eurostat 조사방법론 컨퍼런스 참가 결과

New Techniques and Technologies for Statistics(NTTS) 2011

2011. 5.

통계개발원
STATISTICAL RESEARCH INSTITUTE

Ⅰ 개 요

1. 출장개요

○ 참가회의 : Eurostat 조사방법론 컨퍼런스(NTTS* 2011)

*New Techniques and Technologies for Statistics

- 회의기간 : ‘11. 2. 22.(화) - 24.(목) / 3일간
- 회의장소 : 벨기에 브뤼셀 샤를마뉴 빌딩

○ 출 장 자 : 조사연구실 안다영 주무관

2. 출장목적

○ 최근 들어 서베이를 통한 통계자료 수집에 있어 응답률을 최대화하고 무응답 오차를 최소화하는 데이터 품질에 관한 논의가 확장되고 있는 상황에서 통계작성기관의 연구동향을 파악할 필요가 있음

○ 선진통계청의 새로운 자료수집방법 및 신규작성통계 현황을 파악하고 에디팅, 임퓨테이션 등의 기법과 연동표본의 활용 현황 등을 참고함

○ ‘10년 조사데이터 품질관련 과제의 후속 연구로 ‘11년 상반기 「지역별고용조사에서 무응답 가중치 조정기법 연구」과제를 수행함에 있어 선진연구동향을 전달하고 과제의 핵심이 되는 무응답추정, 가중치조정, 표본설계 등에 신규통계기법 및 각국 통계청의 방법론 활용사례를 참고하여 조사방법론 연구의 질적 향상을 도모하고자 함

Ⅱ 주 요 내 용

1. 회의내역

❒ NTTS(New Techniques and Technologies for Statistics) 소개

○ 유럽연합통계청(Eurostat)에서 주관

○ 일반 및 공식통계의 방법론·기법에 관한 최신연구동향을 교류하고 통계의 장기적인 요구와 발전방향을 논의하기 위한 컨퍼런스

○ 1992년 최초 개최 이후 2-3년마다 비정기적으로 개최됨

2. 회의세션

○ 다음의 21개 세션 하에서 EU국가를 중심으로 한 각국 통계청 업무담당자 및 대학교수 등 전문가들의 논문 발표

Day1	session 1 :	Data and Text Mining
	session 2 :	Data Editing, Imputation
	session 3 :	New Survey Types
	session 4 :	Standardisation
	session 5 :	Statistical Disclosure Control
	session 6 :	Geo-Information for Statistical Analysis
Day2	session 7 :	Aggregation, Estimation, Calibration
	session 8 :	Integration of Multiple Data Sources(1)
	session 9 :	RISQ* project (*Representative Indicators for Survey Quality)
	session 10 :	Remote Access
	session 11 :	Integration of Multiple Data Sources(2)
	session 12 :	BLUE-ETS* project (*BLUE-Enterprise and Trade Statistics)
	session 13 :	Use of Registers
	session 14 :	AMELI* project (*Advanced Methodology for European Laeken Indicators)
	session 15 :	Small Area Estimation
Day3	session 16 :	Well-being Indicators
	session 17 :	Methods to Synthesize and Extract Knowledge
	session 18 :	SAMPLE* project (*Small Area Methods for Poverty and Living Condition Estimatest)
	session 19 :	Electronic Data Collection
	session 20 :	POINT* project (*Policy Influence of Indicators)
	session 21 :	ESS*net project (*European Statistical System)

3. 발표내용

개회사 Research directions in Official Statistics : First results of a survey on Research needs ✍ Daniel Defays, Eurostat

○ 오늘날의 자료는 점차 집약적 정보를 갖게 되고 자료의 품질 및 환경에 의한 정보변화 등의 문제가 대두되고 있는 상황

○ 따라서 현재까지 활발하게 활용되고 있는 기존의 조사 방법론(대용량데이터처리, 행정자료연계, 메타데이터, 측정오차, 새로운 자료수집 방법, 비밀보호기법, 베이지안 추론 등)에 대한 리뷰 및 관련 기술의 변화(클라우드 컴퓨팅, 웹3.0기반, 모바일 애플리케이션, 가상화, 무료공개 프로그램 등)에 대한 논의가 중요한 시점을 강조

기조연설 EU Funded Research on Statistics and Indicators
✍ Dr. Ian Perry, DG R&I(Research and Innovation)

