Practical Applications of Big Data for Official Statistics

Peter Struijs and Barteld Braaksma

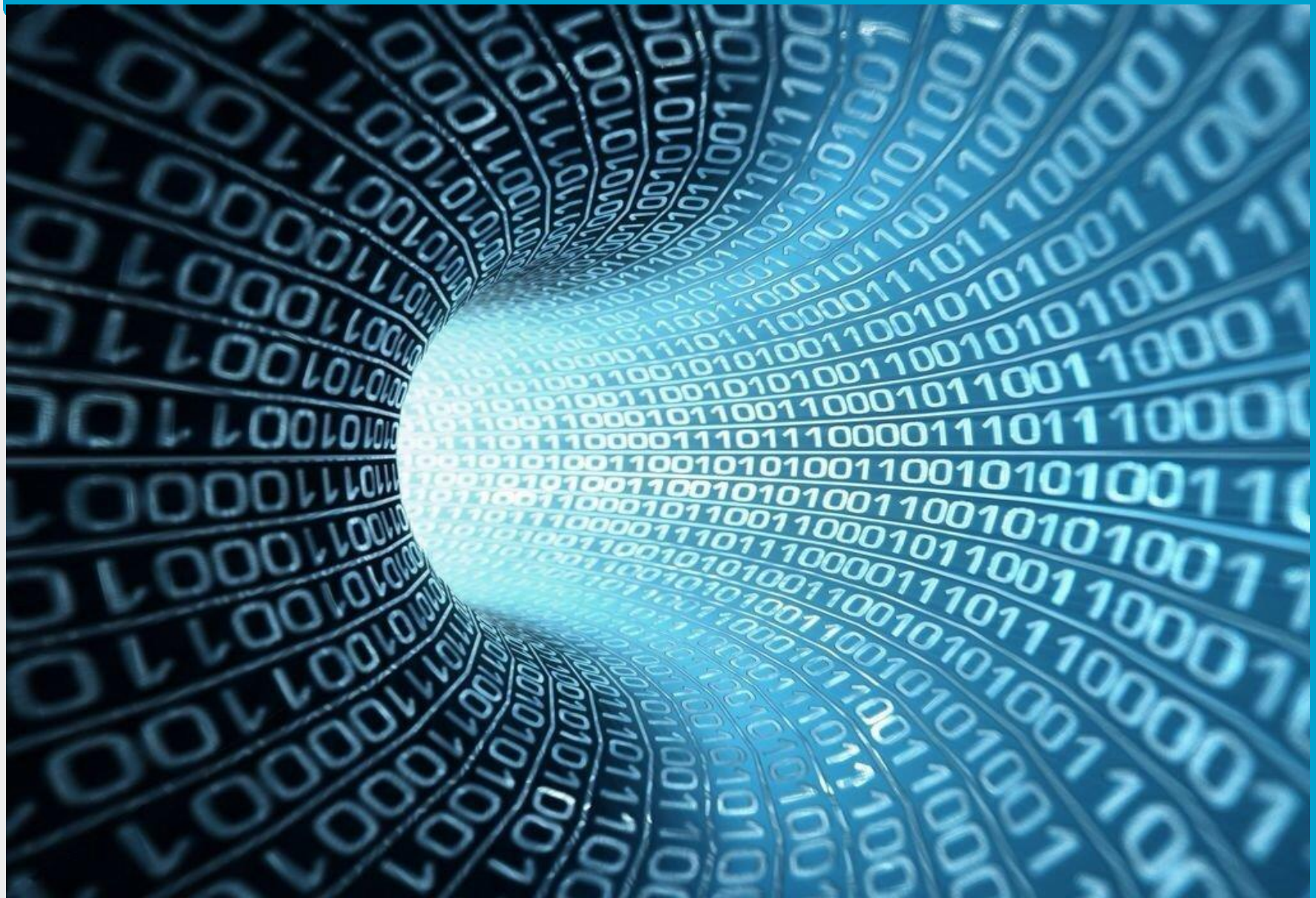
ICES V, Geneva, 23 June 2016

Statistics Netherlands

Outline

- Big Data and official statistics
- Experiences at Statistics Netherlands with:
  - use of road sensor data
  - use of public social media messages
  - use of mobile phone location data
- International examples
- The ESSnet Big Data
- Issues and solutions
- Strategic, policy and organisational challenges

What is Big Data?



