

「2012년 통계자료에디팅(SDE) 국제회의」

국외출장 결과보고

2012. 11.

통계개발원

I. 출장 개요

가. 출장배경

- UNECE가 주관하고 개최하는 통계자료 에디팅(Statistical Data Editing) 회의에서는 국가통계기관에서 생산되는 통계자료의 에디팅과 관련된 경험 교환, 새로운 기법 개발, 방법론적 지침 문제를 다루고 있어 에디팅 분야의 선진 동향을 파악하고 이 분야 전문가와 의견을 교류할 수 있는 좋은 기회가 됨

나. 출장목적

- 본 회의에 참석하여 에디팅 업무 및 연구에 대한 최근 동향을 파악하고 에디팅 업무의 개선과 에디팅 방법론 연구의 질적 향상을 도모하고자 함

다. 출장자

- 연구기획실 이의규 사무관

라. 출장기간

- 2012. 9. 23. ~ 9. 28. (4박 6일간)

마. 출장지

- 노르웨이 오슬로

바. 출장일정

구 분		시 간 대	주요내용
1일차	9.23(일)	13:15~17:15 19:30~22:40	- 이동(인천공항→히스로) - 이동(히스로→오슬로) - 호텔 이동
2일차	9.24(월)	09:00~17:15	- 회의 참가: 선별적 에디팅/매크로 에디팅 등
3일차	9.25(화)	09:00~18:00	- 회의 참가: 복합자료의 에디팅과 임putation 등
4일차	9.26(수)	09:00~15:00	- 회의 참가: 최신 에디팅 방법 등
5일차~ 6일차	9.27(목)~ 9.28(금)	15:05~16:25 19:35~14:20	- 귀국(오슬로→히스로) - 귀국(히스로→인천공항)

II. 회의 개요

가. 회의명

- 통계자료에디팅에 관한 UNECE 국제회의

나. 회의목적

- 국가통계기관에서 생산되는 통계자료의 에디팅과 관련된 경험교환
- 새로운 기법 개발 및 방법론적 지침 제시

다. 회의일시 및 장소

- 2012. 9. 24. ~ 26. (3일간)
- 노르웨이 오슬로

라. 참가자 : 약 60명

- 호주, 오스트리아, 아제르바이잔, 캐나다, 덴마크, 에스토니아, 핀란드, 프랑스, 독일, 헝가리, 아이슬란드, 이탈리아, 일본, 리투아니아, 멕시코, 네덜란드, 뉴질랜드, 노르웨이, 한국, 러시아, 슬로바키아, 슬로베니아, 스페인, 스웨덴, 스위스, 아랍에미리트, 영국, 미국 등

마. 주요의제

- 선별적 에디팅과 매크로 에디팅
- 에디팅에 대한 글로벌 솔루션
- 다양한 자료와 혼합모드에 따른 데이터 통합에서의 에디팅과 임퓨테이션
- 에디팅 프로세스의 효율성을 분석하기 위한 메타데이터와 파라데이터 활용
- 데이터 에디팅과 임퓨테이션을 위한 소프트웨어 및 도구
- 최신방법
- 센서스 자료의 에디팅과 임퓨테이션

III. 주제별 발표내용

가. 개회사

- Steven Vale(UNECE)은 참가자를 환영하고 회의를 주최한 노르웨이 통계청과 조직위 회원에 감사를 표함
- UNECE Statistical Information and Methodology Unit과 UNECE의 역할에 대해 설명
- Claude Poirier(캐나다)는 Work Session의 의장으로 선출되었음
- 다음 사람은 조직위원회에 참가하여 토론 및 세션 기획자로 활동
주제(I): Jeron Pannekoek(네덜란드), Pedro Revilla (스페인)
주제(II): Li-Chun Yhang(노르웨이)
주제(III): Orietta Luzi(이탈리아), Felipa Zabala(뉴질랜드)
주제(IV): Rudi Seljak(슬로베니아), Philippe Brion(프랑스)
주제(V): Claude Poirier(캐나다), Matthias Templ(오스트리아)
주제(VI): Maria Garcia(미국)
주제(VII): Daniel Kilchmann(스위스)

