제3장

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

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 노출위험을 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, 노출위험으로 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 3-1>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 3-1〉 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 3-2>에서는 최종 제공하는 자료 형식에 따라 노출제어 방법들을 분류하였다.

〈표 3-2〉 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 *t*분포(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 3-3>과 같으며 이는 보도자료 형태로 공표된 2012년 집계자료의 민감변수와 동일한 것들이다. 각 상위변수는 하위변수들의 합으로 구성되므로 (예. 자산총액 = 실물자산 + 금융자산) 본 연구에서는 <표 3-3>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 3-3〉 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

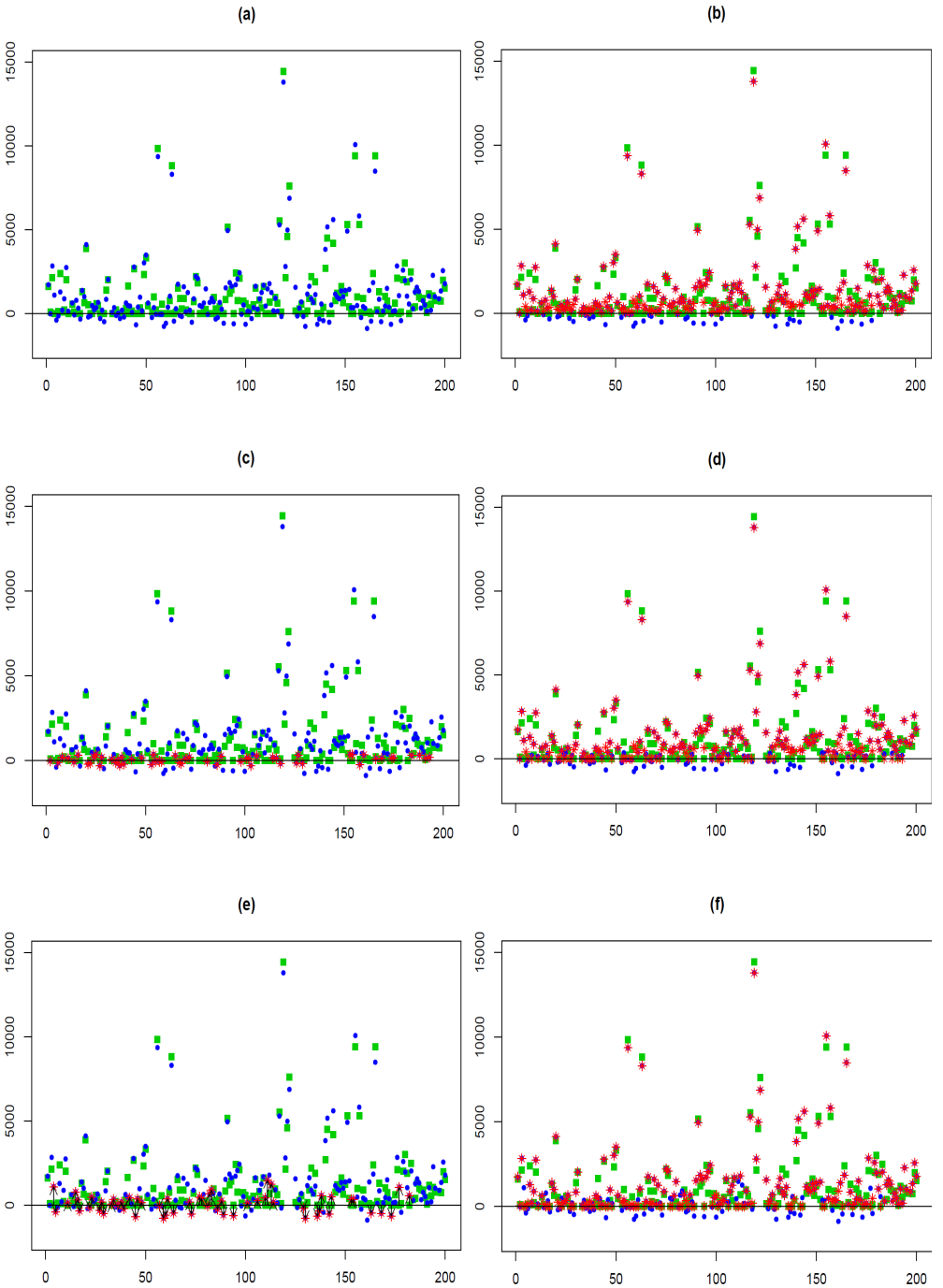
의해 음의 값의 소득을 가질 수 있게 된다. 따라서 매스킹 기법별로 노출위험과 정보손실을 측정하기 이전에 음의 값 발생 문제를 어떻게 처리할지 고민해야 한다. 본 연구에서는 음의 값 처리에 대해 다음의 세 가지 시나리오를 고려하였다.

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 3-1]은 이러한 세 가지 시나리오를 그림으로 표현한 것이다. 이해를 돕기 위해 한 변수에 대해서 시나리오별 매스킹 결과를 200개의 표본[9]을 이용해 나타내었다. [그림 3-1]의 (a)에서는 원래 마이크로데이터를 초록색 네모로, 잡음을 추가하여 매스킹한 결과는 파란색 원으로 표현하였다. [그림 3-1]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 3-1]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 3-1] 잡음 추가 기법 적용과 무응답 정보 처리에 대한 세 가지 시나리오

두 번째 시나리오는 [그림 3-1]의 (c)와 (d)에 나타나있다. 원래 무응답의 개수를 k개라고 하면, 잡음 추가 후에 절대값이 작은 순서대로 k개의 자료가 별 모양으로 [그림 3-1]의 (c)에 그려져 있다. 이들을 모두 0으로 처리하고, 나머지 자료에 대해서 매스킹 처리 후의 절대값을 별 모양으로 표현한 것이 [그림 3-1]의 (d)이다. 즉, [그림 3-1]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 3-1]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 3-1]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 3-4>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 3-4〉 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 3-5>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 3-5〉 세부 알고리즘별 노출위험과 정보손실 측도 결과

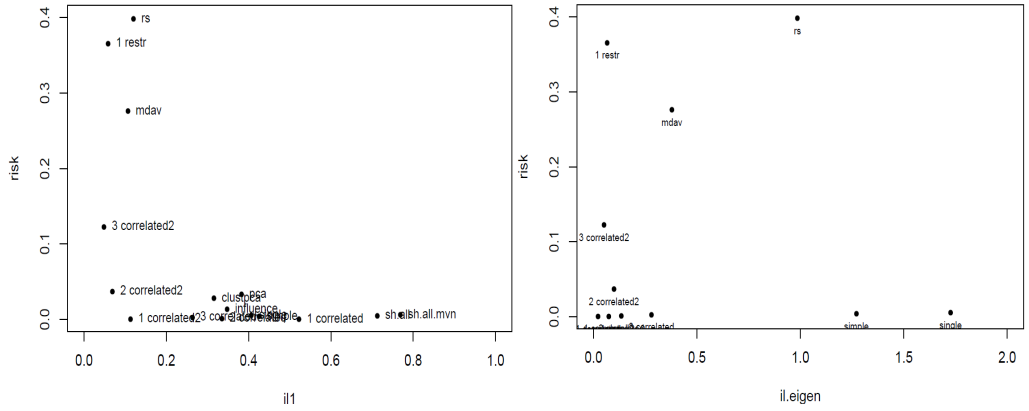
매스킹 기법	세부 알고리즘	노출위험 (risk)	정보손실1 (il1)	정보손실2 (il.eigen)
국소통합 (microaggregation)	single	**0.01**	0.41	1.73
	simple	**0.00**	0.43	1.27
	onedims	0.99	**0.00**	0.12
	mdav	**0.28**	**0.11**	0.38
	influence	**0.01**	0.35	2.18
	pca	**0.03**	0.38	3.75
	clustpca	**0.03**	0.32	3.01
잡음 추가1 (noise addition1)	1 additive	**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 3-5>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 3-2]는 위의 <표 3-5>의 결과에 대한 위험-유용성 지도에 해당된다. 한편, 유용성은 정보손실에 의해서 측정이 되므로 이제 위험-유용성 지도를 위험-정보손실 지도라고 부르도록 하겠다. [그림 3-2]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 3-2] 세부 기법별 노출위험-정보손실 지도

참고로 노출위험과 정보손실의 개념 이해에 도움을 주고자 <부록>의 <부표 3-1>과 <부표 3-2>를 통해 특정 표본에 대하여[12] 각 민감변수별로 매스킹 세부 기법별 평균과 표준편차를 나타내었다. <부록>의 이 표들을 살펴보면 <표 3-5>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 3-2]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.

