2016년 국제조사방법론 심포지엄

【2016 International Methodology Symposium】

참석 결과보고

2016. 4.

통계청
STATISTICS KOREA

통계개발원
조사연구실

목 차

Ⅰ 출장 개요

1. 심포지엄 개요

○ 심포지엄명 : 2016 International Methodology Symposium

; “Growth in Statistical Information: Challenges and Benefits”

- 기 간 : 2016. 3. 22(화)～3. 24(목)(3일간)
- 개최지 : 캐나다, 오타와(Palais des congres de Gatineau)
- 관련 사이트 URL

: http://statcan.gc.ca/eng/conferences/symposium2016/index

* IMS(International Methodology Symposium): 캐나다 통계청이 주관하는 국제 조사방법론 심포지엄으로서 각 국의 통계청 실무자 및 학계 연구진 등이 조사방법론에 대한 최신 동향과 기법을 소개하고 연구결과를 공유하는 자리임.

2. 출장 개요

○ 기 간 : 2016. 3. 21(월)～3. 26(토)(4박 6일)

○ 출장배경 및 목적

- 최신 조사방법론에 대한 이해 및 연구동향 파악
  - 이번 주제는 “통계적 정보의 증가” 로, 이용자들이 활용할 수 있는 정보의 양이 급속히 증가하고 있는 최근 환경이 자료를 수집·분석·공표하는 데 어떻게 영향을 줄 것인가에 대해 주로 논의

3. 출장 주요 내용

○ 조사방법론에 대한 최신 연구동향 및 선기법 파악

- 이번 심포지엄은 총 10개 세션으로 구성되어 있으며 조사방법론과 관련된 다양한 연구주제에 대한 최근 동향을 다룸

○ 조사방법론 연구·개발을 위한 정보 공유 및 자료수집

- 주요 주제는 행정자료 활용, 빅데이터, 공공데이터, 조사과정 데이터(paradata), 자료연계 및 통계적 매칭, 노출제어 등

4. 프로그램 구성

Day 1 – 2016.3.22.(화)

08:00 ~	**Registration**	
08:45 - 09:00	**Opening Remarks**	
09:00-10:00 Plenary Session	**Session 1 :** Keynote Address Methodological Issues and Challenges in the Production of Official Statistics	
10:00-10:30	**Morning Break**	
10:30-12:00 Concurrent Sessions	**Session 2A** **:** Big Data in Official Statistics	**Session 2B :** Applications Related to Growth in Statistical Information
12:00-13:30	**Lunch**	
13:30 -15:00 Concurrent Session	**Session 3A** **:** Total Survey Error	**Session 3B :** Alternative Data Sources to Replace or Complement Survey Data
15:00-15:30	**Afternoon Break**	
15:30-17:00 Concurrent Sessions	**Session 4A :** Open Data	**Session 4B :** Quality of Administrative Data

Day 2 - 2016.3.23.(수)

08:45~-09:45 Plenary Session	**Session 5 :** Waksberg Award Winner Address Toward a Quality Framework for Blends of Designed and Organic Data	
09:45~10:00	**Speed Advertisement for Posters and Software Demonstration**	
10:00-10:30	**Poster Session, Software Demonstration and Morning Break**	
10:30-12:00 Concurrent Sessions	**Session 6A :** New Advancements in Record Linkage	**Session 6B :** Confidentiality
12:00-13:30	**Lunch**	
13:30 -15:00 Concurrent Session	**Session 7A** **:** Non-traditional Methods for Analysis of Survey Data	**Session 7B :** Applications of record linkage and statistical matching
15:00-15:30	**Poster Session, Software Demonstration and Afternoon Break**	
15:30-17:00 Concurrent	**Session 8A :** Paradata	**Session 8B :** Use of Administrative Data

Day 3 - 6.3.24.(목)

08:45-10:15 Concurrent Sessions	**Session 9A :** Scanner Data	**Session 9B :** Health Data
10:15-10:45	**Morning Break**	
10:45-11:45 Plenary Session Sessions	**Session 10A** **:** Data Science for Dynamic Data Systems: Implications for official Statistics	
11:45 –12:00 Closing Remarks	**Session 11 :** Closing Remarks	

Ⅱ 심포지엄 주요 내용

○ 공식통계 생산을 위한 빅데이터 활용, 행정자료의 품질관리, 공공데이터(open data), 자료연계, 비밀보호, 조사과정데이터(paradata), 스캐너 데이터 등과 관련된 연구가 주로 다뤄졌음

- (빅데이터) 다양한 종류의 자료원 증가 등 변화하는 데이터 환경하에서 빅데이터의 활용은 불가피한 상황이며 이에 빅데이터 활용의 장점 및 문제점, 빅데이터를 활용하여 공식통계를 생산하기 위해 필요한 과제 등 설명
- (공공데이터) 공공데이터(open data) 정책이 가지는 의미를 소개하고 캐나다 통계청의 공공데이터의 원칙, 발전 및 현재 공공데이터전략을 소개
- (행정자료 활용) 다양한 행정자료를 활용한 자료연계를 통해서 부가적인 정보를 생산하고, 기존의 조사가 가지는 한계를 극복하기 위해 행정자료를 표본설계나 통계생산의 데이터 보정 등 조사품질개선을 위해 활용

○ 이외에도 비밀보호, 조사과정데이터(paradata), 물가지수작성을 위한 스캐너데이터의 활동 등이 발표되었음

Ⅲ 주제별 내용

주제1 기조연설

❖ **Methodological Issues and Challenges in the Production of Official Statistics**(Danny Pfeffermann, Government Statistician of Israel)

○ (배경) 기술적 진보와 함께 빅데이터의 이용가능성이 매우 증가하였으며 더 정확하고, 더 자세하고, 더 시기적절한 공식통계에 대한 수요도 증가, 공식통계생산을 위한 빅데이터의 수집·관리·접근 부분이 공식통계 생산자들에게는 큰 과제로 남아 있음.