나. 주요 내용

1) 선별적 에디팅과 매크로 에디팅 (Selective and macro editing)

▶ 서론

- 매크로 에디팅 및 선별적(선택적) 에디팅의 목적은 실질적으로 출력 품질을 감소하지 않고 시간과 비용이 많이 드는 수동 에디팅(검토, 처리 또는 재접촉)을 가능한 제한하는 것임
- 매크로 에디팅 및 선택적 에디팅 모두 쌍방향(수동) 검토를 위해 잠재적 영향력이 있는 오류를 갖는 대상을 선택한다는 공통 목적을 가짐
- 이 세션의 논문은 전통적인 선택적 방법의 응용 및 확장을 다루고 구현 문제를 논의

- ▶ 스페인(Optimization as a theoretical framework to selective editing)
 - 비 에디팅 대상의 평균 제곱 측정 오차를 제약식으로 하여 에디팅할 단위의 수를 최소화하기 위해 제안된 선택적 에디팅에 대한 이론적 틀로서 최적화에 대한 스페인의 발표 자료
 - 나머지 측정오차가 한계치 이하에 있도록 에디팅할 레코드의 최소 부분집합을 선택하는 최적화 문제로 선별적 에디팅을 정의
 - 이 방법에 따라 대상을 선별하고 우선순위를 매기기 위해 R 패키지 및 SAS 매크로를 개발 중임

- ▶ 이탈리아(Multivariate selective editing via mixture models: first applications to Italian structural business surveys)
 - 이탈리아는 혼합모형을 통한 다변량 선별적 에디팅을 발표하고 이탈리아 기업 구조 조사에 첫 적용 사례를 논의
 - SeleMix 소프트웨어를 기초로 기업 구조 통계에 하나의 혼합모형에 근거한 스코어링 방법의 적용을 보여줌
 - 가능한 편익을 외부자료의 보조정보로서의 사용으로부터의 정확도와 비용 절감으로 설명

- ▶ 미국 센서스 국(An application of Selective Editing to the U.S. Census Bureau Trade Data)
 - 해외 무역 자료에 선별적 에디팅의 적용 사례 발표
 - 해외 무역 자료에 대한 스코어링 방법과 가상적 편향의 평가를 포함
 - 전통적인 선별적 에디팅 방법에서는 이전 주기의 자료가 스코어 함수를 구성하기 위해 사용되었는데 이는 주기간 큰 변동으로 인한 무역자료에는 가능하지 않음
 - 현재 주기의 자료만을 이용함으로써 변수의 기대 예측치를 계산하는 스코어 함수를 제시

- ▶ 스웨덴(Tree Analysis – A Method for Constructing Edit Groups)

- 의심스러운 자료를 검출하는 데 사용하는 소프트(soft) 에디팅 규칙 하에서 에디팅 그룹을 구성하기 위한 하나의 방법인 tree 분석을 소개
- 대용량 자료는 반복적인 절차로 동질적인 그룹으로 분할되고 그 결과 가지(branches)와 잎(leaves)으로 분할
- tree는 관측치가 속한 잎에 근거하여 관측치에 대한 값을 예측

▶ 독일(An automated comparison of statistics)

- 기업구조 조사자료 에디팅을 실제 통계와 참조자료(이전주기의 통계)의 자동 비교를 발표
- 이 방법은 총계치의 자동 점검과 의심스러운 레코드의 자동 표식(세부적 검토)을 포함
- 이 방법은 원래의 변수 대신 주성분 변수를 기초로 함

▶ 스웨덴(The Use of Evaluation Data Sets When Implementing Selective Editing)

- SELEKT라 불리는 일반적 틀을 이용하여 한계값(threshold) 설정 관련하여 선별적 에디팅의 구현 문제를 검토함
- 목적은 적절한 수준에서 상대적 가상 편향을 유지하면서 가능한 많이 마이크로 에디팅을 줄이고자 함

▶ 이탈리아(Selective editing as a part of the estimation procedure)

