Big data for official statistics

Strategies and some initial European applications

Martin Karlberg and Michail Skaliotis, Eurostat

Big Data – what is it?

- *Volume*
- *Velocity*
- *Variety (social network data, RFID sensor data, satellite image data, business system data…)*

*Generally "**Organic**", i.e. not **designed** (in the survey design sense) by official statisticians*

*Sometimes **Open**, i.e. freely available to anybody online (but sometimes proprietary)*

Overview

1. **Big data and data science**
2. **Some recent international initiatives**
3. **Actual applications:**
   - Internet as a Data Source (ICT statistics)
   - Mobile positioning data (tourism statistics)
   - Price collection via the Internet (price statistics)
4. **Conclusions**

1. Big Data and data science

- ***Data science vs. statistics***
  multidisciplinary – extracting meaning from data
  (mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing)
- ***Area dominated by computer scientists***
  Official statisticians behind on the learning curve
- ***Who could become a data scientist?***
  - Trained statisticians with programming skills?
  - IT specialists lacking training in statistics?
- ***Learning by doing is key***

2. Big Data and official statistics – some recent international initiatives

- *Big Data Task Team (coordinated by UNECE) project proposal (http://www1.unece.org/stat/platform/display/Collection/Draft+HLG+Project+Proposal+on+Big+Data):*
  - **identify, examine and provide guidance regarding main strategic and methodological issues**
  - **demonstrate the feasibility of efficient production of both novel products and "mainstream" official statistics using Big Data sources**
  - **facilitate the sharing across organisations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.**
- *Scheveningen memorandum (DGINS)*

3. Learning by doing is key

- *The time has come to demystify "Big Data"*
- *No data science skills without practice!*
- *Emphasised in BD Task Team project proposal*
- *Relevant activities launched by Eurostat in the domain (details provided in our paper):*
  - Internet as a Data Source (ICT statistics)
  - Mobile positioning data (tourism statistics)
  - Price collection via the Internet (price statistics)

Using the Internet for the collection of (information society) statistics

- *Current situation: Traditional surveys to individuals and enterprises*
- *Need: rapidly evolving phenomena (moving target) – particularly true for ICT*
- *Opportunity: digital footprint available by definition*

Study commissioned by Eurostat:
"analysis of methodologies for using the Internet for the collection of ICT and other statistics"

IaD study for ICT statistics

- *Approach: Internet as a data source (IaD)*
  - **User-centric, network-centric, site centric**
- *Feasibility study: Which items could be collected over the Internet?*
- *Collection mode:*
  - **Households: Respondents asked to download "monitoring program"; transmissions to NSIs**
  - **Enterprises: Data harvested from enterprise websites (2nd phase: complemented with "monitoring program" or server log files)**

IaD study (*cont.*)

- *Going beyond ICT statistics:*
  - **use of federated open data for official statistics: *evaluation of Big Data repositories regarding their***
    - *usefulness as a supplementary source for traditional official statistics*
    - *capacity of replacing official statistical indicators study*

  steps:
    - *Identify Big Data sources*
    - *Assess potential*
    - *Assess practical feasibility*
    - *Analyse conditions for "opening" the Big Data sources*

Use of mobile positioning data for tourism statistics

- *Example:*
  - **Anonymised signal data from one Mobile Network Operator (MNO)**
  - **recalculation of the sample of signal data to a number of people (based on calibration)**
  - **domestic tourism: based on "home anchor point", according to repeated occurrence in the same cell (night-time)**
  - **International tourism: based on roaming data**

Use of mobile positioning data for tourism statistics (*cont.*)

- *Meets a need:*
  - **a more accurate and in-depth data source**
  - **new aspects and indicators for describing the time-space behaviour of people**
  - **a highly quantitative data source**
  - **better time-space insights**
  - **reduced costs**
  - **possibility to register trips to sparsely populated locations such as natural parks**

Collecting prices on the internet

- *Potential:*
  - **Increased volume of e-commerce**
  - **Prices available in the public domain**
  - **High frequency automated collection possible**
- *Feasibility studies (Statistics Netherlands):*
  - **Generic (country, product) software modules (open source license)**
  - **Collection via internet robots**

Issues

- *Data provision:*
  - **Ethical and image issues**
  - **Technical issues**
  - **Continuity issues**
- *Quality:*
  - **Representativity issues**
  - **Comparability vs. relevance**

Data: ethical, image & legal issues

- *ICT "monitoring tool":*
  - **Similarity with phishing?**
    (*Outweighed by response burden reduction?*)
- *Mobile positioning data:*
  - **Individuals monitored (personal data protection?)**
  - **MNO concerns: cost (**actual, opportunity cost**) risks (**strain on real-time systems; business secrets divulged**)**
- *Price collection:*
  - **Enterprises explicitly prohibiting "bots" combing their websites**

Data: Technical issues

- *ICT "monitoring tool":*
  - **Multiple operating systems, various configurations of each system (correlated with survey variables)**
- *Mobile positioning data:*
  - **High-frequency data taxing on the real-time system on MNOs**
  - **Standardisation needed (but "easier")**
- *Price collection:*
  - **Multiple e-commerce site designs**

Data: Business continuity issues

- *ICT "monitoring tool":*
  - **Evolving operating systems**
- *Mobile positioning data:*
  - **If basis for official statistics:** legislation inevitable?
- *Price collection:*
  - **Evolving webshops**
  - **If basis for official statistics:** legislation necessary?

Representativity and comparability

- *New concepts, new collection mode*
  → time series break inevitable
- *One person – one device?*
  One company – one website?
- *Cell phone penetration – a non-issue?*
- *Online prices different from offline prices!*

Possible solutions:

- *Calibrate using traditional statistics*
- *Parallel measurement (for bridging the gap)*
- *Accept break on grounds of increased relevance*

Lessons learned so far

- *Promising results – but issues remain*
- *Trade-off between price/volume and representativity/quality – however, "big data" alternative not always "cheaper but worse":*
  - **Non-response** vs. Big Data representativity
  - **Narrow scope** vs. Big Data granularity
- *The "variety" is hard to tackle – (obviously, more structured Big Data are easier to analyse)*

Increased ICT, cell phone and e-commerce use
→ increased scope, relevance (and acceptance?)

4. The way forward

Official statistics community – now on the move!

Collaboration

- Big Data goes beyond statistics authorities
  → national/"government" strategies needed
- Public-Private Partnerships (could help build-up)
- Multidisicplinary teams – including legal expertise

Action

- Demystify "data science" → build up skills
- Applications-driven approach → **learn by doing**