○ (주요내용) 이에 공식통계생산을 위해 앞으로 놓인 주요한 새로운 방법론적 과제와 그러한 것들을 처리해야하는 방법 제시. 특히나 다음의 당면과제에 대해 제시: 빅데이터의 사용가능성, 데이터 접근성 증대 이면의 사생활과 비밀보호유지, 웹패널을 통해 얻어진 자료의 사용 가능성, 모드효과에 대한 설명, 미래 센서스를 위해 행정자료와 소지역추정의 통합 문제

- '빅데이터'는 흔히 빠르게 변하는 복잡한 데이터, 구조·출처·형식 면에서 일정하지 않아 정확도에 영향을 줄 수 있는 내재된 불확실성을 포함한 대용량데이터를 의미(예- 수많은 자동차의 이동, 통신회사 정보, CPI예측 등에 활용되는 온라인 판매상품 등)

- 빅데이터의 형태: 1) 특정인과 연관된 데이터(핸드폰, 카메라 등), 2) 비구조적이고 비규칙적인 데이터(소셜네트워크, 전자거래 등)

- 빅데이터의 특징: Big Data → Big Problem →Big Headache
  - 포괄범위와 선택적 바이어스: 특정정보가 많아 예측이 맞지 않은 경우가 많음(예를 들어, 인터넷을 통해 얻은 구매정보는 국민전체가 아니라 특정인과 관련된 것이 대부분)
  - 데이터 접근성
  - sampling error: 전통적인 조사가 가지지 않았던 문제들,
  - 동태적데이터를 샘플링하는 것은 기존의 것과는 다름
  - 빅데이터는 모집단의 일부에 대한 정보만 존재. 따라서 광범위한 연계가 필요
  - 추정: 설계기반 모델추정처럼 새로운 빅데이터기반 알고리즘 추정량을 활용
  - 거대용량으로 저장할 수 없음. 클라우드나 데이터센터 필요
  - 데이터 보호: 사생활과 기밀보호, 침입자로부터의 데이터 보호, 노출제어측면에서 임의데이터(Synthetic Data)와 research(safe) room을 활용한 광범위한 변화 필요

○ (결론) 그러나 빅데이터는 불가피하고 잠재적인 편익이 많음: 적시성, 넓은 범위, 샘플프레임이 불필요, 설문지나 조사원 불필요 등

- 국가통계생산을 위해 웹패널 사용이 가능. 그러나 대상자는 인터넷으로 접근하는 한계가 있으므로 흔히 성향점수와 calibration을 활용하여 문제해결 가능

- 선택적 효과와 측정효과 두 가지 혼합적 효과를 가진 혼합모드 효과 활용 가능

주제2 Big Data in Official Statistics

❖ **Challenges to Methodological Research in Official Statistics**
(Kees Zeelenberg, Statistics Netherlands, Netherlands)

○ (배경) 공식통계는 다음과 같은 과제에 직면: 품질유지 및 재고, 이용자 수요, 불응답, 예산제약 등

- 네덜란드 통계청에서 최근 채택한 공식통계에 있어서의 전략적 방법론적 연구프로그램(Strategic Methodological Research Program)의 주요토픽 소개

> · 국민계정의 품질, 특히 GNI 성장률
> · 빅데이터 관련하여 어떻게 대표적인 추정을 할지?
> · 웹조사에서의 불응
> · 복잡하고 일관된 현상에 대한 통계적 분석
> · 공공데이터와 빅데이터에 대한 노출제어
> · 주요 통계적 추정에 대한 적시성의 증가
> · 노동시장에서의 개인활동의 동태성
> · 개발이 분권화(decentralized) 되어 있을 때의 통계개발의 체계

○ (주요 내용) 빅데이터에 대한 품질평가가 중요. 품질평가를 위해 베이지안기법, 벤치마킹, 시계열데이터인 경우 주로 모델기반 추정을 많이 활용되는 것처럼 빅데이터기반에도 알고리즘 추정이 활용될 수 있음. (예를 들어 빅데이터와 GNP, 세금 데이터베이스; 급여나 매출액을 통한 수집. 빠르지는 못함, 기업회계계정에 직접 접근; 개발 중, 기업들의 금융거래; 빠른 데이터수집 가능하지만 사생활문제 상존)

- 따라서 빅데이터기반 모델추정을 활용

- 이용자수요에 따라 국가 통계기관은 단순한 데이터가 아닌 더 적시의 더 좋은 정보를 생산해야함. 이러한 측면에서 빅데이터는

정보생산을 위한 새로운 원천. 그러나 빅데이터는 사람이나 기업에 대한 것이 아니라 사건에 대한 것. 따라서 다른 데이터와 매칭하여 연계하는 것이 쉽지 않음. 따라서 빅데이터는 결국 선택적이어야 함.

- 또한 빅데이터 사용에 있어 주요 고려사항 중 하나는 큰 바이어스 가능성. 바이어스와 단위문제해결을 위한 간접방법: 성향모델 활용
  · 모조설계기반 방법
  · 기계적 학습 기술(예- 회귀트리, k-nearest neighbor 등)
  · 통합; 부차데이터로서 빅데이터를 원조사데이터를 이용하여 추정

○ (결론) 이처럼 빅데이터 활용에 있어서 방법론적 전략이 필요

  · 품질 · 분석 · 대표성부분에서 알고리즘적 모델기반 추정 필요
  · 바이어스와 포괄범위 등 빅데이터가 가진 문제해결을 위해 매칭, 연계 및 통합(즉, 부차데이터로서 빅데이터를 활용하여 원조사 데이터를 추정하는데 활용 가능) 활용 필요
  · 빅데이터에 대한 총체적인 생산프로세스 관리 필요

❖ **Profiling of Twitter data: a Big Data selectivity study**
(Joep Burger, Quan Le, Olav ten Bosch and Piet Daas, Statistics Netherlands, Netherlands)

○ (배경) 인간행동과 경제활동에 대해 증가하는 양의 데이터는 자동적으로 소셜미디어, 센서, 핸드폰등과 같은 매체에 의해 기록됨. 이러한 소위 빅데이터는 공식통계에 대한 잠재적인 자료원임.