- 예측의 일부분으로 SeleMix 방법을 이용한 선별적 에디팅에 대한 이탈리아의 두 번째 발표 논문
- 스코어에 비례한 확률로 비 에디팅 대상의 표집은 편향 추정과 보정을 가능하게 함
- 샘플수가 충분하다면 작은 오류가 많은 경우의 편향 보정은 추정치를 개선할 수 있음

2) 에디팅에 대한 국제 솔루션(Global solution to Editing)

▶ 서론

- 여러 나라의 모든 국가통계기관이 직면하는 자원 제약의 관점에서 통계 생산의 효율성을 개선하기 위한 국제 협력 개발의 중요성을 논의함

▶ 뉴질랜드(Review of the UNECE Glossary of Terms on Statistical Data Editing)

- UNECE 통계자료 에디팅에 대한 용어집의 재검토 작업을 발표
- 자료 수집, 처리, 공표 등 통계자료에디팅과 관련된 200개 이상의 용어를 현재 포함
- 제안된 첨삭 및 수정에 대한 피드백을 참가자에게 요청

▶ 네덜란드/노르웨이(General flow of editing)

- 에디팅의 이론과 실무에서 최근 개발을 고려한 일반적 에디팅의 절차에 대해 발표
- 업무 흐름도는 전체 에디팅 절차를 하부 절차로 분해됨
- 각 하부 절차는 3개의 일반적 에디팅 활동 형태를 포함하며 각각은 하나 또는 몇 개의 통계적 기능으로 구성
- 네덜란드와 노르웨이의 기업구조통계의 에디팅 실무를 이용하여 일반적 에디팅 흐름에 대한 예시 및 논의

▶ UNECE(Update on the development of GSIM)

- UNECE는 현재 진행중인 GSIM(Generic Statistical Information Model) 개발에 대한 update를 제시
- GSIM v.0.8이 곧 공표 예정이며 올해 말경 버전 1.0 도래 예상

▶ 스웨덴(Two Paradigms for Official Statistics Production)

- 스웨덴은 공식통계 생산을 위한 두 개의 패러다임을 발표
- 이 접근방법은 공식통계 안에서 다양한 토의에 의해 영감을 받았으며 에디팅에 적절함

- 패러다임 1은 특별한 목적을 위한 에디팅을 갖는 stovepipe 접근 방법이며 패러다임 2는 일반적 사용을 위한 자동 에디팅을 갖는 통합 시스템과 연관

▶ 캐나다 통계청(The development of a data editing and imputation tool set)

- 통계적으로 수집된 자료와 행정자료 모두에 대한 공통 필수조건을 감지하는 에디팅과 임퓨테이션 도구 세트의 개발을 발표
- 잠재적 솔루션을 평가하기 위해 품질의 차원이 정의됨
- 도구 세트의 바람직한 특성: 기능성, 품질 기준(적절성, 접근성, 해석성, 일치성, 정확성, 시의성), 중요한 소프트웨어 특성(호환성, 신뢰성, 유지성, 상호작동성)
- 제안된 기본 툴 세트는 BANFF, CANCEIS 및 SELELKT, 다른 툴은 기능성에 약간의 틈을 채우기 위해 노력해야 할 것으로 평가

▶ 뉴질랜드 통계청(On tap: developments in statistical data editing at Statistics New Zealand)

- 뉴질랜드 통계청은 통계자료 에디팅에서의 개발 사례를 발표
- 비용대비 효율적으로 목적에 부합하는 통계를 생산하기 위해 새로운 인프라가 요구됨
- 이에 따른 도전과 배운 교훈 소개: 일반적/특정 요구의 균형유지, 일반 서비스로서 구현되는 데 적절한 프로세스 사항 결정, 절차 문화에서 건설적인 혁신 문화로 이동, IT 개발 프로젝트를 위한 신속한 프로젝트 관리의 채택
- 향후 과제로 GBPM의 검토, 개발로부터의 이득에 대한 측정과 보고 프레임워크의 구현, 자료수집 절차의 변형

3) 다양한 자료원과 혼합 모드에 따른 데이터 통합에서의 에디팅과 임퓨테이션 (Editing and Imputation in the context of data integration from multiple sources and mixed modes)

