Journal of the Korean Data & Information Science Society
2013, **24**(5), 1029–1041

http://dx.doi.org/10.7465/jkdi.2013.24.5.1029
한국데이터정보과학회지

정보공개 환경에서 개인정보 보호와 노출 위험의 측정에 대한 통계적 방법

이용희[1]

[1]서울시립대학교 통계학과

접수 2013년 7월 3일, 수정 2013년 8월 23일, 게재확정 2013년 9월 4일

요 약

최근 빅데이터의 등장과 정보 공개에 대한 급격한 수요 증가에 따라 자료를 일반에게 공개할 때 개인 정보를 보호해야 하는 필요성이 어느 때보다 절실하다. 본 논문에서는 마이크로 자료와 통계분석 서버를 중심으로 현재까지 제시된 개인정보 노출제한를 위한 통계적 방법, 정보 노출의 개념, 노출 위험을 측정하는 기준들을 개괄적으로 소개한다.

주요용어: 개인정보 보호, 노출위험, 노출제한 방법, 데이터베이스, 마이크로 자료, 빅데이터, 통계분석 서버.

1. 서론

1960년대부터 표본조사에 대한 통계적 방법이 발전하고 1980년대부터는 정보기술의 혁명이 시작되어 자료의 수집, 저장 및 처리 방법이 비약적으로 발달하였다. 최근에는 통계학과 정보기술의 융합으로 정부, 기업, 개인등이 수집된 자료와 이로부터 생성되는 유용한 정보를 매우 편리하게 이용할 수 있는 시대가 되었다. 하지만 최근까지 수집된 자료나 생성된 정보를 자유롭게 이용할 수 있는 권한과 범위는 상당한 제약을 받았던 것이 현실이다. 이는 많은 자료들이 개인 사생활이나 기관에 관한 민감한 정보를 포함하고 있기 때문에 정보 보호를 위하여 접근이 제한되어 왔으며 실제로 공공에게 제공되는 정보는 자료를 단순하게 요약한 정보가 대부분이었다. 하지만 현재는 과거와 달리 공공기관이나 기업이 보유하고 있는 방대한 자료를 이용하여 새로운 가치를 창출할 수 있다는 인식이 확대되고 있는 상황이다. 이러한 정보 공개에 대한 요구와 더 나아가 빅데이터의 등장이라는 새로운 환경에 따라 정보 공개의 범위와 사용 권한의 확대에 대한 요구가 크게 증가하고 있다.

공공 이용자에게는 자료를 요약한 형태로 제공되는 것이 일반적인 일이나 필요한 경우에는 자료의 단위 또는 개체 (unit)에 대한 정보를 포함한 마이크로 자료 (micro data 또는 raw data)형태로도 제공된다. 마이크로 자료를 공개하는 경우에는 개인이나 기관의 신분이 노출될 위험이 항상 존재하기 때문에 민감한 정보가 노출되는 일이 발생할 가능성이 높다. 마이크로 자료를 공공에게 공개하는 경우에는 공개되는 변수들을 적절히 축소 또는 변경하여 개인의 신분이나 민감한 정보가 유출되는 가능성을 제거 또는 축소해야 한다. 하지만 비밀 보호를 위하여 자료를 변형하거나 축소하는 경우 공개된 자료의 유용성 (utility)이 떨어질 수 있으므로 자료의 변형 전과 후의 분석결과가 최대한 유사하도록 변형 방법을 설계해야 한다. 이렇게 개인정보 보호를 위하여 자료를 변형시키는 것을 노출제한 또는 개인정보 보

[1] (130-743) 서울특별시 동대분구 서울시립대로 164, 서울시립대학교 통계학과, 부교수.
E-mail: ylee@uos.ac.kr

호 (statistical disclosure limitation, masking, privacy-preserving data publishing)라고 한다. 사실 1980년대부터 마이크로자료 제공에 대한 요구가 점점 높아지고 있는 추세를 반영하여 정보 노출의 위험을 낮추기 위하여 현재까지 많은 통계적 방법이 제시되어 왔다. 또한 데이터베이스에 의한 자료의 관리 및 이용에 관련하여 암호학을 포함한 정보기술과 관련된 여러 분야에서 개인정보 노출방지를 위한 다양한 방법들도 제시되었다.

현재 통계학과 정보기술이 당면한 정보 공개에 대한 환경은 최근까지 제시된 노출위험에 대한 개념과 노출을 방지하는 방법들이 고려하였던 범위를 크게 넘어서고 있다. 빅데이터의 분석과 활용에 대한 요구에 따라 현재까지 고려된 자료의 양과는 비교를 할 수 없는 방대한 크기의 자료가 공개될 수 있는 환경이며 더 나아가 이러한 빅데이터들이 서로 결합되어 새로운 정보와 가치를 창출할 수 있는 환경은 개인 또는 기관의 신분이 노출될 위험을 급격하게 증가시킬 뿐만아니라 현재까지 고려하지 못했던 새로운 위험들이 생겨날 수 있다. 많은 빅데이터가 마이크로 자료의 형태이기 때문에 현재까지 개발된 개인정보 보호 방법들을 적용할 수 있지만 자료의 크기와 구조, 분석 환경이 현재까지 제안된 방법들의 범위를 벗어나 새로운 위험을 초래할 수 있다. 따라서 이러한 변화된 환경에 대응하는 정보노출 위험에 대한 개념과 노출제한 방법들이 진화해야한다.

빅데이터의 등장과 정보 공개에 대한 수요 증가에 따라 이에 대처하는 새롭고 유용한 방법을 개발해야 하는 필요성이 증가하고 있지만 현재까지 제시된 정보 보호의 개념과 방법들도 매우 유용한 것도 사실이다. 따라서 본 논문에서는 마이크로 자료를 중심으로 현재까지 제시된 노출제한를 위한 통계적 방법과 개념들을 개괄적으로 소개하고 정보공개 시대의 도래와 빅데이터라는 새로운 환경에 대처하는 방향에 대하여 논의하고자 한다. 2절에서는 마이크로 자료에 대한 통계적 노출제한 방법을 알아보고 3절에서는 통계분석 서버와 정보노출의 개념을 알아보고 4절에서는 노출 위험을 측정하는 방법에 대하여 알아보고 5장은 본 논문의 결론이다. 참고로 Matthews 등 (2011)은 최근까지 제안된 다양한 통계적 노출제한 방법과 개념들을 자세히 리뷰하였다.