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

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

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

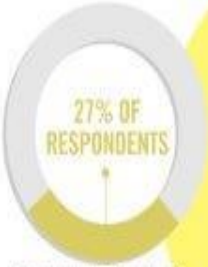
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 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



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

Big Data Characteristics

Definition:

- Volume
- Velocity
- Variety
- (Veracity)

Data characteristics:

- Unstructured data
- Selectivity
- Population dynamics
- Event data
- Organic data
- Distributed data

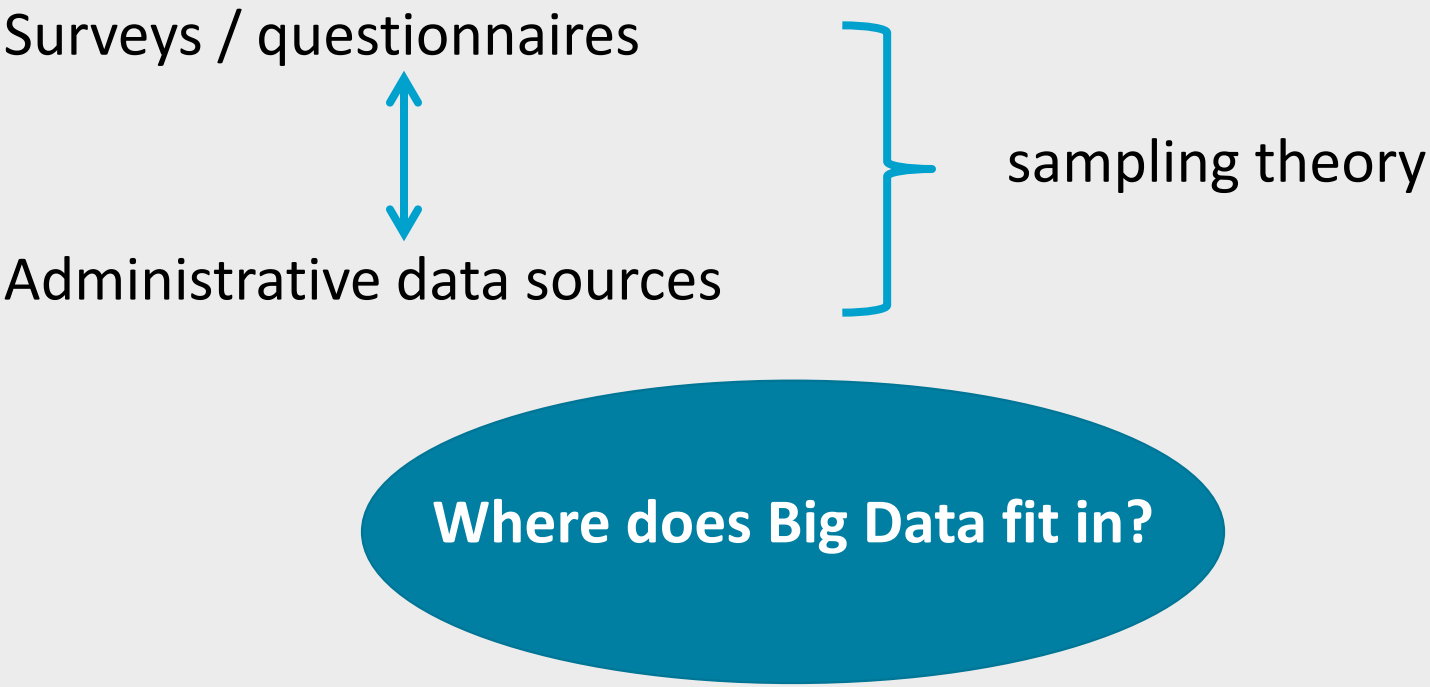
Data use:

- Other ways of processing
- Fundamentally new applications

Potential Opportunities

- New statistics
- More detailed statistics
- More timely statistics
- Nowcasts and early indicators
- Quality improvement
- Response burden reduction
- Cost reduction and higher efficiency

