Big Data and Official Statistics

Game-Changer for National Statistical Institutes?

Peter Struijs, 21 November 2016 – Part 1
PART 1

– What is Big Data?
– The International Context
– Examples:
  - Road Sensor Data
  - Mobile Phone Data
  - Social Media Data

PART 2

– Issues and Approach
– Process, Methodology and IT
– Access and Partnerships
– The Strategic Perspective
– More Examples and Conclusion
Part 1 – 1 What is Big Data?
The FOUR V's of Big Data

Volume: Scale of Data
- 40 Zettabytes (43 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005
- 6 billion people have cell phones
- World population: 7 billion
- Most companies in the U.S. have at least 100 terabytes (100,000 gigabytes) of data stored

Variety: Different Forms of Data
- 30 billion pieces of content are shared on Facebook every month
- 420 million wearable, wireless health monitors
- 4 billion hours of video are watched on YouTube each month
- 400 million tweets are sent per day by about 200 million monthly active users

Velocity: Analysis of Streaming Data
- The New York Stock Exchange captures 1 TB of trade information during each trading session
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure
- By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States
- By 2016, it is projected there will be 18.9 billion network connections – almost 2.5 connections per person on earth

Veracity: Uncertainty of Data
- 1 in 3 business leaders don’t trust the information they use to make decisions
- 27% of respondents in one survey were unsure of how much of their data was inaccurate
- Poor data quality costs the US economy around $3.1 trillion a year
A Definition for Statistical Purposes

Big Data are **data sources** that can be – generally – described as: “high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making”.

(UNECE, 2013)
Big Data Characteristics

Definition:
- Volume
- Velocity
- Variety

Data characteristics:
- Unstructured data
- Selectivity
- Population dynamics
- Event data
- Organic data
- Distributed data

Data use:
- Other ways of processing
- Fundamentally new applications
Examples of Possible Big Data Sources

- Road sensor data
- Mobile phone location data
- Public social media messages
- Websites
- Google Trends
- Satellite information
- Etc...
Potential Opportunities

- New statistics
- More detailed statistics
- More timely statistics
- Nowcasts and early indicators
- Quality improvement
- Response burden reduction
- Cost reduction and higher efficiency
Part 1 – 2  The International Context
UNECE Big Data Activities

- Definition and classification of Big Data sources in 2013
- Big Data project in 2014, with three Task Teams:
  - Partnerships
  - Privacy
  - Quality
- Sandbox in 2014 and beyond
- Big Data survey, together with UNSD, in 2014

Results:
http://www1.unece.org/stat/platform/display/bigdata/2014+Project
The UNECE Classification of Big Data (1)

1. Social Networks (human-sourced information)
   1100. Social Networks: Facebook, Twitter, Tumblr etc.
   1200. Blogs and comments
   1300. Personal documents
   1400. Pictures: Instagram, Flickr, Picasa etc.
   1500. Videos: Youtube etc.
   1600. Internet searches
   1700. Mobile data content: text messages
   1800. User-generated maps
   1900. E-Mail
2. Traditional Business Systems (process-mediated data)

21. Data produced by Public Agencies
   2110. Medical records

22. Data produced by businesses
   2210. Commercial transactions
   2220. Banking/stock records
   2230. E-commerce
   2240. Credit cards
The UNECE Classification of Big Data (3)

3. Internet of Things (machine-generated data)

31. Data from sensors
   311. Fixed sensors
       3111. Home automation
       3112. Weather/pollution sensors
       3113. Traffic sensors/webcam
       3114. Scientific sensors
       3115. Security/surveillance videos/images
   312. Mobile sensors (tracking)
       3121. Mobile phone location
       3122. Cars
       3123. Satellite images

32. Data from computer systems
   3210. Logs
   3220. Web logs
UN Big Data Activities

- Global Working Group on Big Data for Official Statistics, active in several areas:
  - Mobile phone data
  - Satellite imagery
  - Social media data
  - Access / partnerships
  - Advocacy / communication
  - Big Data and SDGs
  - Training / skills / capacity building
  - Cross-cutting issues