2. 마이크로 자료의 노출제한 방법

마이크로 자료는 단위 자료들이 모여서 만들어진 집합이다. 각 단위 자료는 다양한 형태의 변수들의 값으로 구성되어 있다. 예를 들면 자료의 구성 단위가 사람이라면 각 단위 자료는 사람에 대한 특성을 나타내는 변수인 나이, 성별, 직업등을 포함한다. 마이크로 자료는 이용자들에게 매우 유용한 정보를 제공하지만 동시에 개인의 신분이 유출되어 사생활 침해의 가능성도 동시에 존재한다. 따라서 마이크로 자료에 포함된 자료 중 개인의 신분과 민감한 정보를 포함하고 있는 변수들을 축소 또는 변형하여 개인정보가 노출되는 가능성을 줄이는 노력이 필요하다. 이러한 정보 보호를 위한 노력은 정보의 유용성을 저하시키므로 비밀 보호와 자료의 유용성을 동시에 고려해야한다.

2.1. 기본적인 보호 방법

마이크로 자료에는 개인이나 기관의 신분을 직접적으로 알아낼 수 있는 식별변수 (identifier)가 있다. 식별변수의 예는 성명, 상호, 주민등록번호, 법인번호, 주소 등이 있다. 익명화 (anonymization)는 노출제한 방법 중에서 가장 기본적인 방법이며 마이크로 자료가 공개되기 전에 개체를 식별할 수 있는 명백한 식별변수를 자료로부터 제거하는 것이다.

이렇게 직접 신분이 노출되는 변수도 있지만 개인이나 기관의 특성을 나타내는 보조식별변수 (quasi identifier, key variable)도 존재한다. 보조식별변수의 예는 생년월일, 직업, 설립일, 업종, 거주 또는 사업 지역 등이 있다. 보조식별변수가 노출되면 개인이나 기관의 신분이 노출되지는 않지만 외부의 정보와 결합 (matching)하거나 또는 자료에 존재하는 유일한 개체 (uniqueness)임을 이용하여 실질적인

신분 노출이 발생할 수 있다. 최근에는 정보기술의 발전과 여러 가지 레코드 연결 프로그램 (record linkage program)의 상용화로 인하여 마이크로 자료를 공개하는 경우 직접적인 식별변수를 공개하지 않더라도 외부이용자가 공개된 자료와 외부자료의 결합을 통하여 개체의 신분을 쉽게 파악할 수 있다. 보조식별변수는 이러한 위험성을 가지고 있지만 자료를 이용하는 측면에서 매우 유용한 정보이기 때문에 완전하게 제거하는 익명화 방법을 사용하면 자료의 유용성이 크게 감소한다. 따라서 마이크로 자료를 공개하는 경우에는 보조식별변수를 변형하는 방법을 적절하게 사용해야 한다. 사실상 지금부터 소개되는 대부분의 통계적 노출 제한방법들은 보조식별변수의 조합이나 결합을 통하여 정보가 누출되는 위험을 줄이는 방법들이다.

코딩 (coding)은 노출제한 기법의 가장 기본적인 방법 중의 하나로 연속형 변수 (예를 들어 연령)를 범주형 변수 (예 10-19세, 20-30세,...)로 변환하는 방법을 말한다. 코딩의 범위가 클수록 노출의 위험성이 작아지지만 자료의 유용성이 작아진다. 어떤 변수의 값을 순서대로 정열하였을 때 가장 작은 값 또는 가장 큰 값 근처에 있는 개체는 노출의 위험성이 크므로 (예를 들어 최고령자 또는 최대 매출액) 정렬된 자료에서 극단 자료가 포함되어 있는 처음과 끝 부분만 코딩하는 방법이 있다 (예를 들어 90세 이상 또는 매출액 10억 이상). 이를 상단코딩 (top-coding) 또는 하단코딩 (bottom-coding) 이라고 한다.

표본추출 (sampling)은 전체 자료의 일부분을 확률추출하여 공개하는 방법이다. 전체 자료가 아닌 표본의 공개는 비밀보호를 위한 매우 효과적이고 손쉬운 방법이다. 자료의 이용자에게 모수 (parameter)를 추정할 때 필요한 표본추출 설계에 대한 정보 (모집단과 표본의 수, 가중치 등)를 제공할 수 있다. 특별히 표본추출 설계에 대한 정보를 제공할 경우 노출의 위험이 증가하는 경우도 있으므로 이러한 점을 유의하여 적절한 정보를 제공해야 한다. 예를 들어 학교에서 일부의 학생을 표본추출하여 소속 학교의 정보를 감추고 자료를 제공할 때 설계에 대한 충분한 정보를 제공하면 학교의 전체 학생 수를 추정할 수 있고 이를 통하여 학교를 식별할 수 있다. 이러한 경우 학교와 학생이 연결되어 노출의 위험성이 증가한다.

2.2. 잡음첨가방법

잡음첨가방법 (noise addition)은 자료의 값에 잡음 (noise)을 더하거나 곱하여 원래 자료에 약간의 변형을 가하여 공개하는 방법이다. 중요한 개인 정보를 포함한 변수에 잡음을 첨가하여 변형된 자료를 공개하면 이용자가 가지고 있는 외부 정보와 차이가 있기 때문에 외부 자료와 결합하여 신분을 알아낼 수 있는 가능성이 감소한다. 더 나아가서 개인의 신분이 노출되어도 만약에 민감한 정보에 대한 변수들에 잡음첨가방법이 적용되었다면 외부이용자가 알아낸 개인의 정보가 실제 값과 차이가 있음으로 직접적인 정보의 노출 피해를 줄일 수 있다.

잡음을 더하거나 곱하는 경우 편이 (bias)를 피하기 위하여 자료의 값에 더해주는 잡음의 평균을 0으로 하고 곱해주는 잡음의 평균은 1이 되도록 한다. 일반적으로 잡음첨가방법은 변수가 연속적인 값을 가지는 변수에 적용되는 방법이지만 특별하게 잡음의 형태를 조종하면 범주형 변수에도 적용이 가능하다. 잡음첨가방법은 실제 자료에 잡음을 더해주거나 곱해주는 방법이므로 구현하기 쉽지만 잡음을 생성할 때 변수의 분포를 고려해야 하기 때문에 변수의 통계적인 특성을 이해해야 하는 어려움이 있다. 잡음을 원래 값에 더해주거나 곱해준 후에 변형된 자료의 평균은 크게 변하지 않지만 분산이 증가하는 중요한 단점이 있다. 하나 이상의 변수에 잡음을 첨가하는 경우에는 여러 개의 변수들의 상관관계를 왜곡시킬 위험이 있다. 이러한 단점을 보완하기 위하여 상관이 있는 잡음들 (correlated noise)을 여러 개의 변수에 각각 첨가해주면 원래 자료의 상관관계를 유지할 수 있다. 하지만 무상관 잡음첨가 (uncorrelated noise)나 상관 잡음첨가 모두 요약 통계량의 분산을 증가시킨다.