○ EU국가들은 통계를 기반으로 생산된 지표를 활용하여 득책을 수립하고 있으므로 정책수립을 위한 지표의 활용문제, 새로운 지표개발, 공식통계 방법론 연구 등에 관한 방법론적 연구를 탄탄히 할 필요가 있음

○ FP7 프로젝트[1]를 넘어 보다 다양한 통계 및 지표를 생산하기 위하여 유럽연합의 지속적인 지원이 요구됨. 이와 관련하여 DG R&I에서는 '11년 9월 2일부터 연구 및 혁신을 위한 일반적인 전략 프레임워크와 관련하여 주요 상담을 시작할 예정임을 시사

○ '유럽 2020'에 당부

- 유럽 2020이란 지속가능하고 포괄적인 국가적 성장을 이루기 위해서는 핵심요인인 연구와 혁신에 의존한다는 내용으로, 유럽연합의 미래연구와 혁신자금지원이 기여해야 할 목표와 프레임워크를 결정함
- 이는 유럽연합의 연구자금지원 프로그램이 유럽 2020의 과제에 우선순위를 두기 때문에 우선 과제가 아닌 경우 어떻게 예산 지원을 지원받을 수 있는지 등의 사항을 파악할 필요가 있음을 강조

1) 유럽연합의 가장 큰 연구개발(R&D) 지원장치로, 정책적 우선권 및 기업·사회의 필요에 완벽하게 부응하여 현재 과학적 우수성과 기술개발을 진흥하기 위해 필요한 중요한 장치로 자리매김

기조연설

Official Statistics Production:the past, the present, the future
✍ Carl-Erik Särndal, Statistics Sweden and Örebro University

○ 오늘날 공식통계생산의 현실은 통계학 원칙과 불일치가 존재함

○ 공식통계방법론은 수학적 통계방법론의 이론에서 명백하게 설명되지만 공식통계 및 조사연구의 통합된 이론이 존재하지는 않음

○ 이에 일부 관리자는 공식통계 생산을 몇 가지 과학적 이론을 조합하면 얻어지는 것으로 여김. 즉 통계학자들을 계산을 전담하는 기술자로만 보는 경향이 있음

○ 앞으로 공식통계방법론의 이론적 접근은 통계정보를 생산하는 각 단계마다 방법을 제공하는 것을 목표로 해야함을 강조

주제

New Survey Types

❒ Self-Rotating Sampling Design

✍ Andris Fisenko, Central Statistical Bureau of Latvia

○ 2002년에 라트비아 노동력조사에서 처음 소개됨

○ 센서스 자료에서 지역을 추출하고, 추출된 지역을 8분기 간 패널로 사용하여 가구표본을 추출하게 되는데, 이는 가계자산조사 및 국내여행자조사의 표본으로도 활용됨

○ 연동규칙은 2-(2)-2방식으로 추출된 지역은 8분기동안 샘플로 유지됨

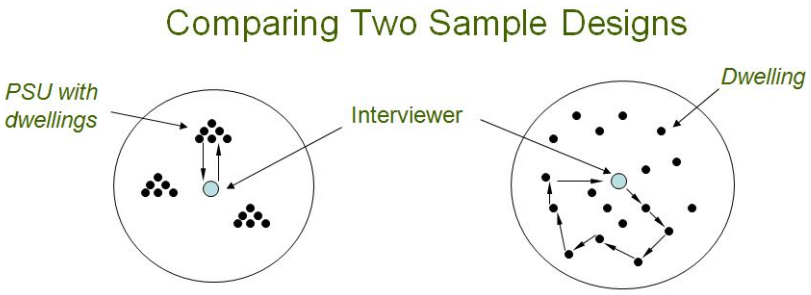
- 단계1. 센서스 자료에서 지역을 추출
- 단계2. 추출된 지역에서 2개의 가구 표본 추출
- 단계3. 2분기 간 조사 - 2분기 간 조사중지 - 2분기 간 조사