○ (주요 내용) 빅데이터의 주요 특성에 대해 설명

- (생성) 주로 다음에 의해 생성
  · 사람근원데이터(human-sourced data): 소셜미디어나 인터넷 검색을 통해
  · 프로세스에 의해 가능한 데이터(Process-mediated data)
    : 스캐너데이터나 온라인이체 등
  · 기계적 학습기술(Machinery-learning technology)

- (주요 과제) 빅데이터에는 다음과 같은 과제들이 존재

· 대표성
· 기타: 측정, 처리, 프라이버시, 연속성

○ 가장 큰 과제중 하나는 빅데이터로부터의 중립적인 추정

- 표본조사와는 달리 빅데이터로부터 생성된 것은 확률표본이 아님. 그래서 빅데이터는 목표모집단의 선택적인부분만을 포함. 이러한 놓친 부분을 설명하는데 부가정보 자료는 수정하는데 이용될 수 있음. 부가적 변수는 행정자료로 연결될 수도 있음.

- 그러나 종종 빅데이터의 단위는 행정자료의 단위와 연계되기 어려움

- 그래서 프로파일링(빅데이터 그 자체로부터 부가정보 도출) 방법 제시

  · 예) 트위터를 활용하여 트위터계정, 이름, 생물학적 정보, 프로필사진 등을 활용하여 성별을 도출 / 교육시작시기, 졸업연도, 교육정도, 현재연도 등을 활용하여 나이 추정

○ (결론) 부가정보는 평가하고 정확한 선별작업을 할 필요, 행정자료를 활용하여 연계하거나 그 자체를 활용하여 부가정보 도출

❖ **The Alternative Data Solution – Experience from Statistics Canada's** Producer Prices Division (Gaétan Garneau and Mary Beth Garneau, Statistics Canada)

○ (배경) 지난 10년간 캐나다통계청은 제품 및 건설생산자가격지수의 품질제고를 위해 노력하면서 서비스생산가격지수로 확장. 생산자가격지수의 대부분은 전통적인 조사방법에 기반을 두고 있는 반면, 응답자의 부담을 줄이기 위해 행정자료 및 대안적 자료 사용을 증가하기 위한 노력을 꾸준히 해옴.

○ 대안적 자료로는 웹상의 가격리스트, 다른 자료수집기관으로부터의 통계, 제3자 데이터파일, 행정자료, 비가격조사의 마이크로데이터와 같은 다양한 형태로부터 수집

- (장점) 응답자부담 및 비용절감, 풍부한 데이터 정보(더 자세한 정

보, 모든 거래 수집, 높은 빈도(특정 날이 아닌 하루 종일 자료 수집) 활용 가능

・예) 주택건설 및 철거허가 데이터; 새로운 주택가격지수를 위한 샘플 프레임으로 사용 가능

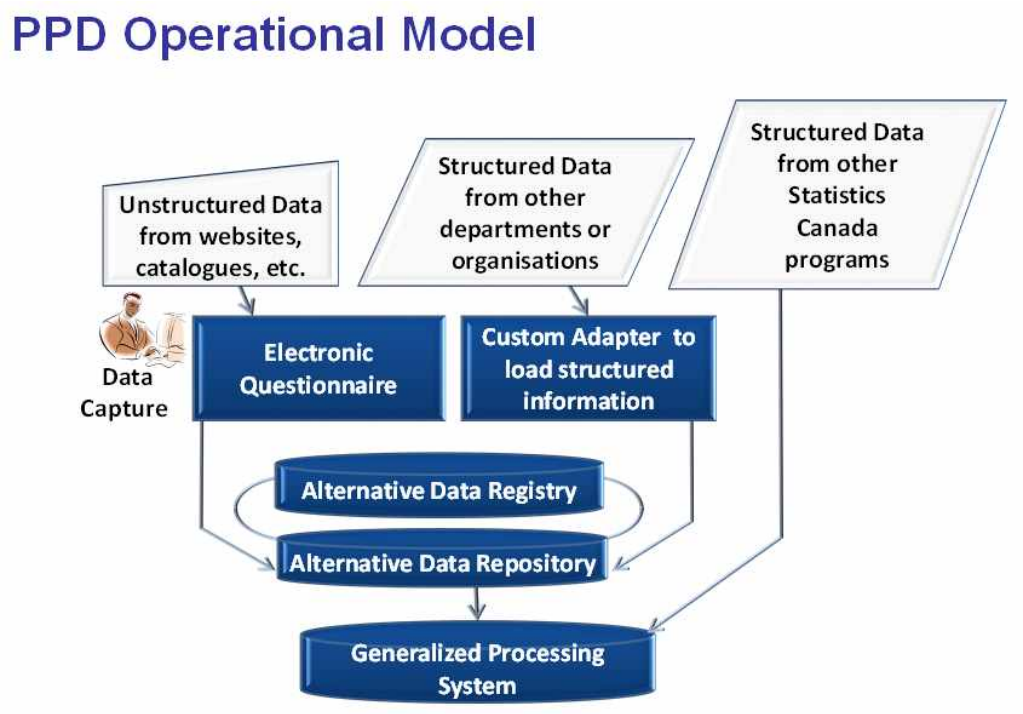
- (단점 및 과제) 데이터품질, 증명프로세스, 데이터 이용가능성, 재정적 고려사항, 데이터조작의 위험, 시간에 따른 변화요소 등

・예) 인터넷상의 가격리스트 사용에 있어서 가격목록과 실제거래가격과의 차이, 웹사이트의 잦은 변경 가능

○ 대안적 데이터 평가; 데이터출처의 신뢰성, 데이터의 품질 및 유용성, 메타데이터의 존재 및 접근성

○ 대안적 데이터 인터페이스

- 생산자가격부서는 현재 약 225개의 다른 자료원으로부터 대안적 데이터를 사용



- 시스템 필요조건; 다른 자료원을 현재 시스템과 프로세스에 통합, 다른 형식과 변수를 적합하게 하는 유연성, 중복방지를 위한 다른 부서와의 조정, 정해진 시간에 데이터수집을 놓치지 않게 하기 위한 데이터 수집전략, 적시에 갱신될 수 있도록 정기적인 모니터링 등

