에 따라 노출위험 정도가 달라질 수 있으므로, 이를 수치적으로 확인한다는 의의를 가진다. 즉, 데이터에 따라 단순히 범주 개수로 짐작하는 것과는 다른 노출위험 비교 결과가 나타날 수 있다. 가이드라인 프로그램은 향후 통계적 비밀보호 분야의 발전에 따라 업데이트 되어가는 것이 바람직하다.

본 연구의 활용도는 어쩌면 보고서의 두 번째 부분인 '프로그램 형식의 가이드라인 작성 방안'에 더 크게 있을 수 있다. 잘 알려지지 않은 엑셀과 R을 직접 연결하는 기술을 제시했고, 이를 구현했기 때문이다. 일반적인 업무 담당자들이 특정 주제에 대한 이론적인 내용까지 숙지하기에는 업무의 종류가 다양한 경우가 많다. 향후 문서로 작성된 복잡한 가이드라인을 대신해서, 손쉽게 따를 수 있도록 프로그램 형식의 가이드라인을 만드는 데 본 보고서가 도움이 되기를 바란다.

23) 대표적인 연구 주제로 최근 재현자료(synthetic data) 및 차등정보보호(differential privacy)가 활발히 연구되고 있다. 통계적 비밀보호를 위해 새로운 기술이 완성되고, 활용되기까지는 시간이 필요할 것으로 보인다.

참고문헌

관계부처 합동(2016),「개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내 -」. 국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부.

박민정, 김경미(2013), “종단자료 비밀보호의 국제 연구동향 및 향후 추진방향”, 통계개발원, 내부 보고서.

박민정, 권순필, 심규호(2013), “가계금융·복지조사 마이크로데이터 매스킹 방안”, 통계개발원.

박민정(2014), “가계금융·복지조사 마이크로데이터 매스킹 방안 평가”, 통계개발원.

박민정(2015), “마이크로데이터 매스킹 기법 실무 활용 안내서”, 통계개발원.

박민정(2016), “통계적 노출제어 최신동향 검토 -차등정보보호를 중심으로-”, 통계개발원.

박민정, 김정연(2017), “재현자료 작성 방법론 검토”, 통계개발원.

박민정, 김항준(2016), “마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰”, 응용통계연구, 39(6), pp. 1041-1059.

박민정, 이용희, 권성훈(2018), “차등정보보호에 관한 연구”, 통계개발원.

신봉섭(2017),「VBA로 통계분석시스템 만들기」, 그린.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P. P.(2012), Statistical Disclosure Control, John Wiley & Sons Ltd.

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

Statistics Netherlands(2008), μ-Argus User’s Manual ver. 4.2.

<부 록>

1. R과 R패키지 sdcMicro가 사용하는 객체

R을 사용하기 위해 알아야 하는 내용은 너무나 방대하다. R과 관련된 많은 문헌이 있으므로, 여기서는 엑셀과 연동된 프로그램 작성과 관련해 몇 종류의 데이터 구조에 관해서만 간단히 설명하도록 한다.

먼저 R이 다루는 데이터 유형에는 논리형(logical), 숫자형(numeric), 문자형(character), 요인형(factor) 등이 있다. 보통 엑셀에서 R로 직접 보내는 인수는 문자형을 취한다.

다음으로 R에서의 데이터 구조(data structures)를 정리해 보자. 가장 흔하게 사용되는 구조는 값(value), 벡터(vector), 행렬(matrix), 데이터 프레임(data.frame), 리스트(list)가 있다.

