「통계자료의 개인정보보호(PSD) 세미나참가」

해외출장 결과보고서

2008. 10.

통계개발원

Ⅰ. 출장개요

□ 필요성

○ 통계청은 응답자의 개인정보 노출을 제한하기위해 통계자료의 외부제공을 제한하고 있으나, 갈수록 이용자들의 자료제공 요구가 증가하고 있음

○ 이에 **개인비밀보호를 강화함과 동시에 이용자들의 요구를 충족시킬 수 있는 방법**으로 최근 선진국에서 널리 활용하고 있는 통계적 보호방법에 대한 연구가 절실한 상황

□ 목적

○ 통계자료의 비밀보호기법에 대한 통계선진국들의 연구사례를 발표하는 2008 Privacy in Statistical Databases(PSD) 세미나에 참가

○ **최신 연구동향을 파악하고 선진 통계기법을 습득**함으로써 통계연구의 질적 향상을 도모코자 함

□ 출장기간 및 출장지

○ 기　간 : 2008. 9. 23.(화) ~ 28.(일), 6일간

○ 출장지 : 터키 이스탄불

○ 출장자 : 정동명 사무관(통계개발원 연구기획실)

□ 주요일정

일시	방문지	비고
9.23.(화)	・출발 (인천 → 이스탄불)	
9.24.(수)~26.(금)	・회의참석 - 장소 : Larespark Hotel 회의장	
9.27.(토)~28.(일)	・귀국 (이스탄불 → 인천)	

□ 소요예산

○ 약 3,000천원

Ⅱ. 회의개요

□ 회의명

○ 2008년도 통계자료의 개인정보보호(Privacy in Statistical Databases; PSD)를 위한 세미나

□ 회의목적

○ 마이크로(micro)자료 및 매크로(macro)자료의 비밀보호를 위한 최신기법을 소개

○ 실제 자료를 활용하여 비밀보호를 적용한 국가별 사례 소개

○ 통계작성기관의 자료제공 방안 및 통계이용자의 자료활용 방안 등에 대한 토의

□ 회의일시 및 장소

○ 일시 : 2008. 9. 24. ~ 26.(3일간)

○ 장소 : 터키 이스탄불

□ 회의주기 및 참가규모

○ 2004년부터 2년 주기로 유럽지역에서 개최

○ 올해는 16개국의 60여명이 참가하여 31개 주제를 발표

□ 발표주제

가. Tabular Data Protection

○ 개요

- 매크로자료(집계표자료)의 작성에 따른 빈도자료 및 양적 자료의 비밀보호에 대한 주제 발표

○ 발표주제 : 9건

- Using a Mathematical Programming Modeling Language for Optimal CTA
- A Data Quality and Data Confidentiality Assessment of Complementary Cell Suppression
- Pre-Processing Optimisation Applied to the Classical Integer Programming Model for Statistical Disclosure Control.
- How to Make the τ-Argus Modular Method Applicable to Linked Tables?

- Bayesian Assessment of Rounding-based Disclosure Control
- Cell Bounds in Two-Way Contingency Tables Based on Conditional Frequencies
- Invariant Post-Tabular Protection of Census Frequency
- Synthetic tabular data preserving the observed conditional frequencies
- How to prevent cell perturbation procedures from becoming data falsification procedures

나. On-Line Databases and Remote Access

○ 개요

- 온라인 상에서 자료의 제공방법과 이용자들의 원격 접근법에 대한 주제 발표

○ 발표주제 : 3건

- Auduting categorical SUM, MAX and MIN
- Reasoning under uncertainty in on-line auditing
- A Remote Analysis Server - What Does Regression Output Look Like?