- 대안적 데이터 인터페이스를 위한 솔루션;
  - 대안적 데이터 등기소: 캐나다 통계청의 행정자료 저장소를 포함하여 데이터제공자·수집정보·메타데이터 등 완벽한 정보를 포함하는 중요한 모듈
  - 저장소: 전자파일을 전송시켜 프로세스 시스템에 적재시킬 수 있도록 해주는 특화된 솔루션
  - 캡쳐도구: 다양한 자료원(출판, 인터넷, 카탈로그, 비정형파일 등)으로부터 대안적 데이터를 캡쳐, 수정이 가능하게 하는 도구

주제3 Total Survey Error

❖ **Using Administrative Records to Evaluate Survey Data**
(Mary H. Mulry, Elizabeth M. Nichols and Jennifer Hunter Childs, US Census Bureau, USA)

○ (배경) 행정자료는 조사 응답의 오류를 평가하는데 사용할 수 있는 데이터 원천을 제공.

- 이러한 평가프로세스는 조사데이터를 이용하여 데이터를 수집 및 추정설계에 도움이 될 가능성이 있음.

- 뿐만 아니라 조사자료와 행정자료 혼합이나 조사자료에서 행정자료원으로 전환에 필요한 추정방법설계를 위한 이해를 제공

○ (사례 연구) 그러나 행정자료를 활용하여 조사자료를 평가하는 것은 항상 생각만큼 쉬운 것은 아님.

- 주거지 이동데이터, 센서스자료, 행정자료의 이사날짜 등을 유사항목의 서로 다른 자료원을 비교 검토. 주소와 사람을 연계하는 것이 관건. 뿐만 아니라 사건 매칭과 업데이트 시점 간 차이로 인해 연계의 어려움

○ (결론) 행정자료를 활용하여 조사자료의 오류를 검토하고 평가할 때는 데이터 자체에 대한 오류검토가 우선되어야하며 모든 데이터

에 대한 항목 정의 및 한계점을 사전에 검토할 필요가 있음

❖ **Big Data: A Survey Research Perspective**
(Reg Baker, Marketing Research Institute International, USA)

○ (개요) 빅데이터는 다른 사람들에게 다른 의미를 부여할 수 있는 용어 중 하나. 흔히들 말하는 빅데이터의 특성은 용량(Volume)이 크고, 속도(velocity)가 빠르며, 비정형화된 시각화(visual), 다양성(variety)의 특성을 가진 데이터를 의미.

○ (의미) 대상에 따른 빅데이터의 의미

- 어떤 사람들에게는 단순히 기존데이터의 장점을 활용하여 새로운 것을 생산할 목적으로 그것을 융합하는 것을 의미할 수도 있고,

- 또 다른 사람들에게 빅데이터는 너무 큰 데이터로서 전통적인 프로세스와 분석시스템이 더 이상 적합하지 않을 수 있음을 의미

- 빅데이터는 다음의 주요한 자료원으로부터 데이터를 통합하는 것과 관련: 1) 회사나 기관과의 거래 시 발생하는 데이터 2) 소셜미디어의 대부분 비구조화된 데이터 3) 인터넷을 포함한 정보를 측정하고 전송할 수 있는 핸드폰, 기기장치, 자동차, 교통스캐너 등과 같은 상호 연계된 기기사용

○ (주요 특징) 빅데이터의 주요 특성에 대해 설명: 1) 빅데이터는 비정형의 특성, 2) 주요 과제- 데이터 품질 및 분석적 어려움, 3) 조사 미래에 빅데이터가 미칠 잠재적 영향

- 빅데이터는 기존조사를 위협할 수 있음. 따라서 스캐너 데이터, 우버(uber) 데이터, 소셜미디어 등의 데이터 연계를 통해 그 가치 제고할 필요

- 프로세스: 다양한 자료원으로부터 필요한 정보를 추출하고 변형 · 정제 · 재코딩 · 연계 · 통합하여 자료저장소에 적재하여 활용

○ 빅데이터 활용을 위해 적극적이고 지속적인 관리가 필요

주제4 Open Data

❖ An International Overview of Open Data Experiences
(Timothy Herzog, World Bank, USA)

○ (개요) 공공데이터(open data) 정책은 정부와 기타 공공기관이 상호작용과 서비스를 제공하는 방법의 변환을 의미.

- 공공데이터는 어떤 목적이든 누구에게나 제한 없이 무료로 재사용, 재배포 가능

- 시민들에게 투명성과 가치를 재고시키고, 정보제공의 비효율성과 장벽 제거, 공공서비스 전달 향상을 위한 데이터기반 프로그램 사용, 혁신적인 사업기회를 고무시킬 수 있는 공공데이터 제공

○ (발전 개요) 현재와 미래 경험, 기회와 과제와 함께 국제적 수준에서의 공공데이터의 전반적인 발전 개요 소개

- 국가통계기관의 목적은 고품질 통계제공, 신뢰할만한 승인통계 제공, 통계사용 확대인 반면, 공공데이터의 목적은 향상된 공공서비스 제공

- (기회) 통계사용증가, 비용절감, 품질향상, 명성 증가

- (공공데이터에서의 국가통계기관의 역할) 수요 증대에 맞는 고가치의 통계 생산, 기술적 전문성 제공, 기준설정 및 관리·감독, 교육훈련제공, 국제사회 동참

- 정부는 개방성, 투명성, 책무성, 공공서비스 향상, 혁신 및 경제적 가치 제고, 공공부문 효율화, 정부정책 지원 등을 위해 공공데이터 지향

- 성공을 위한 핵심교훈: 작은 규모에서 시작하여 학습을 통해 성장, 공공과 시민사회의 적극적 참여/ 이와 더불어 데이터 혁명관점에서 파트너쉽 형성, 데이터갭 줄이기, 데이터 능력강화,

표준화 설정, 데이터의 접근성 등이 필요

❖ **Open data at Statistics Canada**
(Bill Joyce, Statistics Canada)

○ 공공데이터(open data)라는 용어는 비교적 새로운 것이라 할지라도 그 개념은 오랜 역사를 가지고 있고, 데이터 해방운동에 그 뿌리를 두고 있음.

