공식통계 분야에서 BIG DATA 활용방안

Ⅰ 배경 - 왜 BIG DATA를 활용하는가?

□ (배경) 정부정책이나 민간 **의사결정**을 위해, **시의성 있는 다양하고 세분화된 통계가 필요**하나, **공식통계 생산은 비용이 많이 들고, 실시간으로 활용하기에는 한계**

○ 이러한 한계를 극복하기 위해, 우리 **실생활에서 쉽게 접할 수 있는 자료(빅 데이터)**[*)]**를** 활용할 필요성이 제기됨

*) 신생아 집중치료실에서 나오는 수백만의 데이터를 분석하여 **체온 및 심박 수** 등의 **요인변화를 질병에 대한 조기 경고**로 활용('14.5월 미국 백악관 보고서)

○ 우리가 특정한 **문제에 직면**하여 새로운 **해결방안을 모색할 때 혁신**을 하듯이, **빅 데이터 활용**도 **공식통계의 한계에서 시작**하는 것이 바람직

□ **(논의 방향)** 빅 데이터 활용은 세계적 흐름, 우리도 빅 데이터 활용가능성을 검토할 필요가 있다고 판단됨

○ 빅 데이터의 개념, 외국의 활용 사례, 장·단점을 파악한 후, **어느 분야에 어떻게 활용할 것인가에 대한 구체적인 전략 수립** 필요

Ⅱ 빅 데이터(BIG DATA)란?

□ (개념) 일반적으로 빅 데이터는 "우리 실생활에 활용되고 우리의 삶을 향상시키는 모든 데이터(정형 또는 비정형 포함)"를 의미함

* 미국 정부는 빅 데이터를 다음과 같이 정의('14년 5월)

- "미국인 및 전 세계 사람들이 **살아가고, 일하고, 소통하는 방법을 근본적으로 발전시켜 나가는 자료**"(fundamentally reshaping how Americans and people around the World live, work, and communicate)"

○ 빅 데이터는 공식통계자료, 각 부처의 행정자료, 개인생활 및 사회구조에서 발생한 각종자료*)를 포함 **(⇒ 광의의 빅 데이터)**

*) ① 인터넷상의 온라인 거래 자료, ② 개인 간 통신과 관계를 나타내는 모바일, 페이스 북, 트위터 자료 및 위성정보, ③ 기타, 사회구조기능의 흐름을 나타내는 자료 등

□ **(빅 데이터의 특성)** 빅 데이터는 **규모(high volume), 속도(velocity), 다양성(variety) 그리고 복잡성(Complexity)**을 가진 데이터 원천

○ 빅 데이터 자료를 수집하여, 관리하고, 처리하기 위해서는 기존 공식통계 작성과는 **다른 접근방법**과 **시스템이 필요**

⇒ 이곳에서는 빅 데이터의 개념을, **개인생활 및 사회구조에서 발생한 자료(협의의 빅 데이터**로 정의)로 **한정하여** 살펴봄

Ⅲ 외국의 빅 데이터(BIG DATA) 활용현황

□ 빅 데이터 워크숍에서 제기된 외국의 빅 데이터 활용사례를 3가지 유형*)으로 구분하여 제시

*) (유형 1) 인터넷상의 온라인 거래 자료, (유형 2) 개인 간 통신과 관계를 나타내는 모든 자료 (유형 3) 기타, 사회구조기능의 흐름을 나타내는 자료

Big data 원천		활용분야	주요내용
유형 1 : 온라인 거래자료	구글, Baidu 등 인터넷 정보	• 노동통계 • 물가통계 작성	• 구글 트렌드를 활용한 노동시장 예측 조사(이탈리아) • web-crawler 기술을 이용한 소비자물가지수(중국), Baidu 서칭 자료를 활용한 집값예측 (중국)