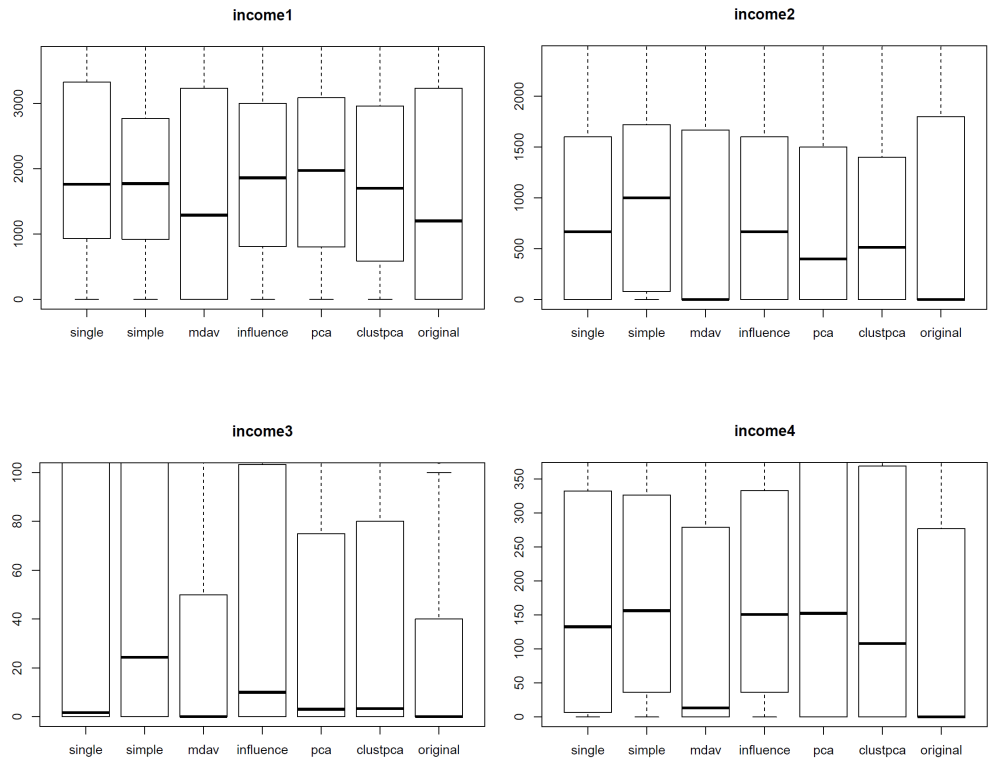
2. 국소통합 기법 적용 결과

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 3-3]과 같이 나타난다. 원래 자료에 대한 상자그림은 각 그림들 오른쪽 끝에 original로 표현되어 있다. 자산 및 부채 변수에 대한 상자그림은 <부록>의 [부그림 3-1] 및 [부그림 3-3]에 수록하였다. 노출위험-정보손실 측도에서도 mdav[16] 기법이 가장 좋은 결과를 보였었는데

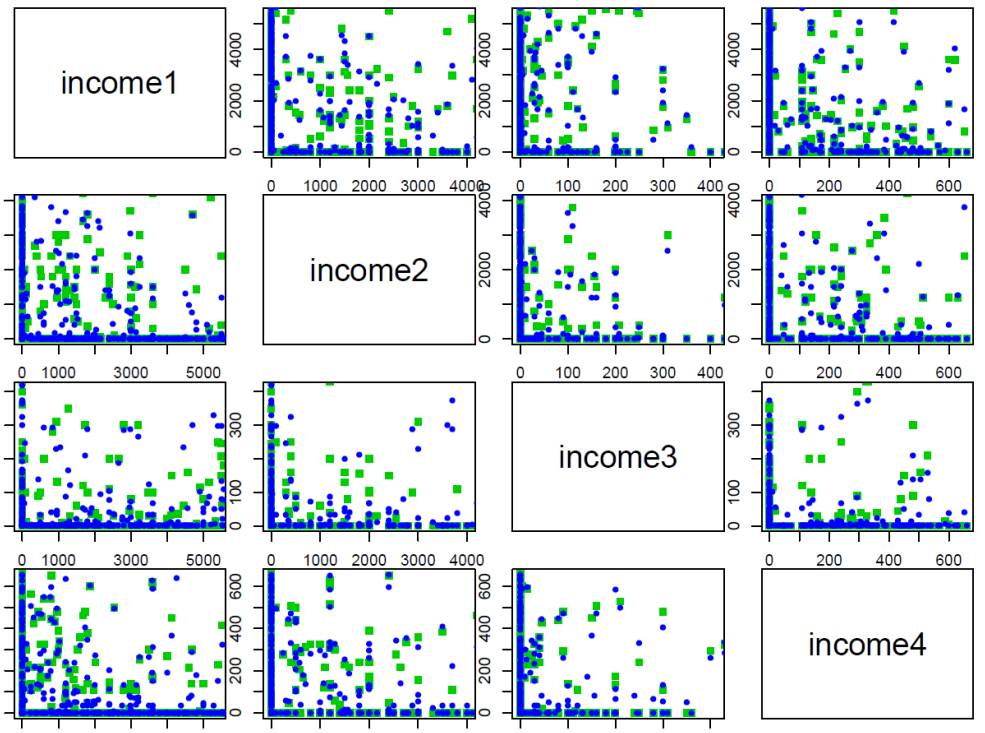
13) 세부 기법 single, simple, influence, pca, clustpca, additive, correlated, sh, sh.mvn 등
14) 세부 기법 onedims 등
15) 각종 상자그림, 산점도 등은 부록의 <부표 3-1> 및 <부표 3-2>의 평균과 표준편차를 구한 동일한 표본 S를 사용하였으나, 극단 이상점들로 인해 부록 표의 수치들과 큰 차이가 있어 보임을 참고하기 바란다.
16) 국소통합 세부 기법 mdav는 자료 공간에서 가장 멀리 있는 것들부터 그룹화해가는 알고리즘(Statistics Netherlands, 2007)을 나타낸다.

