【 2009년 통계자료의 비밀보호에 대한 】

국제공동회의 출장결과 보고

2009. 12.

통 계 청

목 차

Ⅰ 출장개요

1. 출장목적

○ 우리청의 마이크로데이터 서비스 및 SDC기법 연구를 전문가에게 소개하고 이에 대한 의견청취를 통해 향후 나가야할 방향에 대한 정보를 얻고자 함

※ 논문(2개) 제출

① 한국 통계청의 마이크로데이터 제공에 대한 쟁점
- 작성자 : 신윤수, 백동훈

② 매크로데이터의 민감한 셀 결정방법에 대한 사례연구
- 작성자 : 김경미

2. 출장기간

○ 2009 12. 1 ~ 12. 6(4박 6일간)

3. 출장자 및 출장지역

소 속	직 급	성 명	출장지역
통계정보국 전산개발과 통계개발원	국장 통계주무관 통계주무관	신 윤 수 백 동 훈 김 경 미	스페인 (빌바오)

4. 주요일정

일 시	방 문 지	비고
- 12월 1일	서울(인천공항) → 스페인(빌바오)	
- 12월 2일~12월 4일	회의참석	
- 12월 5일~12월 6일	스페인(빌바오) → 서울(인천공항)	

Ⅱ 회의개요 및 내용

1. 회의개요

⑴ 회의명

○ 통계자료 비밀보호에 대한 국제공동회의

⑵ 회의목적

○ 통계자료 비밀보호를 위한 기법 소개

○ 실제 자료를 활용한 비밀보호 적용 사례 및 기법 소개

○ 통계이용자의 자료활용 및 통계작성기관의 자료제공에 대한 토론 등

⑶ 회의일시 : 2009. 12. 2~ 12. 4(3일간), 스페인 빌바오

⑷ 회의주기 : 2년 주기로 개최

⑸ 참가자 : 88명

○ 24개국 69명

○ UNECE 등 국제기구 19명

2. 주요 정책적 시사점

⑴ 마이크로데이터 등 통계활용을 위한 적극 지원할 필요

○ 통계비밀보호는 응답자의 사생활(privacy) 보호 차원에서가 아니라 자료 활용을 최대화한다는 차원으로 발상의 전환 필요

○ 우리의 마이크로데이터 접근(Access to Microdata)이라는 소극적인 용어 보다는 연구개발지원(Research Data) 등 적극적 용어 사용 등 인식전환

※ 통계자료제공심의회에 민간인을 참여(공무원, 민간인을 반반)시키고 위원장을 민간위원 중에서 호선하는 방안 검토

○ 연구자의 연구를 위해 적극적 자료지원을 하되, 산출물을 엄격히 체크하고, 비밀보호문제 등 위반사례 발생시 엄격히 처벌하는 방안 검토(2007년에 영국은 법 개정으로 대응)

※ 통계법의 비밀보호조항 세분화 및 위반시 처벌강화

⑵ 비밀보호기법에 대한 지속·체계적 연구 필요

○ 자료의 제공범위를 확대하고 그 활용도를 높이기 위해 선진국의 비밀보호기법에 대한 지속적인 연구 필요

○ 미국과 유럽을 중심으로 다양한 비밀보호기법이 개발되고 있고 현재 개발된 기법들의 장단점에 대한 토론이 활발함, 우리청의 각각의 통계에 적용할 수 있는 기법에 대한 체계적 연구 필요

⑶ 최신 국제동향을 면밀히 파악하여 정책에 반영필요

○ 우리나라도 매년 여러 국제회의에 참가하고 있으나, 회의결과의 정리 및 정책활용이 부족

○ 현재 일본, 호주 주재관이 있으나, 국제사회에서는 UNECE 통계국의 활용도가 큰 것으로 보여 주재관 파견 검토 필요

⑷ 기타

○ 미국은『2000 센서스』에서는 short, long form을 모두 직접 조사, 그러나『2010 센서스』는 short form은 10년 주기 센서스로 유지하나, long form은 5년주기 ACS(American Community Survey)로 전환, 관련 예산은 종전대로 U.S Census Bureau가 부담

○ MSIS 국제회의 관련, UNECE 담당자(Steven Vale, Diane Sericoff)와 접촉한 바, 현재 UNESCAP와 Asia Session Organizing에 대한 조율문제가 있고,『2010 MSIS 한국회의』에 대해 우즈벡, 키르키스탄, 몽골 등이 재정적 어려움을 토로하였다 함

○ 국내 통계작성기관에게 통계비밀보호기법 전파 필요

3. 향후 계획

○ 통계법 개정 검토

- 통계를 이용자가 효율적으로 이용할 수 있는 기반을 마련해 주고, 이를 위반할 때 강력하게 제재할 수 있도록 통계법 개정 필요

○ 비밀보호기법 연구 및 적용

- 해외에서 개발·적용하고 있는 비밀보호기법에 대해 연구하고, 향후 우리청 자료에 적용할 수 있는 방안 모색
- 완료되지 않은 논문내용(Working Paper 20, 22)에 대해 모니터링 및 결과물에 대한 검토

○ 마이크로데이터 이용자에 대한 교육 실시

Ⅲ 세션별 발표내용

(1) 세션 1 : 통계자료 비밀보호의 조화 - 법과 방법론적인 측면 (Harmonization of statistical data confidentiality-legal and methodological aspects)

Working Paper 2	데이터 통합 관점에서 비밀보호에 대한 원칙 및 지침 (Principles and guidelines on confidentiality aspects of data integration)
- 2009년 6월 유럽통계인 회의(CES)에서 채택된 “통계 또는 관련 연구목적을 위한 데이터통합 시 비밀보호 측면에 관한 원칙과 지침”(부록 2. 참고)을 소개 - 이 지침은 데이터 통합시 비밀보호측면(원칙6)을 기본으로 하여 원칙과 지침을 정함 ※ 부록 1.1. 참고	
Working Paper 3	EU 수준에서 통계 비밀보호 방법의 조화 (Harmonization of statistical confidentiality methods as the EU level)
- 유럽통계시스템(ESS)에서 통계 비밀보호에 대한 방법적인 조화에 대해 논의하였으며, ESS 통계의 보급은 SDC와 관행의 조화부족으로 방해가 되고 있음. 이러한 것들을 법률, 행정, 방법론적인 측면에서 검토	
Working Paper 7	부분 합성파일을 통한 익명화의 조화 (Harmonization of anonymization practices through partially synthetic files)
- 유럽 여러나라의 자료통합할 때 나라별로 제공범위가 다르므로 어려움이 있음. 이것의 대안으로 국가의 데이터를 부분적으로 합성하는 절차에 대해 소개	
Working Paper 5	독일 연방정부의 통계 비밀보호의 조화 (Harmonization of statistical confidentiality in the Federal Republic of Germany)
- 독일의 연방 통계사무소와 지방 통계사무소 사이의 통계 비밀보호 조화 방법에 대해 설명하고 있다. 마이크로데이터의 익명화에 대한 지침뿐만 아니라 결과물의 비밀보호, 결과물 검토, 통계표의 비밀보호에 대해 논의함 ※ 부록 1.2. 참고	

Working Paper 8	마이크로데이터를 기본으로한 조화된 노동력과 이주통계 (Harmonised Labour Force and Migration Statistics Based on Microdata)
- OECD 국가의 마이크로데이터를 기본으로 한 노동력과 이주통계의 조화에 대한 프로젝트를 설명하고 있으며, 잠재적으로 발생할 수 있는 법과 기술적인 문제를 논의함	
Working Paper 6	한국 통계청의 마이크로데이터 제공에 대한 최근 쟁점 (The current Issue on Microdata Services in Statistics Korea)
- 한국통계청의 마이크로데이터 서비스 방법을 소개하고 - 최근 일부 이용자의 요구사항, 대책, 고민 등을 기술	

⑵ 세션 2 : 합성 자료와 하이브리드 자료
(Synthetic and hybrid data)

Working Paper 10	독일 IAB 설립 패널에 대한 합성데이터셋 (Synthetic datasets for the German IAB Establishment Panel)
- 독일에서 부분 합성자료를 이용하여 새롭게 자료를 생성하였으며, 원본과 새로 생성한 자료 간의 신뢰구간을 비교함	
Working Paper 11	집계자료를 기본으로한 숫자형태의 하이브리트 데이터 (Microaggregation -based numerical hybrid data)
- 원본자료와 합성자료를 혼합하여 만드는 하이브리드 자료에 대해 설명함	
Working Paper 13	마이크로데이터 보호에 편비제약조건 다루기 (Dealing with edit constraints in microdata protection : microaggregation)
- microaggregation에서 편집제약조건을 어떻게 효율적으로 처리할 수 있는지 논의한다. 다양한 편집제약조건을 살펴보고, 이런 제약조건을 수용하는 마이크로데이터집계(microaggregation)를 살펴봄	

(3) 세션 3 : 연구자료센터와 가상 실험실
(Research data centres and virtual labs)

Working Paper 18	영국 기밀데이터 서비스 : 잠제적인 기밀보호데이터 사용에 대한 설명과 도전(UK Secure Data Service: specifications and challenges of using potentially disclosive data)
- 데이터서비스 제공자에게 직면한 몇 가지 도전과 제안된 사양서의 새로운 UKDA를 소개함 - 데이터의 비밀보호는 안전한 사람, 프로젝트, 이용환경 및 결과물에 의해 결정 - 이용자에 대한 교육, 훈련을 실시하고 있으며, 비밀보호에 대해 불이행 시에는 법적으로 강력하게 제재를 받음 ※ 부록 1.3. 참고	
Working Paper 19	캐나다 통계청의 실시간 원격접근 인프라 개발 (Development of a real time remote access infrastructure at Statistics Canada, Invited paper)
- 캐나다는 마이크로데이터 접근 향상을 위해 2007.1월에 연구그룹을 설립하였으며, 실시간 원격접근(RTAT) 구축에 대해 아래 3단계에 걸쳐서 진행하고 있음 · 1단계 : 사업요구 사항 정의 · 2단계 : 보안요구 사항 정의 · 3단계 : 프로그램 테스트 - 현재 1단계가 완료되고 2단계를 진행 중이며 이 논문에서는 이에 대한 진행상황을 기술함 ※ 부록 1.4. 참고	
Working Paper 20	결과물 검토에 대한 가이드라인(Guidelines for output checking)
- 결과물 검토에 있어서 에러유형을 아래와 같이 2가지로 구분 · 비밀보호 에러(Confidentiality Errors) : 불안전한 결과물 제공 · 비능률적 에러(Inefficiency Errors) : 안전한 결과물 미제공 - 결과물 검토에 대한 2가지 모델 소개하고 장단점을 비교 · 원칙기본모델 : 비밀보호 에러와 비능률적 에러 최소화 · 경험적 규칙모델 : 비밀보호 에러 최소화 - 결과물에 대한 내용과 질적인 문제는 연구자에게 책임이 있으며 국가통계기관은 노출위험만 관리함 - 결과물에 대한 검토자는 최소한 1명 이상이 있어야함 ※ 부록 1.5. 참고	

Working Paper 15	효과적인 연구자 관리(Effective researcher management)
- 자료 제공자와 연구자에게 적용되는 데이터 비밀보호에 대한 협력적인 접근방법에 대해 논의함. 특히 훈련에 대해 강조를 하고 있음	
Working Paper 16	정보 인프라의 향상 : 독일의 원격접근에 대한 방법 (Improvement of the informational infrastructure - On the way to remote data access in Germany)
- 독일은 원격접근 방식을 서비스할 목적으로 시행한 프로젝트에 대해 소개하고 있음	
Working Paper 17	ESSnet 프로젝트 - EU 마이크로데이터를 위한 분산된 접근 (The ESSnet project - Decentralized access to EU microdata sets)
- 법, 기술, 행정적으로 실현 가능한 범위에서 EU 각 나라의 마이크로데이터를 접근할 수 있는 방법에 대해 논의함	

⑷ 세션 4 : 도구와 소프트웨어의 개선 (Tools and software improvement)

Working Paper 22	데이터이용자가 사용하는 센서스 표를 자동으로 비밀보호하는 방법의 도구화 (Implementing a method for automatically protecting user-defined Census tables)
- 반복쿼리의 문제에 대한 해결책, 셀 키의 이용방식, 추가 모듈의 적용 등에 관한 내용을 중점적으로 다룸 ※ 부록 1.6. 참고	
Working Paper 23	재현성 있는 마이크로데이터 보호를 위한 대화형 그래픽 사용자 인터페이스 (An interactive graphical user interface for microdata protection which allows reproducibility)
- R progarm을 이용한 오픈소스 프로젝트를 구현, 마이크로데이터의 변조(perturbation)기법 중심으로 함.	
Working Paper 24	통계적 정보노출 제한을 위한 오픈 소스 소프트웨어 (On open source software for statistical disclosure limitation)
- 오픈소스, 프리웨어의 기본 개념을 정의 - 여러 소프트웨어를 비교하여 사례연구	
Working Paper 25	L1 CTA를 위한 패키지 (A package for L1 controlled tabular adjustment)
- 매크로데이터의 비밀보호기법 중의 하나인 CTA(Cotrolled Tabular Adjustment)방법을 위한 응용프로그램의 논의	

⑸ 세션 5 : 다음 센서스를 위한 통계 비밀보호 방법
(Statistical disclosure control methods for the next census round)

Working Paper 29	2010년 센서스와 미국 커뮤니티조사(5년)의 집계자료에 대한 공개 회피 기술 (Disclosure avoidance techniques for Census 2010 and American Community Survey Five-year Tabular Data Products)
- 이 논문은『2010년 센서스』및 미국 커뮤너티조사(ACS)의 자료에서 사용하는 통계적 노출 회피 기술을 설명한다. 또한, 많은 통계표는 소규모의 지역에 대해 공개하며, 통계적 노출회피 기술은 swapping, rounding, collapsing categories, applying thresholds, table suppression과 합성데이터 등이 포함 - 미국은『2000 센서스』에서는 short, long form을 모두 직접 조사, 그러나『2010 센서스』는 short form은 10년 주기 센서스로 유지하나, long form은 5년주기 ACS(American Community Survey)로 전환, 관련 예산은 종전대로 U.S Census Bureau가 부담	
Working Paper 27	균형있는 위험과 유용 - 2011년 영국 센서스에 대한 통계적 노출기법(Balancing risk and utility - statistical disclosure control for the 2011 UK Census)
- 2011년 영국 센서스에 적용될 통계적 노출기법이 개발되었으며, 2001년 영국 센서스와 다른 레코드 교환방법을 적용	

⑹ 세션 6 : 사례연구
(Case studies)

Working Paper 32	합성데이터 구조 파일: 개발 및 정보노출제한 (Synthetic data structure files: development and disclosure control)
- multiple imputation을 이용한 합성데이터를 생성하는 방법에 대한 사례연구 - IVEware 패키지를 사용하여 수행, 레코드 연결 방법과 관련한 노출위험도 고려 ※ 부록 1.7. 참고	

Working Paper 32	위험 측정을 위한 분할표의 스무딩 (Smoothing contingency tables to estimate a global risk measure)
- 마이크로데이터 파일의 재식별 위험에 대해 측정, 패널티가 가해지는 최대 우도적(maximum likelihood) 접근에 관해 설명함 - 주로 2차원 분할표의 스무딩의 문제를 다룸, 추가적으로 더 높은 차원으로 확장가능 - 이탈리아의 인구센서스 및 노동력조사에 단계적으로 적용	
Working Paper 34	EUSTAT의 가정환경조사에서 범주형 변수의 목표 다변량 관계를 보호하는 마이크로데이터 변조(Microdata perturbation preserving multivariate relations of target categorical variables in the household environmental survey of EUSTAT)
- 스페인 통계청의 경우, 좀 더 심층적인 분석을 위해 데이터 뱅크와 마이크로데이터의 사용을 원하는 전문적인 이용자가 증가 - 현행 방법은 식별위험이 높은 변수의 범주를 좀 더 넓게 조정하고 지역적인 공표단위를 상향해 작은 지역 단위의 공표를 제한하는 방법 등의 전통적인 방법 - 범주형 변수에 변조기법(연속형 데이터의 비밀보호기법 중의 하나로 특정분포의 값을 원데이터에 가감하는 등의 방법으로 데이터의 값을 조정하는 방법)을 적용한 새로운 접근방법에 대한 사례연구 - 변수의 종속성을 확인하고 조합할 변수의 선택을 변조 - Mu-Argus 소프트웨어를 이용한 PRAM방법, SAS의 임의할당 함수로 실행	
Working Paper 32	민감한 셀 결정방법 적용사례 (Example of application of linear sensitivity rules in Korean statistical data)
- 민감한 셀 결정방법을 한국통계청의 광업·제조업조사에 적용한 사례연구 - (n,k) dominance rule, p% rule, p/q ambiguity rule을 적용하고 장단점 및 적용가능성 분석	

⑺ 세션 7 : 통계적비밀보호방법의 장단점 분석 및 새로운 방향 (Risk/benefit analysis and new directions for statistical disclosure limitation)

Working Paper 39	통계표데이터의 기밀성을 보호하기 위한 향상된 프레임워크 및 결정 시스템(An enhanced framework and decision system for protecting the confidentiality of tabular data)
- 공공이용파일의 정보노출위험을 완화할 수 있는(그 중에서도 매크로 데이터) 실용적인 프레임워크를 개발 - 그러한 프레임워크를 OptShield라고 명명 - 마이크로데이터 수준에서 민감한 레코드의 최적 스위칭, 변조, 감추기 방법 등을 결합하여 데이터의 무결성을 유지하면서 비밀을 보호하는 방법을 얻고자 함	
Working Paper 40	통계적 정보노출사건을 이해하기 위한 게임 이론의 응용 (An application of game theory to understanding statistical disclosure events)
- 통계적 노출 이벤트(그것이 어떻게 발생하는지를 포함해)를 이해하기 위한 게임이론의 적용과 그것의 결과	
Working Paper 41	통계적 정보노출제한에서 투명성의 역할 (The role transparency in statistical disclosure limitation)
- 통계적 비밀보호기법에서의 투명성의 역할 즉, 적용된 비밀보호기법의 과정이 알려짐으로써 주어진 데이터로 원래의 데이터를 추정할 수 있는 가능성에 대해 논의 - 임의 데이터 교환이 이를 설명하는데 사용 - 이 과정에서 합법적 데이터 사용자와 악의적 목적을 가진 침입자(intruder)를 수학적으로 구분	
Working Paper 42	수치적 데이터 비밀보호 테크닉에 관한 유사성 공동 지표 (A common index of similarity for numerical data masking techniques)
- 수치적 데이터의 비밀보호기법은 기본적인 잡음추가를 통한 방법에서부터 시작하여 상관관계 잡음, 일반적인 가법적 변조, 다중적 무응답 대체, 데이터교환 등의 여러 가지 방법으로 발전 - 이러한 많은 모형들을 기반으로 비밀보호된 데이터와 원데이터 사이의 유사성을 비교하는 것은 점점 힘들어지고 있음 - 이에 따라 동일한 기준으로 각각의 방법에 대한 평가를 하기위한 새로운 측정방법들의 발전이 필요 - 모든 비밀보호기법에 적용할 수 있는 Common Index of Similarity (CIS)를 개발 - 비밀보호기법이 적용된 데이터와 원데이터가 유사하지 않은 것을 0으로 하고 원데이터와 동일한 것을 1로 정했을 때, 각각의 방법을 비교하기 위한 비밀보호이전의 데이터를 0과 1의 범주로 측정하여 수치적 데이터를 비밀보호 하는 것이 그 기본 개념 ※ 부록 1.8. 참고	

부록 1. 주요논문(번역본)

부록 1.1. Working Paper 2

통계적 비밀보호에 대한 유럽 통계인 지침 회의

요약 :이 짧은 페이퍼는 2009년 6월 컨퍼런스 유럽 Statisticians(CES)에서 새롭게 채택된 "통계 또는 관련 연구목적을 위한 데이터통합 시 비밀보호 측면에 관한 원칙과 지침"을 소개한다. 이는 또한 CES와 다른 기관의 통계 데이터 비밀보호의 영역에서 이전의 표준 지침을 요약하고 있다.

1. 서론

2006년 유럽 통계인 컨퍼런스(CES)에서는 통합된 데이터셋과 관련하여 비밀보호와 프라이버시 우려를 검토하는 태스크포스를 설정할 것을 요구하였다. 이 호주 통계청에 의해 주도된 통계데이터통합의 비밀보호와 프라이버시 측면 관련 태스크포스는 통계 및 관련 연구 목적을 위한 통합 데이터셋의 생성과 이용과 관련된 법적 접근(assessing) 및 완화(mitigating), 기타 비밀보호 이슈 등에 대한 일반적 프레임워크를 수립하는 것을 목표로 일련의 원칙과 지침을 설정하였다.

이 통계 또는 관련 연구 목적의 데이터 통합 착수의 비밀보호 유지 측면에 대한 원칙과 지침은 2009년 CES 회의에서 승인되어, 이후 UNECE에 의해 발간되었다. 이 원칙과 실천 지침은 이제 업무에 적용되어 2011년 CES에서 재검토할 것이다.

2. 원칙 및 지침의 내용과 적용

간행물은 도입 부분에 원칙과 지침의 개발 배경을 설명하고 다른 국제 용어집과 일치되는 11가지 주된 정의를 담고 있다. 다음에는 일련의 세부 지침과 함께 8가지 원칙이 나오고, 마지막으로 업무 사례(business case) 개요가 부록으로 실려 있다.

이러한 원칙과 지침은 국가 통계 조직에서 수행하는 데이터 통합 작업에 적용된다. 데이터 통합은 다른 행정기관이나 조사 출처의 단위 레코드로부터 저작권을 갖고 공표할 수 있는 새로운 공식 통계를 편집/통합하는 것과 관련이 있다. 또한, 이 통합 데이터 셋은 기존 조사출처를 통해서는 불가능했던 경제 및 사회 연구의 범위를 지원하는 데 사용될 수 있다. 이 원칙과 지침은 통계 등록(register)의 작성 및 유지 관리에 관련성이 있지만, 이들 작업을 다루지는 않는다.

3 관련 UNECE 표준(standard) 및 간행물

3.1 통계 비밀보호 관리 및 Microdata 접근 - 바람직한 실행을 위한 원칙과 지침

이 지침은 Conference of European Statisticians (CES)의 요청으로 호주통계국에 의해 주도된 태스크 포스에 의해 마련되었다.
이 책자는 비밀 보호가 주된 국가적 이슈가 되어오고 있으나, 인터넷을 통해 증가하는 데이터 배포가 점차 국제적 이슈화 되는 것을 인식하고 있다. 연구자들은 국가마다 다른 액세스 규칙과 기준을 사용하는 것을 비판적으로 생각한다. 종종 연구자들은 비밀 우려 때문에 다른 나라의 마이크로데이터 접근이 불가하다. 국제기구도 특별히 다국적 비교를 위해 연구 목적의 마이크로데이터 사용에 관심이 증가하고 있다.