2.3. 자료교환

자료교환 (data swapping)은 정해진 영역내의 개제들의 자료를 서로 교환하는 방법을 말한다. 자료교환은 Dalenius와 Reiss (1978)에 의해 처음 소개되었고, Dalenius와 Reiss (1982)에 의해 발전되었다. Dalenius와 Reiss는 범주형 자료에서 결합 도수분포 (joint frequency distribution)가 변하지 않도록 하는 자료교환의 절차를 제안하였다. 초기의 연구는 대부분 범주형 자료에 대한 방법이었으나 Reiss (1984)와 Dalenius (1986)가 연속형 자료에 대하여 자료교환 방법을 확장하였다. 자료교환 방법에 대한 자세한 개관은 Fienberg와 McIntyre (2004)에 잘 설명되어 있다.

자료교환은 자료의 값 자체를 변화시키지 않고 그 위치를 바꾸는 방법으로 원래의 자료와 교환된 자료에 의해 계산된 통계량이 동일하게 얻어지도록 구현된다. 자료교환의 방법을 잘 설계하면 자료의 교환 후에도 중요한 요약 통계량들이 원래 자료의 통계량과 같아지게 또는 매우 유사하게 할 수 있다. 자료가 교환되는 정도와 범위가 커지면 민감한 정보의 유출 가능성이 줄어들지만 자료 교환한 후에는 변수들의 관계에 대한 통계량 (예를 들어 상관계수)이 변하기때문에 자료의 유용성이 감소한다. 따라서 자료교환의 방법을 구현하는 경우 정보유출의 위험성과 자료의 유용성에 대한 상호적인 고려를 하여 주어진 목적을 달성하도록 교환방법을 설계한다.

일반적으로 자료를 공개할 때 이용자가 가지고 있는 외부의 정보와 공개된 자료의 정보가 결합되어 개인의 신분이 노출되고 민감한 정보가 유출된다. 자료교환을 적용한 후 공개를 하면 노출된 응답자의 신분에 대한 불확실성이 커지기 때문에 외부 이용자들의 비밀누출 행위에 대한 시도를 차단하는 효과가 있다. 하지만 자료교환은 다른 정보보호 방법들보다 비용이 많이 드는 단점이 있다. 자료에 대한 전반적인 구성이나 변수의 특성 및 분포를 알고 이를 이용하여 자료교환을 하기 때문에 자료의 일부 또는 전체를 간단한 규칙에 따라 변형시키는 방법보다 시간이나 비용이 많이 든다. 또한 구현을 하는 경우 컴퓨터 작업을 하기 위하여 공간과 시간이 필요하다. 따라서 빅데이터에서 나타나는 비정형 또는 실시간 자료에는 적용하기에 무리가 있을 수 있다. 하지만 이러한 비용의 지출에 비하여 원래 자료와 매유 유사한 유용성을 갖는 자료를 공개할 수 있기 때문에 통계기관들이 중요한 표본조사의 마이크로자료 공개 시 자주 사용하고 있다.

2.4. 인위자료

인위자료 (synthetic data)는 앞에서 살펴본 통계적 노출제한 기법과 매우 다르다. 주로 전통적인 노출 제한 기법은 원래 자료를 바꾸거나 변형시켜서 노출의 위험성를 줄인다. 인위자료는 원자료를 생성하는 가상의 통계적 모형 (statistical model)을 가정하고 원래 자료를 통계적 모형에서 발생한 모의자료 (simulated data)로 대체하는 방법을 말한다. 인위자료는 원래 자료와 직접적인 관련이 없는 자료이지만 원자료를 생성하는 가상의 모형에서 추출된 자료이기 때문에 원자료가 지닌 통계적인 특성들을 가진다. 따라서 이러한 인위자료에 기반한 분석은 원자료에 기반한 분석과 매우 유사할 것이다. 이러한 인위자료의 공개는 노출의 위험성을 크게 줄일 수 있다. 이러한 장점이 있음에도 불구하고 자료를 생성하는 적절한 통계적인 모형을 찾는 것은 매우 어려운 작업이므로 아직까지 널리 쓰이지 못하는 상황이나 근래에는 많은 연구가 활발히 이루어지고 있다 (Duncan 등, 2011).

3. 통계분석 서버와 정보노출의 개념

3.1. 마이크로 자료에 대한 접근과 통계분석 서버

마이크로자료에 노출제한 방법을 적용하여 대중이 제한없이 이용할 수 있도록 공개하는 경우 자료의 유용성이 상대적으로 감소하여 이용자들이 다양하고 세밀한 분석이 불가능한 경우가 많다. 이러한 경우

를 위하여 노출제한 방법을 최소한으로 적용한 원래 자료에 매우 가까운 마이크로자료를 이용할 수 있는 환경을 제공하는 경우가 필요하다. 이렇게 마이크로자료에 대한 제한적인 접근을 허용하는 경우에는 이용할 수 있는 자격요건, 제공하는 수단과 방법 (제한된 공간 또는 온라인 접속 등), 분석할 수 있는 범위 (전체 자료, 일부 자료 또는 요약 자료) 등을 고려하여 그에 대한 공개 정책을 수립하고 시행해야 한다. 노출제한 방법이 최소한으로 적용된 마이크로자료를 이용할 수 있는 자격을 가진 사람들은 자료에 포함된 정보를 누출할 위험성이 매우 작은 신뢰받는 개인이나 조직이어야 한다. 또한, 마이크로 자료를 이용하려는 사람들에 대해서는 자료유출에 대한 도덕적 또는 법적인 책임에 대하여 이용 전에 구속력 있는 서약서를 받는 것이 중요하다.