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



[그림 3-3] 국소통합 세부 기법별 소득 변수들의 상자그림

참고로 mdav에 의한 매스킹 결과를 보기 위해 하위변수별 산점도를 비교하면 [그림 3-4]와 같다. [그림 3-4]는 소득변수에 대해 표본 S를 이용한 결과를 보인 것이다. 자산 및 부채 변수에 대한 하위변수별 산점도는 <부록>의 [부그림 3-2] 및 [부그림 3-4]에 수록하였다. 소득 하위변수들 사이의 상관계수를 구해보면 <표 3-6>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 3-4] 국소통합(mdav) 기법 적용 전후 소득 변수들의 산점도

〈표 3-6〉 국소통합(mdav) 기법 적용 전후 소득 변수들의 상관계수

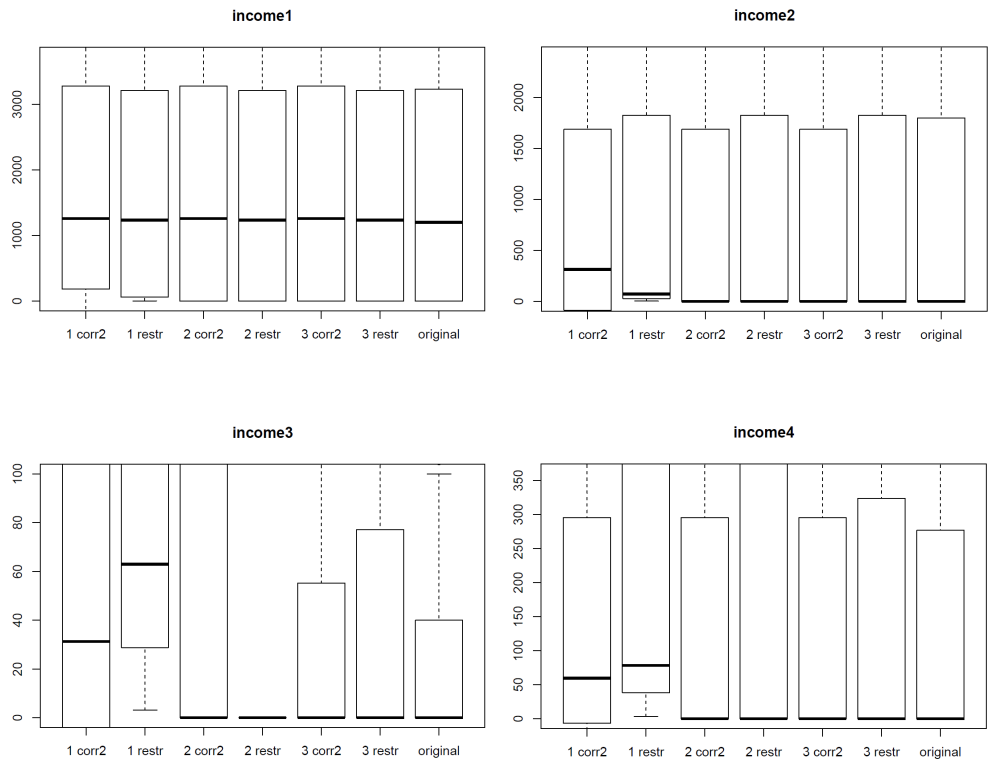
	original				국소통합(mdav)			
	income1	income2	income3	income4	income1	income2	income3	income4
income1	1.00	-0.23	0.02	-0.21	1.00	-0.22	0.05	-0.20
income2	-0.23	1.00	0.43	-0.10	-0.22	1.00	0.19	-0.12
income3	0.02	0.43	1.00	0.04	0.05	0.19	1.00	0.02
income4	-0.21	-0.10	0.04	1.00	-0.20	-0.12	0.02	1.00

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

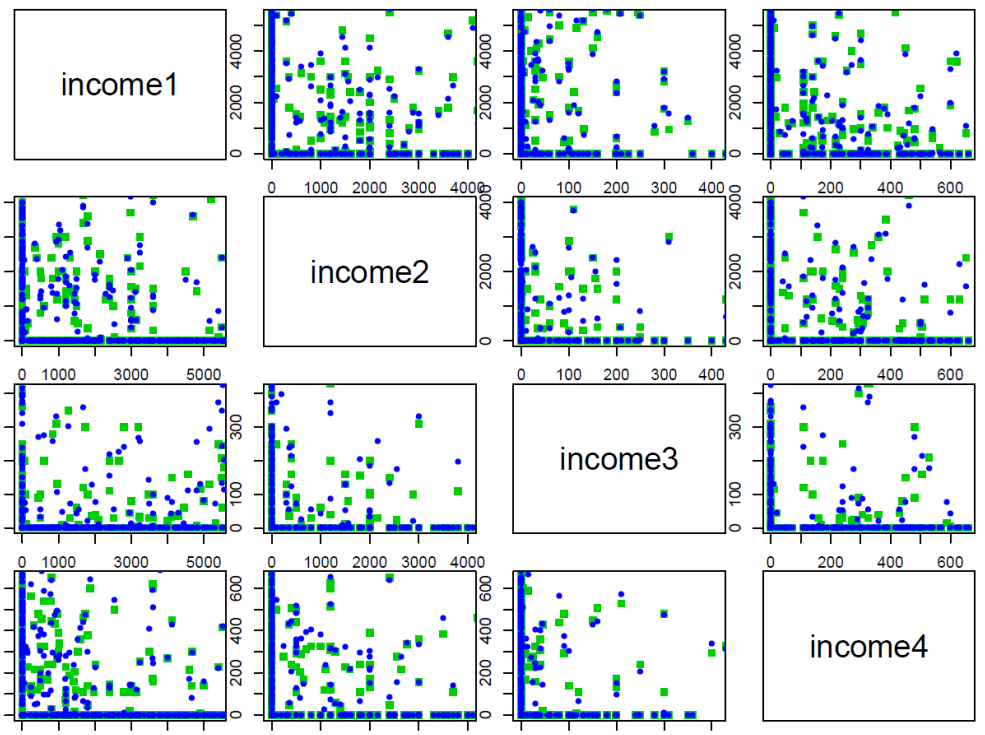
17) 잡음 추가 세부 기법 correlated2는 자료의 공분산 구조를 고려하는 상관 잡음 기법의 로버스트한 알고리즘(Templ과 Meindl, 2008)을, 세부 기법 restr은 자료의 크기를 고려하는 알고리즘을 나타낸다.

restr에 대해 상자그림을 그려보면 소득 변수에 대하여 [그림 3-5]와 같이 나타난다. 자산 및 부채 변수에 대한 상자그림은 <부록>의 [부그림 3-5] 및 [부그림 3-7]에 수록하였다. 노출위험은 다소 증가하나 앞의 정보손실 측도에서도 시나리오 3의 correlated2 기법([그림 3-5]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 3-5] 잡음 추가 시나리오별 소득 변수들의 상자그림

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



[그림 3-6] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 소득 변수들의 산점도

〈표 3-7〉 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 3-8>과 같다. <표 3-8>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 3-8>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 3-8>의 결합안 1, 2, 3의 결과를 얻게 된다. 시나리오별로 잡음 추가에 비해 노출위험 측도는 낮아지고 정보손실 측도는 높아졌음을 알 수 있다.