○ 캐나다 통계청의 공공데이터의 원칙 발전과 현재 공공데이터전략을 소개

- 단순히 공짜라는 의미 보다는 개방된 데이터로서의 의미

- 접근성 증대를 목표로 웹을 통해 다년간 발전해옴

- Sir Tim Berners-Lee의 5스타 랭킹* 중 캐나다통계청의 공공데이터는 3스타에 랭크

* Sir Tim Berners-Lee의 5스타 랭킹시스템:
☆: 허가를 가진 형식에 관계없는 웹상의 모든 것
☆☆: 기계가 읽어들일 수 있는 형식(예-엑셀)
☆☆☆: 비전매특허의 기계가 읽어들일 수 있는 형식(예- CSV, XML)
☆☆☆☆: 의미 있는 웹에 대해 연결될 수 있도록 적합하게 구성된 형식(예- RDF, SPARQL)
☆☆☆☆☆: 데이터셋과 상호분석 가능하도록 실제로 연결된 상태

- 2012년 2월부터 웹상의 모든 표준 집계 데이터는 무료, 공공데이터허가 채택, 심지어 고객맞춤데이터에 대해서도 허가나 사용료를 제거, 비전매특허의 기계가 읽어들일 수 있는 형식으로 전환

○ 자료기여자로서의 캐나다 통계청
- open.canada.ca를 통해 인구주택총조사, 농림어업총조사, 국제무역, 요약표, 산업 및 직업분류, 지역참고자료 등 제공

- 마이크로데이터접근을 확대시키는 것이 목적

- 앞으로 공공데이터베이스 중심 모델을 통한 데이터 공표, 더 일관성 있고 표준화된 웹 데이터 이용 및 서비스 도입, 캐나다 통계 데이터의 공공데이터 저장소 마련

○ 서비스제공자로서의 캐나다 통계청

- 국가 재정위원회와 업무협약을 채결하고 차세대 공공데이터 포털(open.canada.ca)을 개발

- 2013년 봄에 시작, 공공정부플랫폼(OGPL)을 사용, 이용성과 접근성을 향상, 모바일 기기 지원, 향상된 검색기능, 통합전산시스템(Shared Services Canada)이 제공하는 확실하고 유연한 인프라 활용

- 공공데이터에서 공공정부로의 전환

주제5 Waksberg Award Winner Address

❖ **Towards a Quality Framework for Blends of Designed and Organic Data**
(Dr. Robert Groves, Georgetown University, USA)

○ (배경) 확률표본은 가구나 인구 대상에 대해 거의 보편적인 틀, 표준화된 측정수단이며 다변량 데이터를 얻어 통계적 절차로 분석되어온 75년 동안 경험적 사회과학의 초석이 되어왔음.

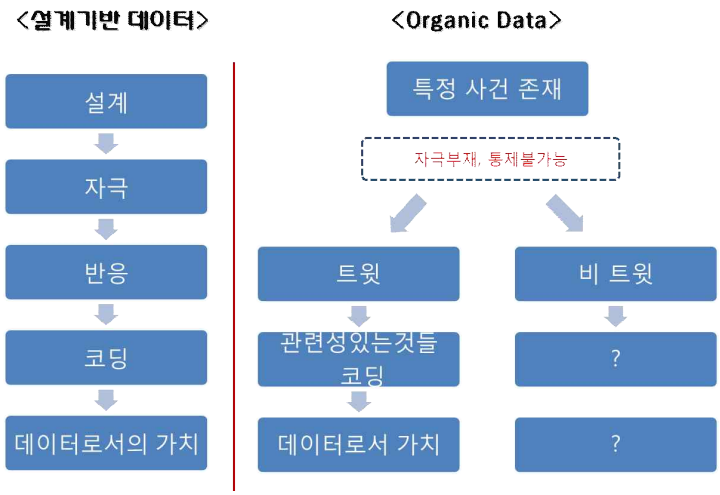
- 그러한 측정구조는 선진국에 사회와 경제에 대해 알고자하는 거의 모든 지식을 제공하고 있음.

○ (데이터 환경 변화) 그러나 우리는 현재 통계기관이나 사회과학이 주도했던 것과는 다른 데이터 세상에 살고 있음.

- 고차원의 데이터가 인터넷검색활동, 모바일, 소셜미디어, 센서, 소매점 스캐너 및 기타 기기 등을 통해 도처에서 생산. 일부는 이러한 데이터원들은 매년 40% 비율로 크기가 증가할 것이라고 예상. 그 규모와 함께 확률기반표본조사는 궁지에 빠지게 됨.

○ organic data: 거래데이터(스캐너, 카드사용, 의료서비스 등), 소셜네트워크 데이터, 모바일(GPS), 인터넷검색데이터 등

- 장점: 지금까지 없던 새로운 행동 측정 가능, 실시간 데이터, 공간데이터 확보 가능, 네트워크 측정 가능
- 단점: 모든 인구가 포함되지 못함, 자료제공자 식별 어려움, 데이터구조가 분석하기 어려움, 데이터 접근과 내용이 데이터 소유자에 의해 통제될 수 있음
- 특징: measurement stimulus 부재 및 통제 불가능



- 이러한 데이터에도 적용할 수 있는 통계품질에 대한 개념 진화 및 대안적 용어 필요

○ 게다가 선진국에서의 표본조사 상태는 그리 왕성하지 못함. 정보에 대한 수요는 증가함에도 불구하고 조사참여률 하락과 예산압박으로 인해 공식통계기관은 어려움에 있음.