최근에 들어 마이크로 자료 자체를 공개하기 보다는 데이터베이스에 연결된 통계분석 서버 (data analysis server, statistical analysis server)에 인터넷으로 접속을 통하여 마이크로 자료를 분석하여 요약된 정보를 제공받는 서비스가 증가하고 있다. 통계분석 서버는 마이크로 자료의 공개가 불가능한 빅데이터 분석을 위하여 사용되는 경우가 많다. 이러한 인터넷을 이용한 자료분석 서비스는 실제로 마이크로 자료 자체에 접근할 수 없어도 정교하게 설계된 연속 접속을 통하여 민감한 정보를 파악하는 것이 가능하다. 이러한 정보 노출의 위험은 과거에는 흔히 볼 수 없는 새로운 형태의 위험이기 때문에 이에 대응하는 정보보호 기법이 필요하다. 대표적인 정보 누출의 위험에 대한 경우는 이용자가 통계분석 서버에 계획된 쿼리 (targeted queries)를 지속적으로 요청하여 필요한 민감한 정보를 획득할 수 있는다는 것이다. 예를 들어 분석의 결과 (평균, 히스토그램, 집계표, 회귀계수 등)에 잡음을 첨가하는 방법이 정보보호 기법으로 사용되는 서버에 지속적으로 동일한 쿼리를 요청하여 얻어진 많은 통계량들의 평균을 구하면 잡음의 효과를 상쇄시켜서 (대수의 법칙을 이용) 노출보호 방법을 무력화 시킬 수 있다. 이렇게 동일한 쿼리가 지속적으로 요청되는 경우 이를 공격으로 탐지하여 첨가되는 잡음의 분산을 증가시키는 방법을 적용할 수 있다. 새로운 통계분석 서버 환경에 대한 정보 보호의 개념과 방법은 Gomatam 등 (2005)에 소개되어 있다.

3.2. 정보 노출의 개념

통계분석 서버를 제공하거나 마이크로 자료를 공개하기 위하여 노출제한 방법을 적용하려면 먼저 정보노출 (disclosure of information)에 대한 개념과 정의를 구체화해야 한다. Duncan과 Lambert (1989)은 정보노출에 대한 정의를 다음과 같이 4개의 범주로 분류하였다.

- 신분노출 (identity disclosure)

  신분노출은 공개된 자료에 포함된 개인이나 기관의 신분이 노출되는 것이다. 마이크로 자료를 공공이 이용할 수 있게 공개할 때 신분노출의 위험이 어떤 위험보다 가장 피해가 크다. 신분노출이 되면 개인의 사생활이 침해되고 법적인 분쟁을 일으킬 수도 있으며 자료를 공개한 기관의 공신력이 크게 떨어져 이후의 자료의 수집이나 공개에 나쁜 영향을 미칠 수 있다. 따라서 마이크로 자료를 공개하는 경우 신분노출의 위험성을 최소로 하는 방법의 선택이 최우선적으로 고려되어야 한다.

- 특성노출 (attribute disclosure)

  특성노출은 개인이나 기관의 신분이 노출되지는 않지만 공개된 자료에 포함되어 있는 민감한 특성 (예를 들어 소득, 매출액, 범죄기록 등)에 대한 정보가 자료의 이용자에게 누출되는 경우를 말한다. 특성노출은 대부분 공개된 자료와 외부의 자료를 결합하는 경우에 주로 일어나며 그 결과로 신분노출이 일어날 수 있다. 개인정보 보호를 위한 방법을 적용할 경우 외부 자료와의 유사성이나 특성들의 민감성을 고려하여 특성노출이 일어나도 그에 따른 파생 위험이 최소화 되도록 해야한다.

- 추정노출 (inferential disclosure)

  추정노출은 공개된 자료에 포함된 개인의 자료와 외부자료를 연결 (matching)하지 않았더라도 개인정보가 노출되거나 새로운 정보를 추론할 수 있는 위험을 말한다. 예를 들어 이용자가 "A기업의 순이익이 동종업종의 평균 순이익보다 20% 낮다"라는 사실을 알고 있다고 하자. 더 나아가 업종의 평균 순이익에 대한 정보를 포함하는 공개된 자료가 있다면 이용자는 A기업의 순이익을 공개된 정보를 이용하여 추론할 수 있을 것이다. 이러한 예는 개인이 공개된 자료에 포함되어 있지 않더라도 관련된 정보의 공개와 추론에 의하여 개인정보가 노출되는 상황이 가능하다는 것을 암시한다. 이러한 추정노출의 위험은 공개되는 자료가 다양하고 많아지는 빅데이터의 환경에서 크게 증가할 수 있다. 하지만 자료를 공개하는 입장에서는 이미 공개된 외부 자료의 특성과 관련성을 완전하게 파악하기 어려우므로 추정노출의 위험을 수량화하여 측정하는 것은 어려운 일이다.

- 모형노출 (model disclousre)

  모형노출은 모집단노출 (population disclosure)라고도 불리며 이는 특정집단에 대한 민감한 정보가 공개된 자료에서 추론된 모형으로부터 노출되는 경우를 말한다. 예를 들어 근로자 개인의 임금이 노출되는 것이 아니라 근로자의 특성과 임금의 관계 (relation)가 자료로 부터 유추되는 경우 모형노출이다. 이러한 경우 여자와 남자 근로자의 임금 격차에 대한 상세한 정보가 모형노출에 의해 공개될 수 있다.

현재까지 주로 제안된 통계적 노출제한 방법들은 신분노출이나 특성노출을 방지하기 위한 방법들이 대부분이다. 추정노출과 모형누출은 마이크로 자료의 공개보다는 주로 요약 자료 (교차표, 기술통계)에 의하여 발생되어 왔으며 이에 대한 노출제한 방법은 오래 전부터 개발되어 왔다 (Duncan 등, 2011, 4장). 하지만 이 두 위험은 최근에 데이터베이스에 연결된 통계분석 서버의 이용이 확대되면서 다시 큰 문제로 떠오르고 있다. 자료분석을 제공하는 서버에 정교하게 설계된 쿼리들을 지속적으로 요청하여 얻어진 정보를 계속 결합하면 특정 개인의 민감한 정보를 추정하여 노출시킬 수 있다. 빅데이터의 이용과 정보공개 환경에 대한 요구가 증가하면 추정노출과 모형노출에 대한 새로운 위험이 증가할 것이므로 이에 대응할 수 있는 정보보호 방법의 개발이 필요하다.