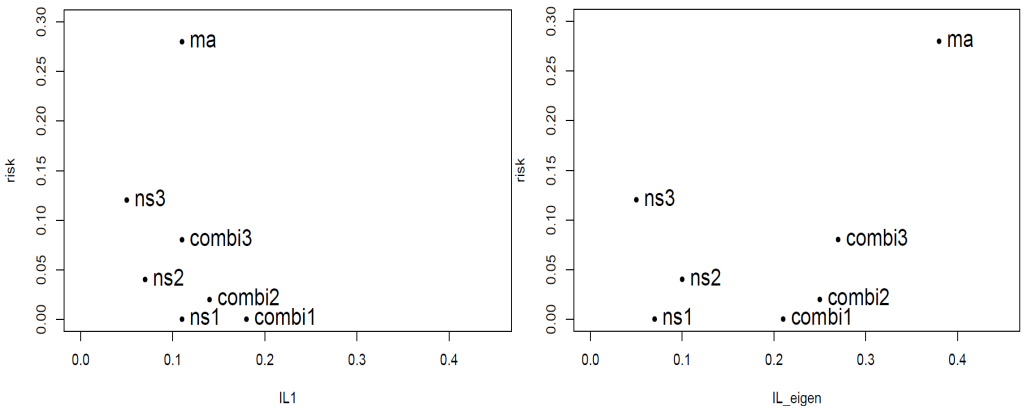
〈표 3-8〉 국소통합, 잡음 추가 및 결합안의 노출위험과 정보손실 측도 결과

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 3-7]은 <표 3-8>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

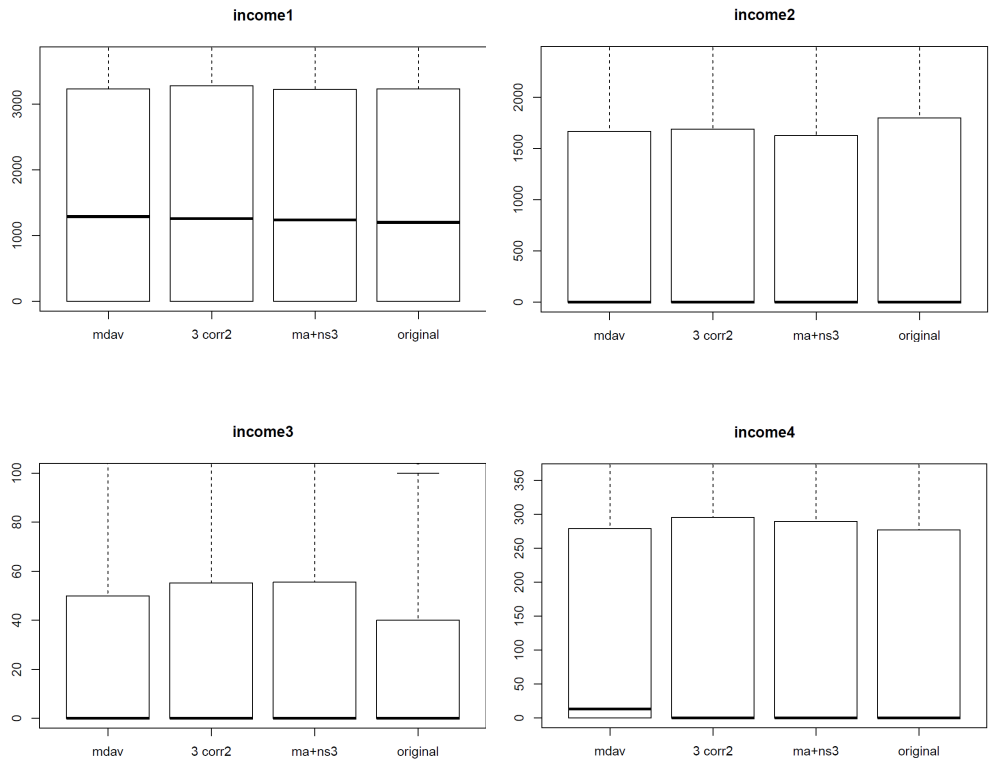
절대적으로 원점에 가깝게 위치해 있다. 그러나 시나리오별로 잡음 추가 기법과 결합안은 각각 보이지 않는 효용이 동등한 곡선 위에 있을 수 있다. 가령 ns2와 combi2 혹은 ns3과 combi3이 각각 노출위험-정보손실 측면에서 동등할 수 있다는 것이다. 따라서 각 시나리오별로 잡음 추가 기법과 결합안은 어느 것이 더 효율적일지 추가적인 엄밀한 검토가 필요하다고 할 수 있다.

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.



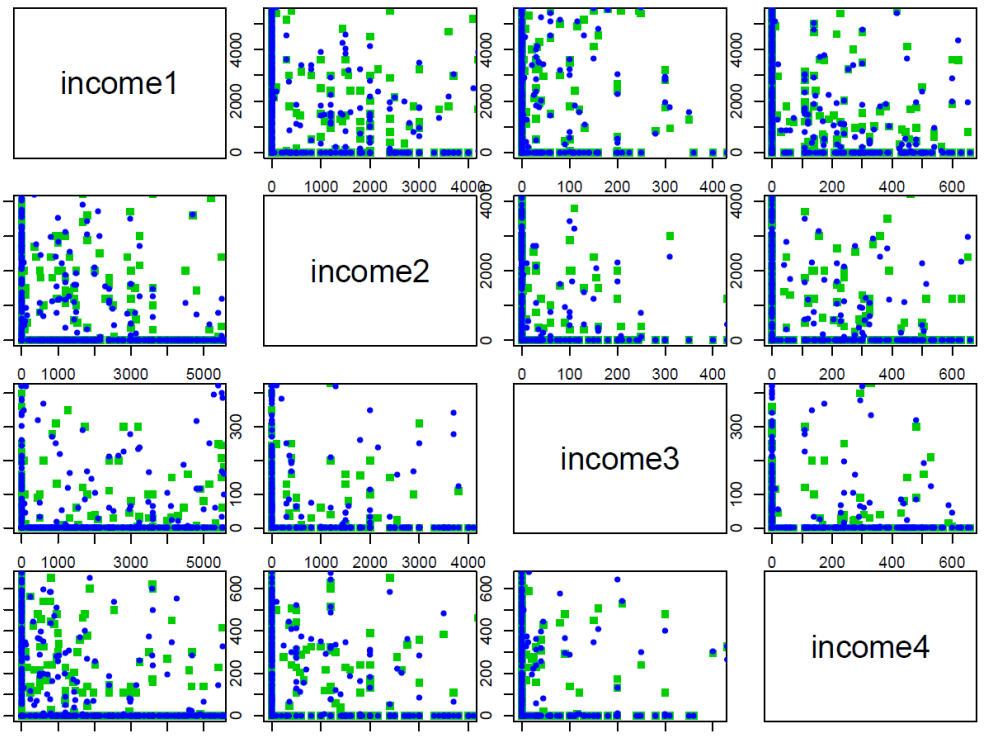
[그림 3-7] 국소통합, 잡음 추가 및 결합안의 노출위험-정보손실 지도

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 3-8]과 같이 나타난다. 자산 및 부채 변수에 대한 상자그림은 <부록>의 [부그림 3-9] 및 [부그림 3-11]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 3-8] 국소통합, 잡음 추가3 및 결합안3에 대한 소득 변수들의 상자그림

시나리오 3의 결합안 의한 매스킹 결과를 보기 위해 하위변수별 산점도를 비교하면 [그림 3-9]와 같다. [그림 3-9]는 소득변수에 대해 표본 S를 이용한 결과를 보인 것이다. 자산 및 부채 변수에 대한 하위변수별 산점도는 <부록>의 [부그림 3-10] 및 [부그림 3-12]에 수록하였다. 소득 하위변수들 사이의 상관계수를 구해보면 <표 3-9>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 3-9] 결합안3 적용 전후 소득 변수들의 산점도

〈표 3-9〉 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 3-10>과 같다. 빈도수가 적은 구간이 발생하지 않도록 재코딩 기준을 결정하였다.