▶ 서론

- 비용과 응답부담을 줄이는 반면 품질에 대한 정보를 국가통계기관에게 제공하기에 행정자료를 포함한 다양한 자료원으로부터의 통합자료는 공식통계에서 더욱 일반화됨
- 목표 변수에 대한 정보가 다양한 틀(e-조사표, 종이 조사표, 직접 회사 시스템 접근, 행정자료의 사용 등)을 통해 수집되는 혼합방식의 자료 수집의 특별한 사례로서 간주될 수 있음
- 다양한 소스/혼합방식 자료의 품질에서의 차이가 주요 문제임. 각 소스/혼합 방식의 자료를 다루는 데 필요한 통합된 자료의 일관성을 보증하기 위한 에디팅 전략 관련

▶ 노르웨이(Micro integration of register-based census data for dwelling and household)

- 주거 및 가구에 대한 등록 기반의 센서스 자료의 마이크로 통합에 대한 논문을 발표
- 2011년 노르웨이 등록기반 인구주택 센서스에 의해 요구되는 자세한 표와 분석을 위한 요구를 만족하는 주거와 가구 연계의 완전한 센서스 자료 파일을 생성하기 위한 통계적 방법과 접근방법을 제시
- 이중 최근방 대체법은 직접 매칭이 불가능한 경우 다양한 형태의 대상간 마이크로-연계 문제에 대한 일반적인 접근방법을 제공

▶ 뉴질랜드(All answers? Statistics New Zealand's Integrated Data Infrastructure)

- 다양한 통계 부분을 커버하고 사회경제 현상의 복합적 다변량 분석을 가능하게 하는 통합 자료 인프라구조(IDI)의 개발 관련 문제를 발표
- IDI에서의 에디팅 전략은 다양한 자료원으로부터의 동일한 대상에 대한 비일관성을 다루고, 하나의 레코드에서 오류나 결측을 처리하고, 하나의 시간 주기와 시계열에서 하나의 레코드에 걸친 변수들에서의 일관성을 보증하기 위함

- ▶ 영국(Editing challenges for new data collection methods)
 - 다양한 자료수집방법 사용의 에디팅 도전과 전자 조사표에서의 점검을 어떻게 다룰 것인지에 초점을 둠
 - 어떻게 내용 점검을 제시할지 그리고 사용 정도에 관해 전자조사표에서의 에디팅 규칙을 테스트하기위해 사용된 실험설계에 대해 논의

- ▶ 프랑스(Methodological Questions Raised by the Combined Use of Administrative and Survey Data for French Structural Business Statistics)
 - 이전 워크 세션에서 발표한 기업구조통계를 위한 다양한 소스 장치인 Esane에 대해 발표
 - 이는 행정자료와 조사자료의 결합 사용에 근거
 - 항목의 정의 문제와 추정에 있어서 다양한 자료원의 사용의 중요성에 대해 초점을 두고 발표

- ▶ 헝가리(Studying the Options of Substituting a Regular Statistical Survey with Administrative Data)
 - 보건통계 영역에서 정기 통계조사를 행정자료로 대체하는 그들의 경험을 발표
 - 연간조사는 행정자료와 비교되었으며 자료제공자의 목표 집단에 대한 행정적 부담을 줄이고 자료품질에서의 전체적 개선에도 불구하고 편향 효과의 도입 위험이 남음

- ▶ 아부다비 통계센터(Evolving data processing in the statistics centre)
 - 자료의 품질을 향상시키기 위해 혼합조사의 자료수집과 자동 오류 탐색을 개발하기 위한 계획을 제시
 - 자동 에디팅이 조사의 품질과 효율을 향상시켰음을 확인하였으며 사업체 조사에서 오류와 이상치를 다루기 위한 방법을 연구 중

▶ 영국(Editing and imputing VAT for the Purpose of Producing Mixed-Source Turnover Estimates)

- 경제통계에서 행정자료를 좀 더 활용하기 위한 방법을 소개
- 혼합 자료원 단기 매출액 추정치의 생산에서 사용된 VAT 매출액 자료에서의 오류를 찾아내고 대체하는 방법을 소개하고 평가
- 행정자료에서의 오류를 찾아내고 수정하는 방법의 선택은 복합 통계를 생산하기 위해 사용된 자료에 궁극적으로 달려있음. 특히 커버리지, 시의성과 정확성과 관련