○ (결론) 이러한 상황은 사회 및 경제과학 추론의 기본적인 패러다임에 전례 없는 위협이 되고 있음. 이런 상황에서의 대안적인 방향 제시

- 기존의 표본조사 패러다임 하에서는 표본조사를 통해 모집단 추정 가능, 그러나 빅데이터 패러다임에서는 산재해있는

organic data(페이스북, 트위터 등)를 통해 모집단을 추정하기 어려움

- 손실데이터에 대해서는 다른 자료원을 혼합하여 추정
- 결국 미래는 조사 자료와 organic data의 혼합을 수반

주제6 Confidentiality

❖ **Enhancing data sharing via "safe designs"**
(Kristine Witkowski, University of Michigan, USA)

○ (배경) 데이터수집의 사회적 가치는 연구파일의 광범위한 확산과 과학적 생산성 증가에 의해 상당히 향상됨. 현재 대부분의 연구는 어떻게 공유될 것인가에 대한 사전 고려 없이 분석이용이하고 정확한 정보수집 중심으로 설계됨.

- 문헌이나 실무에서도 노출분석은 자료수집 이후에 일어날 것이라고 가정

○ (주요 내용) 그러나 상당수의 일반사용자들을 위한 분석적으로 유용한 공공사용데이터를 생산하기 위해서 노출위험은 연구과정 초반에 고려되어야만함.

○ 이론적 틀 및 조사방법연구와 같은 경제적이고 통계적인 의사결정을 도출하는데 노출위험이 “안전설계(safe design)”와 “노출시뮬레이션(disclosure simulation”)의 형태로 연구 초반에 다루어질 수 있는 방법 설명. 적용된 통계적 접근은 다음과 같은 경우에 받아들여질 수 있음

- 1) 다른 표본설계하에서 조사데이터의 구성을 예측하는 모델을 개발하거나 검증, 2) 노출위험, 분석적 유용성, 표본평가 및 데이터베이스 설계에 가장 적합한 노출조사비용 평가에 사용되는

측정방법을 선택하거나 개발, 3) 위험, 유용성, 광범위한 표본 및 데이터베이스 설계 특성 연구비용에 대한 추정을 위해 시뮬레이션을 수행

❖ **Privacy and Security Aspects Related to the Use of Big Data – progress of work in the European Statistical System (ESS)**
(Pascal Jacques, EUROSTAT, Luxembourg)

○ (개요) 데이터 보호 및 개인정보보호는 공식통계생산에 있어 빅데이터 사용을 가능하게하기 위해 우선적으로 해결해야할 과제임. 이것은 2013년 유럽통계시스템위원회(ESSC; European Statistical System Committee)의 국가통계기관장들에 의해 강조된바 있음.

○ 2014년 Riga회의에서 유럽통계시스템위원회는 유럽통계청 빅데이터 TF에 의해 제시된 Big Data Action Plan and Roadmap 1.0(BDAR)을 승인하고 이를 ESS Vision 2020에 통합하는데 합의

- 또한 유럽통계청은 UNECE와 같은 외부 파트너들과도 이 분야에서 협력하기로 함

○ (UNECE 2014 빅데이터 프로젝트) 통계생산의 현대화에 있어서 빅데이터의 역할에 관한 프로젝트

- 공식통계와 관련된 빅데이터의 다른 측면을 강조한 4개 'task team'으로 구성; 사생활보호, 파트너쉽, 샌드박스, 품질

- 사생활보호팀은 2014년 작업을 끝내고 사생활관련 위험관리를 위한 기존의 툴에 대한 개요, 빅데이터 특성관련 식별위험, 국가통계기관 대상 권고초안을 제공. 빅데이터 사용과 관련된 사생활 위험을 처리하기 위해 필요한 새로운 기술사용을 포함한 기존의 틀 확장

○ BDAR: 많은 빅데이터 자료원은 공식통계를 위해 사용될 때 일반국민이나 기타 이해관계자들에 부정적인 인식을 가져올 수

있는 민감한 정보를 포함하고 있고, 이러한 위험은 중·단기 내에 최소화되어야만 한다는 것을 인식

- 그래서 국가통계기관의 역할과 활동을 관리하는 윤리적 원칙에 대한 적합한 검토 및 강력한 전달전략과 같은 복합적인 조치에 착수할 것을 제안

주제7 Application of Record Linkage and Statistical Matching

❖ **Linking 2006 Census data to the 2011 mortality file**
(Mohan Kumar and Rose Evra, Statistics Canada)

○ (사례연구) 2006~2011사망자데이터와 2006센서스 데이터를 연계하는 작업은 기대수명과 이민자 및 토착민과 같은 관심 있는 인구에 대한 사망률 계산과 같은 분석을 위해 수행됨. 연계데이터는 달리 이용할 수 없는 주요한 정보를 제공

- 이 프로젝트는 1991센서스 데이터와 사망파일과의 데이터연계 후속작업. 그 당시에는 이름과 토착민 신원변수는 이용가능하지 않았음. 새로운 연계는 이러한 정보를 포함하고 가장 최신의 인구인 2006년도에 캐나다에 살고 있는 최신의 인구로 처리됨.

- 데이터 연계는 변수의 전처리와 연관블럭생성을 포함한 다각적 단계로 수행됨. 결정적 계층방법(ID변수 연계를 통해)을 사용하여 수행. MixMatch연계 소프트웨어 사용

- 이 연계분석을 통해 사인분석, 사회적 변수의 영향, 기대수명 산출 등 분석에 활용

- 품질평가는 자동검증과 수동검증(각 층으로부터 약 10%의 랜덤 샘플을 추출하여 검증)

❖ **An Overview of Business Record Linkage at Statistics Canada: How to link the "unlinkable"**

(Javier Oyarzun and Laura Wile, Statistics Canada)

○ (개요) 사업체 레코드 연계는 현재 사업체부문 이슈의 통계적 데이터 개발, 생산, 평가 및 분석에 활용되는 중요한 기술. 그러나 데이터 연계는 누군가의 사생활을 침해할 수 있기 때문에 캐나다 통계청은 공익성이 분명하고 침해를 능가할 때만 이용

○ (과제) 사업체연계에서의 과제

- 대다수의 행정자료의 공통의 식별인자 부재
- 비표준화된 형식으로 기록된 정보(길이, 양식)
- 부적합한 설립당시의 이름과 주소
- 인쇄 및 입력오류를 포함한 정보
- 대용량의 행정자료 파일
- 사업체등록부에 다른 번호를 가진 동일사업체 존재
- 사업체번호를 포함하지 않는 사업체등록부와 연계할 필요 (예-농장등록부)

○ (연계표준화) 이러한 레코드연계와 관련된 중요성과 해결과제 때문에 캐나다 통계청은 사업체레코드연계프로세스를 최적화하는 것을 돕기 위해 레코드연계표준화를 개발해왔음.