따라서 이 지침은 마이크로데이터 보급에 대한 몇 가지 일반적인 원칙을 제공하려고 한다.
(i) 연구 기관(공동체)에 의한 마이크로데이터 접근 편의를 위한 국가단위 접근의 동질성 촉진, 그리고
(ii) 국가가 마이크로데이터 접근을 제공하기 위한 그들의 준비(arrangement)를 개선할 수 있도록 사례 연구 지원

이 지침은 마이크로데이터 접근에 대한 세부적인 장치(arrangement)[기술, 준비]가 법률, 사회적 분위기, 연구 공동체를 지원하는 역량(capacity)에 따라 다른 것임을 알고 있다. 이 지침이 접근의 더 큰 동질성을 목표로 하지만, 각 나라가 완전히 동일한 조정(arrangement)을 할 것은 기대하지 않아야 한다.

3.2 통계적 비밀보호 및 마이크로데이터 액세스(2003), 2003 CES의 세미나 세션의 진행

2003년 CES 세션은 통계적 비밀보호와 마이크로데이터 접근의 이슈에 관련된 세미나를 포함하였다. 이 세미나는 CES국과 협동한 의장국인 스웨덴 통계청에 의해 조직되었다. 이 세미나는 매우 유익했던 것으로 평가되어, UNECE와 스웨덴 통계청은 세미나의 회보를 함께 발간하기로 결정했다.

통계적 조직(구성)을 위한 중요한 도전은 마이크로데이터에 대한 개선된 접근은 통계 비밀보호 원칙을 해치지 않는다는 것을 보장하는 것이다. 이 책자의 도입부는 세미나에서 나온 토론의 간략한 요약을 담고 있으며 그 이후로 다음 4가지 주제에 대한 세미나에서 나온 페이퍼를 담고 있다.
(1) 개요 및 마이크로 데이터의 사용; (2) 데이터 비밀보호 유지; (3) 마이크로데이터의 법적 측면; (4) 마이크로데이터 접근

3.3 공식통계의 기본 원칙

"기본 원칙"은 1992년 UNECE에 의해 채택되었고, 이후 유엔 통계위원회에 의해 국제 표준으로 채택되었다. 10개의 원칙은 공식 통계 생산을 위한 전문적인 프레임워크를 제공한다. 원칙 6은 통계적 비밀보호를 다루고 있다:

> 통계 작성(compilation)을 위해 통계 기관에 의해 수집된 개인정보가 포함된 데이터는, 자연인 또는 법인이건 간에 엄격한 비밀보호 사항으로 통계적 목적에 한정적으로 이용된다.

부록 1.2. Working Paper 5

독일연방공화국의 통계비밀보호의 조화/융화

서문

연방통계청[FSO-The Federal Statistical Office]과 주통계사무소[Statistical Offices of the Federal States in Germany]는 연구수행을 위한 다양한 통계정보와 정보서비스, 마이크로데이터로의 접근수단을 제공한다. 공식적 자료의 일관된 표준을 보장하기 위해 통계비밀보호를 고려하여 연방통계청과 주통계사무소간 다양한 형태의 서비스에 대한 조화가 필요하다. 일반적인 법적규제 외에도 원시자료에 대한 접근방법의 조화에 초점을 맞춰 마이크로데이터 이용센터, 결과의 비밀보호, 결과점검, 공표자료의 비밀보호 뿐만이 아니라 마이크로데이터 파일의 익명화에 대한 가이드라인이 예시를 들어 논의될 것이다.

1. Federal statistical system in Germany(독일 연방통계시스템)

유럽통계시스템(ESS-European Statistical System)과 유사하게 연방제도로 인해 독일정부는 다수의 국립 통계자료생산자를 유지하고 있다. 각 주정부와 독일연방의 공식통계조사 이른바 연방통계조사[federal statistical surveys]가 연방통계청과 16개주 통계사무소간 협약에 의해 조직되었다.

이러한 이유로 인해 독일내 통계조사의 대다수는 각주에서 분산형 조사에 의해서 수행된다. 이러한 방식은 자료수집, 처리, 공표, 제3자에게 전파될 때 일관성있는 표준을 보장하기 위해 통계조사 방법이 정의되고 조화된 공통의 가이드라인을 요구한다. 따라서 연방통계청과 주통계사무소간 연락사무소가 필요하다.

다양한 통계행정 운영에 대한 책임이 연방통계법(the German Law on Statistics for Federal Purposes(FSL))에 명시되어 있다. 연방통계청의 의무는 주로 연방통계프로그램의 준비와 고도화에 있다. 보다 밀접한 협력을 위해 FSO의 각부서와 주통계사무소는 공식통계조사의 방법적이고 기술적인 요구사항을 정의하였다. FSO의 자료수집과 분산된 통계조사 심사는 주로 각주통계사무소에서 수행된다. 따라서 각주 통계사무소는 통계법에서 요구된 사항을 각 지역실정에 맞게 현실화하여 FSO에 의해 통합된다.

이러한 협업에 있어서 FSO의 주된 기능은 연방통계조사의 상호간 조정이다. 이러한 일은 중복없이 정의된 표준에 맞춰 시간계획에 의해 수행되어야 한다. FSO는 FSL 8장에 의해 통계자료처리의 대부분이 위임되어 있다. 각 주에서 통계자료가 수집된 이후 FSO는 (일반된 목적의) 연방수준에서 발간될 수 있는 하나의 연방통계를 생산하기위해 각

지역의 자료를 조정한다. 또한 유럽연합통계청의 구성원 차원에서 FSO는 독일원시자료를 Eurostat에 제공한다.

2 Harmonisation of statistical confidentiality(통계비밀보호의 조화)[1)]

Principal five of European Statistics Code of Practice는 '통계자료제공자(가구, 기업, 정부, 기타 응답자)의 개인정보, 제공된 정보의 비밀, 통계생산만을 위한 목적은 강력히 보장되어야 한다'고 규정하고 있다. 일반적으로 통계비밀보호는 공식통계자료 생산자의 가장 중요한 의무중 하나이다. 따라서 각 주의 통계사무소는 개인정보가 침해되는 것을 방지하기 위하여 기관차원의 절차적인 예방책을 준비하여야 한다. 통계비밀보호의 주요목표는 다음과 같다

- 모든 응답자의 신상과 물리적인 상황의 노출에 대한 보호
- 응답자와 통계사무소간 상호신뢰의 보존
- 응답의사와 응답 신뢰도 보장

통계비밀보호에 대한 이러한 요구사항은 표준화의 확립과 통계사무소간 조화를 필요로 한다. 앞으로 논의될 사항들은 FSO와 각 주 통계사무소간 통계비밀보호의 조화에 있어서 채택된 측면들을 기술할 것이다.

2.1 Legal aspects of harmonisation(융화[조화]의 법적 측면)

비밀보호의 요구사항을 보장하기 위해 독일정부는 FSO와 각 주 통계사무소 모두가 만족시켜야 되는 법적표준안을 마련하였다. 이 표준안은 FSL에 규정되었다. 통계비밀보호는 다양한 영역에서 정의되었다. 수집에서 공표까지 통계비밀보호는 자료생산의 모든 단계에서 발생한다. 예를 들어 지역이나 국가차원의 결과를 발표하기 전 민감한 자료의 공개를 막기위해 통계사무소에 의해 요약테이블에 대한 검사가 수행된다.

무엇보다도 공식통계자료의 운영을 위임받아 수행하는 공무원들은 업무수행에 관련된 특별한 선서를 하게 되어있다.

원칙에 따라 자료수집이후 결론의 타당성과 자료의 완결성 그리고 전화번호나 주소 같은 보조적인 특징이 체크될 것이다. 보조적인 정보들은 수집된 정보에서 분리될 것이며 경상조사일 경우에는 분리되어 저장될 것이다. FSO와 각 주 통계사무소간 개인정보가 포함된 원시자료의 교환은 암호화된 데이터라인에 의해 수행된다.

개인정보가 포함된 자료는 특별조항으로 승인된 경우(예: 전파를 위한 개인정보 또는 응답자가 서명으로 승인한 자료의 공표, 일반적으로 접근할 수 있는 개인정보, 응답자나 관련자의 위치를 추적할 수 없는 개인정보, 다른 응답자와 함께 요약된 개인정보)를 제

1) European Statistics Code of Practice for the national and community statistical authorities,2005

외하고는 비밀보호가 엄격히 요구된다.(FSL, Art. 16, Para. 1) 응답자나 관련자의 위치를 추적할 수 없는 개인정보의 경우, 일반 통계조사에서 PUF[Public Use Files]라 불리는 완전히 익명화된 마이크로데이터를 생성할 수 있는 기회를 제공한다. 이러한 자료는 개인에 의해 수행되는 국내외의 연구 또는 고등통계교육기관에서 교육적인 목적을 위해 사용될 수 있다.

마이크로데이터의 노출이 막대한 시간, 비용, 인력에 의해서 가능하다는 전제조건이 충족된다면 FSO와 각 주 통계사무소는 고등교육기관이나 독립연구가 위탁된 기타기관에 마이크로데이터를 제공할 수 있다.(FSL, Art. 16, Para. 6) 이러한 목적을 위해 통계사무소는 연구공동체의 사용자들에게 off-site 사용을 위한 표준화된 SUF(Scientific Use File) 형태의 사실상 익명화된 마이크로데이터 파일을 생성한다. SUF는 이른바 MUC(Microdata Under Contract)와 유사하다. SUF의 사용자들은 전송에 앞서 비밀보호에 대해 약속하여야 한다. 자료는 기정의된 연구프로젝트를 위해서만 사용되어야 하며 프로젝트가 완료된 즉시 자료는 삭제되어야 한다. SUF를 사용한 기관은 권한이 부여된 사람만이 개인정보가 포함된 자료를 사용한다는 사실이 기관차원의 기술적인 수단을 통해 보장하여야 한다.(FSL, Art. 16, Para. 8.) 통계사무소는 내용, 수령 기관, 제공날짜와 제공목적에 대한 기록을 유지해야 하며 최소 5년간 보존해야 한다.(FSL, Art. 16, Para. 9.)

응답자의 개인정보를 보호하기 위하여 행정당국에 개인정보를 제공하는 것을 금지하고 있다. 또한, 각 주 통계사무소에서 수집된 개인정보를 대조하는 것이나 법적 통계목적이나 연방통계조사를 위해 제공하는 것 외의 개인, 기업, 기관, 지역 단위의 참고자료를 만들기 위해 기타 정보를 결합하는 것을 금지하고 있다.(FSL, Art. 21)

통계사무소에서 생산과정에 있어서 자료의 보안을 보장하는 것은 중요하다. 가장 어려운 부분은 연구를 위해 접근되는 자료에 있어서 공식통계조사의 비밀보호를 보존하는 것일 것이다. 비밀보호를 유지하기 위해서는 데이터접속을 위한 안전한 환경이나 익명화된 자료뿐만 아니라 결과가 일반에 공개되기 전 비밀보호여부가 체크되기까지 안전하게 지켜지는 것까지 요구된다.

2.2 Access to German microdata(독일 원시자료 접근)

독일원시자료 접근을 위해 자료의 비밀보호를 보장하기위해 규정된 규제와 가이드라인이 있다. 자료의 안전을 감시하는 상태에서 가능한 편리하게 원시자료에 접근하는 것을 보장하기 위한 다양한 방법이 있다.

독일에서 공인통계의 원시자료접근은 SUF, PUF, 마이크로데이터 이용센터, 통계사무소 RDC를 통한 원격실행에 의해서 가능하다. 독일은 연방제도로 인해 FSO의 RDC와 각 주 통계사무소의 RDC head office(14개 사무소에서 조직) 두 개가 설립되었다. 상기 RDC 외에도 국가에 의해 설립/명명된 두 개의 RDC가 더 있다.(The Research Data Centre of

the Federal Employment Agency at the Institute for Employment Research (FDZ-BA) and the Research Data Centre of the German Pension Insurance (FDZ-RV)) 상기 RDC는 고용과 연금부문에서 원시자료를 제공하며 통계법외의 다양한 법률에 의해 운영된다.

각각의 원시자료 접근요구 처리는 FSO와 각 주 통계사무소간의 협력에 의해 수행된다.

연구자는 자료제공을 위한 서식을 작성해야 한다

- 요청기관명/자료사용자명
- 요청된 원시자료명
- 자료접근방법
- 수행 연구프로젝트 내용

기재된 서식은 양측 RDC에 전달되고 모든 사무소에서 확인할 수 있게 된다. RDC에서 요구서에 동의한 후 연구자는 서식에 기재한 방법에 따라 이용센터나 원격실행에 의해 원시자료[SUF]에 접근할 수 있다.

연구자는 어디든 희망하는 FSO나 통계사무소의 MD이용센터에서 원시자료를 사용할 수 있다. 위 사항은 연방정부와 각 주 통계사무소가 보안요구사항과 기관의 이슈를 규정한 마이크로데이터 이용센터 가이드라인에 합의했기에 때문에 가능한 것이다. 이용센터에 대한 규정은 2.4에서 기술할 것이다.

2.3 Guideline for anonymisation of microdata files(마이크로데이터 익명화에 대한 가이드라인)

연방주의에 의해 독일의 원시자료는 분산형과 집중형으로 나뉘어져 있다. 그러나 RDC 간 합의의 결과로 연구자들은 사실상 익명화된 원시자료를 Wiesbaden, Bonn, Berlin 이나 14개 각 주 통계사무소 어디에서나 사용이 가능하다.

사실상 익명화는 자료의 익명화뿐만 아니라 통제된 자료접근의 조합에 의해서도 달성된다. 이것이 이러한 자료들이 파일형태의 SUF보다 더 상세한 정보를 제공할 수 있는지를 말해준다.

원시자료에 있어서 자료의 익명화를 해제하는 것이 완전히 배제될 수는 없으나 엄청난 시간, 비용, 인력에 의해서만 익명화 해제가 가능하다면 사실상 익명화 되었다고 말할 수 있다. (FSL, Art. 16 Para 6) FSL에 따라서 사실상 익명화된 자료는 연구기관의 연구프로젝트만을 위해 생성될 수 있다.

사실상 익명화는 통계적 관점에서 정보의 가치를 보존하면서 정보의 축소와 변형에 의

해 변수를 통계단위로 재배치하는 가능성을 줄이기 위한 것이다. 익명화를 해제하는 것의 수익과 비용은 각 조사에서 분석되어야 한다. 따라서 RDC는 각 연구마다 원시자료를 익명화해야만 한다. 이것과 관련하여 지역활동단위는 연구자와 밀접한 연구를 수행하여 익명화 개념을 발전시킨다(?). 사용된 원시자료가 비밀보호요건을 충족시키는 가운데 그들의 연구가 충분히 수행될 수 있는 방향으로 개념은 정립된다. 익명화의 개념이 정립되면 FSO와 각 주 RDC에 의해 승인되고 사실상 익명화의 정의와 지역의 특수한 노출위험의 조화에 관한 재검토를 수행한다.

이후 연구자가 부여된 계정으로 접속할 수 있는 이용센터에서 익명화된 원시자료가 제공된다.

연구사업 수행시 매일 수행되어야 할 업무는 요청자료와 on-site 와 off-site 사용을 위한 SUF 익명화 개념을 조화시키는 것이다. on-site 와 off-site에서 사실상 익명화된 자료에 대한 요구사항 관리를 위한 가이드라인을 촉진시키고 간소화하기 위한 절차가 개발되었다.

가이드라인은 다음 단계를 포함한다

1. RDC에서 원시자료 사용을 위한 요구수용

RDC는 연구사업을 처리하고 관련된 조직단위를 참가시키기 위한 것이며 사용자 접촉을 관리한다.

2. 16Para 6FSL 에 따른 접근통제

새로운 기관의 사용을 위해 RDC는 법무부서를 참가시킨다. 법무부서는 FSO의 변호사들과의 협업에 의해 해당기관이 독립된 연구를 수행하는지 인증한다. 법률상 전제조건이 충족되지 않는다면 RDC는 사용자에게 고지하여 접근을 위한 가능한 조건을 제시한다. (비용기반 원격수행, PUF의 특별평가-remote execution on a absorbed cost bases, special evaluation or Public Use File)

3. 익명화 개념의 생성

RDC는 관련된 부서와 협의하여 자료를 익명화하기 위한 기본 가이드라인의 개념을 생성한다. 이 개념은 익명화조치 수행과 관련된 test를 포함한다. 만약에 하위샘플[subsample]이 익명화를 위해 사용되어야 한다면 가중치의 변경에 관해서는 관련부서와 협의되어야 한다.

4. 각 주 RDC 참여

분산된 통계처리에 있어서 각 주 RDC는 모든 지역의 익명화개념의 협의를 위해 통합되어야 한다. (집중형 통계조사시 공지만 하면 된다)

5. 계약

법률관계부서가 자료사용자와 계약을 체결한다. 계약의 일부는 익명화에 대한 것이다. 완료된 계약서의 복사본이 RDC에 전송된다.

6. 자료이관 및 송장

RDC는 off-site에서 사용할 수 있도록 사실상 익명화된 개인정보가 포함된 디스크나 이용센터에서 on-site로 사용될 수 있도록 자료를 제공한다.

조화의 개선에 관한 예시가 경제통계 뿐 아니라 개인이나 가구를 위한 통계에서도 보여질 것이다.

2.3.1 Personal and Household Statistics(개인 및 가구 통계)

통계사무소는 개인과 가구통계부문에서 가장 빈번히 사용되는 census 원시자료에 대한 표준화된 on-site 파일을 생산하기로 결정하였다. on-site 사용을 위한 익명화 개념이 연구자와 밀접한 조정에 의해 개발되며 시간이 소요되고 정교하기 때문이다. 또한 연구자와의 조정 그 자체가 연구자에 대한 원시자료 접근을 연기시킨다. 따라서 표준화된 on-site 파일 생성시 익명화의 개념은 한번만 조정이 필요하며 그 이후로는 이용센터나 원격실행을 통해 다른 어떤 추가적인 조치없이도 사용할 수 있다. on-site파일은 조사의 모든 영역 및 연구의문점에 대한 중요한 특징을 모두 포함하고 있다. 모든 식별자는 제거되고 지역수준의 상세한 내용은 제거된다. 사용자는 모든 분석프로그램 및 결과에 대한주석을 달고 문서화를 해야 한다. 분석프로그램의 구조화를 위해 코드작성에 대한 통일된 가이드라인이 존재한다. 이것은 처리공무원이 결과를 재생산하기 쉽게 도와준다. off-site에서 SUF를 사용하는 관점에서 관련된 통계의 익명화의 개념은 FSO와 각 주간 조정이 필요하다. 표준화된 SUF는 각 조사마다 한번만 수행되면 된다

2.3.2 Business Statistics(경제통계)

SUF 파일을 생산하기 위해서 경제통계 분야에서 FSO와 각 주 통계사무소가 수행한 두 개의 프로젝트가 존재한다. 상기 프로젝트는 두 기관간의 상이한 관점이 이미 존재했음을 반영한다. 경제통계 익명화 프로젝트의 목표는 비밀보호와 자료의 분석가능성이 보장될 수 있는 효율적인 가이드라인을 찾을 수 있냐는 것이었다. 처음 프로젝트는 cross section business statistics를 위해 수행되었고 다음 프로젝트는 수집된 정보를 longitudinal enterprise microdata에 적용하였다. 각각의 프로젝트는 각주의 특성을 고려하고 연방정부는 공통된 산출물을 생성하는데 주안점을 두었다.

FSO에서 익명화 가이드라인이 동의를 얻은 뒤, remarks, feedback, changes를 갖을 기회를 주기위해 각주의 통계사무소에 보내진다. 의사결정에 있어서 각주의 협업의 상당한 장점은 기능적인 책임의 지역화이다. 다시 말해 각주의 통계사무소는 특정한 통계조사에 대해서 책임을 진다는 것이다. 지역사무소는 그들의 지역에 대한 상세한 정보를 갖고 있다. 각 주의 통계사무소는 조사된 기업을 잘 알고 있으며 SUF 에 있어서 어떤 기업이 위험할 수 있는지를 잘 알고 있다. 게다가 각 주는 의사결정의 과정에 관여되어있을 뿐

아니라 비밀보호의 차원에서 통계생산물의 질적인 측면을 개선시킬 수 있도록 중복점검을 할 수 있다.

일련의 과정을 거친 후 표준화된 SUF는 research community에 배포될 수 있다. 이후 발전된 익명화 가이드라인은 다음의 다양한 연도와 계절의 통계조사에서 사용될 수 있다. 이 과정은 기타 다양한 경제통계조사의 SUF파일을 생성을 빠르게 할 수 있다. 축적된 지식은 FSO와 각 주 통계사무소간 공유될 것이다.

2.4 Guideline for Data Laboraties(마이크로데이터 이용센터 가이드라인)

FSO와 각 주 통계사무소는 비밀정보가 담긴 자료에 대한 접근 요청을 공지한다. 따라서 통계사무소는 조화된 이용센터를 구성하고 RDC의 특수하게 구성된 PC를 통해 국내외의 연구자들은 사실상 익명화된 원시자료를 분석할 수 있다.

연구자들은 통계사무소에 대한 내부망에 직접 접근할 수 없다. 이것을 통해 또 다른 민감한 정보에 대한 연구자의 접근 가능성을 배제할 수 있다. 또한 RDC 직원은 분석의 모든 단계에서 감시감독이 가능하다.

이용센터는 차폐된 시설과 연구에 사용되는 자료의 비밀보호가 확보된 자료저장환경으로 구성된다. 이것은 또한 다른 정보와 함께 사실상 익명화된 원시자료를 제공하는 것을 방지한다.

인터넷이 연결된 별도의 PC는 이메일과 인터넷 접속이 가능하다. 관리적인 측면에서 이용센터는 규정된 기준이 요구된다.

- 이용센터는 FSO와 각 주 통계사무소 내에 위치한다.
- 접근을 허락할 때는 법적인 조치가 취해져야 한다.
- 권한이 부여된 사용자만이 이 시설을 이용할 수 있어야 한다.
- 휴대용 노트북, 대용량저장장치, 사진저장장치(디지털카메라, 카메라폰)는 이용센터내 사용이 금지된다.
- RDC 직원은 연구자들의 사용시간동안 언제나 수행내역을 점검할 수 있다.

분석을 위한 컴퓨터는 개인정보 노출을 방지하고 비밀보호를 충족시키기 위해 성능이 제한된다. 아래 사항은 금지된다.

- 인쇄
- 디스켓, USB, CD-ROM, DVD, ZIP Drive로의 자료저장
- 로컬 하드디스크로의 저장
- 저장장치 연결
- E-mail 사용
- 인터넷연결

- PC의 추가반입/반출
- 플로피디스크, CD-Rom, DVD-rom 기타 매체로의 PC 부팅

중간/최종 산출물을 배포하기전 RDC 직원은 비밀보호와 관련된 모든 결과를 체크해야 한다.