- UNSD survey on Big Data for official statistics
GWG Big Data Survey


- Result:
  - 114 projects from 43 countries/organizations (10 Sept. 2015)
  - Exploration vs ‘intended for’ production (about 1:1)
Projects as Reported to the GWG

1. Web data/web scraping (21)
   collect prices, job vacancies, enterprise information ...

2. Scanner data (18)
   for CPI

3. Mobile phone data (15)
   Tourism, border crossings, ‘day time population’

4. Social media/Google trends (8)
   Fast indicators & now-casts: sentiment, unemployment ...

5. Satellite/aerial imagery data (6)
   Land use, crops, ‘poverty’ ...

6. Other (46)
   Smart meters, transport (land, water), health, credit-card, patents .... Incl. admin data
ESS Big Data Activities

Scheveningen Memorandum

Big Data Action Plan and Roadmap Big Data

ESSnet Big Data
Scheveningen Memorandum on Big Data

- Examine the **potential of Big Data** sources for official statistics
- Official Statistics **Big Data strategy** as part of wider government strategy
- Address **privacy and data protection**
- **Collaboration** at European and global level
- Address need for **skills**
- **Partnerships** between different stakeholders (government, academics, private sector)
- Developments in **Methodology, quality assessment and IT**
- Adopt **action plan and roadmap** for the European Statistical System
Action Plan Themes

- Policy
- Experience sharing
- Methods
- Quality
- Legislation
- Ethics / Communication
- Skills
- IT Infrastructures
- Partnerships
- Pilots
The ESSnet Big Data

Framework Partnership Agreement: 22 partners

eurostat

Two Specific Grant Agreements:

SGA-1: February 2016 – July 2017 1.0 M€
SGA-2: January 2017 – May 2018 1.0 M€
List of pilot projects

- Web scraping
  - job vacancies; enterprise characteristics
- Smart meters
  - electricity consumption; temporary vacant dwellings
- Automatic Identification System (Ships)
  - vessel identification data
- Mobile phone data
  - Preparing for access to data
- Scenario for using multiple inputs
- Modelling for now-casting statistics
Subdivision of Pilots into Phases

1. Data access
   • Conditions; partnerships

2. Data handling
   • Production criteria; micro versus aggregated data; visualisation

3. Methodology and technology
   • Methodology for long lasting statistics; process design

4. Statistical output
   • Examples of existing and new outputs; potential users; comparison with current estimates (quality, timeliness, level of detail)

5. Future perspectives
   • Applicability in ESS; future production process; exploration of further possibilities of using and combining (big) data sources
Examples of Activities in Pilots (1)

Target Population: All job vacancies

- Advertised on enterprise website
- Advertised on a job portal
- Employing business is identifiable
- Advertised through an agency

‘Ghost’ Vacancies
Examples of Activities in Pilots (2)
Examples of Activities in Pilots (3)

- Estonian data structure: 4 main tables
  Metering data – main table with hourly consumptions
  Metering points – location
  Agreements – contract info
  Customers – contract holder information

17.11.2016    Maiki Ilves
Examples of Activities in Pilots (4)

https://maartenpouwels.carto.com/viz/8d319f16-8195-11e6-afo4-oecd1babdde5/public_map
Examples of Activities in Pilots (5)

- Big Data sources
- Administrative data
- Statistical data

may enrich statistical output in domains:
Big Data and SDG Indicators

Agreement at UN level:
– 17 Sustainable Development Goals
– 169 Targets
– 230 SDG Indicators

Problem:
– No data available for about a third of the indicators
– Can indicators be based on Big Data?
Part 1 – 3 Example: Road Sensor Data
Road Sensor Data