4. 정보 노출의 위험 측정

마이크로 자료를 공개하는 경우 개인의 정보보호를 위하여 노출제한 방법을 적용한다. 하지만 노출제한 방법을 강하게 적용하면 자료의 유용성이 떨어져서 공공의 자료 이용에 의한 새로운 가치 창출이라는 자료 공개의 목적를 이루지 못할 수 있다. 따라서 개인정보호와 자료의 유용성을 동시에 고려하여 마이크로 자료를 공개해야 한다. 이러한 경우 원래 자료와 공개 자료에서 개인 정보의 노출에 대한 위험성 (risk of disclosure)을 측정할 수 있는 측도 (measure)가 필요하다. 이러한 위험성의 크기를 수량화할 수 있는 측도는 노출제한 방법의 효과를 과학적으로 측정하는 중요한 도구가 된다. 또한 서로 다른 방법들은 효율을 비교하는 경우에도 유용하게 사용된다. 본 절에서는 정보 노출의 위험을 측정하는 전통적인 방법에서 최근에 제안된 차등정보보호 (differential privacy)까지 개괄적으로 소개한다.

4.1. 신분노출의 위험에 대한 측도

Paass (1988)는 공개된 자료의 보조식별변수와 외부 자료의 연결로서 개인정보유출이 일어날 수 있다고 가정하고 노출 위험의 측도로서 공개된 자료의 보조식별변수의 조합에 의해 생성된 경우의 수를 이용하여 전체 자료 중 신분노출이 일어날 수 있는 자료의 비율을 확률적으로 추정하는 방법을 제안하였다.

따라서 이러한 위험의 측도를 이용하는 것은 결과적으로 공개하는 보조식별변수의 개수와 범주의 수를 결정하는 방법과 동일하다. 이러한 방법은 직관적이며 사용이 간편하고 또한 앞에서 살펴본 통계적 노출제한 방법의 적용 범위를 결정하기 쉽다.

Duncan과 Lambert (1989)은 신분노출의 위험에 대한 측도를 의사결정이론 (decision theory)의 개념을 적용하여 제안하였다. 공개하기 전의 원자료를 X라고 하고 자료 공격자 (intruder)가 자료에 속한 특정한 개인 ($t_0 \in X$)의 신분를 알아내려고 할 때 노출제한방법이 적용된 공개 자료 Y에 기반한 예측분포 (predictive distribution) $p_Y(s)$를 계산하여 사용한다고 가정하자. 여기서 공개 자료 Y는 원자료 X의 부분집합이며 포함된 개체의 수는 n이라고 가정한다. 공개자료 Y를 보고 자료에 속한 임의의 개체 s가 공격자가 찾는 개인 t_0와 같은 예측확률을 $p_Y(s)$라 한다. 더 나아가 의사결정 이론에서 사용되는 손실함수 $L(t,s)$를 고려한다. 손실함수 $L(t,s)$는 공격자가 찾는 개인을 t로 결정하였을 때 발생하는 손실이다. 공격자는 실제 목표인 s를 알 수 없으므로 예측분포를 이용한 평균손실함수를 사용한다.

$$\int L(t,s)p_Y(s)\ ds$$

공격자가 선택할 수 있는 최적의 선택 t는 위의 평균손실함수를 최소화하는 선택 (불확실성을 최소화)이며 따라서 공격자가 선택을 하였을 때 가지는 손실 (불확실성)은 다음과 같다.

$$U(Y) = \inf_t \int L(t,s)p_Y(s)\ ds$$

예를 들어 자료의 이용자가 찾고자 하는 특정한 개인이 x_0라 하자. 자료 공격자의 선택은 x_0가 공개자료에 포함되지 않다고 판단하여 연결을 포기하는 경우 ($link = \emptyset$)와 공개자료에 속한 임의의 개인을 선택하는 경우 ($link = y_i$)가 있다. 이러한 경우 손실함수를 다음과 같이 정의할 수 있다.

$$L(x_0, link) = \begin{cases} 0 & \text{if } link = y_i = x_0 \in Y \text{ or } link = \emptyset,\ x_0 \notin Y \\ l_1 & \text{if } link = \emptyset,\ \text{but } x_0 \in Y \\ l_2 & \text{if } link = y_i \text{ for some } i,\ \text{but } y_i \neq x_0 \end{cases}$$

위와 같은 손실함수에서 자료의 이용자의 최소평균손실, 즉 불확실성은 다음과 같다

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

위의 식에서 만약 x_0가 공개자료에 포함되지 않다고 판단하여 연결을 포기하는 경우 특정한 개인 x_0가 공개자료에 포함될 예측확률이 $\sum_{i=1}^n p(y_i)$이므로 손실은 $l_1 \sum_{i=1}^n p(y_i)$이다. 또한 특정한 개인 x_0가 공개자료에 포함되었다고 판단하여 그 중 가장 가능성이 높은 개인을 선택했을 때 목표한 개인과 일치하지 않아 발생하는 손실은 $l_2[1 - \max p(y_i)]$이다. 따라서 만약 $l_1 \sum_{i=1}^n p(y_i) > l_2[1 - \max p(y_i)]$이면 공개자료 중에 가장 가능성이 높은 개인을 선택하고 반대로 $l_1 \sum_{i=1}^n p(y_i) < l_2[1 - \max p(y_i)]$이면 연결을 포기하는 결정을 내리게 된다.

위와 같은 의사결정의 틀에서 자료를 공개하는 기관이 $\sum_{i=1}^n p(y_i)$와 $\max p(y_i)$를 동시에 작게 하는 방법을 적용한다면 자료 공격자의 의지를 약하게 하여 개인정보 누출의 가능성을 낮출 수 있다. 이러한 의사결정에 기반한 개인정보 누출의 위험측도을 이용하여 Reiter (2005)는 정보보호 방법들을 비교하였다.

Marsh 등 (1991)은 원자료의 부분집합으로 익명화된 자료를 공개할 때 모집단 유일성 (population uniqueness; PU)의 개념을 소개하였다. 일련의 가정하에서 개인 정보가 노출되었을 경우 노출된 개인

이 모집단에서 유일하게 존재할 조건부 확률을 노출위험의 측도로 고려하였다. Skinner 등 (1994)은 모집단 유일성뿐만 아니라 표본 유일성 (sample uniqueness; SU)의 개념을 고려하였다. 노출위험의 측도로서 $P(PU|SU)$, 즉 표본에서 유일한 개인이라는 조건하에서 그것이 모집단의 유일한 개체가 될 조건부 확률을 위험의 측도로 제안하였다. 하지만 자료의 이용자는 공개 자료로부터 표본 유일성만을 알 수 있기 때문에 실제로 이러한 조건부 확률은 외부의 정보 또는 통계적 추론에 의해 수량화 될 수 있는 한계가 있다. Skinner와 Elliot (2002)과 Skinner와 Shlomo (2008)은 공개자료에서 표본 유일성이 발생했을 때 실제로 모집단 유일성일 비율을 노출 위험의 측도로 제안하고 이를 추정할 수 있는 통계적 추론을 제시하였다.