▶ 네덜란드(Imputing missing values when using administrative data for short-term enterprise statistics)

- 단기 기업체 통계(STS)에 대한 행정자료 사용시 결측치 대체에 대한 몇 개의 유럽 통계 조직과의 공동 작업 소개
- 이탈리아, 독일, 핀란드의 분석은 주요 STS-행정자료 추정치에서의 주요 수정사항은 인구에서 대상의 불확실성에 의해 야기됨을 보임
- 연구 및 개발은 활동 인구에 가장 적합한 추정방법을 선택함에 초점을 두어야 함을 권고

▶ 이탈리아(Improvement of the Timeliness of the Italian Business Register via Imputation of Missing information in Administrative data)

- 행정자료의 결측 대체를 통해 이탈리아 기업 명부의 시의성 향상을 발표
- 결측을 대체하기위한 충분한 메커니즘이 있다면 최종자료보다 임시 자료가 사용될 수 있음
- 사용된 접근방법의 유용성을 평가하기 위해 두 개 버전의 행정자료가 비교되고 상당한 정확성을 갖게 됨

4) 에디팅 절차의 효율성을 분석하기 위한 메타데이터와 파라데이터 (Using metadata and paradata to analyse the efficiency of editing)

▶ 서론

- 메타데이터와 파라데이터는 자동화 적용과 방법론 발전과 함께 통계 에디팅 절차에서 더욱 더 중요해지고 있음
- 절차의 적합성과 효율성을 분석하고 약점을 찾아내고 지속적 개선을 지원하기 위해 정보의 한 소스로서 특별히 사용됨

▶ 프랑스(Assessment and improvement of the selective editing process in Esane)

- 선택적 에디팅 절차의 결과와 관련된 메타데이터와 파라데이터에 근거한 에디팅 절차에서의 개선을 발표
- 에디팅 직원으로부터의 경고 메시지, 의심스러운 사업체의 수나 피드백에 대한 정보를 포함
- 메타데이터와 파라데이터로부터 나온 정보는 조사표 내용과 구조의 변경과 같은 방법론 개선에 사용될 수 있음
- 자료의 사용자에게 절차를 설명할 때 또한 유용

▶ 핀란드(Outlining a Process Model for Editing With Quality Indicators)

- 절차를 모니터링 하는데 도움을 주는 품질 indicator에 대한 아이디어와 함께 통계자료 에디팅을 위한 절차 모형을 발표
- 50개 이상의 indicator가 고려되고 일부는 특정 통계에서만 적합
- 에디팅 프로세스에 대한 몇 개의 표준 indicator, 에디팅 비율 등을 소개

▶ 캐나다(An Embedded Experiment to Test Non-Response Follow-Up Strategies when Using Electronic Questionnaires)

- 전자 설문조사에 대한 무응답 재추적의 효율성을 개선하기 위한 다양한 접근방식을 테스트하기 위한 실험의 결과를 발표
- 연간조사에 대한 첫 몇 개월 수집에 대한 전화 요청 없이 유의미한

응답이 얻어질 수 있음을 보여줌

- 더 자주 e-메일 재알림은 응답을 얻는데 유익함
- 자료 수집 초기에 매 2주마다 e-메일 재알림 전략은 작은 비용으로 45%의 응답을 얻는 데 성공
- 적절한 시기에 전화 재접촉을 시작하는 것이 필요하나 늦게 시작하는 것은 최종조사 응답률에 오히려 부정적인 효과를 나타냄

▶ 스웨덴(Editing staff debriefings at Statistics Sweden)

- 에디팅 직원의 보고를 들음으로써 에디팅에 대한 질적 정보를 수집하기 위한 접근 방법을 발표
- 이 방법은 질적 파라미터를 제공하기 위해 에디팅 직원과 면접을 이용
- 이러한 보고를 들음은 측정 도구와 에디팅 절차에서 여러 가지 문제를 찾을 수 있으며 더 나은 개선을 위한 기초를 제공
- 에디팅 직원은 문제나 오류를 일으키는 것이 무엇인지에 대한 유용한 아이디어를 종종 가짐
- 다른 방법에 의해 얻어진 정보를 보완하면서 적절한 비용으로 상당한 정보가 얻어짐