- **Measurement points:** 20,000 traffic loops on Dutch motorways; 40,000 on provincial roads
- **Variables:** number and average speed of passing vehicles, for three different length classes
- **Frequency:** per minute (24/7)
- **Volume:** around 230 million records a day
- **Source:** National Data Warehouse for Traffic Information (NDW)
The Main Roads

Highways per NUTS3
A Special Dike
Road Sensors in the Dike
Minute Data of One Sensor for 196 Days
Researching the Data

Cross correlation between sensor pairs
- Used to validate metadata

Trajectory speed vs. point speed
- Average speed is 98 Km/h
Sensors in a Road Segment

sensor
- Road sensor

Road segment
- Main route
- Exit ramp
- Entrance ramp
- Other
Small, Medium-Sized & Large Vehicles
Road Sensor Data: Results

Top 5 traffic intensities on Dutch motorways

A13
A10
A12
A16
A4

Average number of vehicles per hour

2014 2013

Bron: CBS, NDW
Frosted Roads at the Beginning of January...
... and a Press Release on 8 January!

Traffic in the North of the Netherlands, first three working days of 2016

![Bar chart showing average number of vehicles per hour for different roads in 2016 compared to 2012/2015. The chart includes roads A28, A31, A7, A37, A32, and N33. The chart is sourced from NDW, CBS.]
Road Sensor Data: Issues and Non-Issues

Non-issues:
- Privacy
- Data acquisition

Issues:
- Methodology
  - Selectivity
  - Quality
- Processing needs
- Other issues
  - Skills needed
  - Transition from research to regular statistics
Part 1 – 4 Example: Mobile Phone Data
Possible Uses of Mobile Phone Data

- Daytime population statistics
- Mobility statistics
- Tourism statistics
- Other uses
Mobile Phone Activity as a Data Source

- Nearly every person in the Netherlands has a mobile phone
  - Usually on them
  - Almost always switched on
  - Many people are very active during the day

- There is a grid of antennas with good coverage

- Data of a single mobile company was used
  - Hourly aggregates per area
  - Threshold of 15 events
Daytime Population Based on Mobile Phone Data

2013-05-07 00:00:00

- Very sparsely populated
- Sparsely populated
- Normally populated
- Densely populated
- Very densely populated
Issues When Using Mobile Phone Data

- Privacy
- Data acquisition
- Methodology
  - Representativeness
  - Selectivity
  - Quality
- Other issues
  - Infrastructure
  - Skills needed
Part 1 – 5 Example: Social Media Data
Possible Uses of Social Media Data

- Sentiment indicators
  - e.g. consumer confidence index

- Social indicators
  - e.g. social coherence indices

- Other uses
Social Media

- Dutch are very active on social media!
  - Around 60% according to a survey

- Steady increase in smartphone ownership!
  - Possible information source for:
    - Which topics are current:
      - Number of posts and sentiment about these topics
    - As a measurement tool for:

Map by Eric Fischer (via Fast Company)
Social Media Data

- All social media messages:
  - That are written in Dutch
  - That are public

- Data collection: systematically and instantly
  - Collected by the Dutch firm Coosto
  - Some value is added by Coosto on sentiment
  - Paid subscription

- Dataset of more than 3.5 billion messages:
  - Covering June 2010 till present
  - Between 3-4 million messages added per day
Research Question

Can we replicate the consumer confidence index by only using social media data, while reducing production time?
Sentiment Determination

- ‘Bag of words’ approach
  - list of Dutch words with their associated sentiment
  - added social media specific words (‘FAIL’, ‘LOL’, ‘OMG’ etc.)