4.2. k-익명성과 l-다양성

Sweeny (2002)는 Dalenius (1986)에서 논의된 보조 식별변수의 조합을 이용한 정보노출의 위험성을 k-익명성 (k-anonymity)으로 구체화하였다. 공개 마이크로 자료에 흔히 포함되는 보조 식별변수는 성별, 나이, 직업, 거주 지역 등으로 이러한 변수들의 가능한 조합을 고려할 때 각 조합에 속하는 개체들의 수가 유일하다면 노출의 위험성이 증가한다. k-익명성은 보조 식별변수의 조합에 속하는 개체들의 수가 최소한 k개 이상으로 하는 기준을 의미한다. Sweeny (2002)는 선거 기록과 보험자료를 결합하여 미국 메사츄세스 주지사의 의료기록을 알아내는 실제 사례를 보여주었다. 일반적으로 적절한 코딩과 자료 숨기기를 통하여 k-익명성을 보장하는 공개 자료를 만들 수 있다. Table 4.1은 공개된 보조 식별변수가 성별과 거주 지역인 경우 3-익명성을 만족하는 자료를 나타낸다. 즉 성별과 거주 지역의 조합을 고려했을 때 해당하는 개체의 수가 최소한 3명임을 알 수 있다. Table 4.1에서 (A)의 경우는 3-익명성을 만족하며 진단에 대한 노출도 일어나지 않는다. 반면 Table 4.1에서 (B)의 경우는 3-익명성을 만족하지만 개인 정보의 누출이 발생하는 것을 알 수 있다. (B)에서 남자이며 서울에 거주하는 사람은 3명이지만 진단명은 모두 간염이다. 이렇게 3명에 대한 병명이 동일하므로 서울에 거주하는 남자라면 진단명이 간염임을 알 수 있다. 따라서 이런 상황은 비록 신분노출이 일어나지 않을 수 있으나 자료의 이용자는 특정한 보조 식별변수를 가지는 사람들에 대한 민감한 정보를 알 수 있다. 또한 자료 공격자가 외부 정보를 통하여 부산에 사는 남자 중 한 명은 유일하게 당뇨를 가지고 있는 사람의 신분을 알고 있다면 주어진 자료와 연결하여 나머지 부산에 사는 다른 남자들이 간염을 가지고 있다고 추론할 수 있다. 이러한 상황 또한 신분노출이 일어나지 않지만 개인의 민감한 정보는 누출될 수 있다.

위에서 논의된 k-익명성의 단점을 보완하기 위하여 Machannavajjhala 등 (2007)은 l-다양성 (l-diversity)라는 개념을 제안하였다. l-다양성은 식별변수의 조합이 동일한 집단 안에서 특성의 종류가 다양하여, 예를 들어 l개 이상이어서, 집단내의 개인들이 적절하게 분리되어 나타나는 기준을 말한다. Table 4.2 은 공개된 보조 식별변수가 성별과 거주 지역인 경우 3-익명성과 3-다양성을 만족하는 자료를 나타낸다.

Li 등 (2007)은 l-다양성의 단점에 대하여 지적하고 t-근접성 (t-closeness)의 개념을 제안하였다. Li 등 (2007)은 비록 l-다양성이 만족되더라도 정보의 노출이 일어날 수 있는 두 가지 상황을 고려하였다. 첫 번째는 같은 집단에 속하는 개체들의 특성이 전체 자료의 개체들이 가지고 있는 특성과 매우 다르다면 해당 집단에 속한 개체들의 정보가 쉽게 노출될 수 있다는 점을 지적하였다 (disclosure by skewness). 두 번째로 같은 집단내의 개체들이 특성이 다르지만 유사한 특징을 가지고 있다면 집단내에 속한 개인들에 대한 정보를 추론할 수 있어 정보 누출의 위험성이 존재한다는 점을 지적하였다 (disclosure by similarity) Table 4.2의 (B)에서 부산에 거주하는 남성들의 진단병이 소화불량, 위염, 위궤양으로 위장에 대한 질병을 가지고 있다는 유사성을 추론할 수 있으므로 개인 정보 노출의 가능성이 크다. 이러한 단점을 보완하기 위하여 식별변수의 조합으로 만들어진 각 집단이 가지고 있는 특성의 분포가 전체 집단의 분포와 크게 다르지 않아야 한다는 것이 t-근접성 (t-closeness)의 개념이다.

Table 4.1 An example for k-anonymity

(A) 3-anonymity			(B) 3-anonymity but disclosure		
sex	region	diagnosed by	sex	region	diagnosed by
male	Seoul	hepatitis	male	Seoul	hepatitis
male	Seoul	diabetes	male	Seoul	hepatitis
male	Seoul	hypertension	male	Seoul	hepatitis
female	Seoul	diabetes	female	Seoul	diabetes
female	Seoul	diabetes	female	Seoul	diabetes
female	Seoul	hypertension	female	Seoul	hypertension
female	Seoul	hyperlipidemia	female	Seoul	hyperlipidemia
male	Busan	hepatitis	male	Busan	hepatitis
male	Busan	diabetes	male	Busan	diabetes
male	Busan	hypertension	male	Busan	hepatitis

Table 4.2 An example for k-anonymity and l-diversity

(A) 3-anonymity and 3-diversity			(B) 3-anonymity과 3-diversity but disclosure		
sex	region	diagnosed by	sex	region	diagnosed by
male	Seoul	hepatitis	male	Seoul	hepatitis
male	Seoul	diabetes	male	Seoul	gastritis
male	Seoul	hypertension	male	Seoul	hypertension
female	Seoul	diabetes	female	Seoul	diabetes
female	Seoul	diabetes	female	Seoul	diabetes
female	Seoul	hypertension	female	Seoul	hypertension
female	Seoul	hyperlipidemia	female	Seoul	hyperlipidemia
male	Busan	hepatitis	male	Busan	indigestion
male	Busan	diabetes	male	Busan	gastritis
male	Busan	hepatitis	male	Busan	ulcer