5) 데이터 에디팅과 대체를 위한 소프트웨어와 도구들(Software & tools for data editing and imputation)

▶ 서론

- 이 세션에서의 대부분의 발표는 프로그램 R 기반에서 개발된 도구에 기초함
- 프로그램 R은 다른 플랫폼과 데이터베이스와 함께 연동가능

▶ 미국 센서스국(TEA, a generalized system designed to unify and streamline demographic survey processing)

- 원자료에서 공표까지의 최신의 인구조사절차에 대한 일반화 시스템

TEA에 대해 소개

- 비공식적인 R 패키지로서 사용가능하며 미국 센서스 국에서 사용되어왔음
- TEA는 에디팅, 임putation을 위한 플랫폼을 제공하며 이러한 절차를 별도로 진행

▶ 네덜란드(Innovative visual tools for data editing)

- 데이터 에디팅을 위한 두 개의 시각적 도구, treemap과 tableplot를 소개
- treemap은 경제활동계층의 다양한 수준을 갖는 경제자료와 같이 계층 구조를 갖는 자료에 대한 시각화 방법임
- tableplot은 대용량자료에서 이상치와 범상치 않은 패턴을 검출하고 데이터 에디팅과 임putation 시 자료품질을 모니터링하는데 사용될 수 있음
- 이 도구는 R에서 구현되며 CRAN(Comprehensive R Archive Network)에서 무료로 이용 가능

▶ 오스트리아(Interactive Adjustment and Outlier Detection in time dependent data in R)

- R을 이용하여 시계열자료에서의 쌍방향적인 계절 조정과 이상치 탐색에 대해 발표
- 미국 센서스 국의 X12-ARIMA 계절 조정 소프트웨어가 R로 개발된 GUI(graphical user interface)로 포장됨
- GUI를 사용하는 데 R이나 패키지에 대한 사전지식이 크게 요구되지 않음

▶ UNIDO(Screening Methods and Tools for UNIDO Industrial Statistics (INDSTAT) Databases)

- UNIDO(United Nations Industrial Development Organization) 데이터베이스에서 screening 방법에 대해 공동 작업을 발표

- 자료의 논리적 관계에서 매우 복잡한 이상치 탐색 절차까지 다양하게 토의됨
- 페이퍼에서는 방법에 대해 자세하게 논의하고 있지만 발표는 시각화에 초점을 둠
- 때로 매우 간단한 screening 방법이 매우 복잡한 방법보다 매우 효과적일 수 있음

▶ 네덜란드(Automatic data editing with open source)

- 공개 소스의 자동에디팅에 대해 발표
- R 패키지 editrule과 deducorrect의 사용가능성 소개
- editrule 패키지는 에디팅 규칙 관리와 운영에 대한 도구모음으로서 정리된 것임
- 향후 실제 통계 생산 절차에서 이러한 패키지의 응용에 초점을 둘 예정

▶ 리투아니아(Application of developed SAS Macro for editing and imputation in Statistics Lithuania)

- 에디팅과 임퓨테이션을 위해 SAS 매크로 프로그램을 개발
- 오류 탐색, 이상치 탐색, 최근방 대체법, 모형기반 대체, 분포 이용 대체의 5개의 부분으로 구성
- 추가 기능이 확대될 예정

▶ 일본(Multiple Imputation of Turnover in EDINET Data: Toward the Improvement of Imputation for the Economic Census)

- 경제 총조사에서 자료의 품질을 향상시키기 위해 매출액의 다중 임퓨테이션에 대한 작업을 소개
- 표준 단일 임퓨테이션은 여러 가지 제약을 가지므로 다중 임퓨테이션을 이용해 이를 극복할 수 있다고 함
- R 패키지 amelia11(일반적인 목적의 다중 임퓨테이션 도구)는 대용량 자료에서의 다중 대체를 구현함에 성공적임을 보임

- 대체에서 이상치의 영향은 향후 연구에서 고려될 예정

6) 최신 방법(New and emerging methods)