다. Privacy-Preserving Data Mining and Private Information Retrieval

○ 개요

- 개인의 비밀을 보호하면서 자료수집 방법과 개인정보 검색에 대한 주제 발표

○ 발표주제 : 4건

- Accuracy in Privacy-Preserving Data Mining Using the Paradigm of Cryptographic Elections
- A Privacy-Preserving Framework for Integrating Person-Specific Databases
- Peer-to-Peer Private Information Retrieval
- A Protocol for Privacy Preserving Neural Network Learning on Horizontally Partitioned Data

라. Legal issues

○ 개요

- 유럽의 통계시스템에서 비밀성에 대한 법적, 정치적, 방법론적인 현안문제에 대한 주제 발표

○ 발표주제 : 1건

- Legal, Political and Methodological Issues in Confidentiality in the European Statistical System

마. Microdata Protection

○ 개요

- 마이크로자료의 제공을 위해 개인의 정보노출을 제한하는 방법 등에 대한 주제 발표

○ 발표주제 : 14건

- A Practical Approach to Balancing Data Confidentiality

and Research Needs: the NHIS Linked Mortality Files

- From t-Closeness to PRAM and Noise Addition via Information Theory
- Robustification of Microdata Masking Methods and the Comparison with Existing Methods
- A Preliminary Investigation of the Gaussian Versus t-Copula for Data Perturbation
- Anonymisation of Longitudinal Enterprise Microdata - Survey of a German Project
- Making public use, synthetic files of the Longitudinal Business Database
- Extensions of the Re-Identification Risk Measures Based on Log-Linear Models
- Assessing Disclosure Risk for Record
- Robust Statistics Meets SDC: New Disclosure Risk Measures for Continuous Microdata Masking
- Parallelizing Record Linkage for Disclosure Risk Assessment
- Towards a More Realistic Disclosure Risk Assessment
- Use of Auxiliary Information in Risk Estimation
- Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data
- How Protective are Synthetic Data?

Ⅲ. 회의결과

□ 주요 주제별 내용

가. 보조 셀 감추기에서 자료의 질과 비밀성에 대한 평가

- 보조 셀 감추기(cell suppression)는 수십 년 동안 경제센서스와 같은 큰 자료의 노출을 제한하는데 사용되어져 왔음
- 이 연구에서는 셀 감추기의 데이터 품질 및 데이터 비밀성에 대해 설명하고 있는데, 특히 셀 감추기가 적절한 수학적 모델을 사용하여 수행할 수 없을 때는 실패할 수밖에 없다는 것을 보여줌
- 또한, 기본적인 노출의 개념 하에서 적절하게 실행한 셀 감추기가 때로는 심각하게 외부의 공격에 취약해질 수 있음을 나타내고 있음

나 정보노출통제(SDC) 하에서 반올림의 베이지안적 평가

- 이 연구에서는 통제되지는 않지만 임의 반올림 하에서 얼마나 노출제한체계의 비밀성이 베이지안(Bayesian) 방법으로 평가받을 수 있는지를 고려하였음
- 그리고 마르코프 연쇄 몬테카를로(Markov chain Monte Carlo)를 기반으로 반올림하여 공표한 자료가 주어진 경우 원래 셀의 빈도수에 대한 조건부 확률분포를 추정하는 방법을 제시하였음
- 정보노출의 통제하에 효과적인 반올림은 참값에 대한 사후

불확실성의 결과이며, 반대로 한 값으로 집중된 사후분포는 비효율적인 정보노출통제의 증거로 제공

다. 레코드 연결을 통한 정보노출위험의 평가

- 외부 침입자(intruder)는 레코드 연결기법을 사용하여 외부 파일과 제공되는 마이크로자료 파일을 일치시키려고 하며, 식별의 위험은 일치가 정확한지에 대한 확률로 정의됨
- 이러한 확률의 자연스러움과 그것의 추정은 조사되어지며, 몇몇 연결들로 인해 모집단의 유일성 개념하에 정보노출의 위험이 보고서로 작성

라. 온라인 청강의 불확실성에 대한 설명

- 이 연구에서는 통계적 데이터베이스의 온라인 청강(on-line auditing)의 불확실성에 대해 증명하기 위해 베이지안 접근법을 제시하였음
- 제공되는 자료로부터 얻을 수 있는 확률적 추론하에서 베이지안 네트워크 주소를 노출하고, 특히 온라인상에서 최대와 최소 청강을 다루고 있음
- 또한, 제시된 모형이 어떻게 질문의 대답을 거부하는 것에서 유도되는 정보의 암시적인 전달을 다룰 수 있고, 이용자의 사전지식을 관리할 수 있는지를 보여줌