2.5 Confidentiality of output(산출물의 비밀보호)

연구자는 각 주와 FSO의 다양한 지역의 RDC에서 통계자료의 접근이 가능하기 때문에 산출물 점검에 대한 기준이 필요하다. 결과점검에 대한 기준은 연구자들이 서로 다른 지역에서 동일한 자료의 접근이 가능하기 때문에 비밀보호의 측면에서 다른 연구자들에 대한 동일한 산출물과 분석에 대한 서로 다른 처리를 방지하기 위해 매우 중요하다. 예를 들면 FSO와 각 주 통계사무소간 파라미터가 조화되어야 하는 P%-rule을 들 수 있다. 게다가 동일한 테이블 내에서 최소 세가지의 경우에서 최소주기에서 동의된다. RDC 뿐만이 아니라 관련 부서에서의 중복점검도 실행해야 한다.

각 부서의 표준화된 공표를 위해 FSO와 각 주 통계사무소간 테이블의 비밀보호 조정은 반드시 이루어져야한다. 언급한 복잡한 계획을 관리하기 위해 조직실행위원회는 turnover tax 대한 비밀보호 규제를 조화시키기로 결정하였다. 이는 각주와 연방의 표준 테이블이 각주 수준의 집중형 협업모델 차원에서 조정됨을 의미한다. 테이블데이터에 대한 1/2차 셀서프레션[열압축?]을 동기화 하는 것은 복잡한 문제이다. 각 주에서는 각자의 테이블을 생산산하고 공표해야 하며 비밀보호를 관리해야하며, FSO는 연방수준의 테이블을 집계하고 공표해야 한다. 만약 어느 주의 셀이 이미 서프레스되었다면 연방수준의 결과를 공표하는 것이 거의 불가능하다. FSO에서 특정 주의 수치를 서프레스하는 것을 통해 2차셀의 서프레스를 수행한다면 주단위의 테이블테이터를 생산하는 것이 불가능하다. 이 문제는 연방과 각 주에서 공히 자동화된 τ-Argus 테이블자료보호법을 사용하는 것을 통해서 해결될 수 있다. 연방과 각주수준의 이차셀 서프레션의 채용은 FSO와 각주 통계사무소의 각각의 공표에 대해 의무적이다. 어코모데이션 통계조사와 같은 다른 통계조사에서도 이러한 개념을 발전시키고 이용할 계획이다

3 Summary and Outlook(요약과 전망)

FSO와 각 주 통계사무소간 통계비밀보호의 조화에 관한 예시를 고려하는 것은 위에서 살펴본바와 같으며 이미 업무에서 적용되어 결과가 나오고 있다. 그러나 아직도 개선의 여지가 많이 남아있다. 분산형 통계시스템에서 조사수행시 조화를 고려해야할 뿐만 아니라, 각과별 비밀보호 규제, FSO와 각 주 RDC에서 각 연구결과물이 동일하게 처리되는지 연구자들의 다양한 산출결과를 기록할 때 조화될 필요가 있다. 독일내 FDZ-BA와 FDZ-RV 같이 상이한 법률에 의해 운영되는 RDC에게 있어서는 통합된 접근법이 더욱

중요하다고 할 수 있다.

개선을 위해 각 RDC의 직원들이 연구자들이 무슨 이유로 어떤 데이터를 사용하는지 항상 감시할 수 있는 공유되는 사용자관리 DB를 사용할 수 있다.

분산형 통계시스템에서의 조화는 동등하고 고품질의 결과를 생산하기 위해 필연적인 것이다. 그러나 이 과정은 절차를 조정하고 통합하기 위해 많은 자원을 필요로 하며, 관련기관간 인터페이스를 최적화 하는 것이 중요할 것이다. 법률 하부구조가 다를지라도 유럽 상황에 대한 일종의 축소판으로 독일의 분산형 통계시스템은 유럽통계시스템을 위한 모델이 될 수 있다. 향후 양시스템 모두 조화를 위한 개선된 방법과 도구가 필요할 것이다.

부록 1.3. Working Paper 18

영국의 안전한 데이터 서비스
: 가능적으로 노출시킨 데이터 사용의 사양과 도전

개요

안전한 데이터 서비스는 ESRC에 의해 설립된 안 안전한 환경이며, 연구원에게 연구원의 방, 사무실 또는 UKDA.사이트에서 공개된 마이크로데이터 접근을 제공하기위함이다.
이 작업은 통계목적을 위해 비밀보호된 데이터접근을 위한 2007통계법에 의해 법적으로 설립되어졌다. 이 짧은 논문은 데이터서비스 제공자에게 직면한 몇 가지 도전과 제안된 사양서의 새로운 UKDA를 소개한다. SDS인프라가 데이터보안모델의 요구 사항을 충족하실 수 있을 것입니다.

1. 소개

개인 정보의 공개는 해로울 수 있다. 이것은 시민 서비스(응답자)를 부정하고 명성과 신뢰의 상실과 당혹함을 만들 수 있다. 간접적으로, 노출 데이터를 기반의 연구결과는 어느 개인이 속한 단체에 대한 인식에 영향을 미치는 원인이 될 수 있다.

안전한 데이터 서비스(SDS)에서 핵심적인 문제는 데이터 이용자의 합법적 요구 및 비밀보호의 균형이다. 개인 정보의 비밀 보장은 제공되는 정보를 제한, 공개 테이블 및 통계 출력에서 데이터를 조정, 데이터(제한된 접근) 접근에 대한 조건을 부과, 또는 이러한 몇 개의 조합에 의해 보호받을 수 있다.

UKDA안전한 데이터 서비스는 자격조건, 사용목적, 보안, SDS 데이터 액세스와 관련된 기타 기능에 의해 몇몇의 이용자(승인된 연구자)에게 허용할 수 있는 더 상세한 마이크로데이터 파일을 만드는 제한된 접근 절차를 제어할 수 있는 새로운 서비스이다

이 짧은 논문은 데이터서비스 제공자에게 몇몇의 도전과 제안된 설명서의 새로운 UKDA서비스를 소개한다. 제시된 SDS인프라는 데이터 보안 모델의 사항을 충족할 수 있다.

세계적으로 성공적인 다른 안전데이터 장소를 만들고, 군사와 금융 분야에서 이용된 안전 기술을 채택으로, SDS는 훈련된 연구자에게 UK Data Archive의 중앙SDS 서버에 데이터에 보관된 데이터를 원격으로 접근하는 것을 허용한다. 서비스의 목표는 승인된 학

계에 제공하는 것입니다. 이 서비스의 목적은 데이터를 보증하는 모든 필요한 안전장치에서 승인된 연구를 위해 연구자의 연구소에서 연구를 위한 가치 있는 데이터에 안전하게 접근하고 처리하도록 허용하는 것이다.

SDS는 데이터의 안전사용으로 안전 프로젝트의 요소, 안전 장치, 안전 출력(Ritchie, 2006. 그림 1)로 설정된 데이터 안전 사용를 제안한 모델를 사용한다. 위의 목표를 달성하기 위해, 데이터 보안은 기술, 계약, 법률, 교육을 포함하는 여러 요소의 행렬에 의존한다. 이 논문은 SDS가 어떻게 이러한 요구를 충족할 수 있는 필수의 인프라를 설정할 것인가를 보여준다. 이용자와 그들의 연구소에서 법적, 계약적 책임과 함께 기술적인 기능 및 시스템 사양 토론한다. 데이터 접근 전에 이용자의 교육과 훈련 같은 문제를 검토한다. 또한 SDS에서 지원을 목표로 하는 노출가능 데이터의 종류의 대해 토론한다. 마지막으로, SDS 기능이 직면하고 있는 도전을 검토한다.

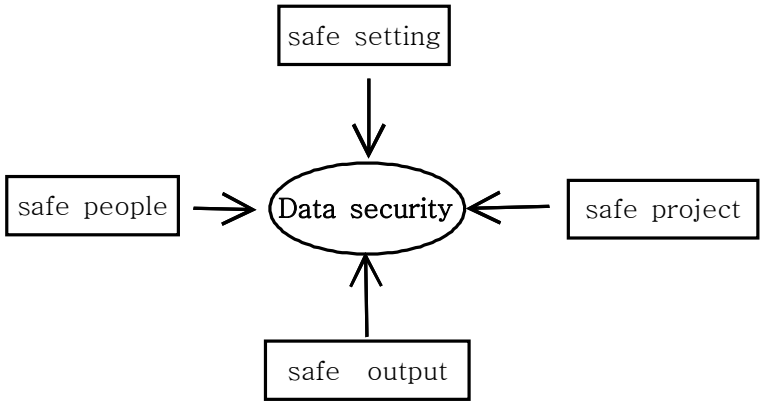

그림 1 : 데이터 안전 보안

2. 시스템 사양

SDS에서 사용되는 기술은 안전하고, 시스템은 최고이고, 가장 투명한 품질수준을 고수해야한다. 기술적 모델은 ONS VML과 ONS VML의 많은 유사성을 공유한다. 이용자의 컴퓨터에 데이터, 통계 소프트웨어에 대한 액세스 권한을 부여하는 '원격 터미널'로 변환하는 시트릭스의 인프라가 기본이다. 중앙 안전 서버 공동공간은 UK Data Archive에서 유지한다. 시스템은 접근을 특별한 사용자 (safe people)와 위치(safe rooms/machines)로 제한하도록 데이터 보관자에 의해 조정한다. 모든 데이터 조작은 서버에서 발생하기 때문에 안전하다.일반적인 보안 정책을 넘어, 보안 서버 자체가 부가적인 보안 대책과 제어를 할 것이다. 승인된 연구자는 연구자의 컴퓨터와 호스트 네트워크사이의 데이터 전송을 암호화하는 VPN(Virtual Private Network/thin-client) 기술로 SDS를 접근할 것이다. 서비스는 두개 네트워크에 참여한 Citrix XenApp server 회사를 고용할 것이다.

시스템이 작동하는 방법?

모든 응용 프로그램 (SPSS, STATA 등) 및 데이터가 UKDA/SDS의 중앙서버로 운영되더라도 승인된 연구자는 여전히 Windows graphical user interface로 상호 실행한다. 연구자는 승인된 연구자의 웹브라우저에 요구되는 어플리케이션을 설치해야한다. 또한 UKDA는 데이터 아카이브로부터 로컬 컴퓨터에 어떤 데이터도 전송하는 것을 막을 수 있다. 예를 들어, Citrix는 데이터파일이 원격서버로부터 이용자의 로컬 PC를 다운로드 할 수 없음을 인정할 수 있다. 마찬가지로, 승인된 연구자는 Citrix 세션으로부터 로컬 컴퓨터에 설치된 엑셀 시트에 데이터를 옮기는 윈도위의 특징인 '자르기와 붙임'을 사용할 수 없다. 승인된 연구자는 원격으로 웹안전(HTTPS) 브라우저를 통해 SDS 시스템에 로그 온 한다. 모든 데이터 처리는 중앙 보안 서버에서 수행된다. 모든 요청은 중앙에서 처리하고 결과에 대한 정보를 반환한다. 최종의 결과물이 통계적 노출제한에 검증된 후 암호화된 이메일에 의해 통계적결과가 중앙 서버로부터 원격장치로 보내지는 것 이외에 네트웍으로 이동되는 데이터는 없다.

주요 특징

- 클라이언트가 데이터를 제거할 수 없다
- 절대 웹페이지에 액세스할 수 없다
- 클라이언트가 데이터를 가져올 수 없다
- 데이터 전송이 기록된다.
- 모든 트래픽이 암호화된다.
- 감사
- 중요 보안 업데이트가 매일이 적용된다.

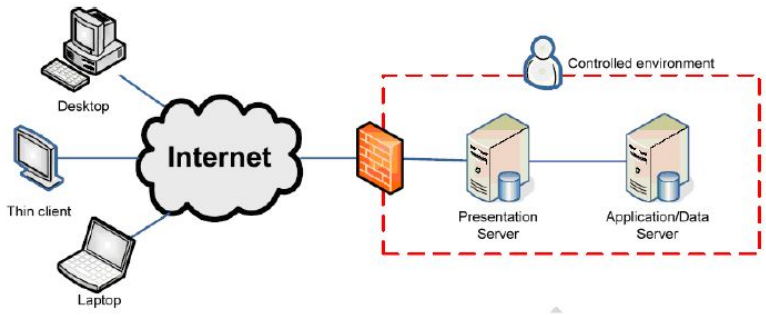

그림 2 : SDS 시스템 아키텍처

3. 법적 및 계약 프레임 워크

SDS의 이용자는 "ONS 승인" 또는 "ESRC 인증 연구원"으로 요구된다. 이것은 첫째, 이사

회가 통계적 연구목적으로 승인한 연구원에게 법에 의해 개인정보에 접근할 수 있다는 통계와 등록 서비스 법령 2007통계에 의해 정의된다.
현재 "ESRC 연구원" 정의는 없지만, "ESRC 연구원" 은 정부조직 또는 다른 데이터 제공자에 파급된 ESDS/UKDA/University of Essex의해 통계연구목적으로 개인정보를 접근을 승낙받은 ONS연구원과 비슷한 지위라고 가정한다. 두 타입 모두 적절한 훈련이 없다면 SDS를 사용할 수 없다. 필수 교육은 UKDA가 이용자가 위반하면 받 을 수 있는 어떠한 처벌들을 충분히 알 수 있도록 교육하는 것이다. 우리는 위반에 대한 어떠한 처벌도 이용자에게 허용할 수 있다. 이러한 처벌들은 사회과학 연구자들이 가장 일으키기 쉬운 부주의한 자료노출을 피하기위해서 합리적이고 필수적이다.

2007년법은 경우에 따라 의회 동의하에 ONS와 다른 부서 사이의 데이터를 공유를 허용한다. 동시에 법은 개인정보 비밀을 보호하도록 되었다. 법은 개인정보를 누설한 사람은 (a) 기소장에 대한 유죄 판결, 2년이내의 징역 또는 벌금, 둘 모두(b)유죄 판결, 12개월의 징역 또는 법정최대 범위내의 벌금, 둘 모두와 같은 위반에 대한 벌칙과 납입해야 한다.

SDS는 만약 이용자가 SD 안전위반 또는 SDS기밀유지 합의서에 작성된 위반을 시도한다면 서비스에 대한 접근을 정지한다. 모든 연수는 따를 것이다.

이용자는 개인과 조직에 상세부분, 제안된 데이터 사용, 연구가 담당하고 보안하는 방법을 설명하는 수행에 대한 능력과 전문 기술을 증명할 수 있는 정보를 온라인 양식에 작성한다.
만약 이용자가 아직 준비되지 않았다면 표준UKDA End User License에 동의하고 접근하기를 희망하는 수단에 적용되는 Special License조건에 동의해야 한다. 이 신청은 먼저 UKDA 직원에 의해 정확성, 완전성을 점검한다. 그런 후에 액세스 권한을 위해 데이터 소유자에게 전달한다. 승인받은 이용자는 적당한 교육(아직 교육 받지 않았다면) 위해 통보될 것이다. 교육 수료후 이용자는 만약 접근을 원하는 데이터 소유자가 허가한다면 자신의 데스크탑 또는 조직의 안전 데이터 접근 공간으로부터 안전 데이터 서버에 접근 허가를 승낙받는다. 조직이 안전데이터 접근 공간을 가지고 있지 않다면 이용자는 근처의 조직의 안전공간으로부터 접근 협상을 선택할 수 있다(SDS는 'matchmaking' 소개하지만 구체적인 협정은 보안감사 시도는 이용자의 책임이다라는 것처럼 조직의 호스팅 공간 통제하에 있어야 한다)

4. 교육 및 훈련

데이터보안에서 사람들이 가장 위반을 어기기 쉽다고 알려져 있다. 사람들에 대한 훈련과 교육은 이용자가 범죄하지 않도록 막는다. 교육은 위에서 언급된 더 엄격한 법적 보호와 비밀보호에 부과되어 있는 잠재적인 효율적 의미 즉 비밀노출의 가능성은 규칙을

지키는 것이 연구자에게 비용을 부과 없이 위반을 감소할 수 있다라는 효과를 부여시킨다.
SDS의 이용자가 되기 전에, 이용자는 먼저 이용자의 법적 및 윤리적 책임, SDS를 사용하는 방법, 원격접근 셋팅에 무엇을 해야하는지, 가능성 있는 공동 공간에 초점을 맞춘 훈련에 의무적으로 참석해야한다. 두 번째 부분은 통계적 노출, 결과물 평가, SDS에서 부분적 데이터셋의 분석부분에 초점을 맞춘다.

SDS의 접근은 이용자가 SDS훈련에 참가한 후에만 승낙된다. SDS 직원에 의해 노출문제에 대한 결과물 분석으로 데이터를 검증하고 어떤 것도 안전 데이터세팅에 벗어나지 않음을 인정한다. 훈련의 목적중 하나는 연구자에게 SDS를 벗어나도 통계결과와 구분하여 데이터의 비밀보호를 인식시켜주는 것이다.

훈련은 이용자가 이성적으로 데이터 노출하지 않도록 한다. 벌칙은 이용자가 알게 된 것을 막는 효과적인 방지책이다. 그리고 우리는 어떠한 벌보다 자료 보호에 대해서 더 염려하고 있다.

5. 장점

이 시스템은 이용자가 로컬 소프트웨어와 네트웍크자원으로된 EUL데이터에 익숙한 것에 비교해서 불편하다. 이용자는 간단하게 모든 것이 접근되지 않고, 접근을 위해 ONS VML옮기고 ONS에 이동하고 로컬접근에 기꺼이 가격을 지불할 것이다.

SDS는 이용자에게 도움이 된다;

- 자신의 개인 작업 영역 또는 공유 영역에서 작업 가능
- 용량증가와 환경적 보호를 통해 메타데이터와 연관하여 매우 민감한 이용가능 데이터에 향상된 접근
- UKDA의 자료 또는 다른 행정자료 링크
- 조사와 문서 라이브러리, SPSS/STATA 라이브러리, 지식 저장소, 공개 검토 및 기술지원을 포함한 공동 상관관계
- 유연성 및 접근은 특정 사용자 (safe people) / 특정 위치 (safe rooms/machines)로 제한
- 분석환경에 친숙한 보호서비스
- 성장 및 확장을 위한 기능
- 계획적으로 데이터보호 및 안전을 구축한 통합 환경
- 감사, 변경 제어, 모니터링, 경보 알림을 포함한 서버관리 과정

6. 데이터

이용자는 SDS에서 다양한 데이터를 이용 가능할 수 있다. SDS는 중요한 자원의 유용과 이용을 넓히고 모든 법적 도덕적 안전 요구를 충족하면서 어떻게 도와줄 것인가를 위해 민감한 데이터 원천의 키를 토의하고 있다. 데이터 사양서는 다음을 포함한다.

- ESRC 데이터로부터 자세하고 다양함.
- 정부사회조사로부터 미리 이용할 수 없는 상세하고 다양한 데이터
- 온사이트에서만 사용가능한 국가자료 또는 학문연구자가 사용할 수 없는 국가자료
- 상용 감도를 가지고 있는 비즈니스 데이터
- 행정자료; SDS는 연구자의 공간에서 데이터를 제공하는 기술 부족한 연구자에게 데이터 연결활동을 위한 안전한 환경을 제공
- 장기적 데이터 또는 의료 데이터

또한, 이용자의 요청으로 KDA 표준과 End User License holdings에서 공개되지 않은 데이터를 가져갈 수 있다.

7. 도전

안전한 데이터 서비스의 두 가지 주요 목표; 개인의 응답의 비밀보호를 하면서 연구목적을 위해 마이크로데이터 이용의 극대화이다. 비밀데이터의 접근은 응답자의 개인정보를 위반하는 위험에 대항하여 연구의 공공 선익에 조합됨에 따라 정당화 될 수 있어야 한다는 비공개법칙에 제외된다. 그러므로 SDS는 다음을 따른다. 노출 위험을 최소하면서 최대한의 데이터 효용. 그것은, 투명하고 간단해야한다.

바람직한 서비스 품질 완성에는 두 가지 문제가 있다. 첫째, 안전한 데이터가 무엇인가의 정의에 합의가 없다. 둘째, 더 많은 논쟁으로 어떤 정보를 손실한다. 모든 노력으로서 비밀보호는 몇 가지 정보손실이다. 데이터의 효용과 노출위험의 균형을 깨는 최대한 허용 위험을 정의하는 것이 중요하다.

비밀유지는 일관성 유지 및 일관된 접근이다. 대체로, 우리는 연구자를 믿어야한다. 예를 들면, 원격으로 노출된 데이터를 접근한 연구자가 데이터를 포토그래프를 사용하거나 복사하는 것을 어떻게 막을 수 없다. 대부분 연구자들은 데이터 접근편익에 위반하고 악의는 아니다. 틀림없이 원격으로 데스크탑 데이터접근은 데이터가 사라진다는 유혹이 있다. 노출가능성은 항상 있다. 법적 프레임과 훈련, 교육은 이용자 즉 승인된 연구자는 비밀보호를 위반하는 것을 단념시킨다.

8. 평가 및 결과물의 모니터링

신중한 이용자 인증과 가장 안전한 분석환경으로 데이터는 노출되지 않는다라고 확신할 수 없다. 데이터 일부의 분실은 데이터 시스템으로 들어오는 것이 아니라 시스템 밖으로 나가는 것이다. 안전보장을 위해 데이터 보관인은 결과물 검토의 양식을 제공해야만 한다. 만약 결과물이 노출과 결정된다면 이용자는 결과물을 안전하게 하는 결정해야한다.

SDS는 결과물을 세 개의 주요 분류로 나눈다.

- 안전 : 위험 없음 / 노출이 매우 적은 - 결과물은 즉시 제공한다.
- 확실하지 않은 결과물 : 노출이 낮음 또는 중간 - 결과물은 한개 또는 다양한 데이터 셀을 제거로 인해 주의 깊게 인정되어진다.
- 불안전 : 노출의 위험이 높음 - 결과물은 현재 양식으로 차단하고 제공되지 않는다. 이것은 안전한 결과물 제출에 대한 연구자의 책임이고 위험노출로부터 제외됨을 증명해야한다.

다양한 기법은 절삭법와 표데이터에 셀 잡음에 초점을 맞춘 **-τ -ARGUS을** 이용한다.