- Use overall score to determine sentiment
  - is either positive, negative or neutral

- Average sentiment per period (day / week / month)
  - $(\#\text{positive} - \#\text{negative})/\#\text{total} \times 100\%$
Sentiment per platform

- Facebook (~10%)
- Twitter (~80%)
- Hyves
- Blogs
- News sites
- Google+
- LinkedIn
- Youtube
- Forums

Consumer confidence
Figure 1. Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. In the insert the development of consumer confidence is shown for the identical period.
Results

- High correlation achieved (0.9)
- Changes in consumer confidence precede changes in sentiment by one week
- Short processing time, so time-to-market may be reduced.
- Sentiment index can be produced on a weekly basis
- To be considered:
  - Use model-based figures as early indicators
  - Reduce sampling of consumer confidence index
General Sentiment Indicator (draft version)
Issues When Using Social Media Data

**Lesser issues:**
- Privacy
- Data acquisition

**Main issues:**
- Methodology
  - Selectivity
  - Meaning of the data
  - Validity of methods used

- Other issues
  - Skills needed
Big Data and Official Statistics

Game-Changer for National Statistical Institutes?

Peter Struijs, 21 November 2016 – Part 2
Outline

PART 1

– What is Big Data?
– The International Context
– Examples:
  - Road Sensor Data
  - Mobile Phone Data
  - Social Media Data

PART 2

– Issues and Approach
– Process, Methodology and IT
– Access and Partnerships
– The Strategic Perspective
– More Examples and Conclusion
Data Sources and Approaches

Surveys / questionnaires

Administrative data sources

Where does Big Data fit in?

New methods may be needed, e.g. modeling for nowcasting and other methods not based on sampling theory
Limitations of the established quality frameworks and methodology

Options

What to do in the changing context of making statistics
Sources of Error

- Big data is not perfect

Overview of Issues

- Getting access to the data
- Usability of the data
  - Meaning of the data, stability of the source, reproducability
- Methodological issues
  - Selectivity, representativeness, unknown population, quality and validity
- Privacy, confidentiality and reputation
- IT-infrastructure and security
- Knowledge and skills
- Transition from research to production
- Strategic challenges
The Top Three Issues

1. Population not known
2. Unbalanced coverage
3. Relevance of data not clear
Population not known

Derive background information

Relate population at meso- or macro-level to other information
Unbalanced coverage

Use modeling approaches
Relevance of data not clear

Calibration / fitting

Study correlations

Use Big Data for “stand alone” information
Part 2 – 2 Process, Methodology and IT
Process of Making Traffic Intensities Statistics

- Select sensors on Dutch highways

- **Preprocessing**
  - Remove non-informative variables
  - Remove bad records
  - Exclude bad sensors
  - Quality indicators for daily data per sensor

- **Processing**
  - Reduce dimensions on same road and region
  - Obtain number of vehicles for each road and region
  - For each road and region, calculate monthly traffic intensity
  - Use of R-Hadoop

- **Validation and publication**
A Big Data Production Process

1. Frame
2. Clean
3. Transform & Select
4. Aggregate & Estimate
A Big Data Production Process: Volume

Raw Data
~80 TB
2010 - 2014

Transform + Select
Big Data specific

Transformed Data
100 GB

Clean Data
500 MB

Estimation
Traffic Index
6 KB
Data Options

Historical database
- Request data via web interface
- Minute data for all highways (48 variables, Jan December 2014: around 2.5 TB)

Data stream
- Every minute, all data for all active sensors
- Continuously collected
Questions on the Validity of Methods Used

- Is it acceptable, under certain conditions, to base official statistics on correlations?

- If so, what are the conditions?