값, 벡터, 행렬 데이터 구조에 담은 데이터는 모두 동일한 유형이어야 한다. 즉 어느 행렬에 숫자형과 문자형 데이터를 함께 담을 수 없으며, 그렇게 담으면 R은 이들을 모두 문자형으로 인식한다. 반면 데이터 프레임은 행렬과 동일한 모양으로 보이지만, 다양한 유형의 데이터를 담을 수 있다.[1] 아래 그림은 벡터, 행렬, 데이터 프레임에 대한 예제를 보여 준다. 그림에서 변수 aa 및 bb는 벡터, cc는 행렬, dd는 데이터 프레임이며, is.은 괄호 안의 변수가 어떤 데이터 유형 또는 데이터 구조인지 알려주는 명령문이다.

```
> aa <- c('a', 'b', 'c')
> bb <- c(1, 2, 3)
>
> is.numeric(aa)
[1] FALSE
> is.numeric(bb)
[1] TRUE
> is.character(aa)
[1] TRUE
> is.character(bb)
[1] FALSE
```

```
> cc <- cbind(aa, bb)
> is.matrix(cc)
[1] TRUE
> is.numeric(cc)
[1] FALSE
> is.character(cc)
[1] TRUE
>
>
>
>
>
```

```
> dd <- data.frame(aa, bb)
> is.data.frame(dd)
[1] TRUE
> is.factor(dd$aa)
[1] TRUE
> is.numeric(dd$bb)
[1] TRUE
>
>
>
>
>
```

[부그림 9-1] R의 벡터, 행렬, 데이터 프레임에 대한 예제

리스트도 데이터 프레임과 마찬가지로 다양한 데이터 구조들을 하나의 객체로 저장할 수 있다. 차이점은 리스트를 사용하면 길이가 다른 다양한 데이터 구조들의 데이터를 한 번에 저장할 수 있다는 것이다. 아래 그림은 변수 ee가 문자형 벡터, 숫자형 벡터, 논리형 값,

1) 데이터 프레임이 여러 유형의 데이터를 입력받으면, 그 중 문자형 데이터 성분을 보통 요인형(factor)으로 처리한다. (Hadley Wickham (2015) Advanced R. CRC Press.)

문자형과 숫자형이 혼합된 데이터 프레임, 총 4가지 종류의 데이터 구조를 포함하는 리스트 변수임을 보여준다.

```
> aa <- c('a', 'b', 'c')
> bb <- c(1, 2, 3)
> cc <- TRUE
> dd <- data.frame(aa, bb)
> ee <- list(aa=aa, bb=bb, cc=cc, dd=dd)
>
>
>
>
>
>
>
>
>
```

```
> ee
$aa
[1] "a" "b" "c"

$bb
[1] 1 2 3

$cc
[1] TRUE

$dd
  aa bb
1  a  1
2  b  2
3  c  3
```

[부그림 9-2] R에서 리스트의 사용

지금까지 R에서 사용하는 다양한 유형의 데이터 형식과 데이터 구조들을 간단하게 살펴보았다. 각종 R 입문서에도 나와 있는 기본적인 내용들을 간단히 정리한 이유는 R패키지 sdcMicro가 새로운 데이터 구조를 만들어 사용하기 때문이다. 매스킹 기법을 적용할 때마다 변화되는 노출위험과 자료 유용성을 편리하게 확인하기 위해, sdcMicro는 sdcMicroObj-class라는 이름의 새로운 데이터 구조[2]를 정의했다. 본 보고서에서는 이를 sdc object로 줄여서 표현하겠다. sdc object는 @로 표현되는 다양한 slot을 가진다. 이는 list가 $로 표현되는 다양한 요소들을 가지는 것과 마찬가지다. list와 마찬가지로 sdc object는 @아래 다양한 성분을 담을 수 있다.