4.3. 차등 정보보호

마이크로자료를 공개하는 경우 최근까지 보조 식별변수와 외부 자료의 결합 (matching 또는 linkage)으로 신분노출이나 특성노출이 생기는 사고가 대부분의 경우였다. 대표적인 예가 미국의 온라인 비디오 대여 서비스를 제공하는 넷플릭스 (Netflix)에서 공개한 회원들의 영화에 대한 평가 자료를 온라인 영화평가 서비스를 제공하는 IMBD의 회원 자료와 결합하여 IMDB 회원들의 민감한 정보가 노출된 사건이다 (Narayanan와 Shmatikov, 2007). 최근까지 정보의 노출은 이러한 자료의 결합을 통하여 발생하였지만 최근에 공개되는 자료와 정보가 많아지고 통계분석 서버의 이용이 증가하면서 자료 공격의 형태가 변하고 있다. 더 나아가 빅데이타가 출현한 환경에서는 추정노출과 모형노출의 가능성이 상대적으로 높아질 것으로 예상된다. 예를 들어 개인의 인구 경제학적 특성을 알고 있고 여러 공개된 자료를 이용하여 소득에 대한 제법 정확한 회귀모형을 추정할 수 있다면 개인의 소득을 모형을 통하여 예측할 수 있다. Dwork (2006)은 다음과 같은 예로 개인이 공개 자료에 포함되지 않더라도 정보가 노출될 수 있는 가능성이 있다는 것을 주장하였다. “Terry Gross는 리투아니아 (Lithuanian) 여자의 평균키보다 2인치 작다” 라는 사실을 알고 있다고 하자. 어떤 자료가 국가별로 남여 평균키에 대한 정보를 줄 수 있다면 Terry Gross의 정확한 키는 노출된다. 공개되는 자료가 증가하고 빅데이터와 같이 자료의 크기가 커지는 경우 전통적인 노출의 도구인 자료 연결 (data matching)과 함께 추론과 모형을 통한 노출도 증가할 수 있다. 따라서 추론과 모형을 통한 노출을 기반으로 하는 노출 위험의 측도가 필요하다.

Dalenius (1977)은 통계적 노출제한 기법의 목표를 다음과 같이 정의하였다.

"자료를 통하지 않으면 알 수 없는 개인의 정보를 자료를 통하여 알아내는 것을 방지한다" ("Access to statistical database should not enable one to learn anything about an individual that could not be learned without access to the database").

Dwork (2006)은 Dalenius의 목표가 실현될 수 없는 명제임을 보이고 차등 정보보호 (differential privacy)라는 개념을 제안하였다. 차등 정보보호의 개념은 한 개의 개체가 자료에 추가로 포함될 때 증가하는 위험을 측정하는 것이다. 따라서 차등 정보보호는 자료의 전체를 보호하는 의미에서 위험을 측정하기 보다는 자료가 변화함에 따라 증가하는 위험을 상대적으로 측정하고 이를 제어하는 방법이라고 할 수 있다. 차등 정보보호의 확률적 정의는 다음과 같다.

정의 4.1 데이터베이스 D_1과 D_2를 고려하고 두 데이터베이스는 자료에 포함된 개체 수의 차이가 한 개라고 하자. 임의의 노출제한 방법을 확률함수 K라고 하면 $K(D)$는 노출제한 방법이 적용된 공개 데이터베이스이다. 이러한 가정하에서 ϵ-차등 정보보호 (ϵ-differential privacy)는 다음과 같은 조건을 만족하는 것이다.

$$P[K(D_1) \in S] \leq \exp(\epsilon) \times P[K(D_2) \in S] \text{ for all } S \subseteq Range(K)$$

위의 ϵ-차등 정보보호의 기준은 마치 통계학에서 자주 사용되는 우도비 (likelihood ratio)형태로 표현할 수 있다.

$$\log \frac{P[K(D_1) \in S]}{P[K(D_2) \in S]} \leq \epsilon \text{ for all } S \subseteq Range(K)$$

위에서 정의된 ϵ-차등 정보보호는 공개된 자료에서 하나의 개인 또는 개체가 제외되어도 자료로 부터 얻은 정보가 유의하게 변하지 않는다는 것을 의미한다. ϵ-차등 정보보호는 매우 강한 정도의 정보 보호를 의미하며 절대적인 개념이 아니라 자료의 크기의 차이에서 발생하는 노출위험의 변화를 측정하는 상대적인 개념이다. 또한 외부 자료의 유무나 가용한 계산 능력을 고려하지 않아도 되는 개념이다.

Dwork (2006)은 ϵ-차등 정보보호의 구현을 위하여 데이타베이스의 쿼리에 대한 민감도 (sensitivity)를 정의하였다. 데이터베이스 D에 대한 쿼리를 함수 f로 정의한다면 $f(D)$는 쿼리에 대한 결과이다.

$$f : D \rightarrow R^k$$

만약 쿼리의 결과가 히스토그램이라고 한다면 $f(D) = (n_1, n_2, ..n_k)$이다. 이때 n_i는 각 구간에 속하는 도수 (frequency) 또는 상대도수이다. 이러한 가정하에서 쿼리 f의 민감도 Δf는 다음과 같이 정의된다.

정의 4.2 D_1과 D_2를 고려하고 두 데이터베이스는 자료에 포함된 개체 수의 차이가 한 개라고 하자.

$$\Delta f = \max_{D_1, D_2} \parallel f(D_1) - f(D_2) \parallel_1$$

여기서 $\parallel x \parallel_1$은 ℓ_1-거리 (norm)이며 벡터 x의 성분 중 최대값이다.

예를 들어 쿼리의 결과가 히스토그램이라고 한다면 민감도는 1이다 ($\Delta f = 1$). 왜냐하면 개체 한 개가 자료에 포함되는 경우 히스토그램은 첨가된 개체가 포함된 도수가 한 개 증가하고 나머지 도수는 변하지 않기 때문이다. 많은 경우에 민감도의 값은 크지 않지만 쿼리의 종류에 따라 그 크기가 상한 (upper bound)를 가지는 경우도 있고 아닌 경우도 있다.