- What to do if there is a shock?
## IT Infrastructural Needs

### Overview of IT platforms used for Big Data analysis

<table>
<thead>
<tr>
<th>Scaling type</th>
<th>Platforms (Communication Scheme)</th>
<th>System/Platform</th>
<th>Application/Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Scalability</td>
<td>Data I/O performance</td>
</tr>
<tr>
<td>Horizontal scaling</td>
<td>Peer-to-Peer (TCP/IP)</td>
<td>★★★★★</td>
<td>★</td>
</tr>
<tr>
<td></td>
<td>Virtual clusters (MapReduce/MPI)</td>
<td>★★★★★</td>
<td>★ ★</td>
</tr>
<tr>
<td>Vertical scaling</td>
<td>Virtual clusters (Spark)</td>
<td>★★★★★</td>
<td>★★★</td>
</tr>
<tr>
<td></td>
<td>HPC clusters (MPI/Mapreduce)</td>
<td>★★★</td>
<td>★★★★★</td>
</tr>
<tr>
<td></td>
<td>Multicore (Multithreading)</td>
<td>★★</td>
<td>★★★★★</td>
</tr>
<tr>
<td></td>
<td>GPU (CUDA)</td>
<td>★★</td>
<td>★★★★★</td>
</tr>
<tr>
<td></td>
<td>FPGA (HDL)</td>
<td>★</td>
<td>★★★★★</td>
</tr>
</tbody>
</table>

Part 2 – 3 Access and Partnerships
Access: Lessons Learned (1)

- Invest in relationship with data provider
  - This includes government data providers and commercial data sources
  - If possible, work on a voluntary basis
  - Try to find a fair balance of interests

- For research purposes, access is less of a problem than for regular use for statistics
  - Start with research and build a relationship with the data provider

- Request only data that is really needed
  - Aggregate data may be sufficient
  - Indirect access may be sufficient
Access: Lessons Learned (2)

- Pay for services provided, not for the data

- Even with public data there may be issues
  - E.g. purpose of use or permissability of webscraping

- Take possible public image effects into account
  - Be transparent

- Work together with partner institutes
  - Make use of international guidelines (UNECE, UN, ...)

21
Big Data Access Principles (UN, draft)

- Social responsibility
- Level playing field
- Equal treatment
- Confidentiality and security
- Transparency
- Respect for business interest
- Proportionality
Part 2 – 4 The Strategic Perspective
Strategic Aspects

- Others start producing statistics
  - there may be quality issues
  - but they are extremely rapid
  - and there is obviously demand

- Need for good, impartial information (benchmark information) will remain
  - without a monopoly for NSIs

- There is a need for validation of information produced by others
Billion Prices Project MIT

Source: State Street, PriceStats

- Official CPI
- PriceStats Index
Annual Inflation Rates in Other Countries

Argentina

Online Index

CPI

Colombia

China

Supermarket Index

Germany

France

Ireland

Developed Food

Developed Fuel

Global - Aggregate

Source: PriceStats – StateStreet Inflation Series
Learn from Others!

Google flu prediction...
Possible Responses to the Issues

- Invest in good relations with data providers
- Invest in methodological research and play with the data to get a grip on quality
- Use only aggregate data if possible
- Explore alternatives to population-based estimation methods
- Keep an open mindset
- Take the strategic challenges seriously
The Roadmap Approach

- Awareness that Big Data is a strategic issue
- Position paper for Board of Directors
- Roadmap Big Data
- External validation of the Roadmap
- Roadmap updated twice a year for Board of Directors
- Roadmap monitor
- Deputy Director General responsible at strategic level
- Coordination group for Big Data
The Scope of the Roadmap

- Identification of outputs to be based on Big Data
- For each output, definition of time target and ownership
- Identification by owner of conditions to be fulfilled
- Commitment by supporting services for fulfilling the conditions (IT, data collection, methodological support, ...)
- Supporting programmes
Supporting Programmes

Big Data features in:

- Innovation programme
- Methodological research programme
## Rolling Planning Products with Big Data