민감한 데이터셀의 정보를 보호하는 다양한 해결법

- 스패닝 변수(표 디자인)의 조합분류. 더 많은 데이터 셀은 개인 정보를 막을 수 있다.
- 민감한(기본의) 데이터 셀의 재계산을 막기 위해 추가(이차적인) 셀의 삭제

τ -ARGUS는 2차셀의 최적계산을 만든다. 전형적인 **τ -ARGUS**은 먼저 이용자가 기본적인 불안전한 셀을 포함하는 테이블을 만드는 것이다. 이용자는 어떻게 이러한 셀을 보호할 지 선택한다. 분류를 조합하고 세계적 기록과 맞아야 한다. 결과표는 덜 불안전한 셀로 테이블을 업데이트한다. 시스템은 기본적인 셀을 보호하는 두 번째 셀을 찾는 것으로 불안전한 셀을 유지할 수 있다.

9. 요약

SDS는 연구자에게 그들의 사무실, 기관의 안전한 장소, UKDA사이트에서 노출된 마이크로데이터를 제공하기 위해 ESRC에서 설립된 안전 환경이다. 그것은 두 가지 목표가 있다: 연구자가 민감한 마이크로데이터의 접근을 향상하고 비밀을 보호하는 것이다. 법적 기능은 통계적 목적으로 비밀데이터를 가능한 접근하도록 만든 2007통계법에 의해 설계되었다.

마이크로데이터에 접근하는 연구자들은 데이터 수집으로 공공의 투자에 존재하는 영향과

과학 분석을 통해 고품질의 과학 향상으로 공공의 이익을 제공한다. SDS는 비밀을 보호하기 위해 가장 안전한 방법을 사용하면서 승인되고 인증된 연구자에게 마이크로데이터를 원격으로 제공한다. 이것은 통계적보호 적용, 법률요건 강화, 연구자를 훈련시는 기술적 보안(Citrix gateway)을 수행하면서 달성하였다. SDS/UKDA는 또한 가치있는 데이터는 DDI를 준수하는 메타 데이터 표준을 사용하여 데이터를 문서화하여 장기간 보존하는 것을 보증한다. 또한 SDS 협력적인 공동의 장소를 이용하는 연구조직이 지리적으로 분산되어 있는 연구자들과 협동하여 정보를 공유하는 것을 목적으로 한다.

부록 1.4. Working Paper 19

캐나다 통계청에서의 실시간 원격 접근 인프라 개발

요약 : 많은 국가통계기관(NSOs)과 마찬가지로, 캐나다 통계청은 상세한 마이크로데이타 접근에 대해 연구자들의 국내 및 국제 요구의 증가에 직면하고 있다. 최근 캐나다 통계청은 응답한 데이터를 비밀보호하면서 동시에 접근하는 방법을 어떻게 개선해야 하는가에 대한 리뷰를 시작했다. 고려할 수 있는 방법 중 하나는 실시간 원격접근방법 (RTRA)을 개발하는 것이다. RTRA은 기본적으로 이용자가 중앙 및 안전한 장소에 보관되어 있는 가볍게 마스킹되어진 마이크로데이터를 실시간으로 실행하여 분석을 허용하는 온라인 원격 접근 시설이다.

1. 소개

캐나다 사회의 모든 분야에서 상세한 마이크로데이터의 이용에 많은 관심을 보인다. 이것은 최근 노인 인구, 이민자 및 사업의 특정 부분에 집중하고 있는 정부의 정책 및 프로그램에 반영에 초점을 맞추고 있다. 이러한 큰 관심은 학계와 국제적인 연구 학회로 확장된다. 기술의 발전은 통계 생산기관이 상세한 데이터를 생산하고 분석할 수 있고 연구자들이 데이터를 찾아내고 분석할 수 있는 상당한 수단으로 개방되었다.

증거기반의 정책발전의 증가하는 수요에 맞추기 위해, 다른 연방 부서의 협의로 캐나다 통계청(StatCan)은 1990년대 수많은 명확한 장기적인 조사를 개발했다. 노동과 소득조사(SLID), 국가인구건강조사(NPHS) 및 어린이와 청소년의 국가 추적조사(NLSCY)는 노동시장과 소득, 건강과 아동 발달과 같은 분야에 데이터를 제공하기 위해 만들어진 몇 개의 조사들이다. 이러한 조사들은 캐나다 통계청이 완전하게 분석할 수 있는 이상의 많은 양의 추적과 교차단면 데이터를 생산하였다. 따라서 학술 연구자들이 데이터를 분석하기 위해 정책에 의해 위임되어졌다. 그러나 추적 분석의 본질에 의해서 연구자들은 보호된 개별 정보에 대한 접근이 필요하다.

이러한 상황은 캐나다가 직면하고 있는 딜레마이다. : 연구자들이 응답자의 비밀을 보호하면서 더 많은 데이터에 어떻게 접근할 수 있을까? 통계 법령에 따르면 캐나다통계청은 수집데이터의 기밀성을 보호한다. 국민의 신뢰 손실은 응답율과 국가통계 프로그램의 전반적인 품질에 매우 해롭게 할 수 있다.

이 논문은 새로은 데이터 구상을 실현함에 있어 현재의 생각과 토론에 대한 논의이다. 섹션 2는 캐나다 통계청의 현재 접근상황에 대한 배경을 제공한다. 섹션 3은 실시간 원격접근(RTRA)의 개발을 둘러싼 설명이다. 섹션 4은 비밀보호측면에 대해 설명한다. 마지막으로, 섹션 5는 미래의 테스트 및 계획뿐아니라 설계된 견본의 특징을 설명한다.

2. 배경

캐나다의통계 법은 기밀 데이터를 통계법의 11, 12, 17(2)의 조항을 제외한 캐나다 통계청 직원 (또는 ""직원)아닌 사람에게 제공되는 허용하지 않는다. 섹션 11 및 12는 지방 통계 기관 및 기타 단체와 정보의 공유를 포함한다. 섹션 17 (2)은 통계청장에 의해 개인의 신상 자료를 특정 유형으로 제공을 허용한다.

이러한 구조를 감안할 때, 캐나다 마이크로데이터는 세가지 방법으로 직접 접근할 수 있다. 첫째, 개인정보의 위험과 비밀을 무시할 수 있도록 충분의 마스킹되어진 기록을 포함된 public-use 마이크로데이터 파일(PUMFs)은 계약 동의하에 제공되어진다. 이 동의는 이용자는 파일의 재식별을 시도하지 않고 단지 통계적 목적으로 데이터 사용을 요구한다.
둘째, 연구데이터센터(RDCs)를 통해서이다. RDCs는 연구자들에게 보다 안전하게 설치되어진 대학에 인구와 가구조사의 마이크로데이터를 제공하는 것이다. 모든 비밀보호규칙의 통계법에 의해 운영되고 통계법으로 “deemed”이라는 고용인으로 통계법아래서 맹세한 프로젝트가 허용된 연구자들만 접근한다. 셋째, 마이크로데이터 파일은 섹션 12통계법령의 권위 아래에 지정하는 특별한 조직에 제공되어진다. 이러한 파일은 조직에 응답공유를 동의한 응답자들의 기록만 포함된다.

대다수 조사들은 원격접근으로 제공한다. 연구자들은 조사를 위한 분석요청을 제출하고, 캐나다통계청의 보안 환경하에 설치된 마스킹되지 않은 데이터 요청을 수행한다. 공개 가능성에 대하여 결과물을 체크하고 연구자들에게 돌려준다. 이 서비스는 반환된 결과물이 비밀기법에 적절하게 점검되었음을 확인하는 기술책임자에 의존한다. 기술자의 가용성, 다른 워크로드 및 결과물의 복잡 / 볼륨에 따라 2-5일 동안 걸린다

2007년 1월에, 연구그룹이 연구목적을 위해 마이크로데이터 접근 향상을 위한 국제적 발전 검토와 새로운 접근을 추구하기 위한 목적으로 설립되었다. 검토범위는 추적조사에서 데이터에 중점을 두고 있는 가계조사 데이터셋으로 제한하였다. 특별한 고려는 교차 국가 연구의 목적을 위한 캐나다이외의 연구자들의 접근을 용이하게 하였다. 특히 조사된 3분야는 합성 데이터 파일(synthetic data files), RTRA 및 데이터 공유이다. 더 상세한 것은 캐나다 통계청(2007)에서 찾을 수 있다.

3. 캐나다 통계청의 RTRA 계획

목적은 3~5년후 기술통계학에서 파생된 메타데이터를 위한 통계표뿐아니라 모델링과 복잡한/섬세하고 프로그램의 원격작업을 풀 패키지 RTRA 프로그램 패키지를 제공하는 것이다. 프로그램은 이용자 친화적 형식이다. RTRA의 개발에는 세 가지 주요 단계가 있다. 이미 완료된 첫째 단계은 보안, 법적 및 기능 요구 사항과 다른 구성 사항에 대한 깊

은 이해를 얻을 수 있도록 에이전시를 허용하는 캐나다 통계청 사업요구를 수집하였다. 현재 수행중인 두 번재 단계는 보안요구사항을 기본적으로 프로토타입을 구체적인 툴로 변환하는 것이다. 세 번째단계는 점진적 접근으로 보안수준을 측정하기 위한 모든 툴과 세밀하게 짜여진 프로그램을 테스트하는 것이다. 첫 번째와 두 번째는 다음 섹션에서 발표하고 세 번째는 섹션5에서 발표한다.

3.1 첫 번째 단계 : 사업요구 사항 정의

첫 번째 단계는 StatCan가 성공적으로 원격 접근시설을 만들 수 있는 사업 요구 사항을 수집했다. 이 과정은 기본적으로 비슷한 NSOs의 과거와 현재의 경험과 계획에 대한 전반전적인 방향수립 포함했다. 연구그룹은 주의깊게 네덜란드통계청, 호주통계청(ABS), 국가통계사무소(ONS), 인스티투트 드 라 Statistique 뒤 퀘벡(ISQ)의 다른 RTRA시설을 조사했다. 장기계획이 조사를 기반으로 개발되었다.

첫 번째 단계는 요구사항 즉 기본적인 기능, 절차, 프로그램의 지배 구조 등을 결정했다. 핵심 요소는 다음과 같다 :

- 범위 : 조사와 행정데이터를 위한 기술 통계와 모델링
- 접근 방식을 정의 : SAS프로토타입 구조의 완성. 프로그램은 연방정부 부서 즉 사전 인증된 파트너가 사용할 수 있다: 그러나 위험의 평가 후 확대할 수 있다.
- 프로세스 모델 개발 : 시스템에 액세스하는 방법을 결정한다. 이것은 정보학, 계약 및 법적 측면을 포함한다. 그램 1은 프로토타입이 개발된 모델을 이다. : 접근을 위한 사전 인정연구자의 요구부터 결과물 수령까지임
- 지배 구조 모델 개발 : 파일의 직접접근, 간접접근, 원격접근, RDC등 서비스로서 이 새로운 접근을 어떻게 관리하고 적용할 것인가의 결정서비스
- 보장하는 적절한 조직의 구현 : 개인 정보와 위험의 개인 영향, 개인정보 취급, 시스템착수를 위한 커뮤니케이션 / 마케팅 계획

3.2 2단계 : 보안 구조 결정

현재 진행중인 두 번째 단계는 비밀과 관련된 요구사항을 구체적인 조치로 변환하는 것이다. 네가지 "보안"관리 포인트 정책이 선택되어야 한다. (1) 보관된 데이터 셋의 보안 (2) 전송시 데이터의 보안 (3) 등록 이용자의 검증 (4) 결과물의 비밀 규칙. 포인트 4는 섹션 4의 유일한 주제이다. 다양한 보안 요소는 사업용으로 만들 수 있는 패키지로서 보여 질 것이다. 예를 들어 더 많은 데이터 마스킹, 덜 엄격한 사용자 인증, 결과물에 대한 덜 제한 적인 통치.

현재 선호된 모델은 호주통계청 원격접근 데이터 연구소와 유사하다. 합법적인 사업에 서명 후, 호주에 거주하는 사용자는 인터넷을 통해 서버에 링크를 이용할 수 있는 이용

자명과 패스워드를 받는다. 그들은 소프트패키지(호주통계청의 경우 SAS, Stata 및 SPSS)를 사용한 작업을 전송할 수 있다. 이용할 수 있는 소프트웨어는 특정 명령의 사용을 막고 결과물의 특성과 크기를 언급하는 규칙에 응답하도록 수정되어진다. 규칙을 준수하지 않는 요청들은 메모로 제출할 수 있고, 직원이 비밀보호를 위한 결과물이라고 진단할때까지 격리로 보관될 것이다. 모든 작업 ADL에 대한 감사를 위한 목적으로 보관된다. 따라서 시스템은 가볍게 마스크된 데이터, 법적책임, 이용자 교육, 사용자 인증, 결과물 제한 및 감사과 같은 여러 계층의 보호에 의존한다.

캐나다 통계청에서, 프로토타입은 현존하는 e-file전송시템(EFT)으로 만들어 질 것이다.: 고도의 보안장치는 두 개의 네트웨크를 개발했다, 그것은 네트웨크 사이의 air gap를 이동 기본 플랫폼으로 이용될 것이다. 다이어그램 2는 EFT를 기반으로 만들어진 프로토티입을 나타낸다. RADL과 마찬가지로 비밀번호 접근이 구현되어다. SAS는 고유의 소프트웨어 패키지로서 인정되어졌다. 법적 측면은 현재 개발중이고, RDC로 신중하게 설계되어지고 있다. 계약철차, 계약 범위, 데이터의 사용, 멀티요청, 연계 문제와 처벌 측면을 포함한다.

4. 비밀 측면

비밀 데이터를 정의하는 절대적인 기준은 없다. 비밀과 비밀이 아닌 것의 경계는 위험을 무시할 수 있나 없나의 임계값으로 해석할 수 있다. 캐나다 통계청은 마이크로데이터의 비밀을 보호할 수 있는 리스크 관리 정책을 가지고 있다. 이것은 지배구조, 보안대책(물리적 및 전자적 보안), 인증 직원 사용을 포함한다. 이것은 초기에 사용되는 기본적인 모델이 될 것이다. 시간이 지남에 따라, 접근범위가 증가할 것이고, 전문지식이 개발되고, 위험관리도 개선되어 갈것이다.
RTRA 구현에서는, 공개 제어를 위한 규칙을 개발하는 것이 필수가 될 것이다. 문헌과 다른 데 NSOs 보고에서 보는 것처럼, 네 가지 다른 면이 자주 고려된다. 네 가지 즉 가볍게 마스킹된 마이크로데이터 파일, 통계표 결과에 대한 자동제공 규칙, 사전스캔 즉 입력(수동과 자동)에 대한 제어규칙, 사후스캔 즉 결과물(수동과 자동)에 대한 제어 규칙이다. 다시 말하지만 , 캐나다 통계청이 선호하는 전략은 네 가지 잠재적 방법의 협정이 될 것이다. 제안된 방법들의 선택하는 결정과정은 비밀의 다른 수준을 고려하는 반면 위험을 다루는 것을 포함한다.

마스킹 또는 비 마스킹

캐나다 통계청의 RTRA경우, 비록 사용자가 등록해야 할 지라도, 어느정도의 감사 제어가 가능하다. 물리적 비밀 체크와 이용자의 서약은 요구되지 않는다. 게다가, 데이터베이스는 만약 이용자가 외부에서 접근해야 된다면 캐나다통계청 안전 네트워크 안에 둘 수 없을 것이다. 결국, 자동적 점검은 다중의 추적과 중복된 요구에 매우 제한된 수단을 제공하므로 2차 공개의 무시할 수 없는 위험을 초래한다. 이러한 이유로, 만약 마이크로데

이터 파일이 약간의 마스킹, 자세한 지역 제거, 응답범주의 그룹화, top-코딩 등으로 되었다면 신중해야 할 것이다. 그러나 PUMF과는 거리가 멀다. 가능한 전략은 일부를 약간 수정하는 것이다. 또는 유일한 존재의 가능성을 규명하거나 높은 위험 기록만 수정하거나 마스킹하는 the Skinner-Elliot 수단 (Skinner and Elliot, 2002)을 이용할 것이다. 마스킹할 것인가 아닌가의 결정은 여전히 평가되어지고 있다.

RTRA의 범위는 기술통계학, 조사에 대한 모델링, 데이터 관리이기 때문에 공개규칙은 표 출력을 개발하는 것이 필요하다. 이러한 목적을 위해서 몇 가지 방법이 현재 고려되고 있다. Tau-Argus (Hundepoolet al , 2009)과 랜덤과 같은 통제된 표 조정 (Cox, 2004), 제한된 절삭 방법과 같은 패키지사용 제한 전략을 포함한다.

평가에 대한 기준은 공개 정보의 풍부함, 다중요구의 위험평가 능력, RTRA 환경에서 자동화/프로그램의 용량, 결과물의 신속성, 많은 데이터 셋과 결과물을 취급하는 용량, 모든 조사프로그램과 RDC 규칙간 조화의 편이성이다. Boudreau, Filep and Liu, (2004)에서 설명한 절삭법통제는 현재 빈도데이터에서 인기가 있다. 수동 검증을 위한 규칙으로서 최소 비가중 계산이 또한 고려되어 지고 있다.

사전 및 사후 검사 규칙

언급한 바와 같이 이러한 규칙은 두 장소사이에 기본적 차이가 있을 지라도 RDC에서 사용하는 것과 매우 유사하다. Tambay 2007은 이러한 유형의 규칙을 생각하도록 잠재적 방법과 논재의 좋은 개요를 제공한다. 그 외에도 작은 진전이 두 곳에서 만들어 졌다.

5. 프로토타입, 테스트, 미래

2009년 4월에, 단지 논리적의 평가를 위한 첫 번째 평가가 EFT를 이용해 완성되었다. 다른 연방 정부에서 일하는 승인된 고용자, PUMF, SAS 프로그램을 완성하였다. 프로세스에 수동적인 개입이 없었기 때문에 매우 성공적으로 평가되었다. 모든 것은 자동적으로 완성됐다. 2010년 봄으로 잡혀있는 두 번째 테스트는 도표 제공절차와 제한된 종단면조사 데이터 파일의 사용을 제공할 것이다. PROC Means, PROC Tabulate, PROC Summary, PROC Freq의 이러한 단계는 단지 허락된 절차에서만 이루어진다. 약간 마스킹된 데이터를 이용할 것인지 아니지 또는 테스트를 위한 표 결과물 이용을 어떤 공개 방법으로 할 것인지의 결정은 아직 최종적이지 않다.

이후 몇 년 동안, 이 계획은 절차를 인정하는 기능을 향상시키고 표준 설정이다.
서비스가 의뢰인 피드백과 요구를 조정, RDC에서 새로운 넓은 지역 네트웨크 구축 고려, 더 많는 종단 조사자료를 미리 접근할 수 있도록 부가하는 것, 진행과정이 인증된 추

적 데이터 개발, 진행과정이 인증된 행정자료, 학계와 민간 부문의 프로세스 모델 확대이다.

6. 결론

캐나다 통계청은 실시간 원격접근을 위한 어플리케이션 개발과정의 초기 단계이다. 기본 모델부터, 모델들은 공개위험이 더 잘 설명되고 관리되도록 확장할 것이다. 시간이 지남에 이 신청은 공개용 마이크로데이터부터 RDC의 비밀마이크로데이터의 범위까지 캐나다 통계청의 마이크로데이터에 대한 접근 연속체의 일부가 될 것이다. 이러한 범위 내에서 RTRA는 학술 및 정책 연구가들에게 마이크로데이터의 효과적이고 시의적절하게 제공할 것이다.

부록 1.5. Working Paper 20

RDC를 통한 결과물 검토에 대한 지침

개요 : 이 논문은 “Guidelines on output checking" ESS(European Statistical System)net 프로젝트의 결과물이다. RDC(Research Data Centre)를 통한 결과물 검토에 대한 어려움을 설명한 후 결과물 검토에 대한 두 가지 모델(경험 모델 : the rules of thumb, 원칙기본 모델 : principles-based model)이 소개된다. 결과물의 모든 형태는 분류되어지며, 각각의 분류를 위해 경험적 모델과 근본적인 원칙이 논의되어 진다. ESSnet 프로젝트가 완료되지 않아서 규칙 및 원칙에 대한 견해만 기술할 것이다.

1. 서론

이 논문에서 “결과물 검토 지침” 프로젝트의 결과가 소개되어진다. 이 프로젝트는 ESSnet 프로젝트 NSIs의 컨소시엄에 의해 수행된 것이다. 컨소시엄은 독일, 이탈리아, 영국, 네덜란드의 통계기관으로 구성되었다. 이 프로젝트는 2009년 말에 완료될 것이다. 그 후에 프로젝트의 최종결과는 http://neon.vb.cbs.nl/case에서 볼 수 있다.

프로젝트는 일반연구자에 의해 분석된 마이크로데이터의 연구결과물에 대한 비밀보호를 다룬다. 많은 국가통계기관들은 외부연구자에게 비밀보호된 마이크로데이터를 접근하도록 한다. 이러한 접근은 국가통계기관의 RDC를 통해 안전한 환경에서 제공한다. 이러한 안전한 환경에서 도출한 모든 결과(result)는 결과물(output)이라고 한다. 이러한 결과물은 비밀보호가 안된 자료가 일반 사회에 공표되는 것을 확인하기 위해 검토가 필요하다. 대부분의 국가통계기관은 결과물 검토에 대해 자신의 경험을 발전시킨다.

결과물 검토의 지침과 경험에 대한 정의를 의해 프로젝트 시작의 많은 이유가 있다.

1. 분배된 경험과 가장 좋은 경험에 의해 경험한 나라들은 각각 다른나로로부터 배울 수 있다.
2. RDC 운영을 원하고 결과물 검토에 대한 경험이 없는 나라는 안전하게 결과물을 제공할수 있는 노하우를 배울수 있다.
3. 국가간 자료접근에 향한 움직임은 ESS내에서 명확하게 볼수 있다. 만약 국가통계기관이 몇몇의 다른 국가통계기관으로부터 결과물 검토를 한다면 국가간 자료접근은 효율적인 방법으로 가능하다. 결과물 검토 경험을 갖은 국가통계기관들의 동의는 이런것을 위해 결정적이다.

2. 결과물 검토의 어려움

2.1. RDC의 특성

결과물 검토의 어려움을 이해하기 위해서는 RDC의 특성을 이해하는 것이 필요하다.

대부분의 국가통계기관은 마이크로데이터의 모든 가능성이 직원과 발간물에 의해 결코 뽑아낼 수 없다는 것은 알고 있다. 조사에서 등록 통계로의 변화는 이용할 수 있는 마이크로데이터의 빠른 증가로 이어졌다. 동시에 IT 발전은 대용량의 자료를 분석할 수

있게 했다. 그러나 많은 국가통계기관은 단순하게 사용가능한 모든 자료를 제공하지는 않는다. 운좋게 대학교, 정책 입안자 등의 자격이 있는 연구자는 이러한 자료를 이용할 수 있다. 그러므로 국가통계기관은 보통 RDC를 통해 마이크로데이터 파일을 연구자에게 제공한다. 모든 국가연구자들간의 균형있는 점을 찾는다.