Big data 원천		활용분야	주요내용
유형 2 : 개인 간 통신 과 관계를 나타내는 자료 및 위성자료	통신자료 (Mobile positioning data)	• 관광통계	• 외국인의 로밍정보를 활용하여 관광객 선호지역, 이동경로 파악(UN 연구)
		• 교통량통계	• 국내휴대폰 이용자의 이동경로를 활용하여 교통 관련 통계작성(이탈리아)
		• 인구통계	• 특정지역의 휴대폰 위치정보를 실시간으로 활용하여 실시간 인구파악 (네델란드)
	Social Media (페이스북, 트위터)	• 보건통계	• 소셜미디어 정보(메시지, 사진 등)를 활용하여 HIV, Flu 만연 관련 통계, 심장병 사망 가능성 통계 작성(미국)
		• 의식조사	• 소셜미디어 정보를 이용한 소비자심리지수작성(네델란드), 트위터 언어를 이용한 삶의 질 만족도 관련 통계 작성 등(멕시코)
	위성정보 (Satellite Imagery)	• 농업통계	• 위성사진을 활용한 경작면적 및 작물 생산량 통계(호주, 중국)
		• 기타, 총조사 및 환경통계	• GIS를 활용한 경제총조사, 환경통계 (멕시코)
유형 3 : 기타, 사회구조, 기 능 을 나타내는 자료	교통영상 (Traffic loops)	• 인구이동통계	• CCTV 등 영상정보를 활용하여 인구이동 통계 작성(네덜란드)
		• 교통량 통계	• 고속도로 통행 자료를 활용한 교통량 통계 작성(네덜란드)

Ⅳ 빅 데이터(BIG DATA) 활용의 장단점 및 고려사항

□ 빅 데이터 활용의 장점

○ **(통계생산 효율성 제고)** 보다 시의성 있고, 세부적인 데이터를 **응답자 부담 없이 낮은 비용**으로 획득하여, **맞춤형 자료생산 가능**

○ **(이용자 중심의 통계 제공)** 정부 정책결정 및 민간분야 의사결정에 필요한 **다양하고 세부적인 통계**를 **실시간(Real time)으로 제공** 가능

□ 빅 데이터 활용의 단점

○ **(자료 획득의 어려움)** 빅 데이터는 주로 사적영역, 특히 이동통신 회사, 구글 등 글로벌 기업이 보유하고 있어 수집하기 어려움

○ **(구조화 및 분류하기 어려운 자료)** 주로 비통계적 목적으로 수집된 자료이므로 **구조화, 분류하는데 한계**

○ **(자료의 대표성 문제발생)** 수집된 자료가 **모집단 전체를 대표하기 곤란**한 경우가 많고, 특히, 모바일 통신자료의 경우 이동통신망이 있는 지역에만 적용되기 때문에 대표성 문제가 발생

□ 빅 데이터 도입 시 고려사항

① 방법론적 (Methodological) 고려

○ **(자료의 대표성 확보)** 통계적 방법으로 추출된 표본이 아니어서, 자료의 대표성에 한계(Sample bias)가 발생 ⇒ **모수 추정방법 검토**

> [예시] 휴대폰 위치정보를 이용한 실시간 인구이동통계 작성 시 휴대폰 미사용자인 유아 및 노인은 제외되고, 청소년, 중장년층이 과대 추출되어 표본의 대표성 (Representative) 저하
>
> ⇒ **모집단(등록센서스) 및 미래부 휴대폰 소유통계를 활용하여, 사후가중치 조정을 통한 추정으로 대표성 확보가능성 검토**

○ (Correlation vs Causation 고려) 빅 데이터를 활용하여 작성된 통계와 기존통계 간 상관관계 분석을 통해 활용가능성 검토

- 그러나 상관관계가 높다고 해서 기존의 공식통계를 대체하는 것은 신중해야 함 ⇒ **상관관계뿐만 아니라 현상에 대한 인과관계도 고려**

> [예시] 트위터, 페이스 북 등 소셜미디어를 활용하여 경기에 대한 **소비자심리지수(Consumer Sentimet Index)를 작성**하여 **기존 통계와 상관관계가 높은 방법론(모델링)을 선택**
>
> ⇒ 빅데이터 적용과정은 **상관관계뿐만 아니라 현상에 대한 인과관계를 고려**하여 연구할 필요가 있음

② 개인정보 보호(Privacy)

○ 소극적동의(Passive consent) vs 적극적동의(Active consent) 필요

- 빅데이터 및 공식통계 특성을 검토하여 정보주체의 동의 필요여부, **소극적동의[1]가 필요한지 또는 적극적 동의[2]가 필요한지를 파악**해야함

1) 소극적 동의: **정보주체(개인)와 빅데이터 제공자(기업 등)간** 개인정보 활용에 대한 동의를 빅데이터 활용 통계생산에 대한 동의로 간주하는 것

2) 적극적 동의: **정보주체와 통계 생산자(통계청 등)간** 개인정보 활용에 대한 별도의 동의가 필요한 것(일종의 참여)