〈표 3-10〉 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 3-11>은 공표 범위에

따른 노출위험 측도 결과를 나타내고 있다. <표 3-11>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 3-11>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 3-11>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 3-11> 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 3-11〉 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

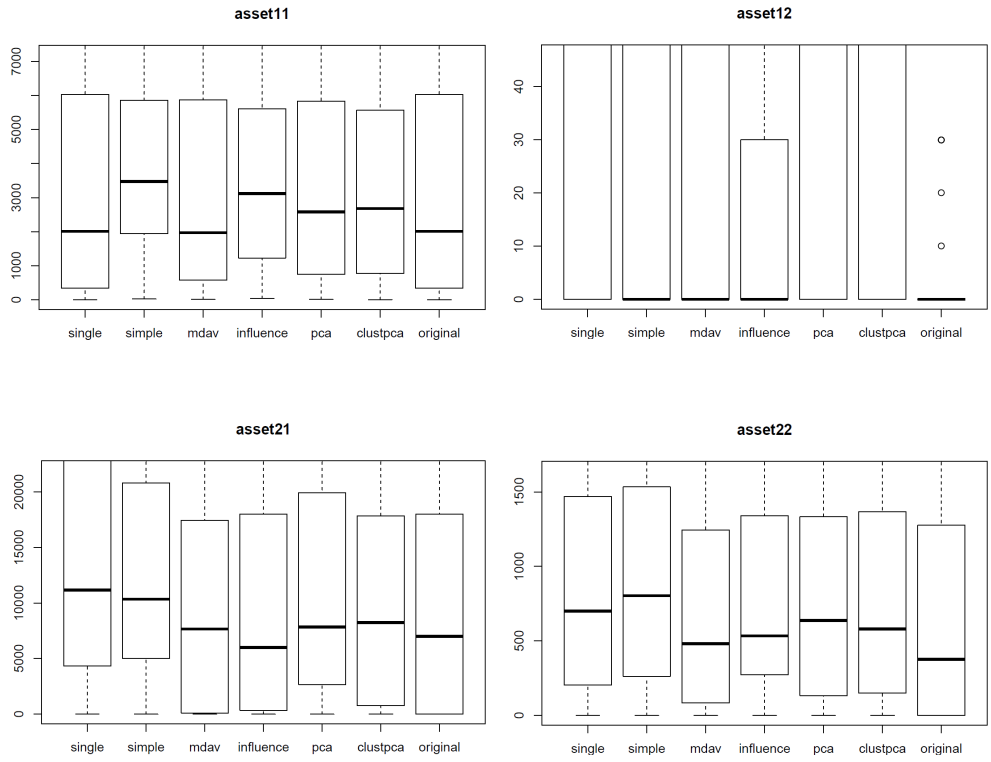
〈부표 3-1〉 매스킹 기법 적용에 따른 표본 S 민감변수들의 평균

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