특히 공인된 연구자는 대부분의 세부적인 원본 마이크로데이터를 안전한 RDC를 통해 이용한다. RDC를 통한 결과물은 국가통계기관의 비밀보호 기법에 의해 검토된다. 간행물은 정의된 형태가 있고 침입자 시나리오는 수를 제한한다. 그런데 RDC 결과물은 어떤 것이 될수 있다. 연구자들은 다른 혹은 복잡한 방법으로 원본자료를 꼬거나, 전환하거나, 연결시킨다. 이런 것들은 결과물 검토를 어렵게 만든다.

2.2. 오류의 두가지 유형

결과물 검토는 마이크로데이터 연구의 사회적 가치를 갖고 균형있는 비밀보호에 간한 것이다. 결과물 검토의 완전한 방법은 자료의 이용도이 커지게 하고 노출위험을 최소화 시키는 것이다. 결과물 검토에 있어서 에러 유형은 아래와 같이 2가지가 있다

1. 비밀보호 에러(Confidentiality Errors) : 불안전한 결과물 제공
2. 비능률적 에러(Inefficiency Errors) : 안전한 결과물 미제공

집계표의 셀값에 대한 단순한 규칙을 예를들어 생각해보자. 만약 셀값을 너무 높게 설정하면 비능률적인 에러가 발생할 것이다. 결과물은 안전하나 정보는 적을 것이다. 다른 경우는 셀값이 너무 적어 노출위험이 높을 것이다.

3. 규칙과 지침

3.1. 두 가지 모델 : 원칙기본 모델과 경험적 규칙 모델

결과물 검토에 대한 두 가지 다른 모델은 이 논문에서 소개할 것이다. 첫 번째는 원칙기본 모델이라 부른다. 이 모델은 비밀보호 에러, 비능률적 에러 둘다 가능한 적게 하는 것이다. 다른 것은 경험적 규칙 모델이라 부른다. 이 모델의 초점은 비밀보호 에러에 예방하는 것이고 비능률적 에러에는 많이 신경쓰지 않는다.

3.2. 원칙기본 모델

원칙기본 모델은 연구자와 RDC 직원 간의 좋은 협조에 중점을 둔다. 또한 이 모델은 비능률 에러를 방지하도록 노력하기 때문에 결과물 검토에 대한 단순한 규칙은 사용할 수 없다. 그 이유는 단순한 규칙은 연구 결과물의 복잡성을 고려할 수 없기 때문이다. 모든 결과물은 결과물의 안정성을 검토하기 전에 전체 상황을 고려해야 한다. 예를 들면, 매우 작은 셀값을 포함한 표는 반드시 불안전하지는 않다. 예를 들면 원본자료가 이미 변환되었다면 정보는 개별자료를 식별할수 없을 것이다. 필요한 것은 노출조절에 의한 적용하는 규칙을 명확하게 이해하는 것이다. 그러므로 연구자와 RDC 직원은 노출관리에 대한 훈련이 필요하다.

원칙기본 모델에서 결과물의 모든 다른 형태는 안전하거나 안전하지 않거나 분류된다. 예를 들면, 회귀분석은 근본적인 자료와 상관없이 안전한 결과물로 간주된다.

만약 결과물 분류가 안전하다면 결과물의 이 유형은 일반적으로 제공될 것이다. 단지

예외적으로 국가통계기관은 제출된 결과물 제공되지 않을 수 있다. 예를 들면, 회귀는 모든 설명변수가 기본적으로 이진표이고 잠제적으로 노출을 할 수 있다. 예외적인 경우 수를 정의하거나 제한해야만 한다.

만약 결과물 분류가 안전하지 않다고 생각되면 연구자가 그것이 노출위험이 없다는 것을 증명하지 않는 한 제공되지 않을 것이다. 자료를 제공받기 위해 연구자는 노출 원칙과 연구자가 사용하는 자료를 잘 이해해야 한다. 그러므로 연구자의 교육은 필수적이다. 안전하지 않은 유형의 결과물을 국가통계기관이 제공하기 위해서는 연구자와 긴밀한 관계가 필요하다.

원칙기본모델은 연구자에게 최대한의 유연성을 주는 장정이 있다. 그러므로 자료는 최대한의 범위로 사용될 것이다. 그러나 이 모델은 또한 단점을 갖고 있다.

- 모델은 국가통계기관 직원과 연구자에 의존한다. 연구자는 그들의 시간과 노력을 투자하는 것이 필요하다
- 기본모델(rules-based)에서 책임은 규칙은 설계한 사람에게 있다. 원칙기본모델에서 책임은 각각의 확인자에게 있다. 확인자는 자신의 경험과 이해 정도에 의존하여 결과물의 제공여부를 결정하기 때문에 엄격하지 않다.

이러한 단점을 보완하기 위해 다른 모델(경험적 모델)을 소개한다.

3.3 경험적 모델

이 모델에서 주요 초점은 비밀보호 에러를 보호하는 것이다. 비능률적 에러는 허용된다. 이것은 매우 엄격한 규칙이다. 이 규칙을 통과한 결과물은 노출위험성이 매우 낮다. 규칙은 연구자와 직원이 제한된 지식과 노출기법을 갖고 자동적으로 적용할 수 있는 이익이 있다.

비록 경험적 규칙인 매우 엄격할지라도 이 모델이 결과물이 노출위험이 없다고 100% 보증할 수 있는 없다. 노출이 있는 결과물이 가끔 제공될 수도 있다. 이러한 것들은 규칙이 융통성이 없고 결과물의 모든 배경을 고려할 수 없기 때문이다.

경험적 모델은 많은 상황에서 유용하다.

- 결과물의 연구자는 보통 노출조절로부터 멀리 있다(예를들며 제한된 세부사항이 포함된 통계 결과물을 원하는 정책자(policy maker))
- RDC를 시작하는 경험이 없는 국가통계기관. 이 경우 연구자와 RDC 직원은 원칙기본 모델을 갖고 일하기에는 너무 경험이 적다. 경험적 모델은 가장 안전한 시작점을 그들에게 제공한다. 경험적 모델을 사용함에 따라 그들만의 경험을 축척할 것이다. 시간이 지남에 따라 그들은 원칙기본모델을 설정 할수 있으며 좀더 복잡한 결과물에 대해 자료를 제공할수 있을 것이다.
- RDC에 대해 자동적인 노출조절 방법이다. 이것은 주로 원격실행(remote execution)과 같은 자료제공유형에 유용하다. 원격실행에서 연구자는 더미자료 사용한다. 그때 그들은 마지막 스크립트를 모든 자료를 실행할 수 있는 RDC 직원에게 보낸다. 결과는 연구자에게 보내진다.
- 원칙기본 모델을 사용하는 RDC 일지라도 경험적 모델은 특별한 결과물을 검토할 때 시작점이 된다. 경험적 규칙 사용은 신속하게 결과물의 포인트를 맞출수 있다. 자

료를 제공할지 안할지에 대해 좀더 신중하게 고려할 수 있다.

3.4. 결과물의 다른 분류에 대한 규칙(Rules)과 원칙(principles)

결과물의 가능한 유형은 연구자 연구형태의 분류되어 진다. 그때 각 분류는 안전한지 안한지를 고려한다. 각각의 분류에 대해 경험적 규칙 및 원리원칙은 정의되어 진다. 이 프로젝트는 아직 완료하지 않았기 때문에 이 논문에서 규칙과 원칙에 대해 논의하지 않을 것이다. 단지 몇가지 예를 들것이다.

결과 분류	경험적 규칙	원리기본 원칙
회귀	제공	제공
빈도표	・가중되지 않은 셀값 >=10 ・행/열 합계의 90%이상 포함하지 않는 셀	아래의 설명에 대해 제공할 수 있다. ・자료가 자신을 노출하는지(변형되었는지;세부수준) ・만들어진 자료가 회사 식별되는지 ・임계값과 이상값 ・응답자의 순위가 알려졌는지 ・셀단위 선택 ・표본선택 ・가중
최소/최대	미제공	노출이 안될때 제공

3.5. 조직과 절차 관점에서의 지침

지금까지 논문은 결과물 검토에 대한 실제 규칙을 다루었다. 이러한 규칙은 떨어져 있어도 질높은 유용한 결과물 검토 절차는 조직과 절차의 이슈가 알맞게 있을 때 존재한다. 프로젝트는 또한 각각의 이슈에 대한 실용적인 지침서를 정의했다. 최소한의 요구는 효율적으로 결과물의 안전을 관리할수 있는 RDC이다. 최고의 숙련은 조직의 충고를 포함하고 RDC의 목적을 표현하는 것이다.

이 논문은 아직 프로젝트가 완료되지 않았기 때문에 지침에 대해 세부적으로 논의하지 않았다. 그러나 아래와 같이 간략하게 언급할 수 있다.

- 결과물에 대한 책임 : 연구자는 모든 결과물에 대한 내용과 질에 대한 책임이 있다는 것은 국가통계기관은 분명히 해야 한다. 국가통계기관의 역할은 결과물에 대한 노출위험을 관리하는 것이며, 그밖에는 아무것도 없다.
- 검토자 인원 : 각각의 결과물에 대해 최소한 1명 있어야 한다. 가장 좋은 것은 결과물의 질과 목적을 위해 각각의 결과물에 대해 2명이다. 만약 한명은 RDC에 있고 다른 한명은 자료관련부서에 있다면 각각은 검토 절차에 있어서 구체적인 지식을 줄것이다. 이러한 경우 최종결정은 RDC에 달려 있다.
- 결과물의 규모 :

4. 결론

이 논문은 ESSnet 프로젝트 “결과물 검토 지침”의 임시결과에 대해 논의했다. 프로젝트는 2009년 말에 완료가 된다. 그 후 프로젝트의 모든 결과는 http://neon.vb.cbs.nl/cased에서 찾아 볼 수 있다.

프로젝트는 조화로는 규칙에 결과물 검토의 유연성 있는 절차를 획득하는 것이 첫 번째 목적이다. 그러나 아직 새롭게 나타나는 세계적인 자료를 접근하는 효율적인 결과물 검토 절차 개발이 필요하다.

부록 1.6. Working Paper 22

데이터이용자가 사용하는 센서스 표를 자동으로 비밀보호하는 방법의 도구화

요약

호주 통계청은 센서스 테이블에 대해 자동으로 보호하는 기법을 개발해왔다. 이 기법은 유사한 테이블과 동일한 테이블에 대한 반복적인 요청에 대한 대비를 할 수 있다. 이러한 기능한 사용자 정의 테이블을 생성함으로써 웹기반 시스템에서의 사용을 가능케 한다.

이 방법은 각 레코드에 영구적인 숫자로 된 키를 할당한다. 이러한 키는 그 테이블에 적용되는 perturbation기법을 위한 일관성있는 값을 생성하는데 사용된다. 동일한 사용자가 동일한 셀을 사용할 때 같은 perturbation기법이 적용된다. perturbation기법은 모든 셀의 차별성을 방지하는데 적용된다. 알고리즘은 추가분에 대한 복원을 하고 총합계의 perturbation을 유지한다.

일련의 제품들이 비밀화된 센서스 데이터에 접근할 수 있도록 개발되어왔다. 이러한 제품들은 사용자들이 웹기반의 시스템을 통해 자동으로 보호된 테이블을 정의하고 다운로드하는 기능을 포함한다. 나아가 가중치가 부여된 것과 대용량 데이터들의 보호에 대한 방법론의 확장에 대한 가능한 방향성을 제시한다.

1. 소개

호주통계청은 통계정보 수집을 위한 권한을 법으로부터 부여받았다. 이 법은 이러한 권한으로 수집된 어떠한 데이터도 개인 또는 조직의 정보를 포함하여 제공할 수 없다는 내용을 담고있다.

호주통계청은 이러한 입법 요구사항을 따르기위한 정책들을 개발해오고 있다. 그중 하나가 아주 작은 부분 모집단에 대한 정보를 포함한 테이블의 제공에 관한 것이다. 호주통계청의 정책에는 아주 작은 값들을 포함한 셀이 있는 자료는 일반적으로 제공하지 말아야함을 명기하고있다. 또한 사용자들로 하여금 다른 제공받은 데이터로부터 이 셀들의 값들을 유추해낼 수 없어야한다.

호주 통계청은 5년마다 인구주택총조사를 실시한다. 총조사를 통한 데이터는 작은 부분에 대한 자세한 분석이 가능하며, 다른 곳에는 사용되어서는 안되는 중요한 정보의 원천이다. 그러나, 총조사 자료를 제공하는데는 법과 정책을 따라야 함은 매우 중요하다.

2006년 인구주택총조사를 위해 호주통계청은 총조사 자료를 자동적으로 보호해주는 방법을 개발했다. 이 기법은 유사테이블, 동일한 테이블에서의 반복되는 요구, 서로다른 테이블에서 같은 셀에 대한 반복적이 요구에 대한 보호를 위해 고안되었다. 이 기법은 부분 모집단 자료에 대한 접근성을 향상시키고 웹기반의 시스템을 통해 사용자들이 그들 자신의 테이블을 정의하게끔 할 생각이다. 이 방법에 대한 목적은 2장에서 자세히 설명하겠다.

이 방법론은 이미 2005년 Fraser와 Wooton에서 발표된 바 있다. 3장에선 이 방법에 대한 간략히 설명할 것이다.
이 방법은 2006년 센서스 테이터에 대한 비밀보호 데이블 자동 생성을 위한 여러 제품들에 적용되었다. 사용자 정의 테이블을 생성하고 다운로드할 수 있는 웹기반 시스템 2개도 여기에 포함된다. 4장에서는 이 제품들에 대해 설명하겠다.
호주통계청은 이 기법의 확장 가능성에 대해 고심중이다. 목적은 유사한 웹기반의 시스템을 통해 다른 자료들에 대해서도 통계 테이블을 생성할 수 있게 하는 것이다. 호주 통계청은 가중 데이터 및 대용량 데이터에 대한 보호 기법에 적용할 수 있는 방법을 모색해 나갈것입니다. 미래에 대한 가능한 방향성에 대해서는 5장에서 논의할 것이다.

2. 이 기법의 목적

이 기법은 사용자들로 하여금 스스로 테이블을 정의할 수 있는 웹기반의 제품에 사용되기 위해 고안된 것이다. 미리 만들어진 비밀 보호가된 테이블에서 자료를 제공하는 시스템과 비교했을 때, 사용자가 테이블을 정의하는 시스템은 추가적인 중대한 위험요소들이 존재한다.
첫번째는 사용자들이 같은 테이블에 대한 반복적인 요청에 대해 보호가 일부 이루어지지 않을 수 있다는 것이다.
만약 비밀보호기법이 perturbation 데이터에서 랜덤으로 자료를 받는 과정을 거친다면, 사용자는 같은 테이블에 대한 다른 자료를 받을 수도 있을 것이다. 이러한 서로다른 자료드에서 각 셀값들을 비교하면 비밀보호가 되지않은 테이블에 대한 원래 값이 드러나게 될 것이다. 예를들어 셀 값들을 평균하면, 보호되지 않은 원래 자료가 드러날 것이다. 웹기반의 제품은 이러한 위험을 방지할 필요가 있었던 것이다.
두번째는 차별성와 관련된다. 차별성은 사용자가 유사한 두 부분 모집단에 대해 더 작은 부문 모집단에 대한 데이터를 찾으려 할 때 발생할 수 있다. 호주통계청은 사용자들에게 부분 모집단을 정의함에 있어 다양한 방법을 사용할 수 있도록 의도하였다. 그러므로 이 기법은 차별성에서 기인하는 위험 또한 방지할 수 있다.
또한 웹 기반 제품에서 유용하게 사용되는 이 기법은 유연성을 제공한다. 사용자는 센서스를 통해 수집된 다양한 자료들에 대해 이와 관련된 특정 테이블을 정의할 수 있다. 표준 지리 구조 또한 다양한 레벨을 가지고 있다. 이상적으로 사용자들은 이러한 다양한 레벨들로 부터 특정 테이블을 정의할 수 있으며, 표준 지리 구조와 조합하여 그들 자신만의 특정 지리 영역을 생성해 낼 수 있다.
총조사를 통한 표준 출력물은 인구와 가족, 주거에 대한 수치도 포함된다. 이 기법은 세가지 다른 레벨에 대한 테이블을 생성하는 것도 가능하다.
이 기법은 분석의 목적에 적절하지 않기 때문에 데이터를 변경하여서는 안되고, 테이블은 일관성있는 호주 인구 사진을 제공해야만 한다.
끝으로, 이 기법은 유효적절한 시간에 비밀보호가 이루어져서 실용적으로 적용되고 사용될 수 있어야한다.

3. 기법에 대한 간략한 설명

이 기법에 대한 발표는 이미 2005년에 실시한 바 있다. 그 자료에는 이 기법에 대한 자세한 설명이 있으며, 호주통계청에 의해 처음으로 실행되었음이 설명되어 있다. 이 기법은 각 단위 레코드의 유한 그룹 요소들을 연결한다. 이 요소들은 인덱스로 조합, 연결되어 각 셀에 적용되는 perturbation을 얻어내는데 사용된다. 호주통계청이 이 기법을 적용할 때, 그 요소들은 '레코드 키'로 알려진 영구적인 숫자값이다. 테이블이 생성될 때, 레코드키는 모듈로 산식을 사용, 조합되어 셀 레벨의 키로 만들어진다. 이 셀 레벨의 키는 셀에 적용되어진 고정된 검색 테이블의 perturbation을 정의하는데 사용되어진다. Zero 셀은 레코드를 생서하지 않고, perturbation도 생성되지 않는다. 내부 셀들과 합계값들은 독립적으로 perturbation되어있기 때문에, perturbation된 데이블은 추가될 수 없다.

이 기법은 각각의 셀들이 어떠한 테이블에서도 동일한 perturbation기법을 얻을 수 있도록 해준다. 그러므로 반복된 요청에 의한 같은 테이블에는 같은 결과가 나온다. 또한 서로 다른 테이블의 아주 작은 셀도 매번 같은 방식이 적용되기 때문에 다양한 결과를 원래 값을 예측하는 것을 방지 할 수 있다. 집계 테이블은 원래 셀의 값과 셀 레벨의 키에 기인해 적용된 perturbation값들을 포함한다. 셀 레벨 키는 집계 테이블의 열을 정의하고, 셀의 값은 집계 테이블의 칼럼을 정의한다. 보다 큰 셀 값을 위해 모듈로 산식이 칼럼을 정의하는데 사용된다. 집계 테이블에 포함된 perturbation기법의 적용으로 perturbation의 법위가 적용되어 질 수 있으며, 평균 규모의 perturbation이 0이 아닌 셀들에 적용이 가능한 것이다.

0인 셀을 제외하고 테이블의 모든 셀은 perturbation될 기회가 주어지는 것이다. 모든 셀에 적용되는 perturbation의 최대값은 고정되어 있다. 이는 보다 큰 셀은 적절하게 적은 양의 perturbation이 적용되게 된다는 것이다.

모든 셀이 perturbation이 가해질 수 있기 때문에 두 테이블 간의 차별에도 동일하게 perturbation이 적용되는 것이다. 이는 작은 셀값의 차별성을 방지한다. 만약 작은 셀이 차별성에 의해 계산되어진다면, 이 셀들은 평균적으로 적절하게 큰 perturbation이 적용되게 된다. 이 기법은 차별성이 실제값에서의 차별을 보장하지는 않는다. 하지만, 실제값의 분산은 식별되어질 수 있는 개인정보를 보호할 수 있는 충분한 불명확성을 제공한다.

이 기법은 사용자가 테이블을 위해 선택한 데이터 아이템들이나 지리적 영역에 의존하지 않는다. 이 기법은 데이터의 다른 레벨들에 적용될 수 있기 때문에, 이들 레벨들간의 정확한 일관성을 보장하지 않는다. 예를들어 같은 지역내의 거주지수와 인구수는 서로다른 방법에의해 perturbation되어진다.

이 모든 이유때문에, 호주통계청은 이 기법이 웹기반의 테이블 생성 제품에 적합할 것이라 판단하고 선택했다.마지막으로 이 기법이 적용되기 이전에 정리해야할 상세한 사항들이 있다. 특히 아래와 같은 사항들이 연구되어져야 한다.

1. 레코드 키를 생성하는데 사용되는 분포
2. 셀 레벨의 키를 생성하는데 가장 적합한 레코드 키의 조합
3. 집계 테이블의 perturbation 값들의 분포
4. 추가적인 알고리즘

이러한 변수들을 선택하는 것은 위험이 노출되는데 상당한 영향을 끼친다. 특히, 집계 테이블 perturbation 값의 분포는 보호기법이 테이블에 적용되는 양을 결정한다. 그리고, 이러한 분포의 소유는 호주통계청의 법과 정책을 만족시키기 위해 선택되어졌다. 2005년 설명한 바와 같이 키와 집계 테이블은 다음 기준을 충족시켜주는 분포의 정수값 perturbation을 생성하기 위해 고안되었다.

1. 평균은 0
2. perturbation값은 음의 셀값이나 매우 작은 셀값에서는 생성되지 않음
3. perturbation값은 고정된 분산을 가진다
4. 어떤 perturbation값의 절대값이든 고정된 양의 정수 보다 작다

2005년에 제안된바와 같이, 추가적인 모듈 역시 전체적인 기법에 포함되어있다. 이 모듈의목적은 추가사항을 perturbation된 테이블에 적용하는 것이다, 따라서 사용자들은 더욱 편리하게 테이블을 사용할 수 있다.이 추가모듈은 다양한 제약 아래 사용되어지는 쌍방향의 기법들이다. 제약은 다음과 같다.

1. 결과 테이블은 음의 정수값을 가지지 않는다
2. 매우 작은 0이 아닌 셀값은 추가 테이블엔 나오지 않는다
3. perturbation화된 테이블의 총계는 계속 유지된다
4. 셀값의 변경의 최소화된다

추가 모듈은 perturbation 기법이 모든 것을 공유하지 않는다. 예를들어, 작은 셀 값이 두개의 서로다른 테이블에 나오면, 셀은 서로 같은 perturbation값을 받게된다. 그러나 이것은 추가적인 모듈에 의해 서로다른 방식으로 변경된다. 이것은 중대한 비밀보호상의 위험이라 생각되어지지 않았다. 이것은 예를들어 많은 양의 테이블에 대한 셀값의 평균 구하기의 경우에서와 같이, 사용자가 추가 모듈에 의해 적용된 변화에 대한 취소가 가능하다. 그러나 사용자는 원래 자료 말고 perturbation된 테이블에 대해서만 복구가 가능하다.

또한 0 셀은 추가적인 모듈에 의해 매우 작은 수의 사용되지 않는 테이블로의 변화가 가능하다. 이것은 논리적으로 어떠한 지원자를 포함하기 있지 않은 특별한 경우에 이상적이진 않게 사용된다. 그러나 또다른 제약을 위해 모듈로 약간의 0 셀을 변경하는 것이 가능하도록 해야할 필요가 있다.