<샘플로 추출된 지역에서의 가구 연동 규칙>

1분기	2분기	3분기	4분기	5분기	6분기	7분기	8분기
A	A	B	B	A	A	B	B

○ 노동력조사에서 2단집락추출과 단순임의추출 간 비교

- 2단집락추출 활용 시 단순랜덤추출에 비해 약 9배의 비용감소 효과를 보임



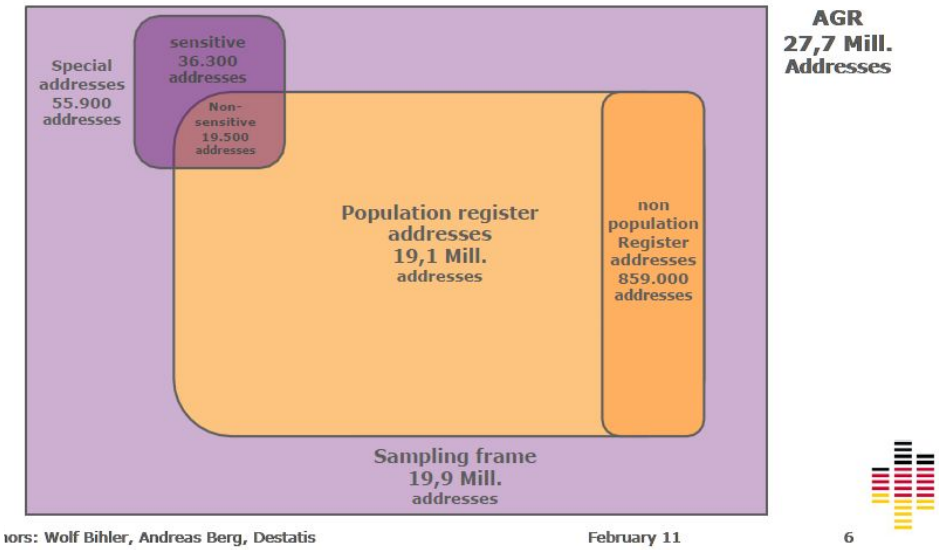
○ 현재 라트비아 통계청의 가정부문 에너지소비조사에 적용하기 위한 논의 중

❒ The household survey in the German Census 2011

✍ Andreas Berg, Wolf Bihler, Destatis

○ 독일의 센서스는 1875년부터 1910년까지 5년마다 실시되었으나 이후 세계전쟁으로 중단됨. 올해 센서스는 독일 통일 이후 첫 센서스임(마지막 1987년 센서스 이후 25년 만)

○ 샘플링 프레임은 2010년 1월 9일 기준 주소 및 건물 등록상태(AGR), 샘플단위는 주소(address)가 됨



○ 품질측정을 위한 사후조사는 모든 층에서 동등하게 5%로 부표본구성

○ 2010년 1월 9일 이후 AGR에서 새로 발견된 주소를 추가로 추출(2011.4)

○ 예비결과 공표시기는 조사종료 후 18개월, 최종결과는 종료 후 24개월

주제 Statistical Disclosure Control

❒ Microaggregation-based Hybrid Data

✍ Josep Domingo-Ferrer, Universitat Rovira i Virgili

○ 하이브리드 데이터는 원자료와 합성자료(synthetic data)를 혼합하여 얻어진 형태로, 비밀보호된 자료와 노출위험이 낮은 합성자료의 장점을 결합하는 노출제어기법에서 제안되었음

○ 제안된 방법으로 micro-aggregation과 합성자료를 결합하여 하이브리드 데이터를 생성할 경우 원자료의 평균벡터와 공분산행렬을 유지하게 됨

○ 하이브리드 데이터셋은 단일매개변수인 정수 k에 따라, 원자료 또는 k가 데이터셋의 레코드수와 같은 경우 완전한 합성자료에 매우 가깝게 얻어짐

○ 일반적인 다변량 micro-aggregation에 비해 제안된 방법은 비밀특성에 대해 분산 및 공분산의 측면에서 더 효율적이며 노출위험을 낮출 수 있는 방법임을 강조

주제 Aggregation, Estimation, Calibration

❒ Multi Stage Indirect Sampling

✍ Hans Kiesl, Regensburg University of Applied Sciences

○ NEPS(National Educational Pannel Study)는 각 나이대별로 구성된 6개의 표본을 갖는 복합 표본설계의 새로운 조사임

○ 이 중 하나는 유치원 또는 보육원에 다니는 2010년 기준 4-5세 어린이 표본으로, 독일에는 표본추출틀로 사용할 유치원 전체리스트가 없는 상황. 따라서 간접적인 표본추출을 통한 대안을 모색(유치원이 의무는 아니지만 약 95%가 유치원에 다닌다는 점에서 착안)