- 4단계: 1) 표준화: 임퓨테이션을 통해 각 변수를 표준화 2) 결정적 연계를 통해 매칭 3) 유사스코어를 부여하여 데이터 정리 4) 최종 스코어 생성(이름, 주소, 행정데이터 연계를 통해 합계 산출)

○ (향후 과제) 앞으로 캐나다 통계청은 연계품질평가 전략을 수립해 나갈 예정

주제8 Use of Administrative Data

❖ **Creating a longitudinal database based on linked administrative registers: An example**

(Philippe Wanner, Université de Genève and NCCR On The Move, Switzerland and Ilka Steiner, Université de Genève, Switzerland)

○ (연구배경) 지난 몇 십년간 국제적 이동은 많은 산업국가의 인구성장의 중요한 역할을 해왔으며 이주자들의 경제 및 사회적 통합에 대해 불꽃 튀는 많은 논쟁이 있어왔음. 그 결과 이주를 문서화하는데 새로운 데이터가 필수적

○ (사례 소개) 이러한 상황에서 거주자등록부, 재외국민명부, 전통적인 센서스를 대체한 구조적조사(Structural Survey)와 기타 보험등록부 등을 기반으로 결정적 연계와 확률적 연계방법을 활용하여 스위스 인구데이터베이스를 생성

- 목적: 연방통계청과 함께 이주에 대한 연구과제의 일부로서 도착에서 그 기간 동안 직업적, 경제적, 인구학적 상태 변화를 확인하면서 출발까지 재외국민을 종단면으로 추적하기 위함
- 현재 1998년과 2013년 사이에 스위스에 살았던 약 4백만 재외국민에 대한 15년간의 추적이 이용가능.
- 다양한 등록부를 연계하여 사회경제적 지위, 주거지의 이동, 귀화관련 조치 등에 대한 정보 제공
- 장점: 수집비용절감, 종단면적 접근 가능, 이민에 대한 새로운 연구 도입, 행정기관과 대학간 협력 가능
- 한계 및 어려움: PIN번호 변경으로 인한 등록부의 품질, 서로 다른 여러 행정자료의 조화

❖ **Use of admin data to increase the efficiency of the sample design of the new National Travel Survey**
(Charles Choi, Statistics Canada)

○ (사례소개) 관광통계프로그램의 재설계의 일환으로서 캐나다통계청은 캐나다인여행자로부터 정보 수집을 위해 국내여행통계(NTS; National Travel Survey)를 개발. 2018년 실시예정

- 이 새로운 조사는 캐나다 거주자여행조사(the Travel Survey of Residents of Canada)와 국제여행조사의 캐나다 거주자부문(the Canadian Residents Component of the International Travel Survey)을 대신할 것임.

- 기존조사의 한계점: TSRC의 경우에는 LFS의 보조조사로서 의무가 아니라 우선순위에서 뒤쳐짐, ITS는 표본틀이 없고 낮은 응답률을 보임

- NTS는 행정자료사용을 최대화하면서 캐나다통계청의 일반적인 표본틀과 처리도구를 이용할 것임.

- 행정자료 최대한 활용: 표본설계, 임퓨테이션, 데이터 보정 및 추정분야에서

- 향후에는 다음과 같은 행정자료에 대해 잠재적 사용예정: 여권데이터, Canadian Border Service Agency/ Canada Revenue Agency 파일을 이용하여 표본설계 개선에, Credit Card data를 통해 국제여행자를 식별하고 임퓨테이션에 활용 예정.

주제9 Scanner Data

❖ **Challenges Associated with Using Scanner Data for the Consumer Price Index**
(Catherine Deshaies-Moreault and Nelson Émond, Statistics Canada)

○ 실제 대부분의 소매상들은 고객과의 거래의 세부적인 내용을 기록하기 위해 스캐너를 사용. 일반적으로 상품코드, 간략한 설명, 가격, 판매량 등을 포함

○ 이것은 캐나다의 가장 중요한 경제적 지표 중에 하나인 소비자 물가지수(CPI)와 같은 통계의 자료원과 굉장히 관련이 있음

○ 스캐너 데이터 이용은 자료수집비용을 낮추면서도 지역적 범위를 확대시키고 판매수량도 포함하면서 계산에서 사용되는 늘어나는 가격 수에 의한 CPI의 품질을 개선시킬 수 있음

○ 그러나, 이러한 데이터 사용에 많은 (과제)가 존재

- 데이터 입수: 수집권한과 전송문제
- 대용량 파일, 편향된 분포(예- 아울렛), 큰 매출액 처리 부분
- 분류체계: 제품이 아닌 상품으로 분류, 기존 보다 더 세분화된 분류체계
- 1년에 걸쳐 제품식별코드의 높은 변화율
- CPI사용을 위한 품질 평가 부재
- 시스템 변화를 위한 높은 초기 비용
- 장기적으로는 대표성과 범위를 높일 필요

❖ **Product homogeneity and weighting when using scanner data for price index calculation**

(Antonio G. Chessa, Statistics Netherlands, Netherlands)

○ 스캐너데이터에서 국제상품식별코드(GTIN; Global Trade Item Number)의 판매가격과 수량에 대한 이용은 더 적확한 가격지수 수집가능성을 제공.

- 그러나 다음과 같은 많은 비판적인 질문들에 대해 결국 대답해질 필요가 있음: 무엇이 개인소비자 상품이고, 어떻게 물가지수 작성을 위해 가중치가 적용될 것인가?