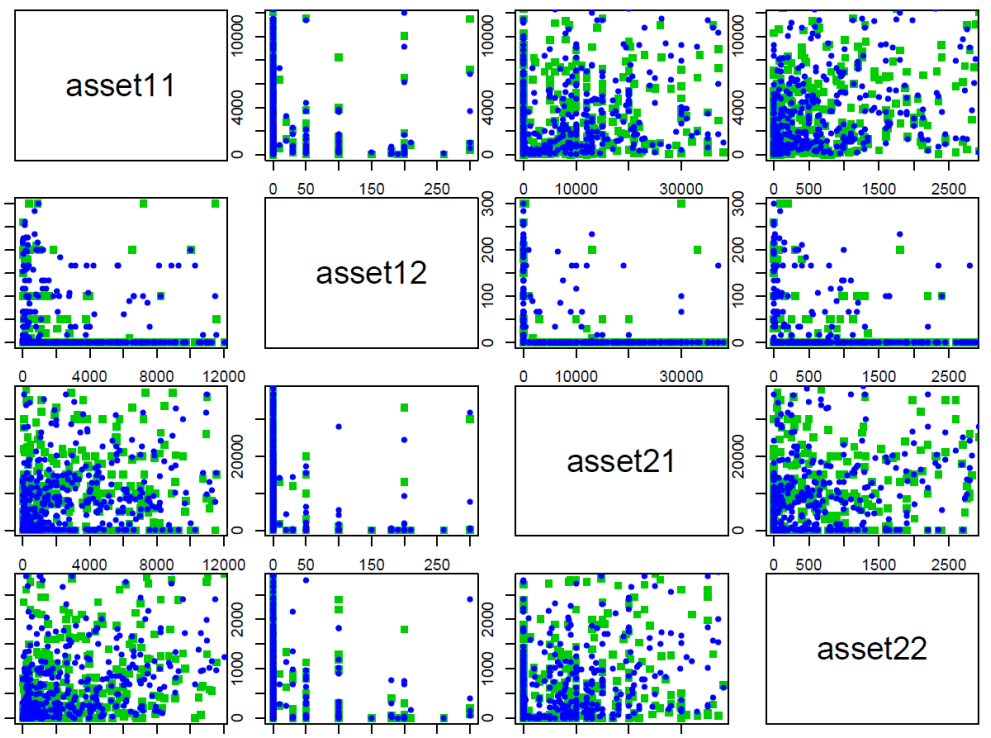
〈부표 3-2〉 매스킹 기법 적용에 따른 표본 S 민감변수들의 표준편차

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



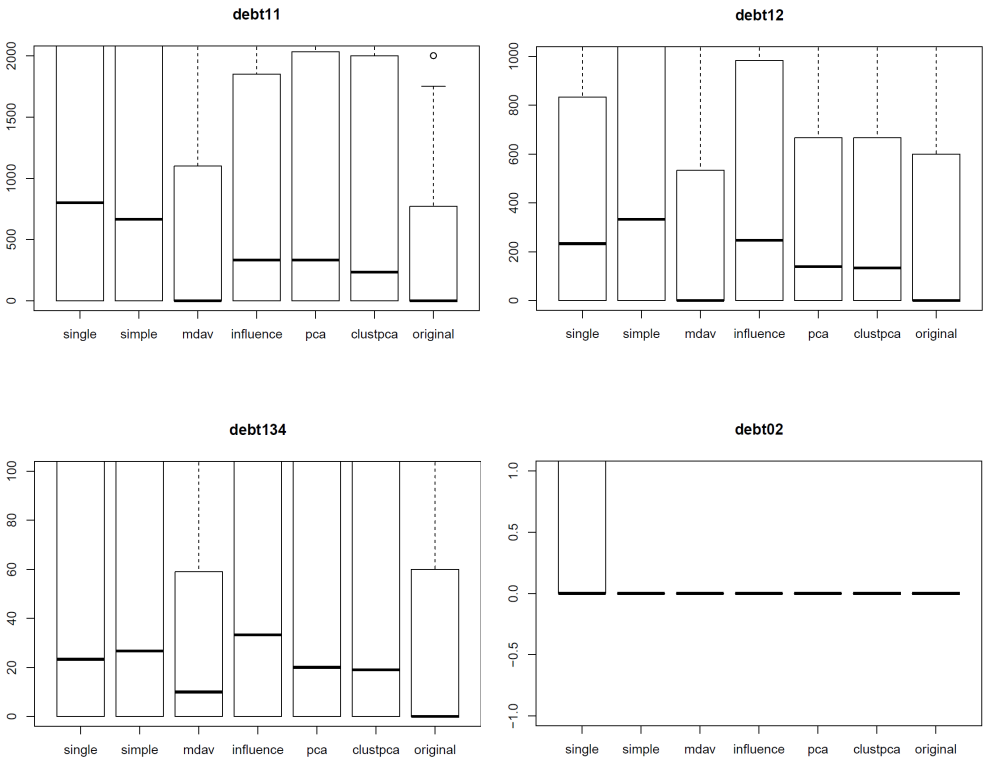
[부그림 3-1] 국소통합 세부 기법별 자산 변수들의 상자그림



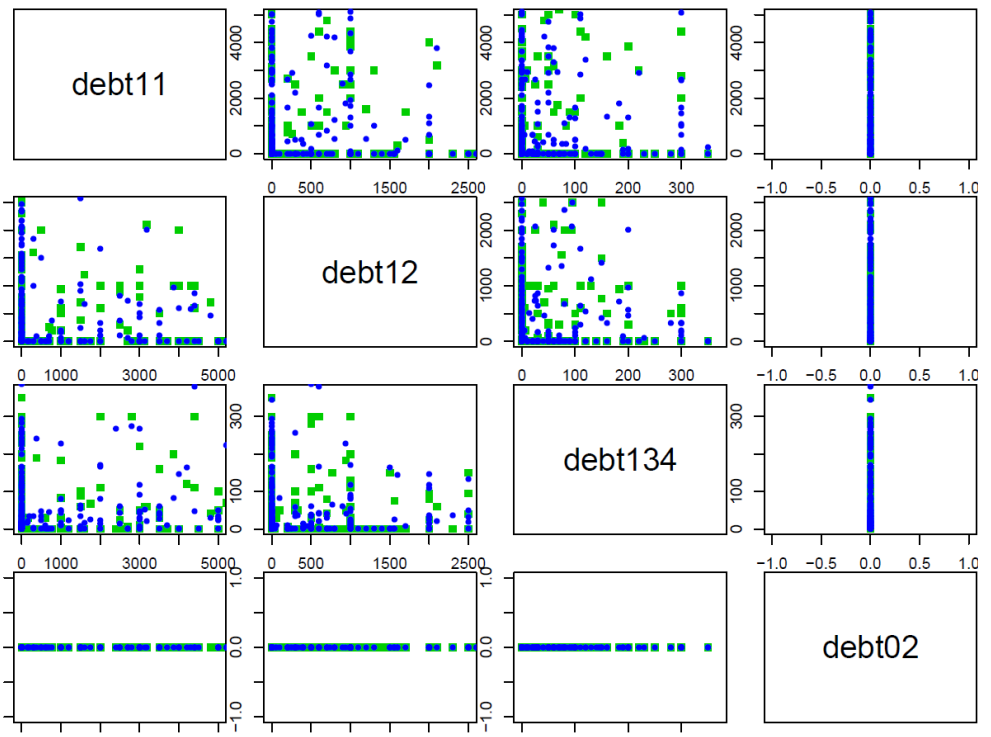
[부그림 3-2] 국소통합(mdav) 기법 적용 전후 자산 변수들의 산점도

〈부표 3-3〉 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 3-3] 국소통합 세부 기법별 부채 변수들의 상자그림

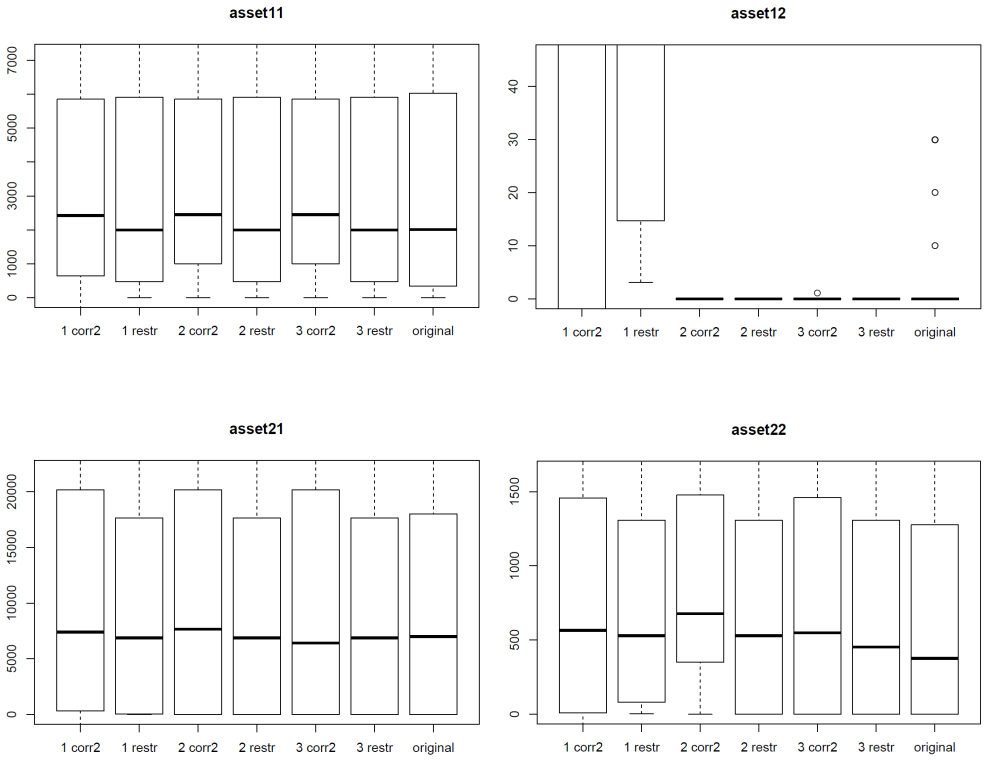


[부그림 3-4] 국소통합(mdav) 기법 적용 전후 부채 변수들의 산점도

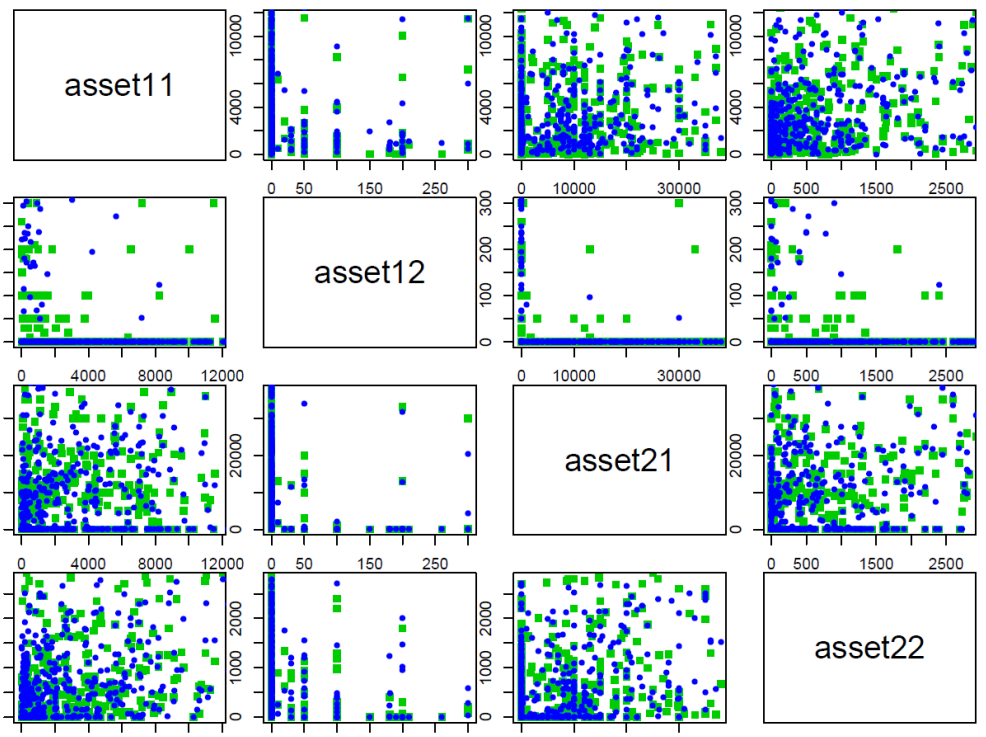
〈부표 3-4〉 국소통합(mdav) 기법 적용 전후 부채 변수들의 상관계수

