Center for Big Data Statistics

Het CBS Center for Big Data Statistics (CBDS)
JADS (April, 2018)
Statistics Netherlands

- Founded in 1899
- 2,187 employees
- 3,900 tables in StatLine
- 675 news articles
- 51,000 times in news media
- >10 million visits to CBS.nl
- 116,000 followers on Twitter
- CBS law
Mission Statistics Netherlands

To be the independent provider in the field of high quality, coherent, public statistical information about - in principle - all sectors of Dutch society.
From primary to secondary to new sources

- Targeted data collection, pre-determined questions and indicators
- Structured data, collected by government, not for statistical purposes
- Surveys and census
- Admin sources
- Big Data

Not necessarily for statistical purposes, volume, velocity and variety

2000BC 20th Century 21st century

Center for Big Data Statistics
1. Statistics based on new data sources

- Dot maps
- Sensor data
- Social tension indicator
- Experimental stats
- Official stats based on BD
- Consumer price index
- Traffic Intensity

Future of official stats

Center for Big Data Statistics

PHENO MENON
2. Better statistics based on large data sources

- Van den Brakel, J. et al. (2016) Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Discussion paper 201601,
3. Answering complex policy questions

- How do we tackle the energy transition?
- Can we predict migrant streams?
- How do we match offer and demand in the labour market?
- How do we make our businesses smarter?
- How do we manage mobility and pollution?
CBDS exploits new data sources in collaboration with partners to obtain current and relevant information for users on societal problems by applying state of the art techniques and methods.
From start up to scale up
CBDS works on various themes; theme covers projects that are directly related to the statistical programme of the divisions (close to the centre) or new project (further from the centre)
Results on www.cbs.nl/innovatie

- Social media en text mining: sentiment analyses
- Webscraping: internet economy
- Visualisation techniques: dotmaps
- New sources: mobility, ship movements, cybersecurity, housing market
Big data research

1. BD Sources
   • Getting access
   • Collect additional sources

2. Discover BD
   • Explorative studies, visualisation
   • Patterns recognition, feature extraction, profiling

3. Control and correct BD
   • Determine quality of large amounts of data
   • Smoothing of large amounts of data
Big data research (2)

4. Combining BD
   - Combine BD with traditional sources
   - Find linkage variables, combine in time and geo-location, micro vs macro

5. Estimate with BD
   - Non-probability approach, use Machine Learning methods
   - BD (sub)population studies, ‘selectivity’ en correction, real-time statistics

6. Processing BD en IT
   - ‘IT research and algorithms (implementation fast methods)
   - Publish at higher level of detail and statistical security
Energy

Sources
TENNET, KNMI data, Solar panels register, taks data for solar panel installation, use of electricity, aireal pictures, smart meter data

Methods
Modelling and predicting techniques, pattern recognition, combining sources
Electricity from renewable sources

- SN makes statistics based on registers and surveys
- Ambition is to make these statistics based on new sources: current, accurate and more detailed
Share of renewable energy in the EU
Example: Solar power based on Big Data

Challenges
- Real power is unknown
  - Number of installations?: registers are incomplete

What do we do?
1. Complete registers with new sources: tax data and aereal pictures
2. Directly determining solar power
1. Complete registers with new sources

- Location of solar panels with aerial pictures
2. Directly determining production
3. Indirectly determining solar power

Public grid demand in MWh (from Tennet)

Solar irradiance in J/m² (from KNMI)

Zonnestroom

Health

1. Looking at the effectiveness of health care: combining large administrative data sources, through deepened analysis determining the effectiveness and quality of care

2. New data sources environment and lifestyle: use new big data sources to evaluate and get a better view on the relationship between health, lifestyle and environment
Traffic intensity

- Origin destination matrix at the city-district level.

- **Result**
  projected traffic intensity per stretch of road for people commuting from home to place of work (assuming for now they would travel by car and take the shortest route)
Labour market

Map offer and demand on the labour market

Measuring unemployment

Defining professional categories

Sources:
Micro-data on vacancies that are offered to job seekers; profiles of jobseekers; Dataprovider (textkernel), webscraped vacancies; POLIS; Randstad; survey on vacancies

Method:
ML (Combination skills, education, profession, and multi-dimensional classification)
Housing market

Machine learning algorithm based on training set using manually classified pictures

On social media make distinction between publicity, intention, and real movers
Smart farming

Some scenario’s for using the InfoBroker:

- Advisor
- Farmer
- Other farmers
- Government

Coöperation
Realtime-Model (expert)
Benchmarking

InfoBroker

Static data
Real time cow-centric data

EMOS webinar, 21 Feb 2018: Snijkers - Business Surveys/Statistics
SDGs

1. Makswell (H2020)
2. Deep Solaris (Eurostat)
3. Trainees on SDG 3.9
   “substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination” by 2030
4. Kostat
   SDG and poverty indicator based on mobile phone and satellite data
   • Literature review on methodology
   • Approaches (CBS, KOSTAT): survey en admin on income
   • Approaches (CBS, KOSTAT): survey en admin on poverty
   • Use mobile phone – satellite data to measure poverty
Partners