Data Sources and Approaches



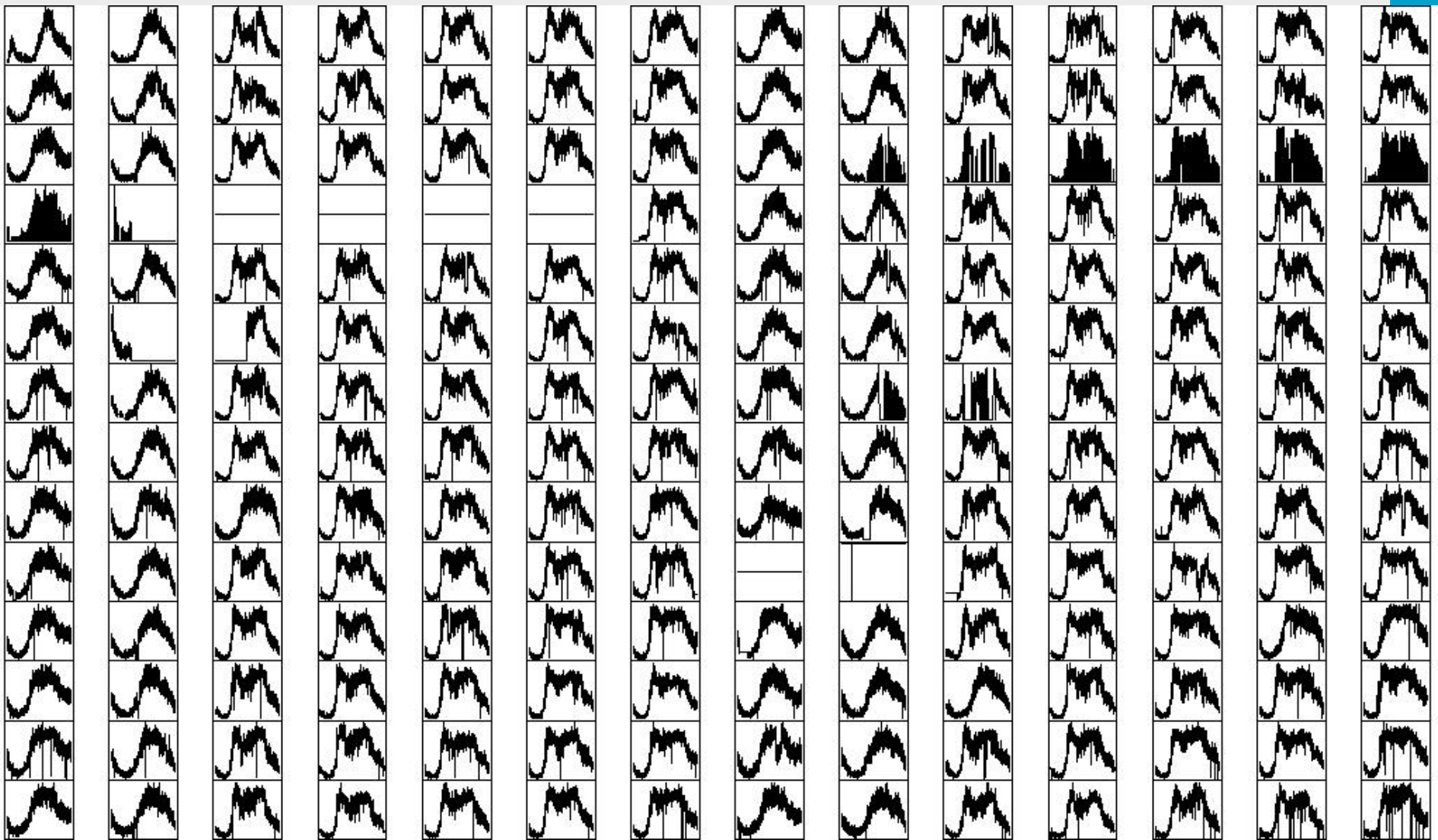
New methods may be needed, not based on sampling theory

1. 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)