	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



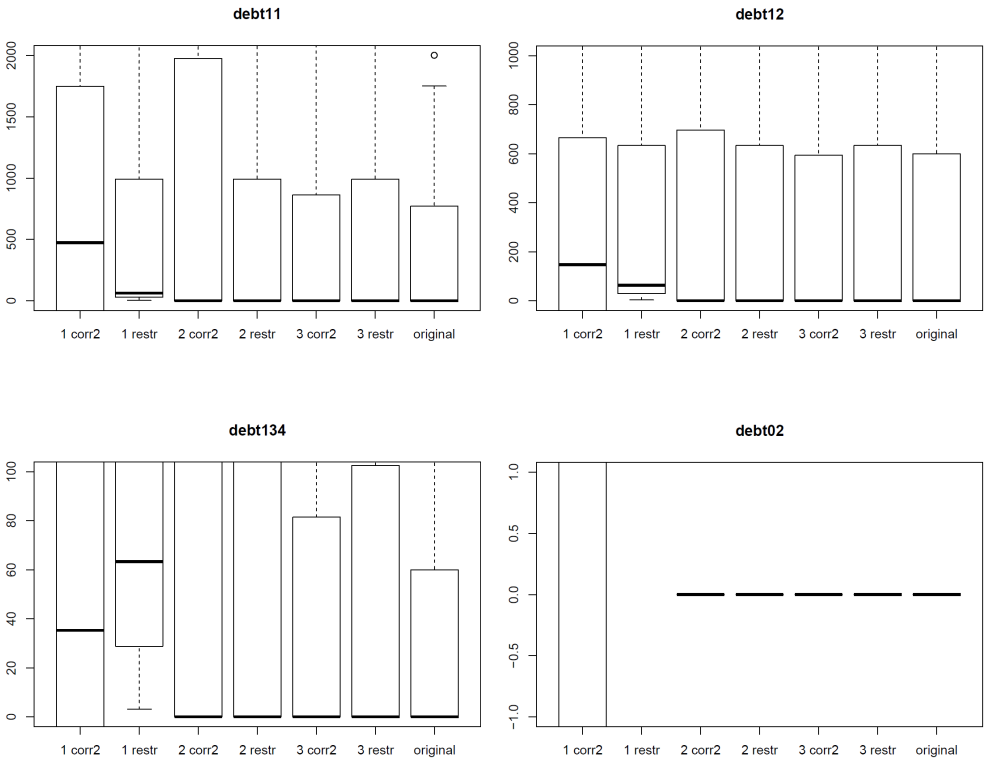
[부그림 3-5] 잡음 추가 시나리오별 자산 변수들의 상자그림



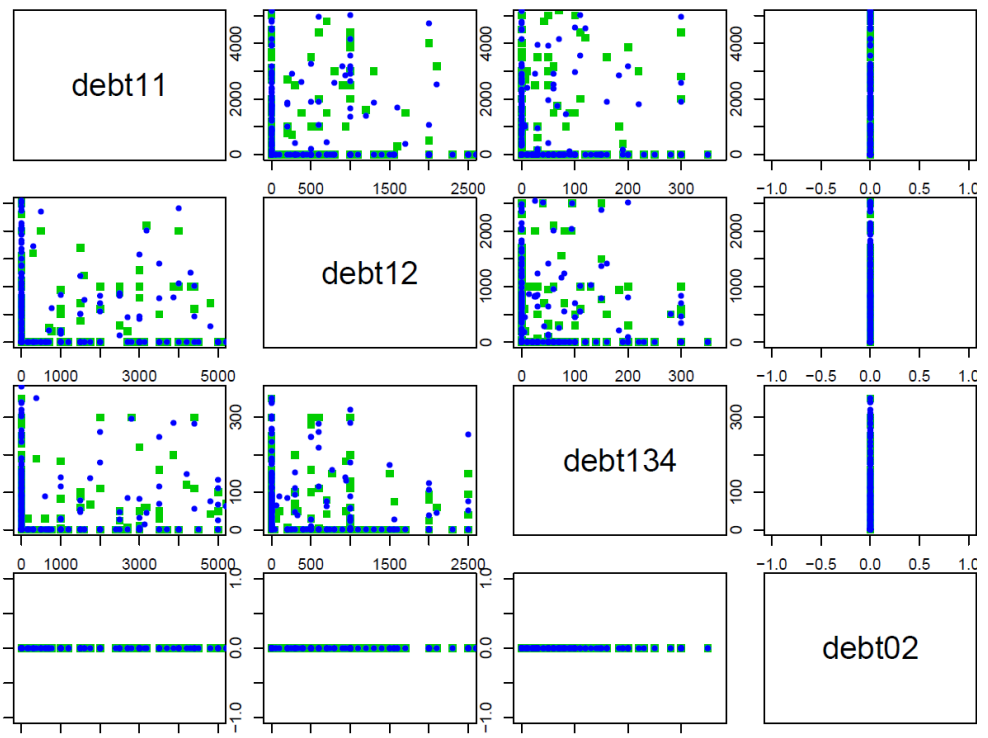
[부그림 3-6] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 자산 변수들의 산점도

〈부표 3-5〉 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 3-7] 잡음 추가 시나리오별 부채 변수들의 상자그림