노출위험 측정 가이드라인 프로그램은 엑셀이 전송한 노출위험 분석 대상 마이크로데이터를 sdc object를 만드는 함수에 입력값으로 사용한다. 또한 sdc object의 출력물인 노출위험이 계산된 결과를 받아서 엑셀에 전달한다. 따라서 R의 입력 형식 및 출력 형식인 sdc object라는 데이터 구조를 이해할 필요가 있다. 엑셀이 전달하는 정보를 R에 맞는 형식으로 바꾸고, R이 생성한 결과물을 담은 데이터 구조 형식을 이해해서 엑셀에 적절히 전달해야 하기 때문이다. 엑셀이 전달한 마이크로데이터를 aa에 저장하고, sdc object를 다음과 같이 생성해서 aa.sdc라고 하자. 노출위험 측정 결과는 @로 표현되는 성분, originalRisk에 저장된다. 이 결과를 적절히 편집하여 엑셀에서 보여주는 것이 가이드라인 프로그램의 주요 역할이다.

2) S4 object에 속한다.

```
aa.sdc <- createSdcObj(aa, keyVars=key, weightVar = weight)
aa.sdc@originalRisk$individual
```

이로써 엑셀과 R의 통신에 대한 이해를 돕기 위한 데이터 구조에 관한 설명을 마친다. R패키지 sdcMicro의 기능에 관한 자세한 설명은 박민정(2015)을 참고하도록 한다.

2. 엑셀에서 R과 통신하기 : 외부 프로그램 활용

R에는 엑셀의 스프레드 시트를 불러오는 다양한 기본 명령어들과 패키지들이 있다. 예를 들면, R패키지 xlsx는 myfile이라는 엑셀 파일에서 Sheet1을 읽어서, 시트에 있는 자료를 data.frame 형식으로 aa에 반환할 수 있다. 다음은 이에 관한 R에서 사용하는 명령문이다.

```
aa <- read.xlsx2("myfile.xlsx", sheetName = "Sheet1")
```

그러나 이러한 방법은 R이 엑셀에 있는 자료를 불러오는 방법으로 본 연구의 관심사가 아니다. 본 연구에서는 엑셀을 주체(front end)로 해서 R을 실행하고 그 결과를 엑셀로 가져오는 방법을 다룬다. 사용자는 외관상으로 R이 실행되는 것을 알 수 없다. 이렇게 엑셀에서 R과 통신하는 방법을 구현하기 위해, 엑셀과 R이 아닌 다른 외부 프로그램들을 사용할 수도 있고, 엑셀에서 R과 직접 통신하도록 프로그램을 구축할 수도 있다.

엑셀과 R을 연결하는 외부 프로그램들로, R 패키지 excel.link 및 BERT 프로그램이 있다. R패키지나 BERT와 같은 외부 프로그램을 사용하면 좀 더 간단하게 엑셀과 R의 통신을 구현할 수도 있으나, 여러 측면에서 제약이 있다. R패키지 excel.link는 현재 베타 버전이며, 속도 및 안정성 면에서 부족한 측면이 있다. BERT 같은 프로그램은 유료화될 가능성을 예측하기 어렵고, 저작권 문제를 해결해야 한다. 반면 R과 엑셀, 두 프로그램의 통신을 직접 구현하는 것은 향후 문제의 소지가 없다.