○ NEPS에서 고안한 3-stage 간접샘플링은 불편추정량을 가짐

○ 초등학교 표본을 추출하고, 추출된 초등학교에 입학한 아이들의 출신 유치원들을 정의하는 것으로 유치원 표본을 만듦. 집락추출법은 아님

- 단계1. 200개의 초등학교 선택
- 단계2. 선택된 초등학교에 입학한 출신 유치원을 파악
- 단계3. 단계2의 유치원 중 최종 250개 유치원 표본 선택, 각 20명씩 추출

○ NEPS의 3-stage 간접샘플링에서 무응답 처리는 Lavallée(2007) and Xu and Lavallée (2009)을 참고하고 있음

❒ Optimal allocation algorithm for a multi-way stratification design

✍ P.D. Falorsi & P. Righi, Italian National Statistical Institute(NSI)

○ 이탈리아 NSI의 표본설계는 최소표본으로 관심 있는 모든 영역에 속하는 모집단 단위를 관측하고자 표본배분 문제에 가장 중점을 두고 있음

○ 일반적으로 표본설계는 일원화층화추출을 사용하는데 그러한 설계는 조사비용이 많이 들고 층이 너무 세분화되어 응답부담이 커지게 되며, 간혹 실제로 조사를 시행할 수 없는 경우가 있음. 이때 다원화층화추출 방법이 대안이 될 수 있음

○ 표본배분 알고리즘은 모집단의 서로 다른 분할에 속하는 부모집단에 대한 표본크기를 결정하는 과정으로, 제안한 방법으로 추정할 경우 부모집단 추정치의 표본오차는 임계치보다 낮게 됨

○ 이 알고리즘은 대규모 설문 조사 시 다변량 할당 문제를 해결하기 위한 도구로 유용함. 이를 시뮬레이션 자료 분석을 통해 증명하고 이를 강조

주제 Integration of Multiple Data Sources

❒ Register-based National Accounts - Survey design for Yearly National Accounts?

✍ Anders Wallgren & Britt Wallgren Statistics Sweden and Örebro University

○ 등록기반 센서스는 수년 간 많은 국가에서 논의되었고, 특히 인구센서스에 있어 완전하게 혹은 부분적으로 등록기반 센서스를 활용하고 있는 상태임

○ 연간 국민계정 통계는 보통 마이크로데이터 조사에서 작성되지만 등록자료를 활용하여 부분적인 변화를 제안함. 이를 통해 커버리지 오차 감소,

통계품질에 있어 비교성·일관성 증가, 데이터 생산시스템의 효율화가 증대됨을 보여줌

○ 조사 간, 국가 내 부처들 간, 넓게는 국가들 간 의외로 많은 조사들이 서로 관련되어 있는데 이 경우 협조를 통하여 등록기반 통계를 생산하고, 이 때 상위 관리자의 역할이 중요함을 강조

- 커버리지 오차 관련: 과소 커버리지와 과대 커버리지는 상쇄되지 않음

	all units in the population	population totals	domain totals, worst cases
과소 커버리지	25 %	1 % - 2 %	10 % - 20 %
과대 커버리지	8 %	0.5 %	5 %

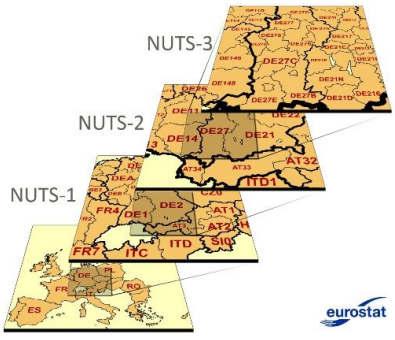
❒ Applying Statistical Matching for Facilitating the Development of New Indicators in the Framework of Social Statistics

✍ Aura Leulescu & Emilio Di Meglio, EUROSTAT

○ 지방 행정기관 정책수립을 위하여 빈곤과 사회적 배제에 대한 조치를 취할 수 있는 신뢰성 있는 통계가 필요함. 이를 계산하기 위한 방대한 양의 행정자료가 지역 단위마다 보유하고 있는 상태

○ EU-SILC*는 유럽 전역에서 소득과 생활실태에 대한 고품질 정보를 제공함. 단, NUTS 3 이하의 레벨에서는 아님

* The European Union Statistics on Income and Living Conditions (EU-SILC)