추가 모듈에 의한 추가적인 변경을 perturbation 기법에 의한 것과 비교해 볼 때 대체로 작다. 이것은 보통 매우 작은 테이블에 추가 모듈이 큰 영향을 끼칠 때 적용된다.

4. 웹 기반 시스템

호주통계청은 센서스 비밀보호 기법을 활용하여 센서스 테이터로부터 테이블을 생성하는 다양한 시스템을 개발했다.

호주통계청 직원들에 의한 인쇄물을 포함한, 2006년 센서스로부터 생성된 모든 테이블은 동일한 기법을 사용하여 보호되어졌다. 이러한 이유로 이 기법의 적용을 위한 내부 시스템이 있다.

또한 통계청 외부 사용자들이 이용 가능한 웹기반의 제품들이 있으니 CDATA Online과 Census TableBuilder이다. 이들 제품들은 호주통계청과 Space-Time Research Pty Ltd에

의해 공동 개발되었다.
CDATA Online은 인터넷과 지원 가능한 웹 브라우저를 통해 무료로 이용이 가능하다. CDATA Online은 사용자들이 생성할 수있는 정형화된 테이블과 지도, 그래프들을 주제별로 다양하게 제공한다. 데이터 세트들이 사용자들이 요청가능한 테이블들을 정의하는 동안 테이터세트들은 비밀처리에 직접적으로 사용되지 않는다.
그것은 테이블들이 테이터세트안에서 결집되어진것들은 비밀적이지 않아서이다. 대신, CDATA Online을 사용하여 생성된 각각의 테이블은 표준방법을 사용하여 비밀화된다. 한명의 사용자가 테이블, 지도, 그래프를 요청하고, 각각의 결과들은 익스포트하거나 다운로드 받을수 있다. CDATA Online은 www.abs.gov.au/CDataOnline 에서 접근할수 있으며, 이 페이지는 또한 유저 매뉴얼에 대한 링크를 포함한다.

Census TableBuilder 는 CData Online과 비슷하다. 하지만 좀 더 유연성을 가진다. 기초적인 주제별 데이터세트들을 포함하는대신에 Census TableBuilder는 사용자들이 센서스 데이터 전체로부터 직접적으로 테이블을 생성하는 것을 허용하다. 이것은 사용자들이 그들의 테이블에 센서스 데이터의 대부분의 혼합들을 포함가능하다는 것을 의미한다. CData Online을 사용함으로써, 사용자들은 지도,그래프 뿐만 아니라 테이블들을 생성할 수 있다. Census TableBuilder 사용자들은 그 제품에 대한 사용을 허가받기전에 등록할 필요가 있다. www.abs.gov.au/TableBuilder 에 Census TableBuilder에 대한 더 많은 정보와 사용자 매뉴얼이 있다.

5. 앞으로의 방향

호주통계청은 2011년 Census에도 데이터를 비밀화하는데 동일한 방법을 적용하려고 하고 있다. 그리고 웹기반시스템의 큰 변화를 기대하고 있지 않다. 하지만 여전히 미래의 일을 위한 배출구이다.

비밀화되어진 테이블에 기반한 정보에입각한 결정을 하기위해 사용자들은 데이터의 품질에 대한 조치를 필요로 한다. 이상적으로, 사용자들이 데이터의 비밀을 위협하는 정보를 드러내지 않으면서 그들의 분석에의한 비밀효과를 조절할수 있도록 허용해주는게 좋다.
호주통계청은 카이스퀘어 테스트를 포함하여 정보손실을 측정하는 몇몇의 조사를 수행하고, 이 일은 계속되고있다.
이것의 목표는 CData Online과 Census TableBuilder의 사용자들을 위한 정보문서를 포함할 수 있는 지표를 개발하는것이다. 이 문서는 최근에 사용가능한 것보다 그 방법에 대한 더 많은 정보를 준다. 그것은 또한 혼란과 가법 또는 표로된 생성물의 품질의 측정과 관련된 평균적인 변화들에 대한 정보를 포함한다.

호주통계청은 최근에 데이터에대한 접근을 국제적으로 하기위한 타협적인 접근을 찾고 있다. 이것은 세계의 다른 통계조사기구에서도Census TableBuilder를 사용하도록하는 것이라고 할수 있다. 그래서 이러한 기구에서 자신의 센서스 데이터를 시스템으로 불러와서 그것을 사용자들이 사용할수 있도록 한다. 여기서 호주통계청의 데이터를 타협없이

대외적으로 사용할수 없는 요소가 몇가지 있다.
하지만, 호주통계청에서는 비밀요청을 받았을 때 이 요소들을 어떻게 해야되는지에 대한 조언을 해줄 수 있다. Census TableBuilder를 적용한 다른 기구에서도 그 시스템이 그들의 메타데이터와 호환이 되는지 증명하기를 필요로한다.

그 방법론은 호주통계청의 입법과 정책, 사용자들의 요청에 의해 설계되었다. 다른 통계기구들은 그들의 입법과 정책, 사용자들의 요구가 다르다. 이러한 이유 때문에 호주통계청의 방법론을 자신들의 관점으로 바꾸기를 좋아한다. 예를들어, 미리정의된 전체데이터를 보호하는 방법으로 레코드키를 할당하는 방법에 대한 기본적인 제안이 있다. 호주통계청은 이 경로를 선택하지 않겠지만, 다른 기구들은 그들의 센서스 데이터들을 정확히 보호하도록 결정할것이다. Shlomo와 Young는 호주통계청의 방법론의 특징을 통합하려고 시도하였지만 그것은 또한 난외적인 사항도 전체적으로 보호하도록 설계되었다.

호주통계청은 또한 방법론을 확대하는데 관심이 있으며, 시스템이 샘플조사 데이터의 테이블들을 자동으로 보호하도록 허용하는것에도 관심이 있다. Survey TableBuilder의 개발의 첫단계는 그 방법으로 샘플링된 가중치를 통합하는 가장좋은 방법을 결정하는 것이다.
그 방법은 테이블들이 더 이상 유용한 정보를 포함하고 있지 않다는 동요가 없도록 호주통계청이 요구하는 충분한 보호를 제공하는게 필요할것이다.

Survey TableBuilder에서 사용할수있도록 한 Survey데이터를 믿으면 추가적인 방법론적 변화가 있을것이다.

만일 호주통계청이 반복적인 조사의 테이블들을 사용하도록 한다면 그 시간대에 다른관점으로 추정이 가능한 시간적인 속성을 보호할 방법을 찾는게 필요하다.

또한, 최근의 방법론은 경제관련조사의 이익과 지출 같은 연속적인 데이터항목을 효과적으로 적용할 수 없다.
불안감과 관련된 장래성있는 수많은 접근이 있다. 연속적인 데이터항목의 보호를 필요로하는 불안감은 수를 필요로하는 불안감보다 비율적으로 크다.
이것은 정보와 효율성의 손실이 연속적인 데이터와 관련해서 더 크다는 것을 의미한다. 통계적 속성을 보호하는 것을 적응시킬수있는 수많은 접근이 있다. 하지만, 통계적 속성중 어느것에 더 높은 우선순위를 두느냐를 결정하는게 필요할 것이다. 이러한 제약들은 그 방법안에서 통합되어질수 있다.

Census 방법 아래 호주통계청은 Suvery 데이터에 증가된 접근을 허용하기위한 방법과 시스템을 개발하기위한 통합적인 접근을 찾는데 관심을 가지고 있다.

부록 1.7. Working Paper 32

합성데이터 구조 파일 : 개발 및 정보노출제한

요약

최근 수년간 연방 통계청 연구 데이터 센터에서, '제어된 원격 데이터 수행 및 안전 센터'는 경제통계 마이크로데이터를 접근할 때 가장 자주 사용되는 방법이다. 그러므로 InfinitE(e-science 시대의 지식 기반 - 2009년 9월 출범) 프로젝트의 목적 중 하나는 계량경제학 모델을 설명하고 문법 오류가 없는 코드를 형식화하는데 활용할 수 있는 노출 제어 데이터구조 파일을 개발하는 것이다. 이 파일을 제공하는 한 가지 방법은 다중 무응답처리에 기반한 합성 데이터셋을 만드는 것이다. 이 방법의 결정적인 장점은 그 보편성에 있다. 모든 제한과 필터구조가 고려될 수 있다. 이 논문은 원본 데이터와 보호 효과를 적용한 통계적 매칭으로 분석 잠재력을 고려한, 2001년 생산부문 월별 지역 리포트의 '부분 합성 데이터셋'을 만들기 위한 초기 접근법을 포함한다.

1. 독일 프로젝트 InfinitE

경제통계조사와 관련된 독일 통계 생산자는 그들의 통계 수요에 대한 근본적인 변화를 알아냈다. 2000년 초, scientific use files(SUFs) - 이곳의 연구원은 통계청이 아닌 그들의 일터에서 일할 수 있다 - 를 비롯한 과학 공동체는 독일에서 공식적인 마이크로데이터에 적절하게 접근하는 실험적인 방법을 한 가지 생각해냈다. SUFs는 선택적이며 강력하게 요구되는 통계에 이용될 수 있다. 하지만 경제 조사의 SUFs가 널리 허용되는 것은 아니다. 그 이유 중 하나는 새로운 데이터 섭동 방법 때문인데 이 방법은 데이터의 기밀성을 보장하기 위해 필요하지만 통계적 추론 가능성의 일부는 파괴한다. 또한 자료 수집과 관련 SUF 생성 사이에 과도한 지연이 있다. 최근 수년간 연방 통계청 연구 데이터 센터에서 '제어된 원격 데이터 수행 및 안전 센터'는 경제통계 마이크로데이터를 접근할 때 가장 자주 사용되는 방법이다. 그러므로 InfinitE(e-science 시대의 지식 기반 - 2009년 9월 출범) 프로젝트의 목적 중 하나는 계량경제학 모델을 설명하고 문법 오류가 없는 코드를 형식화하는데 활용할 수 있는 노출 제어 데이터구조 파일을 개발하는 것이다. 게다가, 연구 데이터 센터(the research data centres, RDC) 직원이 매우 많은 시간을 소비하는 결과물 확인 작업은 가능하면 자동화되어야 한다. 이 논문에서 우리는 데이터구조 파일의 개발과 다중 무응답처리를 통해 이러한 파일을 생성하는 방법을 첫 번째 논점으로 다룬다. 2장에서는 합성 데이터셋을 생성하는 다양한 방법과 우리 프로젝트에 지금까지 적용한 무응답처리 소프트웨어 IVEware를 간략히 소개한다. 우리 방법을 개발하고 테스트하는 데 사용한 첫 번째 데이터는 생산부문 월별 지역 리포트이다. 이 데이터를 3장에서 간략히 설명한다. 범주형 속성 자료에는 특히 난해한 부분이 있는데 이것을 4장에서 자세히 검토한다. 익명화의 한 면은 데이터의 분석적 잠재력을 유지하는 것이다. 5장에서는 우리 방법으로 이것을 얼마나 잘 이룰 수 있을 것인지 원본 자료와 합성 자료(2001년 월

별 리포트)의 비교결과로써 제시한다. 하지만 익명성 이상의 중요성을 갖는 것이 데이터의 기밀성을 유지하는 것이다. 특별히 변형된 합성 데이터의 보호 효과는 적절한 매치 실험으로 설명한다. 이 실험과 2110년 월별 리포트의 첫 번째 결과물에 숨겨진 이론의 개요를 6장에서 제공한다.

2. 데이터구조 파일의 개발

지금까지 데이터구조 파일은 종종 원본 자료의 샘플로 구성되었는데 이 원본 자료는 추가적인 익명화 작업이나 데이터셋 값 범위에서 임의로 만들어진 값으로의 변경이 필요했다. 변수들이 비록 두 가지 접근법 모두로 유지되었지만 변수의 속성과 거기에 의존하는 구조(필터, 편차-공분산 행렬)는 완전히 사라진다. 따라서 연구자는 실제 문제가 적절히 구현될 지에 대한 정보를 얻지 못하더라도 그 프로그램이 실행가능한지 여부를 확인할 수 있다. 이런 이유로 연구자의 분석 프로그램은 원본 자료를 사용하는 다음 단계를 위해 수정이 불가피한 경우가 많다. 그리고 추가 수정은 연구자와 RDC 직원의 몫이다.

2.1 합성 데이터구조 파일

상당히 고품질의 데이터구조 파일을 제공하는 한 가지 방법은 다중 무응답처리 기법에 기반한 합성 데이터셋을 만드는 것이다. 이 방법의 결정적인 장점은 접근방법의 보편성에 있다. 모든 제한과 필터 구조가 고려될 수 있다. 게다가 이 접근법은 범주형 변수와 같은 방법으로 연속형 변수에도 적용될 수 있다. 이 혁신적인 접근법은 높은 유연성과 복잡한 데이터셋에 대한 적용성 때문에 지난 수년간 국제적으로 널리 사용되었다.
다중 무응답처리를 활용하여 합성 데이터셋을 생성하는 방법은 1993년 Robin이 처음으로 제안했고 2003년 Raghunathan, Reiter 그리고 Rubin이 발전시켰다. 기본적인 원리는 개별적으로 분석되는 합성 데이터셋을 매번 여러 개 만드는 것이다. 분석의 결과로 단순한 결합 규칙이 만들어 진다(2003년, Raghunathan 등).
완전 합성 데이터셋과 부분 합성 데이터셋은 원칙적으로 구분될 수 있다. 완전 합성 데이터셋의 경우 샘플에 포함되지 않은 모집단은 결측값으로 처리된다. 이 결측값 때문에 무응답 모델에 포함되는 추가 정보(예를 들어, 연방고용기관의 사업체 등록자료 또는 고용통계)가 필요하다. 반면 부분 합성 데이터셋의 경우 조사 단위의 모든 속성 또는 오직 민감한 속성만 합성 데이터로 대체된다.

2.2 IVEware

이 프로젝트는 모든 종류의 조사에 쉽게 적용할 수 있는 표준화된 익명화 절차를 개발하는 것이 목적이다. 이를 위해 사용되는 소프트웨어는 배우기 쉽고(컴퓨터 과학자에게 뿐만 아니라 다른 이에게도) 저렴해야 한다. 무응답처리 소프트웨어 IVEware는 이 두 가지 조건을 충족한다. IVEware는 Raghunathan, Solenberger 그리고 Van Hoewyk이 개발했는데 무료로 내려받을 수 있다. 이 프로그램은 순차적인 회귀 기법을 사용한다: X_1, ... , X_i를 무응답이 없는 데이터셋 변수라고 하고, Y_1, ... , Y_j를 무응답을 포함한 데이터셋 변수라고 하자. 그리고 Y 변수의 순서는 결측값을 고려하여 오름차순으로 정렬됐다고 하자. 첫 번째 단계에서 X 변수 관측값이 주어졌을 때의 Y_1의 조건부 분포가 추정된다. 그

리고 이 분포로부터 Y_1의 값이 여러 개 뽑힌다. 다음 단계에서 X 변수 관측값과 이전에 추정된 Y_1 값이 주어졌을 때의 Y_2의 조건부 분포가 추정되고 이 분포로부터 Y_2 값이 추정되는 과정을 반복한다.

IVEware는 네 종류의 변수를 구분한다: 연속형, 범주형, 복합형(0을 범주형 값, 나머지를 연속형 값), 그리고 수치형 변수(예를 들어, 기업의 지역 조직 수)가 그것이다. 연속형 변수 추정에 보통의 선형 회귀 모델이 사용되는 반면, 범주형 변수에는 논리 모델 혹은 일반화된 논리 모델이 적용된다. 복합형 변수는 2단계를 거쳐 추정된다: 먼저 0 또는 0 아닌 자료는 논리 회귀로 평가된다; 그 다음으로 0 아닌 자료는 선형 회귀 모델로 추정된다. 수치형 변수는 보통 Poisson 회귀를 사용한다. IVEware는 원본 자료 구조를 유지하고 변수 간의 의존성을 보존하기 위한 몇 가지 가능성을 제공한다. 상한과 하한은 bound문으로 선언할 수 있고 restrict문은 값이 측정 조건을 충족할 때만 추정될 수 있음을 제한하는데 예를 들면 출생아수는 여성에 대해서만 추정될 수 있는 것 등이다. IVEware는 SAS에서 실행할 수 있고 개별적으로 실행할 수도 있다. 아쉽게도 SAS가 Engerprise Guide 환경에서 동작하면 이 프로그램은 동작하지 않는다.

3. 생산부문 월별 지역 리포트

InfinitE 프로젝트 구성원은 1998년에서 2001년까지의 생산부문 월별 지역 리포트 자료를 바탕으로 다양한 익명화 전략을 개발하고 비교하는 것에 동의했다. 이 조사는 연구자들의 수요가 매우 많지만 30개 정도의 직접적인 질문 문항이 포함되어 있다.

이 리포트는 20인 이상을 고용한 생산부문의 모든 경제활동이 대상이다. 소속 사업체를 포함하여 20인 이상을 고용한 소규모 지역 사업체도 포함된다. 이 리포트에는 경제활동, 위치, 직원수, 매출액, 임금 그리고 근로시간 등의 내용이 포함된다. 원칙적으로 1달치 월 자료에 의한 분석도 가능하다. 하지만 지금까지 연단위의 자료만이 RDC 연구자에게 제공된다.

4. 합성 데이터구조 파일 생성

무응답처리와 IVEware 프로그램의 사용법에 대한 경험을 얻기 위해 우리는 먼저 2001년 1개 년도의 조사를 검토했다. 이 해의 데이터에는 50.347 지역단위가 포함되었다.

4.1 전제조건

연속형 변수는 세제곱근 풀이를 사용하여 변형된다. 이 함수는 자연 로그 만큼 증가율이 높지는 않다. 사업체 조사에서 대부분 그러하듯 데이터에 특이치가 있는 경우 이 특성은 이점이 있다.

한번에 하나의 변수만이 익명화될 것이다. 그러므로 그 데이터셋은 중복된다: 원본 데이터 다음에 같은 데이터셋이 한 번 더 추가되지만 이때에는 합성되어야할 변수의 값이 결측값으로 대체된다.

4.2 범주형 변수의 무응답처리

4~6가지 이상의 개별적인 값을 갖는 범주형 변수의 무응답처리 과정에는 난관이 있다고

알려져 있다. 3가지 대안에 시도되었다. 독일의 15개 연방주로 부호화된 지역단위 위치 정보를 사용하여 얻은 결과에 대해 논의한다.

대안1: 연방주 속성을 16가지 값을 갖는 범주형 변수 모델로 고려한다.
프로그램은 매번 10시간 이상 실행한 후 중단되었다. 스크린에는 "비정상 종료" 메시지가 나타났고 로그파일에는 오류 원인에 대한 추가 정보가 없었다. 이 오류는 예정된 백업과 업데이트 등으로 네트워크가 불안정한 결과라고 추측했다. 이 대안은 독립형 PC에서 테스트를 더 해볼 필요가 있다.

대안2: 모든 연방주에 대해 더미 변수를 만든다.
더미 변수에 대한 무응답처리는 지역단위가 연방주에 위치한 순서에 따라 처리된다. 첫 번째로 가장 큰 값을 갖는 지역단위(Nordrhein-Westfalen)를 갖는 연방주 값이 평가되고, 그 다음 두 번째로 큰 값을 갖는 지역단위(Baden-Württemberg)가 평가되며 이 과정이 반복된다. 첫 번째 단위의 더미 값이 1이 되면 그 후의 더미 값은 자동으로 0으로 처리된다. 이 대안의 자료처리 시간은 수용할 만 했다: 회귀분석의 반복횟수를 10으로 설정했을 때 합성 데이터셋 하나 당 30 ~ 40분이 걸렸다. 이 대안은 상당히 잘 작동했다(부록의 표 1-3 참조).
3개의 모델을 테스트했다. 모델2와 모델3에는 보충 설명 변수가 추가되었다. 3가지 모델을 테스트한 결과, 가장 값이 작은 연방주 두 곳인 Bremen과 Saarland에서 합성 데이터와 원본데이터 간의 표본오차[percental deviations]가 크다. 이런 오차는 설명 변수를 추가할수록 더 커진다. Sachsen 자료는 모델1에서 1% 정도로 상당히 작은 오차를 보인다(표 1); 이 값은 모델2에서 13% 이상으로 커진다(표2). 같은 연방주 자료에 대해 모델 간 차이점이 있다는 것은 연방주에 대해 각각의 모델로 평가하는 것은 민감할 수 있다는 것을 말한다. 이 문제는 더 연구할 필요가 있다.

대안3: 3단계 무응답처리: 첫 번째 구 연방주 또는 신 연방주, 그 다음 구 연방주의 지역, 마지막으로 지격 내의 연방주. 독일은 "구 연방주 북부"(Schleswig-Holstein, Hamburg, Niedersachsen, Bremen), "구 연방주 중부"(Nordrhein-Westfalen, Hessen, Rheinland-Pfalz), "구 연방주 남부"(Baden-Württemberg, Bayern, Saarland) 그리고 "신 연방주"로 4곳으로 구분된다. 첫 번째 단계에서 지역단위가 구 연방주에 속하는지 신 연방주에 속하는지 판단한다. 두 번째 단계에서는 서부 독일의 지역단위를 북부, 중부, 남부 중 한 지역에 포함시킨다. 마지막 단계에서는 할당된 지역의 연방주를 무응답처리한다.
신/구 연방주 할당 및 지역 할당은 더미 변수로 계산되고, 각각의 연방주에 대한 할당은 범주형 변수로 계산된다. 이 대안의 착안점은 대안2에 비해 무응답처리 단계의 횟수를 줄이는 것이다. 여기서는 모두 7회가 필요하다: 한번은 구 연방주, 신 연방주 할당에, 두 번은 구 연방주 지역 할당에(남부, 중부; 두 번 안에 할당되지 않은 단위는 자동적으로 북부에 할당), 네 번은 각각의 연방주 할당에(지역 당 한번) 필요하다.
범주의 세 개일 때, 다항 회귀는 논리 회귀에 비해 이진 더미 변수 계산 시간이 상당히 오래 걸리는 것으로 보인다. 특히, 신 연방주(범주 6개)를 실행하는 데는 3~4 시간이 걸

린다. 모든 지역단위의 표본오차는 모두 1% 이하이지만, 개별 연방주의 오차는 부분적으로 매우 커서 Hessen의 경우 414%에 이른다(표 4).
위에 언급한 3가지 대안을 요약한다면, 대안2가 가장 유망하다. 대안1은 계산 시간이 너무 길어서, 대안 3은 다항 회귀에서의 결과가 안 좋아서 폐기된다.