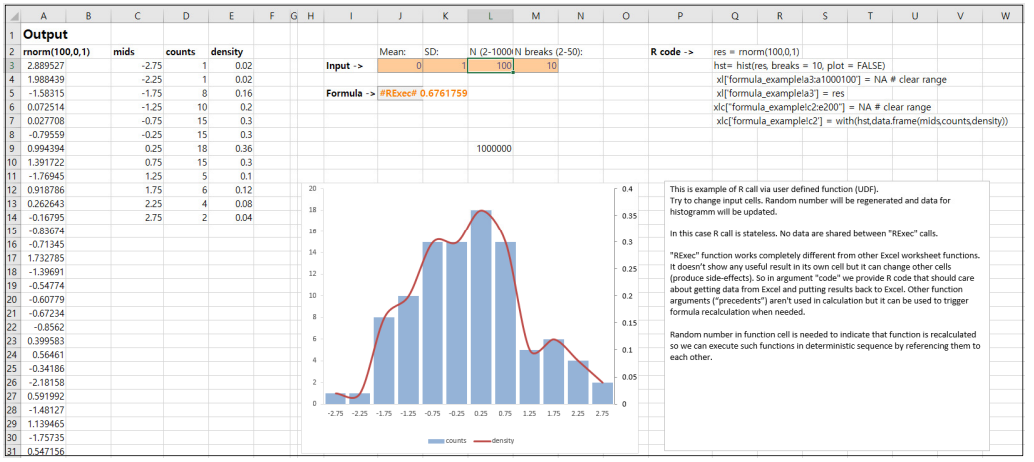
가) R 패키지 excel.link

엑셀에서 R을 호출하고 통제하기 위한 R패키지 excel.link[3]는 2018년 9월 현재 버전 0.9.8-1까지 만들어져 있다. 엑셀 시트의 셀 값들 및 시트 상에 직접 코딩된 R 코드를 R에 보낼 수 있다. 패키지 excel.link를 설치하면 관련 함수를 엑셀 VBA에서 통제함으로써 R을 통제한다. R로 엑셀의 셀 값들을 송신할 수 있을 뿐만 아니라, R로부터 그래픽스를 포함한

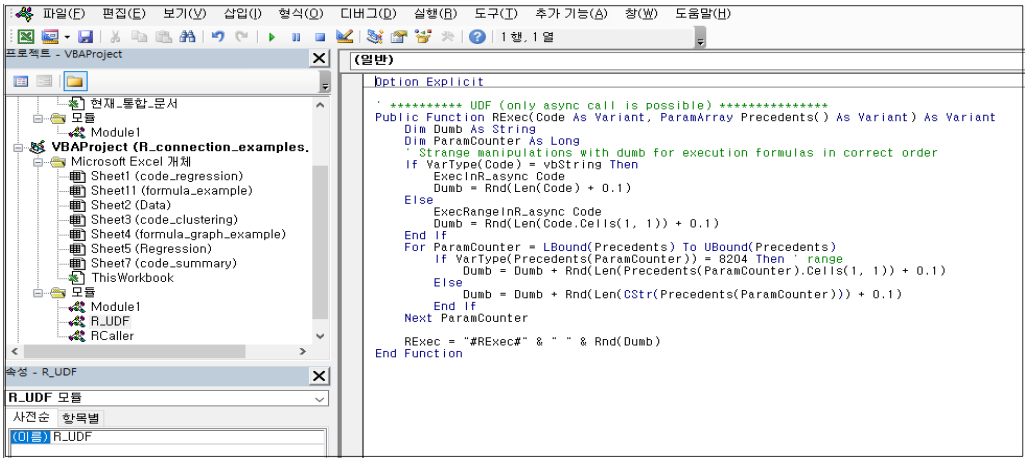
3) https://cran.r-project.org/web/packages/excel.link

실행 결과를 수신할 수 있는 기능이 잘 갖추어져 있다. 다만, 셀의 개수가 많아지면 R과의 통신 속도가 아주 느려져 사용이 불편하다. 아직 개발 중에 있는 R패키지로 판단되며, 향후 개발이 더 진전된다면 사용해 볼 만한 가치가 있다고 보인다.

R패키지 excel.link를 사용하거나 배우기 위해서는 이 패키지의 'doc' 폴더에 제공되는 'R_connection_examples.xlsm' 파일의 예제들을 이해하면 된다. 이 예제들은 몇 가지 경우에 있어서 excel.link를 어떻게 사용하는지를 잘 보여주고 있다. 먼저 시트의 구성과 이를 이용 및 통제하는 엑셀 VBA 구조를 이해해야 한다. 아래의 [부그림 9-3]은 그 예제 중 하나인 시트의 모양을 보여주며, [부그림 9-4]는 관련 엑셀 VBA 코드의 일부를 보여준다.