<table>
<thead>
<tr>
<th>Matrix of statistical products</th>
<th>2017</th>
<th>2018</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety and health</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyber-crime safety</td>
<td>☠️</td>
<td>☠️</td>
<td>☠️</td>
</tr>
<tr>
<td>Real-time safety feeling indicator</td>
<td>⚠️</td>
<td>⚠️</td>
<td>⚠️</td>
</tr>
<tr>
<td>Sources: Coosta, cyber security data, admin sources</td>
<td></td>
<td></td>
<td>Admin data, integrated health data, sensors</td>
</tr>
<tr>
<td>Economic growth and labour market</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet economy</td>
<td>☠️</td>
<td>☠️</td>
<td>☠️</td>
</tr>
<tr>
<td>Sharing economy</td>
<td>⚠️</td>
<td>⚠️</td>
<td>⚠️</td>
</tr>
<tr>
<td>Nowcast economic activity indicator</td>
<td>⚠️</td>
<td>⚠️</td>
<td>⚠️</td>
</tr>
<tr>
<td>Sources: Buckaroo transaction data, internet robot data, google, admin sources</td>
<td></td>
<td></td>
<td>Coosta, satellite data, bank transactions data, admin sources, AIS</td>
</tr>
<tr>
<td>Smart cities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy consumption indicators</td>
<td>☠️</td>
<td>☠️</td>
<td>☠️</td>
</tr>
<tr>
<td>Day-time population</td>
<td>⚠️</td>
<td>⚠️</td>
<td>⚠️</td>
</tr>
<tr>
<td>Sources: Smart meters, admin sources, mobile phone data</td>
<td></td>
<td></td>
<td>Admin sources, sensors, Internet of things</td>
</tr>
<tr>
<td>Mobility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commuting patterns</td>
<td>☠️</td>
<td>☠️</td>
<td>☠️</td>
</tr>
<tr>
<td>Transport over sea</td>
<td>⚠️</td>
<td>⚠️</td>
<td>⚠️</td>
</tr>
<tr>
<td>Sources: Admin sources, AIS</td>
<td></td>
<td></td>
<td>OV-public transport use data, new sources</td>
</tr>
<tr>
<td>SDG’s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 SDG indicator (to be decided) based on Big Data</td>
<td>☠️</td>
<td>☠️</td>
<td>☠️</td>
</tr>
<tr>
<td>Minimum of 3 SDG indicators (to be decided) based on BD</td>
<td>⚠️</td>
<td>⚠️</td>
<td>⚠️</td>
</tr>
<tr>
<td>Sources: Admin sources, big data</td>
<td></td>
<td></td>
<td>Integrated data</td>
</tr>
</tbody>
</table>
Data Scientists Needed!

NERD ALERT!

Register your domain today at www.activeC.net
Open Minds Needed!
Part 2 – 5 More Examples and Conclusion
Potential Opportunities

- New statistics
- More detailed statistics
- More timely statistics
- Nowcasts and early indicators
- Quality improvement
- Response burden reduction
- Cost reduction and higher efficiency
Some basic emotions
First Results

- Angry
- Excited
- Happy
- Sad
- Scared
- Tender
Sad
Credit Card Data: BEA US

Estimates using Monthly Credit Card Data are similar to Retail Trade aggregates.
Use of Satellite data for official statistics on land use and crop estimation. Statistics Australia is one of the countries working on this topic. (however, photo's are not of ABS.)
Satellite Data, Economy

Watch containers and ships move

Watch filling of oil storage tanks change
First estimates were 70% accurate
Traffic Intensity and GDP

Provisional results from data camp with students

- GDP vs Traffic
  3% increase in GDP corresponds to 12% increase in traffic

- Traffic ahead of GDP
  1 quarter

- Correlation
  82% from 2010-Q3 till 2014-Q4
  91% from 2011-Q2 till 2014-Q4
Spring in the Netherlands

Flowering of the wood anemone

2013-02-24

2013 2.5 mean 8 days below zero
2014 8.3 mean 0 days below zero
Conclusion: The Way Forward

- Get to know Big Data
- Use Big Data for efficiency and response burden reduction
- Use Big Data for early indicators
- Use Big Data for filling gaps and new demands
- Use new professional methods where needed
- Create the right environment
- Don’t do it alone!
The Future
Time for Discussion!
Questions?

Thank you for your attention!

p.struijs@cbs.nl