○ 본 발표에서는 이탈리아 피사 지방과 관련된 다양한 자료들(PI-Silc의 표본자료, 국세자료, 고용자료)로부터 통합 자료를 만듦으로써 표본자료와 행정자료를 매칭하여 데이터베이스를 구축함

○ 분석결과 PI-Silc와 행정자료는 정확한 매칭과 레코드연동 절차를 가짐. 즉, PI-Silc의 약 27%는 RA와 매칭되고, 약 63%는 JC와 매칭됨

○ 하지만 이것은 전적으로 PI-Silc의 오버샘플링 문제와 행정자료와 연계 시

언더커버리지 문제가 발생하기 때문에 만족할만한 결과는 아님

○ 또한 키값으로 사용된 변수에 에러와 결측치가 존재함. 예를 들어 JC는 갱신되지 않은 예전 주소가 포함되어 있음

○ 추후 등록자료의 오류를 수정하고 최신 정보로 갱신하여 매칭률을 높이고, 원표본과 매칭자료의 분포를 비교하여 자료의 품질을 높이려는 계속적인 노력을 기울일 것임을 강조함

※ 사용된 자료

- PI-SILC : 피사 지방의 소득과 생활실태 조사
- JC : 지방구직센터 데이터베이스의 사회 및 인구통계학적 정보
- RA : 국세청 데이터베이스의 납세자자료

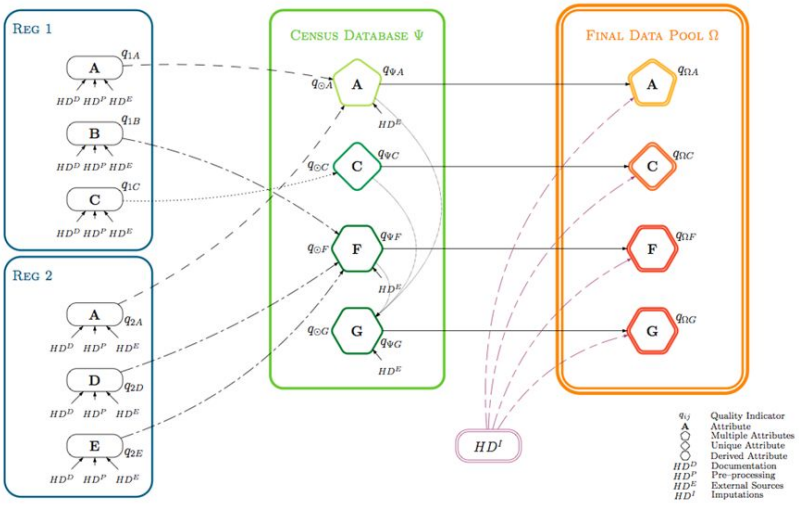
주제 Use of Registers

❒ Quality Measures for Administrative Registers : Preliminary Results for the Austrian Census 2011

✍ Manuela LENK, STATISTIK AUSTRIA

○ 오스트리아는 2011년 센서스에 등록기반센서스 방식을 도입함에 따라 행정자료의 안정성에 대한 심도 깊은 연구가 필요하게 됨에 따라 등록자료의 품질을 평가하기 위하여 자료가공 시 품질을 모니터링할 수 있는 구조적 프레임워크를 제시함

<품질 프레임워크, Quality Framework>



※ Raw data (Registers), Census Databases (CDB), Final Databases (FDB)

○ 이 프레임워크는 원시자료, 센서스 데이터베이스, 임퓨테이션을 반영한 최종데이터셋의 3단계로 구성되며, 다음 3개 정보를 결합하여 품질측정 도구로 사용함

HD(hyperdimensions) ; Documentation / Pre-processing / External Source

ⅰ. 5개 등록자료의 속성변수(SEX, FTHT, EDU)에 대한 품질측정값

등록자료	속성변수	HD_D	HD_P	HD_E
REG 1	SEX	1.000	1.000	0.998
REG 2	SEX	0.792	0.942	0.999
REG 3	SEX	0.444	0.746	0.997
REG 4	SEX	0.792	0.993	1.000
REG 3	FTHT	0.381	0.698	0.847
REG 5	EDU	0.928	0.950	0.800

※ 변수 (FTHT) Full/Half-Time employed, (EDU) highest education on national level

품질측정값 = 사용가능한 레코드수 ÷ 전체 레코드수
HD_D : hyperdimension documentation
HD_P : hyperdimension Pre-processing
HD_E : hyperdimension External Source