5 합성 데이터와 원본 데이터 비교

합성 데이터셋에 유지되는 분석 잠재력을 알아보기 전에 합성 데이터와 원본 데이터의 비교 분석 결과를 살펴본다. 2001년 월별 리포트의 서로 다른 합성 데이터셋을 5개 만들었고 경제활동, 위치, 직원수, 매출액 그리고 직원수 등의 핵심 변수로 분석을 제한한다. 우리 프로젝트의 파트너 중 하나인 IAW(Institute for Applied Economic Research) 직원이 더 종합적으로 분석할 수 있을 것이다. 직원수의 univariate 분포는 5가지 합성 데이터에서 모두 잘 보존된다. 평균은 127.3과 127.9 사이를 오르내리는데 참값은 128.1이다. 표준오차는 원본 데이터에 비해 일반적으로 약간 높았다(합성 데이터는 585.8~598.1; 원본 데이터 575.9). 매출액 분석 결과는 그리 좋은 편이 아니다. 평균과 표준편차 모두 참값에 비해 상당히 낮게 나왔다: 평균은 오차가 8~9%, 표준편차는 오차가 20%에 이르렀다.

표1:직원수와 매출액의 평균과 표준편차

직원수와 매출액의 상관계수는 합성 데이터셋에서 훨씬 더 높게 나왔다. 원본 데이터의 상관계수는 0.80인데 비해 합성 데이터셋의 그것은 약 0.91이다.
지역단위당 직원수가 평균적으로 가장 많은 경제 부문은 자동차 산업이다. 원본 직원수의 평균은 651인 반면 합성 데이터셋의 평균은 370에서 471까지 차이를 보였다. 이는 5개의 합성 데이터셋 중 3개에서 얻은 결과이고 다른 두 경우에는 광업의 평균이 자동차 산업의 그것을 넘어섰다. 원본 데이터에서 광업은 직원수 평균이 502명으로 2위를 차지했다. 모든 데이터셋에서 채석 사업의 직원수 평균이 25.8에서 28.1 범위로 가장 적었는데 참값은 21.0이다.
특히, 매출액의 평균이 합성 데이터셋에서 매우 낮게 나타나는 것을 설명하기 위한 조사가 필요하다. 이를 통해 합성 데이터셋을 개선할 수 있을 것이다.

6 마이크로데이터의 기밀성

마이크로데이터의 기밀성을 위해 항상 두 가지 지침을 따라야 한다. 한 가지는 이전 장에서 이미 언급했듯이 데이터의 분석적 유효성이 광범위하게 유지되어야 한다는 것이다. 다른 한 가지는 잠재적인 데이터 침입자에 의한 기밀 정보의 노출 위험을 가능한 최소화해야 한다는 것이다. 후자는 아래에 설명하듯이 실제적인 매칭 시나리오를 실행함으로써 테스트해볼 수 있다.

6.1 수학적 모델

Elliott과 Dale의 1999년 연구에서 보듯이, 데이터베이스 간의 매칭에서 데이터 침입자는

외부 데이터베이스 A와 전체 기밀 데이터베이스 B를 매치한다. 이를 위해 침입자는 외부 데이터와 기밀 데이터가 공통으로 가지는 변수 즉, 키 변수[key variables]를 사용한다.
가능한 할당의 후보가 되기 위해서 A와 B의 원소의 조합에 포함되는 레코드 쌍 (a,b)이 특정 변수들에 대해 같은 값을 가져야 한다. 이 변수들은 전체 데이터를 공통부분이 없는 블록으로 나누기 때문에 이하 '차단 변수[blocking variables]'라고 부른다.
차단 데이터의 목적 중 하나는 이후의 할당 과정과 배정된 주요 저장 공간의 복잡도를 낮추는 것이고 다른 하나는 매칭되지 않는 자료의 수과 관련이 있다. 노출 제어 부문의 차단 데이터에 대한 구체적이고 경험적인 연구는 Lenz와 Vorgrimler의 2005년 연구에서 확인할 수 있다.
기술적이지 않은 방법으로, 매칭의 개념은 A의 원소 a와 B의 원소 b의 짝을 만드는 것으로 설명할 수 있다. 이때 레코드 a와 b는 매칭되었다고 한다. 이후 매칭되지 않는 짝을 최소화하기 위해 노력하게 된다. 잠재적인 데이터 침입자가 목표 조사의 분석에 관여할 수 있다면, 매칭의 문제는 다음과 같이 수학적 용어로 형식화될 수 있다: 거리 측정값 d:A×B→[0,1](또는 유사한 측정값 w:A×B→[0,1])에 기반하여 매핑 Φ:A→B를 찾는데 이는 A의 모든 레코드를 가까운(또는 유사한) B의 레코드로 매핑하는 것이다.
더 정확하게, 이 매핑은 다음의 할당 문제로 정의될 수 있다.

$$\text{Minimise} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} d(a_i,b_j)x_{ij}, \qquad \textbf{(AP)}$$

$$\begin{aligned}
\text{subject to} \quad & x_{ij} \in \{0,1\} \quad \text{for} \quad i=1,\ldots,n; j=1,\ldots,m, \\
& \sum_{j=1}^{m} x_{ij} = 1 \quad \text{for} \quad i=1,\ldots,n \quad \text{and} \\
& \sum_{i=1}^{n} x_{ij} \leq 1 \quad \text{for} \quad j=1,\ldots,m.
\end{aligned}$$

(AP)의 제약조건은 외부데이터 A의 각 레코드 a가 목표 데이터 B의 각 레코드 b로 일대일로 할당된다는 것을 보장한다. 다시 말해, $x_{ij}=1$ 이라는 것은 a_i가 b_j와 연결된다[conected]는 것을 의미한다. 그러므로 A의 레코드의 수는 B의 레코드의 수보다 작거나 같다고 가정할 수 있다.
계수 $d(a_i,b_j)$가 한번 계산되면, 단순한 방법[simplex method]과 같이 고전적인 검증된 방법을 이용하여 선형 할당 문제를 풀 수 있다. 주로 세금 통계에서 만들어지는 더 큰 데이터 블록의 경우 효율성 문제로 근사 추론이 권장된다. 다행히, 근사 추론을 사용하면 2006년 Lens의 연구에서 나타난 최적의 해결법에 가까운 결과를 얻을 수 있다.

6.2 경험적 결과

우리는 2001년 월별 리포트에 포함된 36000 건의 사업체에 대한 첫 번째 매칭 시나리오를 만들었다. 상업적인 데이터베이스에서 이 자료에 대한 9000건의 추가 지식을 얻었다. 차단 변수로 사용된 것은 지역단위의 위치(신/구 연방주의 경우), 경제활동(두 자리 숫자

로 된 NACE 코드)의 일부 그리고 6개의 범주로 요약된 직원수 등이다. 할당을 위한 키 변수는 직원수와 매출액이다. 지금까지 이 시나리오를 하나의 합성 데이터셋에 대해 수행했다. 9000건 중 18건만이 정확히 매치되었으며 이는 0.2%이다.
이 결과를 신뢰할 수 있는지 확인하기 위해 원본 데이터와 합성 데이터에서 대응되는 값을 부분적으로 계산했다. 두 자리 숫자로 된 NACE 코드는 두 자료에서 23.6%가 일치했고 직원수의 범주 자료는 80.2%, 신/구 연방주의 위치 자료는 77.5%가 같았다. 세 변수를 결합한 경우 14.7%가 일치했다. 이 결과는 합성 데이터셋을 만드는 것이 보호 효과를 고려한 자료의 기밀성에 크게 기여한다는 것을 나타낸다.

부록 1.8. Working Paper 42

통계적 정보노출제한에서 투명성의 역할

(The role transparency in statistical disclosure limitation)

- Alan F.Karr (National Institute of Statistical Sciences, 미국)

요약

숫자 데이터 마스킹 기법은 기본적인 noise addition 기법으로부터 상관된 노이즈를 바탕으로 하는 새로운 접근법, general additive data perturbation과 multiple imputation, micro-aggregation, data swapping, data shuffling, copula based perturbation 방법 등등에 이르기까지 발전되어 왔다. 이러한 다양한 접근법에 대해서 원본 자료와 마스킹 된 자료 간의 유사성을 비교하는 것은 매우 어렵다. 따라서 서로 다른 방법들을 평가할 수 있는 새로운 측정방법의 개발이 요구된다. 이 연구에서는 새로운 유사성공통지수(Common Index of Similarity: CIS)를 개발하였다. 이 지수는 0부터 1까지의 값으로 측정하며, 0에 가까울 수록 원본 자료와 마스킹 된 자료 간의 유사성이 없음을 의미하고, 1이면 두 자료가 동일하다는 것을 의미한다.

1. 소개

숫자 자료 마스킹에 대한 초기 접근법은 대부분 잡음을 원본 값에 추가하는 것과 잡음이 더해진 마스킹 된 값을 공표하는 데 한정적으로 포커스가 맞춰져 있었다. 더해지는 잡음의 분산으로 단순하기는 하지만 효과적인 방법이었기 때문에 서로 다른 마스킹 된 자료의 값들을 비교하는 것은 매우 쉬웠다. 마스킹 된 자료와 원본 자료 사이의 유사성은 더해진 잡음의 분산과 역의 관계에 있었다. 잡음 분산이 클수록 유사성은 낮게 되는 결과를 보인다. 심지어 이러한 경우, 자료 제공자는 다른 변수에 서로 다른 수준의 잡음을 더해주게 되었다. Kim에 의해 제안된 상관된 잡음 추가 방식은 더해진 잡음의 분산이 이 모든 변수에 대해 상수였기 때문에 이러한 문제를 완화시켰다.

1990년 초기에 통계적 노출제한 기법에 대한 새로운 관심이 시작되었다. 이것은 숫자형태 자료의 마스킹에 대한 새로운 기법의 발전을 불러일으켰다. 이러한 기법들은 마스킹 된 자료를 생성하기 위해 사용되는 모형들을 다양화하였다. multiple imputaion, general additive data perturbation, sufficiency based approach과 같은 몇몇 기법들은 선형모형을 이용하여 마스킹 된 자료를 생성한다. copula based methods와 같은 다른 기법들은 자료를 교란하기 위해 좀 더 복잡한 모형을 채택한다. 선형모형을 기반으로 하는 접근법에서 모형의 일부인 잡음추가를 결정하는 것이 항상 쉬운 것은 아니다. copula를 기반으로 하는 접근법에서 추가된 잡음을 분리하는 것은 좀 더 어렵다.

추가하면, 우리는 또한 잡음 추가 방법이나 통계적 모형을 직접적으로 수반하지 않아도 되는 다른 접근법들도 가지고 있다. 이러한 접근법에는 data swapping과 micro aggregation 기법을 포함한다. 데이터 스와핑 기법에서는 특정한 근접범위 내의 값들이

랜덤하게 교환되고 마스킹 된 자료로서 공표된다. micro aggregation 기법에서는 원본 값과 근접한 범위에서 동일 변수의 평균값에 의해 원본 값이 교체된다. 이러한 두 기법 모두에서의 관심사항은 변수들의 rank에 의해 결정된 근접범위 내의 값이 필요하다는 것이다. data shuffling과 같은 다른 방법들은 copula를 기반으로 자료변조를 하고, 마스킹 된 자료를 생성할 때 원본 자료를 이용하기 위해 swap을 역으로 수행한다. data shuffling은 자료변조기법과 스와핑 기법이 결합된 기법이다. 이 방법은 rank를 기반으로 하며, 잡음 추가와 같은 방법을 직접적으로 수반하지 않는다.

자료제공자가 서로 다른 접근법을 비교할 때, 공통적인 기준을 제공하는 측도 없이 그들의 상대적 성능을 평가하는 것은 매우 어렵다. 예를 들어, 전통적인 잡음 추가 모형을 이용하여 생성된 마스킹 된 자료를 micro aggregation기법을 이용하여 생성된 자료와 비교할 때, 우리는 이들 두 기법으로부터 나온 마스킹 결과를 비교할 수 있는 공통된 측도를 가져야만 한다. 두 접근법을 위해 매개변수들이 사용되었기 때문에(잡음 추가에 대한 분산과 micro aggregation을 위해 얼마나 많은 값들이 집계되었는지) 두 방법을 직접적으로 비교하는 것은 매우 어렵다. 두 방법을 합리적으로 비교하기 위해 허용될 수 있는 공통 기준은 무엇인가. 이 연구의 목적은 서로 다른 기법들 사이의 기준을 제공할 수 있는 CIS를 개발하는 것이다.

2. CIS의 주요 특성

CIS를 개발함에 있어 우리는 측도로서 중요하다고 여기는 몇 가지 특성을 확보하고자 시도하였다. 첫 번째 특성으로 측도는 모든 방법에 대해 표준화되어야 한다는 것이다. 잡음 추가방법에서, 잡음의 분산이 유사성의 좋은 측도를 나타낸다. 그러나 이 방법에서 잡음의 분산은 실제로 0보다 큰 모든 값을 가질 수 있기 때문에 표준화 된 것은 아니다. 예를 들어 원본 변수의 분산보다 두 배 더 큰 분산을 추가하는 것이 가능하다. 이러한 경우, 두 변수 간 변조 크기를 표준화 하는 방법이 없다. 이 연구에서 CIS 측도는 0과 1 사이의 값으로 표준화 되었다.

우리는 또한 측도가 유용성이나 노출위험을 직접적으로 대표하지 않길 원했다. 자료의 유용성과 노출위험에 대한 많은 측도들이 있기 때문에 특정한 측도를 선택하여 문제를 해결할 수 있다. 예를 들어, 특정한 방법은 특정한 측도에 의해 정의된 자료의 유용성을 측정할 수 있으나, 다른 방법에 대해서는 그렇지 않다. 유용성이나 노출위험의 특정한 측도를 기반으로 유사성 지수를 선택하는 것은 잠재적으로 편향될 가능성을 갖고 있다. 이 연구에서 CIS 측도는 유용성이나 노출위험을 직접적으로 나타내지 않고, 따라서 어떠한 특정 방법을 선택했느냐에 의해 편향되지 않는다.

3. CIS의 내용

$\boldsymbol{X}$를 비밀보호 되어야 할 숫자 변수의 집합으로 정의하고, $\boldsymbol{Y}$를 동일 변수에 대한 마스킹 된 값들이라 하자. 이 마스킹 된 변수는 어떠한 마스킹 기법에 의해서든 생성될 수 있다. $\boldsymbol{X}$와 $\boldsymbol{Y}$ 사이의 유사성을 측정하는 지수를 개발하는 것이 본 연구의 목표이다.

우리가 제안하는 CIS 측도는 숫자 변수들 사이의 정준상관(canonical correlation)의 개념

에 기초한다. 정준상관분석(CCA)는 통계적 방법론으로서 두 변수 집합 사이의 관계를 밝히고 정량화 할 수 있게 한다. CCA는 변수들의 선형 결합 집합과 가장 상관성이 높은 다른 변수들의 선형결합 집합을 을 밝히게 된다. 마스킹의 경우에서 우리는 원본 변수들의 집합(X)과 마스킹 된 변수들의 집합(Y)을 고려하여 CCA 방법을 사용한다.

CIS 측도는 다음과 같이 계산된다. Σ_{XX}, Σ_{YY}, Σ_{XY}를 각각 X, Y의 공분산행렬, X와 Y의 공분산행렬이라 하고, 다음과 같은 식을 생각해보자.

$$\Sigma_{XX}^{-1}\,\Sigma_{XY}\,\Sigma_{YY}^{-1}\,\Sigma_{YX} \tag{1}$$

이 행렬이 (k×k)차원의 대칭행렬임을 쉽게 보일 수 있다. 여기서 k를 X의 변수들의 개수이다.

수식 (1)로 부터의 행렬은 k개 고유값(eigen value)을 가진다. 위 수식의 주된(가장 큰) 고유값 λ는 다음과 같은 통계적 성질을 갖는다. λ는 X에 기여할 수 있는 Y에서의 변동성 비율의 최대값을 나타낸다. 두개의 선형결합a^TX, b^TY을 고려해보자. 여기서 a와 b는 길이가 k인 벡터이다. 정준상관분석에서는 a^TX와 b^TY사이의 상관관계를 최대로 하는 특정한 a와 b 벡터를 찾는다. λ의 제곱근은 X와 Y의 어떠한 선형결합 사이의 최대 절대 상관계수를 의미한다.

정준상관분석은 실제로 광범위하게 응용되어 왔다. 정준상관분석은 상관관계를 최대로 하는 a와 b를 밝혀내는 데 종종 이용되어 왔다. 정준상관분석은 매우 중요한 다변량 분석 도구로서 복잡한 변수들 간의 잠재적인 관계를 단순화시킬 수 있게 해준다. 고유값들과 그에 대응하는 상관관계는 (1)번 식에 의해 직접적으로 계산할 수 있고, SAS와 같은 대부분의 통계분석 툴은 정준상관분석의 고유값, 상관관계, a와 b, 그리고 다른 자세한 분석들을 제공해준다.

이러한 이유로, 우리는 (1)번 식으로부터 주된 고유값을 CIS로 사용 할 것을 제안한다. 이 측도의 특성은 X의 공분산 행렬이 평가에서 결정적인 역할을 하기 때문에 변수들 간의 관계도 포함한다는 것이다. 예를 들어, 잡음의 분산이 $d\sigma^2$와 같은 하나의 변수에 대해서, (1)번 식은 $1/(1+d)$이 된다. 다변량 변수들에 대해서는 X의 변수들 사이의 관계들 때문에 이러한 축소가 가능하지 않다. 다시 말하면, X의 구조 때문에, 주어진 수준의 잡음이 더해졌을 때, 유사성 결과는 기대했던 것보다 더 높게 나타난다.

대조적으로 상관성 있는 잡음에 대해서는 변수의 개수가 몇 개이던지 더해진 잡음의 분산이 $d\Sigma_{XX}$면 (1)번 식은 $1/(1+d)$으로 축소된다. 단순 잡음 추가 방법과 상관성 있는 잡음 추가 방식을 비교할 때, 비교의 측도는 보통 d의 값에 의해 결정된다. 그러나, X와 Y사이의 유사성을 볼 때, 단순 잡음 추가방법과 상관성 있는 잡음 추가 방법 사이의 커다란 차이가 있다. 상관성 있는 잡음 추가 방법의 CIS는 변수의 개수가 1개보다 클 때, 단순 잡음 추가 방법보다 작게 나타날 것이다. 이들이 유일하게 같아지는 경우는 Σ_{XX}가 비대각원소가 0의 값을 갖는 대각행렬인 경우 즉, X 변수들 간에 상관성이 없을 때이다.

따라서 이 측도를 이용하면 다음과 같은 장점을 얻을 수 있다.
(1) 이 측도는 유사성의 표준화된 값을 제공한다.(0~1사이의)
(2) 유사성을 측정함에 있어서, 노출제한 될 변수들의 구조를 반영한다.
(3) 자료 마스킹 과정에 의해 “variance added"에 대한 의미있는 해석을 제공한다.
(4) 잡음 추가 방법에서 사용되었던 분산 측정과 일치성을 갖는다.
(5) 어떠한 숫자 형 자료에 대해서도 적용될 수 있다.
(6) 표준화된 통계적 분석 툴을 이용해서 간편하게 계산할 수 있다.

5. 결론

이 논문의 목적은 어떠한 마스킹 기법을 적용했는지에 상관없이 원본 자료와 마스킹 된 자료 사이의 유사성을 계량화 할 수 있는 새로운 측도를 소개하는 것이다. 이러한 측도의 개발은 서로 다른 마스킹 기법 사이의 효과적인 비교를 돕게 될 것이다. 우리는 정준상관분석를 기반으로 하는 측도를 소개하였다. 센서스 자료를 이용하여 경험적으로 평가하였을 때, 측도의 수행능력은 좋았으며, 실제로도 효과적으로 사용될 수 있을 것이다.

부록 2. 데이터 통합에 대한 원칙과 지침(번역본)

통계 또는 관련 연구목적의 데이터 통합의 비밀보호 유지 측면에 대한 원칙과 지침

이 원칙과 지침은 2009년 6월 컨퍼런스 유럽 Statisticians(CES) 회의에서 승인되었다.

1. 서론

1. 유엔 유럽 경제위원회(UNECE)는 8번째 회의에서 유럽 경제위원회의 분과에서 이루어진 공식통계의 기본원칙에 관한 결정 C(47)을 채택했다. (http://www.unece.org/stats/documents/e/1992/32.e.pdf 참조). 기본 원칙 6번은 "통계 작성(compilation)을 위해 통계 기관에 의해 수집된 개인정보가 포함된 데이터는, 자연인 또는 법인이건 간에 엄격한 비밀보호사항으로 통계적 목적에 한정적으로 이용된다."는 것을 선언하였다.

2. 데이터 통합은 다른 행정기관이나 조사 출처의 단위 레코드로부터 저작권을 갖고 공표할 수 있는 새로운 공식 통계를 편집/통합하는 것과 관련이 있다. 또한, 이 통합 데이터 셋은 기존 조사출처를 통해서는 불가능했던 경제 및 사회 연구의 범위를 지원하는 데 사용될 수 있다. 몇 가지 경우에서 통합 데이터셋의 사용은 단일 소스 데이터셋의 사용에 비해 추가적 입법(legal)과 정책적 우려(concern)를 초래할 수 있다. 이런 추가적 우려는 일반적으로 개인 정보 보호 정책과 데이터 보호 요구에 관련되어 있으나 이것에 국한된 것은 아니다. 이 원칙과 지침은 통계 등록(register)의 작성 및 유지관리에 관련성이 있지만, 이들 작업을 다루지는 않는다.

3. 이 원칙과 지침은 국가통계조직(NSOs)에 의해 수행되는 데이터통합 작업에 적용된다. 어떤 경우에는 국제 통계조직이 다른 국가들로부터 마이크로데이터 셋을 결합한다. 하지만, 데이터 파일간 공통된 단위가 있을 가능성이 희박하기 때문에 이 상황에서는 비밀보호에 관한 문제가 발생하지 않는다.

4. 이 원칙과 지침의 목적을 위해서, 단일 소스에 대한 임퓨테이션 처리에서 다른 소스의 사용은 비슷한 이슈가 적용될 수도 있지만, 데이터통합으로 간주되지 않는다. 같은 조사의 2가지 사례(instance)는 단일 소스로 간주된다.

5. 이 원칙과 지침의 사용을 지원하기 위해 다음과 같은 정의가 사용된다.