○ 현재 활용 구성비: 조사자료(78.9%)와 스캐너데이터(21.1%)

○ 장점: 무응답증가에 대한 대안 가능, 소비패턴변화 반영 가능, 직접자료수집비용증가의 대안 가능, 풍부한 데이터 획득 가능

○ 문제점

- “relaunch”: 품목의 소비가능 부분은 여전히 동일하지만 바코드의 변화
- 바코드자체는 제품이 아님.
- 바코드의 적합성
- 정확성에 대한 검증 부재, 데이터 품질에 대한 보증 미흡
- 소매업자마다 가능한 할인처리 방안 부재
- 관련 가격 인상을 포착할 수 있도록 GTINs이 결합된 폭넓은 제품 정의 필요. 이를 위해 같은 제품 특성을 가지는 GTINs를 묶어 GTIN 그룹(제품(“product”)) 생성. 특성은 통계적 모델선택방법(정보 기준) 선택(제품의 특성을 추출하기 위해 텍스트마이닝 등을 활용한 데이터 정교화 작업 필요)

○ 제품을 정의할 때 품목특성의 다른 선택에 대한 가격지수의 민감성 예시를 제시. 동질제품으로서 GTINs는 의류와 의약품류에 적합하지 않음을 보여줌.

○ 또한 가격지수는 가중치 유형에도 민감. 통계적 관점에서 판매액이나 판매량에서의 비중으로 가중치를 부여한 것은 분명이 동일가중치보다는 우월. 스캐너데이터를 통해 계산된 가격지수는 기본적인 집계를 위한 전통적인 Jevons 지수 사용은 버려야만 한다는 것을 보여줌.

○ 보다 일반적인 방법을 위해

- 제품 차별화 및 구분
- 데이터 처리방법 개선
- 지수계산: 웹데이터에 의해 수집된 가격을 사용할 때는 추가적인 자료원으로부터 가중치가 필요

주제10 Plenary Session Sessions

❖ **Data Science for Dynamic Data Systems: Implications for Official Statistics**
(Mary E. Thompson, University of Waterloo, Canada)

○ (배경) 현대 데이터사이언스의 동적인 모습: 변화하는 인구, 행정자료양의 증가, 개인과 기업체의 거래자료, 지속적인 데이터의 공급, 실시간으로 준석하고 요약하는 능력

- 센서스 설계가 변화를 겪고 있음: 주소데이터베이스의 발전, 행정자료 및 거래데이터베이스 개발 및 이용
- 데이터 사이클의 속도 증가: 보다 빠른 데이터 수집과 처리, 경제 및 사회통계의 더 빠른 이용
- 틀(frame)은 더 커지고, 더 풍부해지고 끊임없이 업데이트됨
- 자료연계를 통해 데이터베이스는 더 풍부해지고, 복잡해지고, 더 확장됨
- 표본설계과정의 활발한 변화: 적응적 표본설계, 반응적 데이터 수집 및 조사과정데이터를 이용한 표본설계, 복합모드 및 복합프레임 사용, 분석에서도 특정효과 통합
- 부가데이터의 새로운 자료원 출현: 센서데이터, 물가지수작성에서의 스캐너데이터, 온라인구인광고, 신용카드 데이터베이스, 소셜미디어 등

○ (과제) 공식통계생산측면에서의 당면 과제

- 모델의 복잡성: 빅데이터용 적합 모델은 복잡함. Cox(Biometrika, 2015); 빅데이터는 복잡한 구조로 표준통계절차 적용시 오히려 오도될 수 있음. 모델은 평균과 회귀계수 추정에 관해 자료원의 변화성에 대한 복잡한 패턴의 영향을 설

명할 수 있어야 함.

- 변화하는 샘플에 대한 매커니즘: 순환데이터(rolling data)는 변화순간과 순환추론기술이 필요

○ (공식통계의 고려 사항) 이러한 맥락에서 공식통계의 영역에서 인구프레임 유지의 중요성이 강조되어야 함; 여러 프레임의 복합사용, 모집단범위추정에 연계; 추정과 모델보정을 위한 부가정보로서 대규모의 비조사자료 활용; 데이터 처리시기에 연구자・서비스제공자・정책입안자들에 대한 이슈, 통합과 배포; 고차원 시계열자료 예측을 위한 반복적 방법과 규칙화; 변화포착에 있어서 정교한 데이터 시각화도구의 유용성과 한계점

○ (앞으로) 현재는 새로운 데이터자료원의 힘을 이용하여 일대약진을 이루고 있으며 데이터 수집 및 관리방법은 급격히 변화하고 있음. 앞으로 분석방법의 발전은 새로운 일과 통계 및 컴퓨터과학의 기타분야사이의 더 많은 연결 포인트를 선도

Ⅲ 시사점

□ 빅데이터라는 변화된 현실을 인식하고, 국가통계생산자들이 직면한 빠른 속도로 변화하는 다양한 형태의 자료원으로 대표되는 변화된 데이터 환경을 이해하고, 이를 적극 활용하고자하는 다양한 연구 및 사례들을 살펴볼 수 있었음

- 행정자료를 포함한 빅데이터라 불리는 다양한 자료들을 활용하여 단순히 직접 통계생산 뿐만이 아니라 일련의 통계생산 과정 전반에 활용. 특히나 표본설계, 조사자료의 오류검증, 여러 자료를 복합적으로 사용한 자료연계, 무응답 대체 등에 활용

□ 빅데이터를 일종의 자료원으로 인식하고 기존조사의 품질 개선에 활용

- 행정자료 등 다양한 자료원을 활용 및 연계하여 기존 조사의 품질개선을 위한 도구로 활용하고 있었음.

- 그래서 가까운 미래는 기존의 전통적인 조사뿐만 아니라 행정자료를 포함한 빅데이터가 활용된 혼합 형태를 수반할 것이라는 의견이 많았음.

- 더불어 빅데이터를 활용하기 위한 모델기반 추정방법, 연계방법, 모수추정을 위한 가중치 활용 등이 조사 품질개선의 다양한 부분에 활용되고 있었음.

□ 빅데이터는 피할 수 없는 현실이며 다양한 수요자의 요구에 응할 수 있는 각종 통계생산의 한 자료원으로서 조사과정의 다양한 측면에서의 행정자료 활용방안 모색이 필요

□ 빅데이터의 다양한 활용을 위해 행정자료 등의 품질개선을 위한 노력도 빅데이터 활용만큼이나 필요하고 중요한 일