[부그림 9-3] R_connection_examples.xlsm의 예제 시트

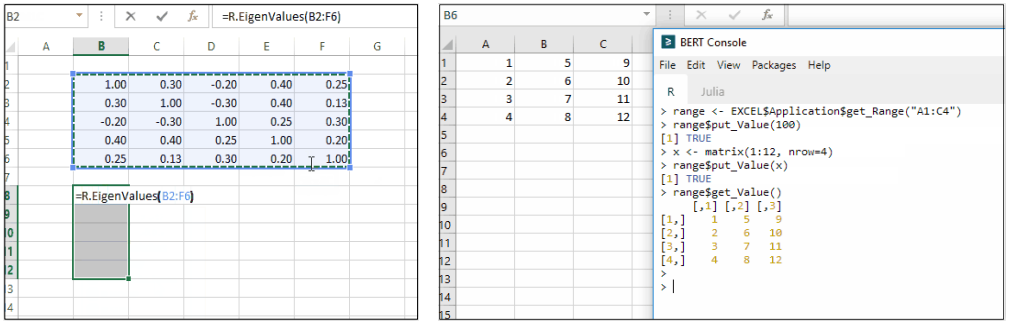


[부그림 9-4] R_connection_examples.xlsm의 VBA 코드

나) BERT 프로그램

BERT[4]는 엑셀과 R의 통신을 도와주는 독립된 프로그램이다. 현재 무료로 사용할 수 있다. 독립된 프로그램인 만큼 excel.link와 같은 R패키지보다 훨씬 사용이 편리하고 쉽다. 이미 제공되어 있는 R함수들을 엑셀에서 사용할 수 있을 뿐만 아니라 직접 코딩한 새로운 R함수들도 엑셀에서 호출할 수 있다.

BERT의 관련 예제 및 사용 방법들은 웹사이트에 자세하게 설명되어 있다. 예를 들면, 아래 [부그림 9-5]에서처럼 R의 EigenValues() 함수를 엑셀 시트 위의 셀 안에서 다른 엑셀 함수를 사용하듯이 직접 사용할 수 있다. 또한 BERT의 console 창을 통해 R에서 엑셀을 직접 제어할 수도 있다.



[부그림 9-5] BERT를 이용한 엑셀과 R의 통신

또한, 아래 [부그림 9-6]과 같이 엑셀 VBA 함수 Application.Run를 사용하면 R에서 정의한 함수 Add를 실행할 수 있다. BERT.Exec를 사용하면 입력하는 R 코드(예. “plot(sort(rnorm(1000)))”)를 실행할 수도 있다.

```
Sub test()

   x = Application.Run("R.Add", 1, 2, 3, 4)
   MsgBox x

End Sub
```

4) https://bert-toolkit.com/

```
Sub SamplePlot()

   Application.Run “BERT.Exec”, “plot( sort(rnorm(1000)))”

End Sub
```

[부그림 9-6] 엑셀 VBA에서 R 코드의 실행

◈ 집필진

• 박민정(통계청 사무관)
• 김경미(통계청 사무관)
• 심규호(통계청 사무관)
• 정동명(통계청 과장)
• 정남수(통계청 과장)
• 정미옥(통계청 사무관)
• 한승훈(통계청 사무관)
• 최유성(통계청 주무관)
• 이의규(통계청 서기관)
• 권순필(통계청 사무관)
• 성병찬(중앙대학교 교수)

개인정보보호와 데이터 활용을 위한 통계적 방법론(Ⅰ)
– 마이크로데이터에 적용한 비밀보호 기법: 이론과 활용 –

■ 발행연월일 | 2020년 4월 21일 인쇄
2020년 4월 22일 발행
■ 발 행 인 | 통계개발원장 전 영 일
■ 발 행 처 | 통계청 통계개발원
35220 대전광역시 서구 한밭대로 713
(월평동 282-1) 통계센터 6~8F
☎ 대표 (042) 366-7100 Fax (042) 366-7123
■ 인 쇄 처 | 강문인쇄사 (042) 226-4722~3