③ 파트너십 구축(Partnership)

○ 빅 데이터 수집 및 처리, 통계 추정방법 등 **절차의 투명성 확보**와 통계작성기관과 정보제공자(빅 데이터 제공자) 간 긴밀한 협력이 필요

○ 비밀보호 및 신뢰성에 관한 명확하고 강력한 규칙 수립

④ 수집된 데이터 관리 및 처리시스템의 구축

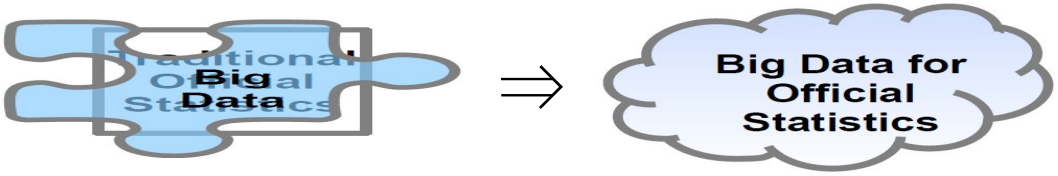
○ 대용량 데이터의 빠른 처리를 위해서는 데이터 **저장 공간(data storage) 및 분석 인프라(analysis infrastructure) 구축**이 요구됨

Ⅴ 향후 발전방안 – BIG DATA 활용을 위한 제언

□ 빅 데이터 활용을 위한 연구와 인식의 전환 필요

○ 전 세계적으로 빅 데이터 활용은 **거대한 흐름**(Big trend), 우리나라도 공식통계에서 빅 데이터를 활용할 수 있는지 연구할 필요

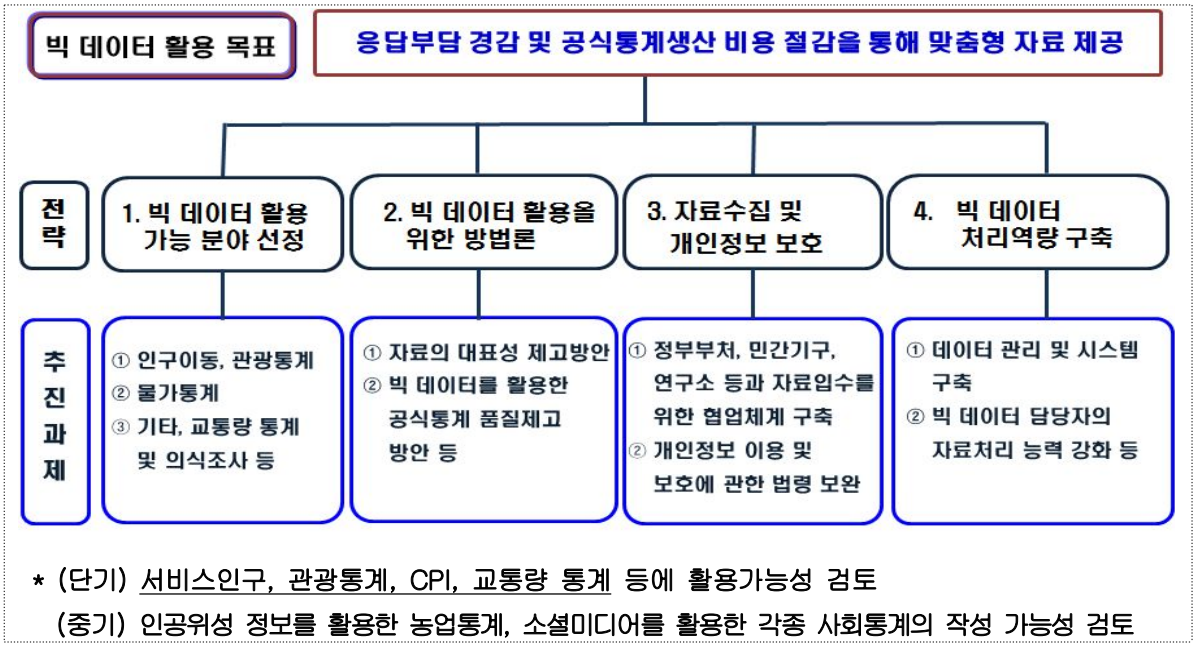
- 공식통계의 체계를 유지하면서 빅데이터를 활용하기 보다는 **빅데이터의 비구조적인 특성을 고려하여 공식통계에 활용하는 방안 검토** 등 유연한 사고로 전환 필요