▶ 서론

- 이 세션하의 발표논문은 자료의 에디팅과 임putation을 해결하고 개선하고 최적화하는 방법과 기술에 있어서 새로운 아이디어와 진전을 발표
- probability editing, machine learning method, 모형기반 대체방법, 수치형 자료의 자동에디팅방법에 대해 발표

▶ 스웨덴/에스토니아(Probability editing)

- 확률표집 프레임워크를 사용하여 에디팅할 대상을 선별하는 방법을 제시
- 이 방법은 모든 자료 유형에 적용 가능하고 이 방법을 사용하여 추정량의 통계적 특성에 접근 가능
- 확률 에디팅을 사용할 때 측정 오류로 인한 편향이 소거됨

▶ Eurostat(Use of Machine Learning Methods for Categorical Data Imputation)

- 신경망 또는 베이지안 신경망 분류를 이용하여 범주형 자료를 대체하는 machine learning method의 사용을 소개
- machine learning method의 결과와 로지스틱 회귀와 다중 대체로부터 나온 결과와의 비교에서 machine learning method가 전통적인 방법에 있어서 쉽게 자동화되고 실질적 향상을 제공할 수 있음
- machine learning method가 대용량 자료에서 확장 가능

▶ 슬로베니아(Implementation of the Bayesian approach to imputation at SORS)

- 베이지안 접근방법의 대체 구현 소개

- 가구조사에서 연간 총수입을 대체하는 방법으로 베이지안 모형과 선형 회귀를 결합
- 임putation으로 인한 분산을 평가하기 위해 반복하여 실행
- 데이터가 자료 분할에 의해 잘 설명되지 않는 자료에서는 그 결과가 좋지 않음

▶ 네덜란드(Automatic editing with hard and soft edits – some first experiences)

- 자동오류탐색문제를 해결하고자 할 때 약한 에디팅 규칙이 고려될 수 있는 새로운 자동 에디팅 방법을 소개
- 약한 에디팅 규칙의 사용은 오류위치탐지를 향상시키지만 복잡성이 추가될 경우 향후 연구가 필요

7) 센서스 자료의 에디팅과 임putation(Editing and imputation of census data)

▶ 서론

- 이 세션 주제는 센서스자료에 적용된 에디팅과 임putation 기법의 방법론적 발전에 초점을 둠

▶ 슬로베니아(Editing of multiple source data in the case of Slovenian Agricultural Census 2010)

- 2010년 슬로베니아 농업총조사에서 다양한 소스 자료의 에디팅에 대해 발표
- 조사와 행정자료에 기초하고 메타데이터 추출 절차와 다양한 소스를 통합하기 위한 데이터베이스 접근방법을 사용하여 관리
- 에디팅 비용이 이전 센서스와 비슷하지만 이 접근방법이 품질 개선에 투명성을 이끌 뿐 아니라 응답부담과 IT 비용을 줄임

▶ 오스트리아(The data imputation process of the Austrian register-based Census)

- 등록기반 인구주택 총조사의 임putation 절차 발표
- 2011 센서스는 완전히 등록부에 기초한 첫 센서스로 이전 센서스와 비교하여 비용을 상당히 절감
- 다양한 소스로부터 사용가능한 변수들을 조정하는 우선순위화 규칙을 개발하는 것이 가장 큰 도전
- 에디팅과 임putation 방법은 결정적 분포 대체(핫덱 대체의 일종)를 포함

▶ 영국(The Practical Implementation of the 2011 UK Census Imputation Methodology)

- 센서스 임putation 방법론의 실무적 구현에 대해 발표
- 에디팅과 임putation 패키지 CANCEIS가 2011년 센서스에서 처음으로 사용됨
- SAS wrapper가 CANCEIS 모듈 사이에서 자료를 옮길 때 사용되었으며 수동 에디팅과 임putation을 줄이고 이전 도구보다 더 유연하고 더 빠른 방법임

▶ 아부다비 통계센터(Edit And Imputation of the 2011 Abu Dhabi Census)