○ 표 ⅰ을 보면 같은 속성변수에 대한 품질측정값이 차이가 있음. 예를 들어 REG 2자료의 SEX를 보면 HD_D일 경우 0.792, HD_P일 경우 0.942, HD_E경우 0.999임. 여기에 표 ⅱ와 같이 4개의 가중치 모형을 적용시키면 전체적으로 품질측정값이 높아지는 결과를 얻을 수 있음

ⅱ. 속성변수(SEX, FTHT, EDU)에 가중치가 적용된 품질측정값

등록자료	속성변수	4개의 가중치 적용 예 (HD_D/HD_P/HD_E)			
		q(33,33,33)	q(25,25,50)	q(20,30,50)	q(20,20,60)
REG 1	SEX	0.999	0.999	0.999	0.999
REG 2	SEX	0.911	0.933	0.940	0.946
REG 3	SEX	0.729	0.796	0.811	0.836
REG 4	SEX	0.928	0.946	0.956	0.957
REG 3	FTHT	0.642	0.693	0.709	0.724
REG 5	EDU	0.891	0.867	0.868	0.853

※ Equally weighted: 0.33/0.33/0.33
HD_E Priority: 0.25/0.25/0.50
Low-Mid-High: 0.20/0.30/0.50
Strong HD_E Priority: 0.20/0.20/0.60

주제 Electronic Data Collection

❒ New data sources for statistics; Experiences at Statistics Netherland

✍ Piet Daas, Marko Roos, Chris de Blois, Rutger Hoekstra, Olav ten Bosch, Yinyima, Statistics Netherlands

○ 응답률감소와 조사예산 삭감 등 실제 열악한 조사환경에서 서베이 및 행정자료 활용 등의 전통적인 조사방법 이외의 새로운 자료수집에 대한 연구가 필요

ⅰ. 인터넷에서의 물가조사

- 이미 소비자물가지수에 활용 중이나 수동적으로 자료를 수집함. 현재 웹 로봇 혹은 스크립트, 툴 등을 이용한 자동적으로 자료를 수집하는 방법을 연구 중
- 10개월 간 매일 6개 웹사이트 조사(4개 항공사, 1개 주유소, 1개 주택)
- 자동화 시스템 구축에 있어,
  · 장점: 조사 대상 웹사이트의 데이터베이스에 접근이 가능한 경우
  · 단점: 잦은 레이아웃 개편, 특히 4개 중 3개 항공사

ⅱ. 휴대폰 위치 자료

- 네덜란드에서 휴대폰 사용률이 92%이상이고, 모든 송수신의 내역이 휴대폰기지국(셀)에 기록됨. 휴대폰자료는 1인 고유ID를 갖는 셈임
- 접근법: 네덜란드 최대 통신사로부터 고유 전화ID, 통화시간, 위치정보가 표시된 14일 간의 550백만 발신자료를 받음. 이를 통해 사람들의 낮시간 동안의 이동, 주중과 주말 동안의 다른 경제활동, 여행(로밍) 정보 등을 얻을 수 있음
- 방법론적 문제: 셀 커버리지 지역이 다르므로 셀마다 발신강도가 다르게 해석될 수 있음. 유선전화를 더 많이 사용하는 사람들이 있고 통신사 선택에서 오는 대표성 문제 등

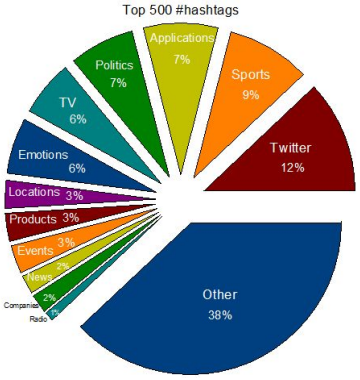
ⅲ. 트위터 메시지

- 소셜네트워킹 서비스 중 하나인 트위터에서 제공하는 개인정보, 견해, 감정 등의 잠재적인 데이터소스를 활용하는 방안
- 데이터수집①: 트위터의 검색기능을 이용하여 위트레흐트로 지역으로부터 반경 200km내 게시글 수집. 하지만 위치정보에 승인하지 않은 이용자는 검색이 안되므로 불완전함