【×】 【○】

□ 빅 데이터를 공식통계에 활용하기 위해서는, **빅 데이터 활용가능 분야, 활용방법, 자료 수집 및 개인정보 보호 및 빅 데이터 처리 방안** 등에 관한 **기본전략 수립**이 필요 (– **행정통계과 주관**)

〈 예, 빅 데이터 전략체계도(안) 〉



* (단기) 서비스인구, 관광통계, CPI, 교통량 통계 등에 활용가능성 검토
(중기) 인공위성 정보를 활용한 농업통계, 소셜미디어를 활용한 각종 사회통계의 작성 가능성 검토

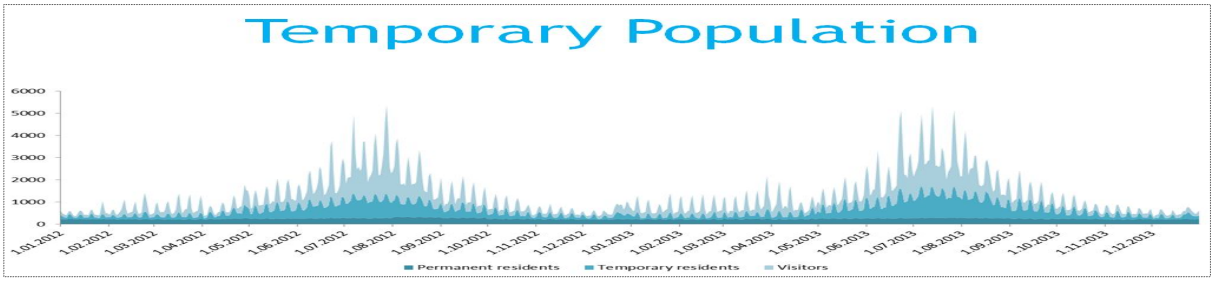
□ 구체적인 사례 제시 - 모바일 정보를 활용한 통계작성

○ (모바일 정보의 정의 및 특성) 시공간에서 모바일 장치의 위치를 추적하는 적극적 또는 소극적 위치 데이터

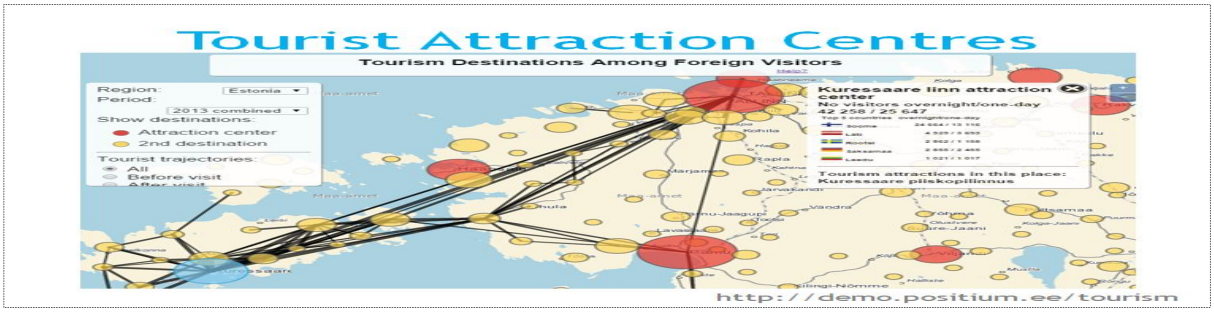
* 적극적 위치정보(Active positioning) : 휴대폰의 위치정보를 실시간으로 수집 ⇒ 소유자 동의 필요
* 소극적 위치정보(Passive positioning) : 이동통신회사로부터 받은 휴대폰 위치 정보 ⇒ 소유자 동의 불필요

○ 주요 적용 분야

- **(인구통계)** 휴대폰 보유자의 시간별 위치정보를 활용하여 지역의 실시간 인구통계(daytime population)를 생산하여 도시설계 정책에 활용



- **(여행관련 통계)** 외국인의 로밍정보를 활용하여 여행객들의 이동경로 및 체류시간 등의 정보를 분석하여 여행관련 통계 생산



- **(교통통계)** 국내 휴대폰 보유자의 이동정보(출발지~종착지)를 활용하여 출퇴근 패턴을 분석하고, 교통량 및 주요이동 경로 등 교통통계 생산에 활용

⇒