마. 암호화선거 패러다임을 사용한 데이터마이닝에서의 개인정보 보호에 대한 정확성

- 데이터마이닝 기술은 특히 시스템에 있는 참가자가 상호 의심할지도 모르는 환경에 상당히 퍼져 있는 경우 민감한 정보의 사용과 취급에 대한 우려가 제기됨
- 이 연구에서는 개인정보보호 데이터마이닝(PPDM) 시스템에서 정확성을 유지하기 위해 규모가 큰 인터넷 선거상의 문헌에서 인용한 암호화의 원조로 잘 알려진 것을 사용하도록 주장
- 또한 제시된 접근방법은 온라인 선거에 대한 고전적인 homomorphic 모형과 특별히 다수 후보의 선거 지원을 위한 확장모형을 바탕으로 하고 있는데, 이에 대한 몇 가지 약점을 설명하고 PPDM 설정에 homomorphic 모형의 변동을 처음 사용한 최근의 계획표를 공격한 것을 보여줌
- 그리고 PPDM이 수평으로 분할된 자료로 구성된 동질의 데이터베이스 상에서무작위의 숲 분류 알고리즘을 얻기 위해 빌딩 블록으로 어떻게 사용되어질 수 있는지를 설명

바. 유럽통계시스템에서 비밀성에 대한 법적, 정치적, 방법론적 현안문제

- 이 연구에서는 과학적 목적을 위해 마이크로자료 파일에 접근해야 하는 연구기관의 요구와 연관된 도전성을 다루고 있음
- 이러한 요구는 응답자의 비밀을 보존하면서 법적 요구와 균형을 이루도록 해야만 하는데, 여기에서는 법적 요구사항을 존중하면서도 연구기관의 자료사용을 높여주는 유럽연합의 유용한 정책과 법률문서를 나타내고 있음

- 구체적으로, 연구과제를 다루고 있는 현재 절차와 통계적 정보노출통계를 위해 유럽통계시스템체제(ESSnet)에서 추진하는 사항을 설명
- 마지막으로 유럽연합에서 최근에 조사한 미래의 동향과 좀 더 구체적으로 원격 액세스 시설의 개발, 노출통제 도구의 강화와 회원국에서 일반적인 정책의 일관된 추진방향 등에 대해 설명

사. 합성자료의 보호방법은?

- 이 연구에서는 통계적 노출제한의 합성과 완전한 합성자료에 의해 제공된 비밀보호를 측정하기 위한 컴퓨터과학 통계자료의 보호방법을 제공
- 완전히 합성자료로부터 유도되어 이용자에게 제공되는 자료파일에서 레코드의 모든 구성요서들은 근사적으로 확률분포로부터 추출되었기 때문에 실제자료라고 표현할 수가 없지만, 여전히 노출위험이 존재함
- 통계적 정보노출제한에서 이러한 위험은 추측노출확률로 요약되어지며, 개인정보 보호 데이터베이스 쿼리에서 이러한 위험은 완전히 다른 개인의 비율에 의해 측정이 가능
- 둘 사이의 관련성과 이에 대한 결과가 주어지고 실제 적용된 예제들이 제공

아. 센서스 빈도수 자료의 불변 사후통계표 보호방법

- 일부 국가들은 사후 통계표 방법으로서 표에서 흩어져 있

는 센서스 빈도자료를 보호하기위해 임의 반올림 방법을 사용하는데, 이러한 방법들은 전형적으로 주변 합계에서 통합된 내부 셀들과 다른 통계표에서 동일한 셀들 사이의 일치하지 않는 결과임

- 임의변조 빈도수에 대한 사후 통계표 방법은 총합을 보전하고 커다란 불일치의 정도를 옳게 해주기 위해 제안되었으며, 이러한 변조는 불변확률전이행렬과 마이크로자료 키변수의 사용에 의해 이루어짐
- 이 방법은 센서스 빈도수를 보호하기 위해 일반화된 사전 및 사후 통계표 방법과 비교

□ 기타사항

○ 주제별 연구내용에 대한 참조

- 보고서 : Privacy in Statistical Databases (Springer, 2008)
  • 통계개발원 자료실에 비치
- 인터넷주소 : http://unescoprivacychair.urv.cat/psd2008/