- 기관에서 처음으로 시행된 2011 센서스에서 에디팅과 임putation의 경험을 발표
- CANCEIS는 성공적으로 이행되었으나 가구 구성에서 사회적인 차이(대가족, 다처제 등)로 인해 소수 공통 에디팅규칙이 완화되거나 재정의되어야 했음
- 기부자(donor) 대체과 결정적 대체의 사용이 수동적 개입에 대한 요구를 최소화 함

▶ 멕시코(Editing Census Data: Mexico's experience)

- 지리정보시스템의 도움으로 센서스 자료의 에디팅 경험을 발표
- 가장 최근 센서스는 6가지 종류의 조사표의 전통적 자료수집방법을

사용하였음

- 이들 중 하나는 도시 환경 문제를 다루었는데 공간적 관점에서 잠재적 오류와 불일치성을 찾아내기 위해 mapping 기술의 사용에 잘 적합

IV. 시사점 및 참가소감

가. 시사점

- 데이터 에디팅은 통계생산 프로세스에서 가장 비용이 많이 드는 부분이고 비용절감을 위한 주된 부분으로서 통계작성기관 고위층의 관심사임
- 에디팅의 목적은 단지 자료를 수정하는 것만이 아니라 오류로부터 조사 개선의 정보를 얻는 것임. 에디팅과 임퓨테이션 전략을 개발하는데 도움이 되도록 원자료와 에디팅된 자료를 포함하는 두 개의 자료를 갖는 것이 필요
- 선별적 에디팅은 표준화하기 어려움. 해당 데이터와 변수에 맞는 방법을 채택할 필요가 있음. 파라미터와 메타데이터를 이용하여 가장 중요한 오류에 초점을 맞추는 것이 필요. 몇몇 나라의 에디팅 직원 역시 선별적 에디팅을 채택하는 것을 꺼리고 있으며 교육과 지원이 필수적이라 논의
- 메타데이터와 파라미터에 대한 요구는 사용자의 유형에 따라 다름. 내부 사용자에게는 생산과정의 효율성을 높이기 위함이 목적이라면 외부 사용자에게는 자료의 품질에 대한 정보를 제공하기 위함
- 소프트웨어 세션에서는 R을 이용한 발표논문이 대다수임. 많은 통계 작성기관 직원이 SAS에 친숙하지만 점차적으로 R을 사용하기 위해 배우는 사람이 늘어나고 있는 추세임. 프로그램 R은 많은 통계 패키지를 무료로 사용할 수 있다는 장점이 있음
- 영국과 아부다비 통계센터는 센서스에서 자동 에디팅과 임퓨테이션을 캐나다의 CANCEIS를 이용한 사례를 발표, CANCEIS에 대한 활용 여부 검토 필요

나. 참가소감

- 각국의 에디팅 업무의 경험을 듣고 다양한 에디팅 방법을 습득하며 에디팅에 관한 최신 동향을 파악할 수 있는 기회가 되었음
- 특히 네덜란드에서 발표한 R 프로그램을 사용한 자동에디팅과 시각화 기법이 흥미로운 주제임
- 이전 John Kovar 캐나다 의장에 이어 Claude Poirier(캐나다)는 Work Session의 의장으로 선출되어 통계 자료 에디팅 분야에서 캐나다의 영향력을 실감함
- 노르웨이의 물가와 유럽 경제위기로 인해서인지 50~60명 인원이 회의에 참석하였으나 열정적으로 각국의 통계업무에 대한 경험 공유와 국제협력을 논의하는 모습이 인상적임

다. 기타

1) 차기 회의 의제

- 선별적 에디팅/매크로 에디팅
- 에디팅의 다른 방법 - 기업 명부와 다양한 소스
- 데이터 에디팅 이행 시의 지원
- 최신 에디팅 방법
- 센서스와 사회 자료의 에디팅
- 국제 협력과 소프트웨어 및 도구
- 센서스 자료의 에디팅과 응답대체: 등록기반 센서스 중심으로
- 매크로 에디팅

2) 차기 회의는 2014년 봄에 프랑스 통계청(INSEE)에서 개최 예정

3) 본 회의에 대한 자세한 정보는 다음 웹사이트에서 얻을 수 있음

www.unece.org/stats/documents/2012.09.sde.html