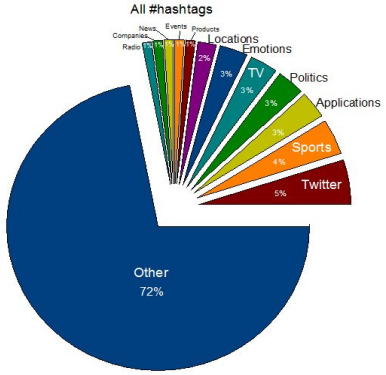
- 데이터수집②: 네덜란드에 위치해있는 팔로워(등록된 친구)가 많은 사용자들 38만명을 수집. 이 중 1천2백만 트윗(게시물)을 수집하여 주제별로 식별함(해쉬태그가 적용되어있는 트윗을 수집)
- 해쉬태그* 별 토픽

* 트위터에서 작성한 게시물의 태그(일종의 키워드인 셈)

<Top 500 #hashtags>



<All #hashtags>



- 향후연구: 해쉬태그 없이 트윗을 수집하는 방법, 토픽 분류에 대한 텍스트마이닝 기법 개발

Ⅲ 시 사 점

❒ EU국의 변화하는 통계연구 및 생산

○ 지난 30년 간 EU국은 지속적으로 감소하는 응답률과 조사예산 삭감 등의 열악한 환경에서 행정자료를 활용한 서베이에 관한 노력을 기울여왔으나,

○ 이제는 정보통신기술을 활용한 데이터의 수집, 특히 자동수집장치에 대한 연구를 하는 추세임

- 등록자료나 조사 등 전통조사를 통한 자료 수집에서 벗어나 웹 혹은 데이터베이스를 통한 자료 수집에 초점을 둠. 특히 자료수집을 위한 자동화시스템 구축에 대한 연구가 활발함

○ 가공통계가 점점 많아지는 추세가 인상적
- 어려워지는 조사환경을 반영하듯 EU국은 새로운 통계작성 시 조사통계를 지양하는 분위기

○ 휴대폰 및 GPS, 소셜네트워킹서비스 등 개인화된 서비스를 통한 자료수집에 대한 연구가 활발
- 새롭고 재미있는 데이터 소스의 유용성은 충분하나,
- 데이터의 대표성 문제와 데이터 수집의 자동프로세스 구축의 문제가 핵심 사항임
- 개인적 성향 및 특화된 계층의 특성을 보는데 유용할 것임

○ 이용자 중심의 통계제공
- 기존의 정책활용을 위한 통계에서 정부 및 이용자 모두를 만족시키는 통계생산 및 제공하려는 움직임
- 여행자에 대한 다양한 통계연구 및 통계생산
- 핀란드의 Accomodation statistics, 에스토니아의 tourism statistics

Ⅳ 기 타 사 항

❒ 자료제공 웹사이트

○ http://www.ntts2011.eu

○ http://www.cros-portal.eu/book/ntts-2011

[부록] 세부 회의 주제 및 일정

일시		세션 1	세션	세션 3
2.22 (화)	09:30-09:50	등록		
	09:50-14:15	개회사 및 기조연설		
	14:30-16:00	Data and Text Mining	Editing, Imputation	New Survey Types
	16:30-18:00	Standardisation	Statistical Disclosure Control	Geo-Information for Statistical Analysis
2.23 (수)	09:00-10:00	기조연설		
	10:30-12:00	Aggregation, Estimation, Calibration	Integration of Multiple Data Sources (1)	RISQ[1] project
	13:15-14:10	기조연설		
	14:30-16:00	Remote Access	Integration of Multiple Data Sources (2)	BLUE-ETS[2] project
	16:30-18:00	Use of Registers	AMELI[3] project	Small Area Estimation
2.24 (목)	09:00-10:00	기조연설		
	10:30-12:00	Well-Being Indicators	Methods to Synthesize and Extract Knowledge	SAMPLE[4] project
	13:15-15:00	Electronic Data Collection	POINT[5] project	ESS[6]net projects
	15:30-15:55	기조연설		
	15:55-16:30	폐회사		

1) RISQ : Representative Indicators for Survey Quality

2) BLUE-ETS : BLUE-Enterprise and Trade Statistics

3) AMELI : Advanced Methodology for European Laeken Indicators

4) SAMPLE : Small Area Methods for Poverty and Living Condition Estimatest

5) POINT : Policy Influence of Indicators

6) ESS : European Statistical System