(a) 복합 마이크로데이터(Composite microdata) - 데이터 통합으로 인한 단위 레코드 데이터
(b) 비밀보호(Confidentiality) - 정보의 비밀을 유지하지 위한 정보 제공자에게의 의무
(c) 데이터 통합(Data Integration) - 두 개 이상의 소스에서 데이터를 결합하는 과정을 통하여 새로운 output을 생산
(d) 데이터 매치(Data Matching) - 공통 특성을 기반으로 서로 다른 소스에서 마이크로데이터의 연계
(e) 데이터 공급자(Data Provider) - 데이터 또는 메타 데이터를 생성하는 조직. 이 원칙 및 지침에서 이 용어는 통계 또는 비통계 소스로부터의 데이터 파일의 공급자를 포함하나 개인 응답자는 포함하지 아니한다.
(f) 자연인 또는 법인(Natural or legal persons) - 법률에서 인정한 개인 및 법적 기관
(g) 공식 통계(Official Statistics) - 국가 통계시스템 내에서 또는 정부 조직간의 통계 프로그램 하에서 수행된 모든 통계 활동
(h) 개인 정보 보호 정책(Privacy) - 자신의 개인적인 문제와 관계 비밀을 유지하기 위한 개인의 권리와 개인정보 보호의 대상(subject)에 대한 정보를 소유한 자의 의무를 포함한다.
(i) 연구 목적(Research Purposes) - 이러한 원칙과 지침의 관점에서, “관련 연구 목적(related research purposes)”은 통계적 output을 결과로 갖는 경제 사회적 현상 조사나 설명을 위한 특별 활동으로 정의된다. 이들 활동은 통계조직(이 경우 결과는 반드시 공개될(published) 필요는 없다), 또는 외부 연구자(CES의 “통계적 비밀보호과 마이크로데이터 접근을 관리하는 원칙과 지침을 따르는)에 의해 착수될 것이다.
(j) 통계 활동(Statistical Activity) - 통계 정보의 수집(collection), 저장(storage), 변환(transformation) 및 배포(distribution)
(k) 통계 목적(Statistical Purposes) - 공식통계의 기본원칙에 부합하는 방식의 데이터 사용은 통계업무처리의 각 단계에 적합하고 공식통계의 생산에 기여하는 것을 말한다.

6. 데이터 통합은 정확한 매칭, 확률적 매칭 그리고/또는 통계적 매칭이 포함될 수 있다. 통합 데이터셋의 잇점은 아래와 같은 사항을 포함할 수 있다:
(a) 신규 또는 개선된 통계의 생산
(b) 어떤 정보가 현재 존재하는 곳을 측정할 수 있는 더욱 해체된(disaggregated) 정보의 생산
(c) 단일 데이터 소스로부터 가능한 것보다 더 큰 숫자 단위의 광범위한 변수를 다루는 복합 마이크로데이터를 사용하는 연구 수행 능력
(d) 기존 데이터 소스를 개선하거나 유효하게 하는 잠재력
(e) 응답자 부담을 경감시키는 잠재력

7. 첨부된 원칙과 관련된 지침은 통계적 연구목적의 통합 데이터셋을 생성, 사용하는 과정의 비밀보호 측면과 법적 평가 및 완화를 위한 일반적인 프레임워크를 제공함으로써 기본 원칙 6를 확장시킨다. 특히 그것들은 공식 통계의 기본원칙들은 다른 소스의 공식통계와 같이 통합된 데이터에 똑같이 적용된다는 것을 인식하고 있다.

8. 이러한 원칙을 개발하면서, 일부국가에서는 공식통계를 생산하기 위해서 일차적으로는 주민등록과 같은 (통합된) 행정자료를 사용하고 사용가능한 행정자료에 중요한 자료가 빠져있거나 품질이 너무 낮을 경우에만 통계조사를 실시한다는 것을 알게 되었다. 이들 나라에서는, 통계 데이터셋의 통합은 국가 통계 기관 업무의 보편적인 부분이다. 이들 나라는 데이터 통합이 다양한 출처에서 이루어진 여부에 관계없이 이미 개인 비밀과 개별 비즈니스 데이터에 대한 보호에 대한 강한 법적 프레임워크와 명백한 규칙을 갖고 있다.

9. 그러나, 많은 다른 나라에서 통계와 연구 목적을 위하여 다른 소스로부터 복합 마이크로데이터를 생산하는 통합 데이터의 개념은 상대적으로 생소하다. 첨부원칙, 관련 지침, 그리고 업무사례 개요는 그러한 작업이 일부 명확성과 응용의 지속성을 제공할 수 있도록 설계되었다.

10. 아래의 원칙과 지침뿐만 아니라, 특정 상황과 관련된 특히 공식 통계 목적의 데이터 통합 활동의 강한 전통이 없는 국가에 대해, 두개의 다른 개념이 발견되었다. 첫번째는 소스 데이터 수집의 무결성을 구체적으로 위협할 때는(예를 들어 응답율 감소 위험에 노출되는) 데이터 통합을 수행해서는 안 되며, 두번째는 연구 목적의 데이터 통합은 오로지 승인 처리가 공공의 이익을 위한 것이라는 것을 증명할 수 있는 경우에만 고려되어야 한다는 것이다. 공식 통계 목표는 공식 통계의 기본원칙을 따른다면 공공의 이익을 위하는 것이라고 가정한다. 이러한 개념은 여기에 완성도를 위해 언급되었지만, 널리 허용되는 것은 아니며 따라서 아래 원칙과 동일한 상황을 갖는 것은 아니다.

2. 원칙과 지침(PRINCIPLES AND GUIDELINES)

○ **원칙 1 : 데이터 통합은 오로지 통계와 관련 연구 목적을 위하여, 국가통계기관(과 국가통계시스템 내부 조직)에 의해 이루어져야 한다.**

- 지침

(a) 위의 원칙은 통계법 그리고/또는 데이터 보호에 관한 법 안에 안치(enshrined)되어야 하고, 정부에 의해 엄격히 존중되어야 한다.

(b) 명시적 입법보호가 없는 상황에서는, NSOs는 자연인과 법인에 관련된 데이터 통합은 삼가해야 한다.

(c) 법률이 규정하지 않는다면, 국가 또는 준국가 정부부서 또는 공공 기관에서 가지고 있는 기존 행정 또는 통계 소스 데이터의 NSOs에 의한 통계 또는 연구 목적 사용은 특정 자연인 또는 법인의 프라이버시에 위반된다.

○ **원칙 2 : 국가통계기관(NSOs)는 오로지 표준 승인 처리(예를 들어 business case) 후에 그들의 공식 통계 위임에 일관되게 데이터 통합 활동을 시작한다.**

- 지침

(a) 통계청이 자연인과 관련된 행정/규제목적의 데이터를 사용하는 것과 같이 통계 또는 관련 연구목적을 넘어서는 권한을 갖는 영역에 대해서는 법에 의해 특별히 권한이 부여되는 경우를 제외하고는 개인정보와 관련된 통계 또는 관련 연구수행을 삼가야 한다.

(b) 통계 목적의 새로운 조사를 착수하기 전에, NSO에서 이미 사용가능한 통합된 자료가 대안으로 사용될 수 있는지 고려되어야 한다.

(c) 새로운 데이터 통합 제안에 대해서는 표준승인절차가 수행되어야 한다. 이것은 공식적인 사업방식(business case)의 형태를 띨 수 있다. 사업방식 개요 사례는 부록에서 주어지지만, 각 국은 데이터 통합 프로젝트를 뒷받침하는 프로세스를 위하여 자신들의 템플릿을 수립해야 한다. 승인 절차는 어떻게 통합 과업이 공식 통계를 생산하거나 개선할지, 또는 관련 연구에 공헌할 지를 식별할 수 있어야 한다.

○ 원칙 3 : 모든 데이터 통합 프로젝트의 공공 이익은 데이터의 사용 그리고/또는 공식 통계 시스템의 무결성 위험에 관한 개인 정보 또는 비밀보호에 관한 우려에 우선할 수 있을 만큼 충분해야 한다.

- 지침

(a) 데이터 통합은 보안 환경에서 공식통계시스템의 무결성에 대한 위험 노출이 없는 방법으로 이루어져야 한다.

(b) 법률에 의하지 않거나 표준 승인 절차를 따르지 않고 공급된 자료에 대해서는 통합될 데이터와 관련된 모든 직접적인 식별자는 통합 과정이 끝나는 즉시 제거되어야 한다.

(c) 표준승인절차의 일부로서 NSO에 의해 고려되는 모든 혜택, 비밀보호, 위험을 확인하는 책임을 가진 적절한 기구에 관한 논의가 필요하다. 혜택의 리스트는 통합된 데이터셋의 의도된 장기간 보유 또는 시간외의 계획된 확장에서 나온 것들을 포함해야 한다.

(d) 어떤 나라에서는, 표준 승인 절차 내에 개인정보 보호 영향 평가를 포함해야 한다는 것이 입법 요구사항이 될 것이다.

(e) 합리적이고 실용적인 곳에서는, 데이터 제공자의 동의를 구해야 한다.

(f) 또한, 개인 정보 보호 및 비밀 유지의 개념은 간접적인 식별(일반적으로 특이한 특성을 지닌 unit에 대해) 위험과 기존 출처의 데이터보다 폭넓은 변수를 포함하고 있는 통합 데이터 셋의 증가된 민감성에 대한 주의 깊은 관리가 요구된다.

○ 원칙 4 : 특별히 응답자에게 데이터통합과 같은 조치를 배제하기로 서약한 경우 데이터는 통합될 수 없다.

- 지침

(a) 표준승인절차는(데이터 통합 사업방식 같은) 응답자가 제공한 자료의 목적과 관련되어 응답자에게 어떤 보증을 해주었는지를 조사해야 한다. 통계청장은 그 제안의 모든 요소가 이들 보증과 양립되지 않는 경우 데이터 통합 제안을 승인해서는 안 된다.

○ 원칙 5 : 통합된 데이터는 승인 통계 또는 연구 목적만을 위해 사용되어야 하고 기존 승인 목적에서 현저한 변화가 있을 때는 새로운 표준 승인 절차를 새로이 거쳐야 한다.

- 지침 : 달리 법에서 제공하지 않는다면, 아래의 경우 언제라도 새로운 승인을 거쳐야 한다

(a) 통합 프로세스에 사용된 소스(데이터 집합)의 주요내용이 변화된 경우(예를 들어, 단위의 범주를 추가하거나 삭제, 또는 다루는 변수의 타입이 바뀐다.) 또는 새로운 소스가 통합절차에 추가되도록 제안된 경우

(b) 통합 프로세스에 의해 다루어진 단위(units)의 수가 유의미하게 확장된 경우(예를 들어 경제의 몇몇에서 모든 지점까지 확장)

(c) 통합의 방법이 변경되고(예를 들어 통계적에서부터 정확한 매칭까지) 이 변경사항이 자연인/법인의 노출 위험을 현저히 변화시킬 수 있는 경우.
(d) 통합으로 생긴 데이터셋이 또 다른 공식통계 목적을 위해 사용되거나 원래의 표준승인절차에서 제공되지 않는 외부 연구자에 의해 제안된 연구목적을 위한 경우.

○ 원칙 6 : 단위 레코드 및 연계될 데이터셋에 포함될 데이터 변수의 개수는 승인목적을 위해 필요한 최소개수를 넘어선 안된다.

- 지침

(a) '승인된 목적(approved purpose(s))'은 데이터 통합 비즈니스 케이스에서 승인된 목적을 말함. 이 목적을 지원하기 위해 필요한 데이터 변수들만이 승인된 데이터 통합 수행을 위한 데이터셋에 포함되어야 한다.
(b) 통합될 단위 레코드의 수는 승인된 목적을 지원하기 위해 필요한 최소한이어야 한다. (예 : 전체 데이터 소스의 표본을 통합할 때 고려해야 한다)

○ 원칙 7 : NSOs는 모든 데이터통합을 개방적이고 투명한 방식으로 실시해야 한다.

- 지침

(a) 데이터 통합에 관련된 NSO의 정책 및 작업의 개요(overview)는 공개(publish)되어야 한다.
(b) 모든 데이터 통합 작업의 주된 통계적 결과는 공개적으로 이용 가능해야 한다. 데이터 통합 작업이 공식통계의 작성을 개선하는 데 사용될 때,(예: 질적 개선을 통해) 그 공식통계의 공개는 이 요구를 충족한다. 복합 데이터베이스로부터 생산된 통계의 메타데이터는 데이터 통합에 쓰인 원천 데이터 소스에 대한 정보를 포함하여야 한다.
(c) 달리 법률로 명세되어 있지 않다면, 행정 기관은 그들의 정보를, 합리적이고 실용적인 곳이라면, 통계 또는 연구 목적에 일반적으로 쓰일 수도 있다는 것을 응답자에게 알려야만 한다.

○ 원칙 8 : 데이터 통합의 결과로 생성된 복합 단위레코드들이 어떤 식별자를 포함하고 있지 않다면, 데이터에 대한 접근은 NSO의 권한있는 직원으로 제한되어야 한다. 통계 마이크로데이터에 외부인의 액세스 권한을 부여하는 모든 제안은 명백한 법적 근거를 가져야 하며 공식통계 의 데이터 사용 목적과 일관되어야 한다. 접근 권한을 부여받은 모든 사람(들)은 마이크로데이터에 대한 그들의 이용이 승인된 접근계획과 일치되고, 권한이 없는 사람이 그 데이터셋에 접근하지 못하게 하겠다는 내용에 대한 법적 강제력이 있는 보장을 해야 한다.

- 지침

(a) NSO에서 수행한 통합에서 나온 복합 microdata는 동일한 국가 통계 시스템 내 다른 공식통계 생산자에 의해 통계 또는 관련 연구 목적으로 사용될 수 있다. 또는 적절한 국가 법률에 의한 지원이 있다면, 초-국가통계 시스템 안에서도, 2(b) 가이드라인을 준수하는 비즈니스 케이스가 NSO에 위해 승인된다면, 사용될 수 있다. 승인 시에는 통계와 관련 연구 활동이 행정 목적의 데이터 수집 또는 자료처리로부터 조직적 측면에서 엄격히 분리되어 있는지 고려해야 한다.
(b) NSO는 통합된 데이터의 변수가 데이터 공급자에 의해 어떤 행정 및 규제 목적으로 이용될 수 있다면, 이 데이터 변수를 제공해서는 안된다.

(c) 외부 연구자는 비즈니스 case가 NSO에 의해 승인 된 경우, CES 지침 "통계 비밀보호 및 Microdata 액세스 관리"에 따라 통합된 데이터셋에서 microdata에 액세스할 수 있다.

(d) 자료 이용자의 의무에 대한 사항이 계약에 포함되어야 하며 비밀보호 규칙을 위반한 경우 법적으로 명시된 잠재적이며 실행가능한 제재를 가해야 한다.

첨부 - 업무사례(business case) 개요 예제

원칙 2에서 표준 승인 절차가 새로운 데이터 통합 제안을 수용하는 것과, 이 작업이 정식 비즈니스 케이스의 형태를 가져야 함을 제시했다. 비즈니스 케이스 방식을 이용할 경우, 다음 주제 영역들이 다루어질 것을 권고한다.

A. 목적
비즈니스 케이스는 통합 데이터가 사용될 목적을 기술하여야 한다.

B. 공식 통계의 이익(Benefit to Official Statistics)
비즈니스 케이스는 제안된 프로젝트가 공식통계를 어떻게 생산하고 개선할 것인지에 대해 기술해야 한다. 공식 통계 개선은 정확성, 신뢰성, 타당성, 시의성, 일관성, 통계의 적용범위, 개념, 통계를 작성하거나 비용 및 응답 부담을 감소하기 위한 방법에 대한 개선을 포함한다.

C. 기타 혜택(Other Benefits)
비즈니스 케이스가 다른 누구에게 도움을 주며, 프로젝트에서 어떻게 도움이 될지를 설명해야 한다.

D. 위험 평가(Risk Assessment)
비즈니스 케이스에는 비밀보호 위험, 데이터 소스의 무결성 위험, 이외 다른 관련 위험에 대한 위험 평가와, 어떻게 이런 위험을 관리할 것인지에 대한 설명을 포함해야 한다.

E. 유지(Retention)
비즈니스 케이스는 사용 목적을 지원하기 위하여 통합 데이터셋을 얼마나 오랫동안 유지해야하는지에 설명해야 한다. 유지는 정기적인 리뷰의 대상이 될 수 있다.

F. Data sources
제안(proposal)은 어떤 데이터소스를 데이터 통합에 사용할 것인지 기술하여야 한다. 여기에는 제안 받은 소스 기관을 명세하고, 각 기관으로부터 받을 데이터에 대해 일반적인 용어로 기술한다.

데이터 통합 프로젝트 하에서 소스 데이터가 수집된 법률적 근거가 나열되어야 한다.

G. 대안(Alternatives)
왜 데이터 통합이 비용, 품질 또는 준수 부담 측면에서 다른 대안들보다 선호되는지에 대한 이유가 명세 되어야 한다.

H. 이해 관계자(Stakeholders)
비즈니스 케이스는 데이터 통합 프로젝트의 모든 주요 이해 관계자(내외부 둘다) 및 그들과의 협의 결과를 나열해야 한다.

I. 이름과 주소의 유지(Retention of names and addresses)
만약 데이터 통합 프로젝트에서 연계를 위해 개인의 이름과 주소를 보유할 필요가 있다면, 이러한 내용을 얼마나 오래 유지할 지 함께 명시해야 한다.

J. 리뷰 빈도(Frequency of Reviews)

비즈니스 케이스는 데이터 통합 검토가 열리는 빈도를 지정해야 한다.

K. 개인 정보 보호 정책 영향 평가 (Privacy Impact Assessment)

법률적 혹은 적절한 NSO 정책에 의한 예외가 없으면, 프라이버시 영향 평가가 완수되어야 한다. 비록 개인 정보 보호 정책은 일반적으로 자연인에 관한 것이지만, 기업체나 사업체의 경우 법인에도 관련이 있음을 함께 언급해야한다. 예를 들어, 일부 국가에서는 농장 같은 특정 비법인 기업에 대한 개인 정보보호를 고려할 수 있다.

부록 3. 비밀보호방법론 정리

Ⅰ. 매크로데이터

1. 민감한 셀 결정방법

○ 임계치 결정방법
- 빈도데이터에 이용하는 방법
- 임계치(threshold value)를 결정하여 그 값 이하는 민감한 셀로 결정하는 방법

○ (n,k) dominance rule
- 상위 n개의 데이터가 전체 셀 값의 k% 이상을 차지하는 경우로 판단
- 보수적인 기준으로 대규모집계표에 주로 이용

○ p% rule
- 자료에서 두 번째로 큰 값에 해당되는 데이터를 이용, 제일 큰 값에 해당하는 데이터를 얼마나 추정할 수 있는가 하는 정도로 판단
- 정보제한량이 적고 널리 이용되는 방법

○ p/q ambiguity rule
- 원시자료의 일부 개별정보를 알 수 있는 경우를 고려하여 p% rule을 좀 더 강화한 방법
- 캐나다 통계청에서 개발한 방법으로 이용자가 알고 있는 원시자료의 정보의 양에 따라 더욱 보수적으로 판다.

2. 비밀보호기법

□ 감추기(Suppression) 방법

○ SODC(Suppressing Only the Disclosure Cells) 방법
- 민감한 셀 결정방법에 의해 민감한 셀을 감추어 제공하지 않는 방법
- 민감한 셀로 결정된 셀만을 감추어 주변합이나 분류구조를 이용해 감추어진 정보가 쉽게 노출됨

○ HCCS(Heuristic Complementary Cell Suppression)
- SODC방법의 문제점을 해결하기 위한 보조셀감추기 방법
- 보조셀을 감출 때 직관적으로 행과 열에 두개 이상의 셀이 감추어지도록 하는 방법
- 선형관련식에 의해 정보가 노출될 수 있음

○ CCS(Complementary Cell Suppression)
- HCCS방법의 단점을 보완하기 위한 방법
- 보조셀 결정시 목적함수를 이용한 네트워크식을 활용하여 선형관력식에 의해 정보가 노출되지 않도록 함
- 감추어진 셀의 수와 감추어진 정보량이 최소화 되도록 보조셀을 결정

□ 변조(Perturbation) 방법

○ 반올림(Rounding) 방법

ⅰ) 전통적인 반올림방법(Traditional Rounding)
- 일반적인 반올림방법
- 기준값을 10으로 할 경우 10 미만은 0으로 10 이상은 10으로 변환

ⅱ) 제어된 반올림방법(Controlled Rounding)
- 모든 셀이 기준값의 배수가 되도록 더하기, 빼기를 반복하는 방법
- 특정한 고리를 가지고 같은 수를 더하고 뺌으로써 주변합을 바꾸지 않으면서 자료를 변환

ⅲ) 50/50 반올림방법
- 기존의 반올림방법에 확률적 개념을 적용한 기법
- 확률적으로 접근하여 랜덤하게 반올림함으로써 반올림되기전 원래의 값을 추측하기가 어려운 것이 장점

○ CTA(Cotrolled Tabular Adjustment) 방법
- 크기데이터(magnitude data)에 적용하는 방법
- 민감한셀 결정방법에 의해 노출위험이 없는 셀이 되기 위한 값으로 변환될때까지 원자료에 임의의 값을 더하거나 뺌
- 선형모형 프로그램을 이용하여 리밸런싱(rebalancing) 과정을 수행하여 셀 값을 조정
- 원자료의 집계결과와 그 주변합이 다름

○ Switching 방법
- 셀 간의 값을 교환하여 정보노출을 제한하는 방법
- 분포를 고려하여 교환할 수 있는 방법에 대한 연구 필요

Ⅱ. 마이크로데이터

○ Swapping 방법
- 동일한 key변수 조합을 갖는 레코드간의 자료값을 상호 교환하는 방법
- 변환 후 Key변수별 집계값은 바뀌지 않음, 변수선정이 중요

○ Coding 방법
- 기준을 정의하고 자료를 코드화하는 방법
- 일반적으로 Top-coding과 bottom-coding을 많이 이용

○ Grouping 방법
- 범주들을 통합하여 노출위험을 줄이는 방법

○ Rounding 방법
- 기준값을 정해 반올림하여 노출위험을 줄이는 방법

○ Micro aggregation 방법
- 기준변수를 선정, 그에 따라 정렬하고 그룹화 하는 등의 방법으로 그룹화 하고 그룹내의 각 자료의 값을 그룹내평균 등으로 대체하는 방법
- 어떻게 그룹화 할 것인지, 그룹내 대표값을 무엇으로 결정할 것인지에 따라 다양한 방법이 있음

○ Perturbation 방법
- 데이터에 특정분포를 이용한 잡음(noise)를 추가하여 노출위험을 줄이는 방법
- 주로 가법잡음, 승법잡음등을 이용
- 이용되는 분포에 따라 다양한 방법이 연구되어지고 있음

※ 감추기(Suppression)기법과 변조(Perturbation)기법의 장단점
- 감추기기법은 정보의 오용 및 왜곡을 막을 수 있는 장점을 가지나 제한되는 정보량이 많을 수 있는 단점을 가짐
- 변조기법은 제한되는 정보량이 감추기기법에 비해 적을 수 있으나 자료이용자가 자료이용시 비밀보호기법에 대한 이해가 필요하며 원데이터와 그 값이 달라 국가통계기관에서 적용에 어려움이 있음