쿼리의 민감도가 계산되었을 때 ϵ-차등 정보보호를 구현하는 방법은 쿼리의 결과에 표준편차의 크기가 $\sqrt{2}\Delta f/\epsilon$을 가지는 이중지수분포 (double exponential distribution; Laplace distribution)에서 생성

된 잡음을 첨가하는 것이다. 예를 들어 히스토그램에 잡음을 첨가할 경우 이중지수분포에서 발생된 k개의 독립인 잡음들 $(Z_1, Z_2, ..Z_k)$을 각 도수에 첨가하는 것이다.

$$\text{histogram output} = f(D) + int(Z_1, Z_2, ...Z_k) = (n_1, n_2, ..n_k) + int(Z_1, Z_2, ...Z_k)$$

이러한 결과는 만약 $f(D)$의 값이 이중지수분포를 따른다고 가정하고 ϵ-차등 정보보호의 기준에 계산하면 쉽게 유도할 수 있다. 이러한 잡음첨가 방법은 모든 쿼리에 적용될 수 있는 것은 아니다. 예를 들어 쿼리의 결과가 문자열, 그림 등이라면 잡음첨가를 할 수 없으며 다른 방법을 사용해야 한다. Nissim 등 (2007)은 ϵ-차등 정보보호가 너무 강한 기준이므로 이를 약화시킬 수 있는 (ϵ, δ)-차등 정보보호 ((ϵ, δ)-differential privacy) 기준을 제안하였다.

$$P[K(D_1) \in S] \leq \exp(\epsilon) \times P[K(D_2) \in S] + \delta \quad \text{for all } S \subseteq Range(K)$$

ϵ-차등 정보보호의 개념은 원격접속을 통하여 자료의 요약 통계를 제공해주는 통계분석 서버에 적합한 개념이다. 이는 전통적으로 공개된 마이크로 자료와 외부자료를 이용한 연결을 통하여 정보가 노출되는 상황과는 다른 조건에 더 적합하다. 사실상 마이크로 자료를 공개하는 경우 노출제한 방법을 적용하는 것은 매우 어려운 작업이다. 자료 자체의 특성, 가용한 외부 정보, 자료 공격자의 성향, 자료 유용성 등을 종합적으로 고려해야 하기 때문에 비용이 많이 들고 한번 자료가 공개되면 수정하기 어렵다. 지금까지는 통계분석 서버를 제공하는 경우 마이크로 자료의 공개보다는 정보 노출의 위험성이 작았다. 하지만 제공 항목이 다양해지고 제공 서비스가 증가하면 자료 공격자가 계획된 쿼리를 지속적으로 사용하여 정보 노출이 가능성이 높아진다. 또한 계획된 쿼리가 지능화하거나 자동화될 수 있는 가능성이 충분하기 때문에 추정유출이나 모형유출의 위험이 증가할 것이다. 이러한 환경에서 ϵ-차등 정보보호는 노출위험을 수량화할 수 있는 기준을 제공한다는 데 큰 의미가 있다. 하지만 실제로 이를 구현하려면 더 많은 연구가 필요하다.

5. 결론

본 논문에서는 마이크로 자료의 공개나 자료의 요약 정보를 제공하는 통계분석 서버에 사용되는 노출제한 방법과 노출위험에 대한 개념들을 개괄하여 살펴보았다. 현재는 과거와 달리 공공기관이나 기업이 보유하고 있는 방대한 자료를 이용하여 가공된 새로운 가치를 창출할 수 있다는 인식이 확대되고 있는 상황이다. 정보공개에 대한 요구와 더 나아가 빅테이타의 등장이라는 새로운 환경의 변화에 따라 정보의 공개의 범위와 사용 권한의 확대에 대한 요구가 급격히 증가하고 있으므로 더 많은 공공자료와 개인별 자료가 공개될 것으로 예상된다. 또한 빅테이타에 대한 관심이 커짐에 따라 빅데이타를 분석하거나 다른 정보와 결합하는 경우도 증가할 것으로 예상된다. 이러한 자료의 양적 질적 팽창에 따라서 개인 정보의 노출 위험도 커질 것이 분명하다. 또한 지금까지 개인정보가 유출된 경우와는 매우 다른 경로로 사생활이 침해받을 수 있다. 특히 통계분석 서버를 통한 정보제공 서비스에 대한 요구가 커지므로 개인정보에 대한 물리적인 보안뿐만아니라 생성된 결과에 기반한 추론에 의하여 개인정보가 유출될 수 있는 가능성에 대비한 노출제한 방법의 연구가 활발하게 이루어져야 할 것이다.

References

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, **15**, 429-444.

Dalenius, T. (1986). Finding a needle in a haystack or identifying anonymous census record. *Journal of Official Statistics*, **2**, 329-336.

Dalenius, T. and Reiss, S.P. (1978). Data-swapping: A technique for disclosure control. *Preceedings of the Section on Survey Research Methods, American Statistical Association*, Washington DC, USA, 191-194.

Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, **6**, 73-85.

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

Duncan, G. T., Elliot, M. and Slazar-Gonzalez, J. (2011). *Statistical confidentiality principles and practice statistics for social and behavioral sciences*, Springer, New York, NY, USA.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006)*, Springer, Venice, Italy, 1-12.

Fienberg, S. E. and McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In *PSD 2004, Lecture Notes on Computer Science,* edited by J. Domingo-Ferrer and V. Torra, Springer, New York, NY, USA, 14-29.

Gomatam, S., Karr, A. F., Reiter, J. P.and Sanil, A. P. (2008). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statistical Science*, **20**, 163-177.

Li, N., Li, T. and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE 2007, IEEE 23rd International Conference on Data Engineering*, 106-115.

Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, **1**, 3-20.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N. (1991). The case for samples of anonymized records from the 1991 census. *Journal of the Royal Statistical Society A*, **154**, 305-340.

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

Narayanan, A. and Shmatikov, V. (2007). How to break anonymity of the Netflix Prize dataset, preprint in http://arxiv.org/.

Nissim, K., Raskhodnikova, S. and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *STOC 07, Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, ACM, New York, NY, USA, 75-84.

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

Reiss, S. P. (1984). Practical data-swapping: The first step. *ACM Transactions on database systems*, **9**, 20-37.

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

Skinner, C. J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, **10**, 31-51.

Skinner, C. J. and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society B*, **64**, 855-867.

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

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, **10**, 557-570.

Journal of the Korean Data & Information Science Society
2013, **24**(5), 1029–1041

http://dx.doi.org/10.7465/jkdi.2013.24.5.1029
한국데이터정보과학회지

Review on statistical methods for protecting privacy and measuring risk of disclosure when releasing information for public use

Yonghee Lee[1]

[1]Department of Statistics, University of Seoul

Received 3 July 2013, revised 23 August 2013, accepted 4 September 2013

Abstract

Recently, along with emergence of big data, there are incresing demands for releasing information and micro data for public use so that protecting privacy and measuring risk of disclosure for released database become important issues in goverment and business sector as well as academic community. This paper reviews statistical methods for protecting privacy and measuring risk of disclosure when micro data or data analysis sever is released for public use.

Keywords: Big data, database, disclosure limitation methods, micro data, privacy protection, risk of disclosure, statistical analysis sever.

[1] Associate professor, Department of Statistics, University of Seoul, Seoul 712-749, Korea.
E-mail: ylee@uos.ac.kr