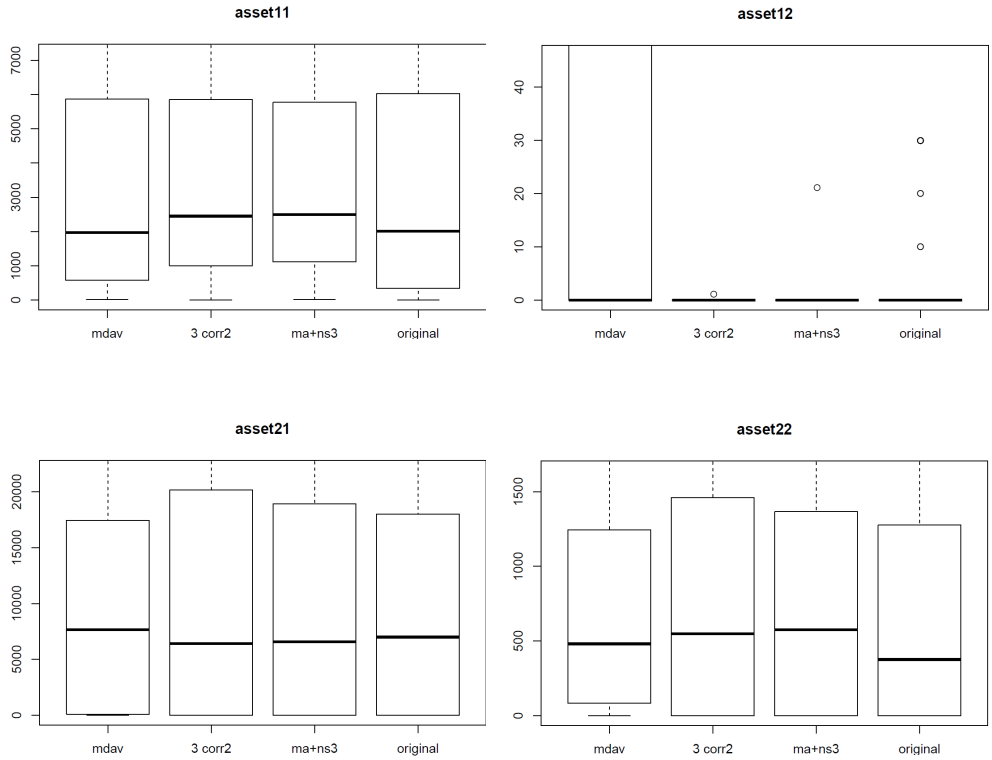


[부그림 3-8] 잡음 추가(시나리오3, correlated2) 기법 적용 전후 부채 변수들의 산점도

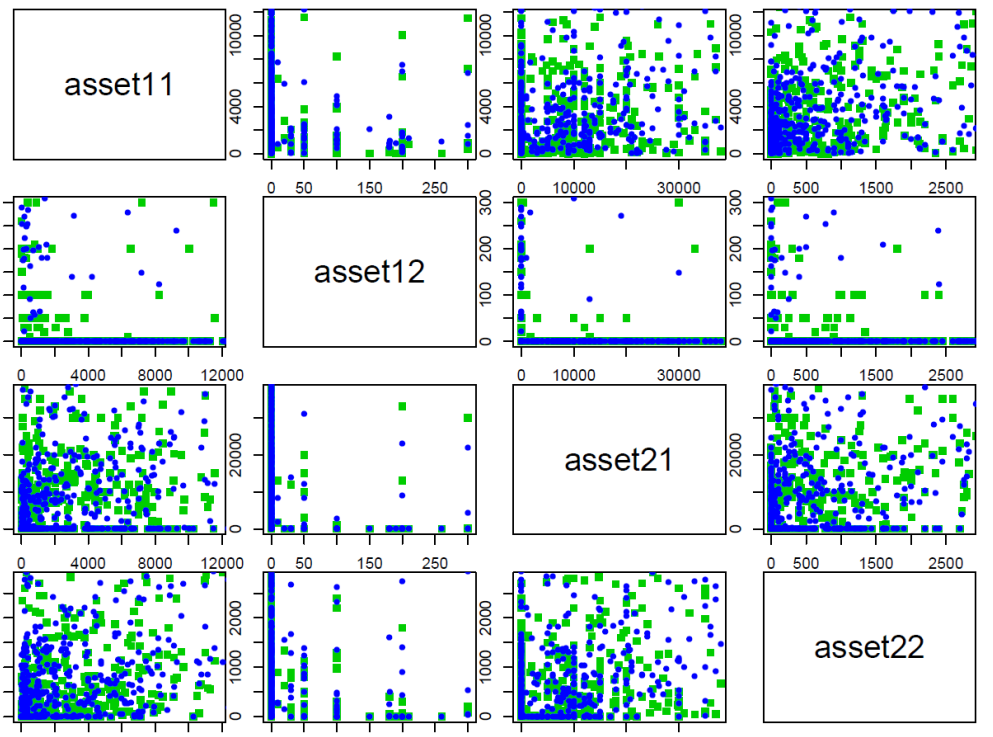
〈부표 3-6〉 잡음 추가(시나리오3, correlated2) 기법 적용 전후 부채 변수들의 상관계수

	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



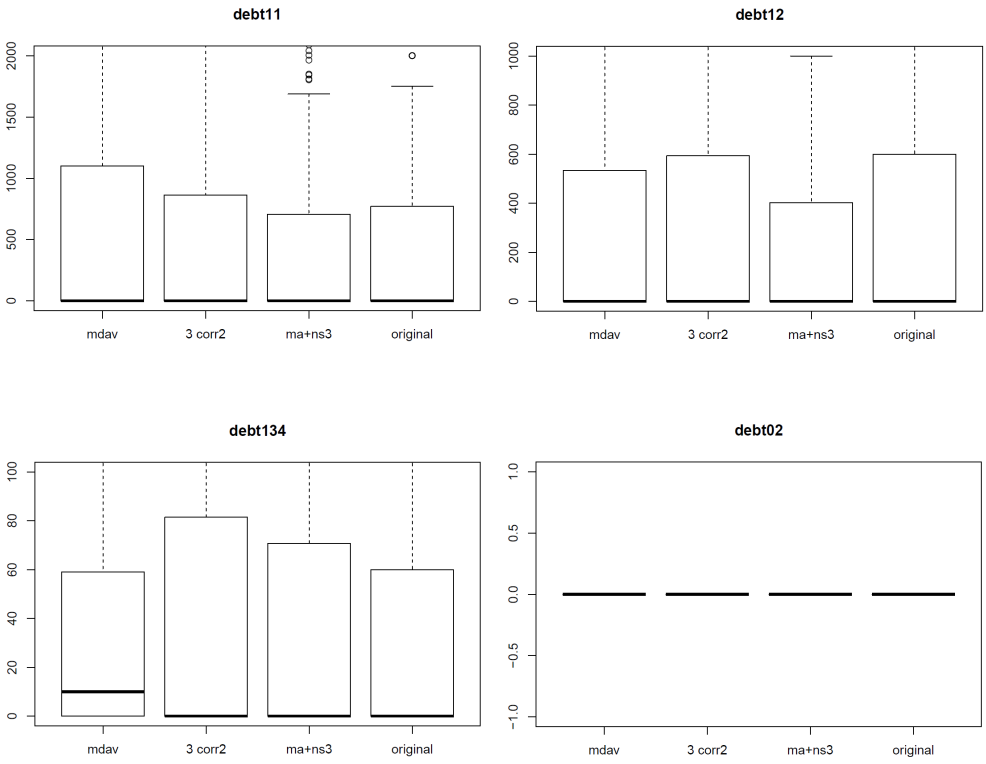
[부그림 3-9] 국소통합, 잡음 추가3 및 결합안3에 대한 자산 변수들의 상자그림



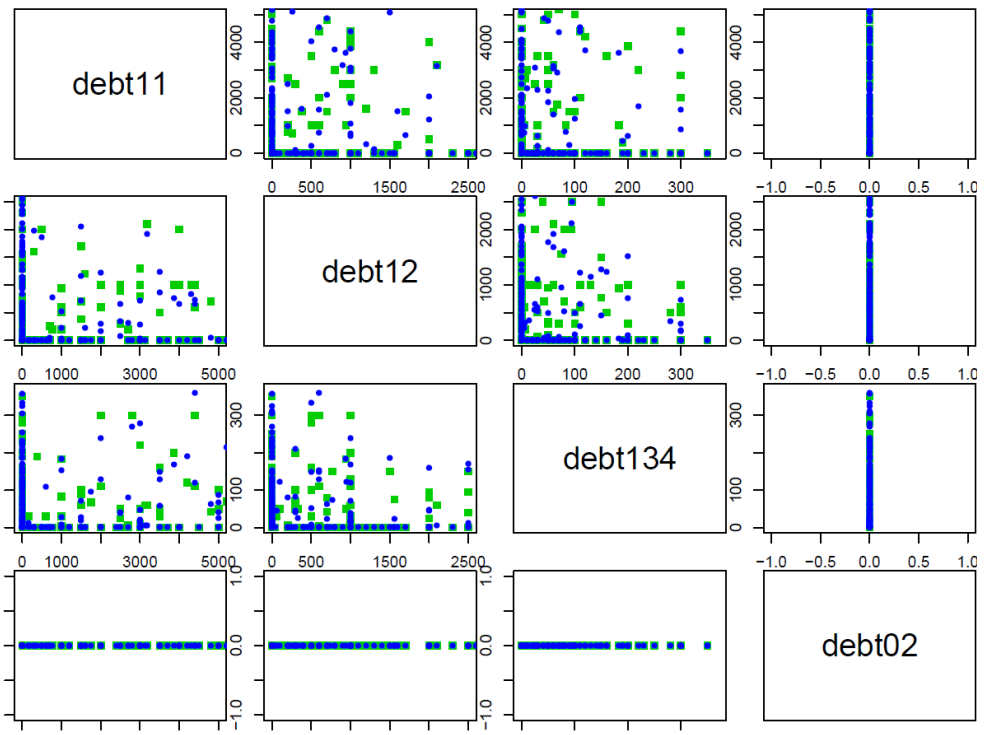
[부그림 3-10] 결합안3 적용 전후 자산 변수들의 산점도

〈부표 3-7〉 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 3-11] 국소통합, 잡음 추가3 및 결합안3에 대한 부채 변수들의 상자그림



[부그림 3-12] 결합안3 적용 전후 부채 변수들의 산점도

〈부표 3-8〉 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