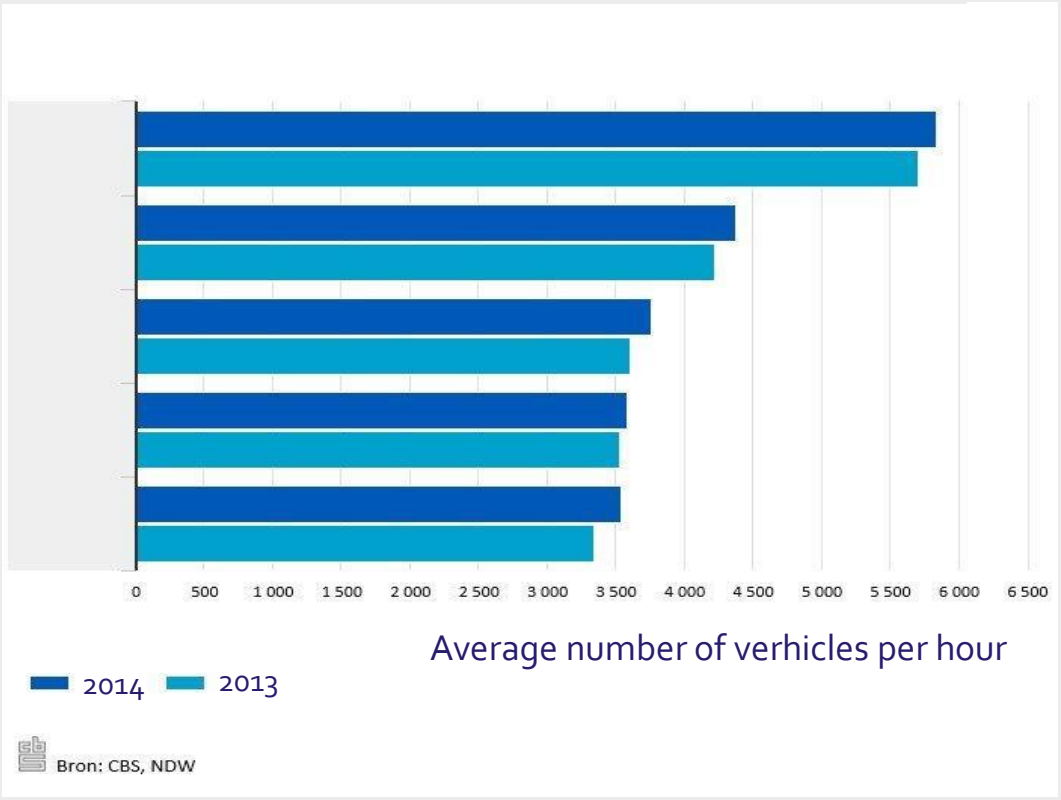
Locations

Minute Data of One Sensor for 196 Days

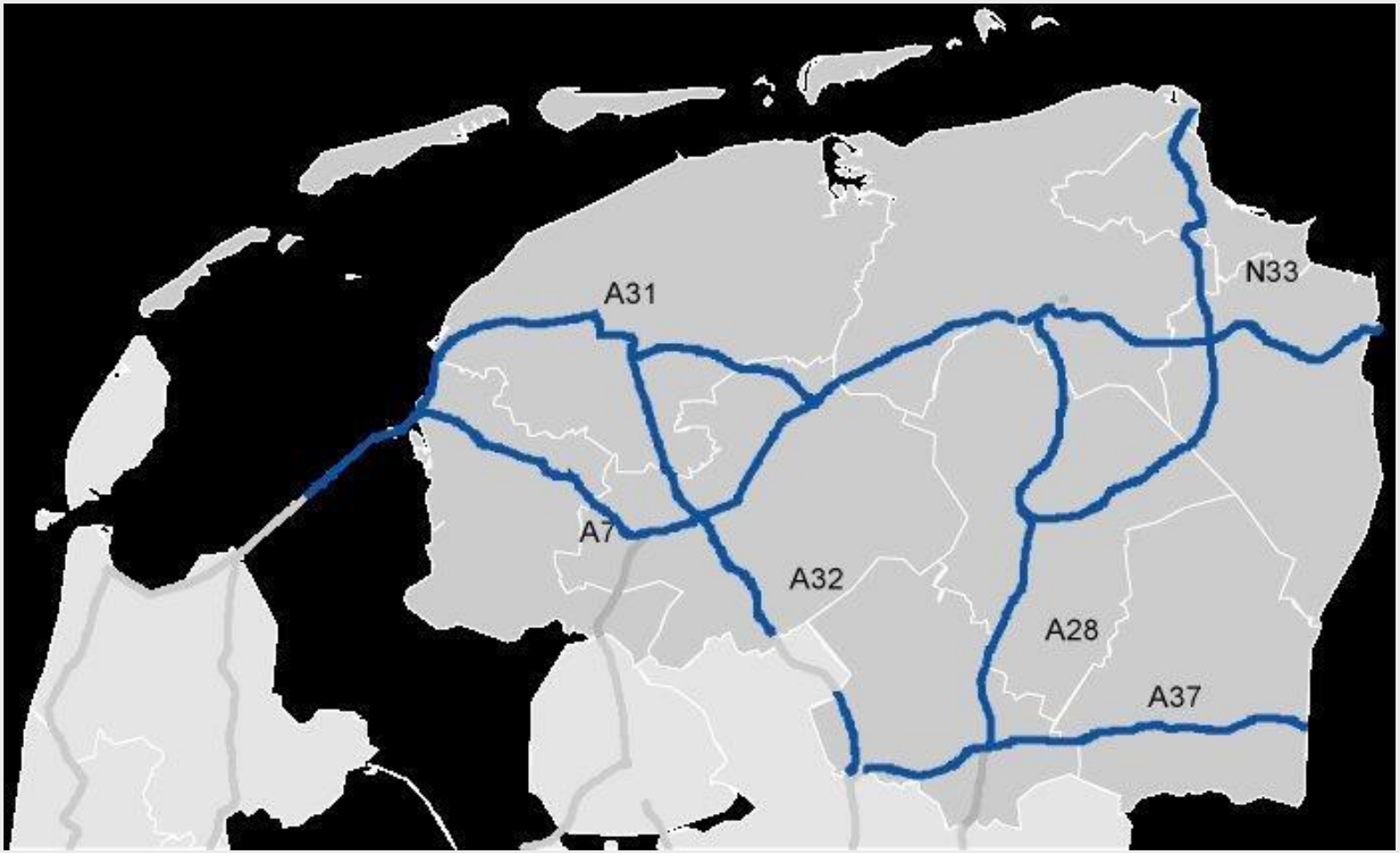


Road Sensor Data: Results

Top 5 traffic intensities on Dutch motorways

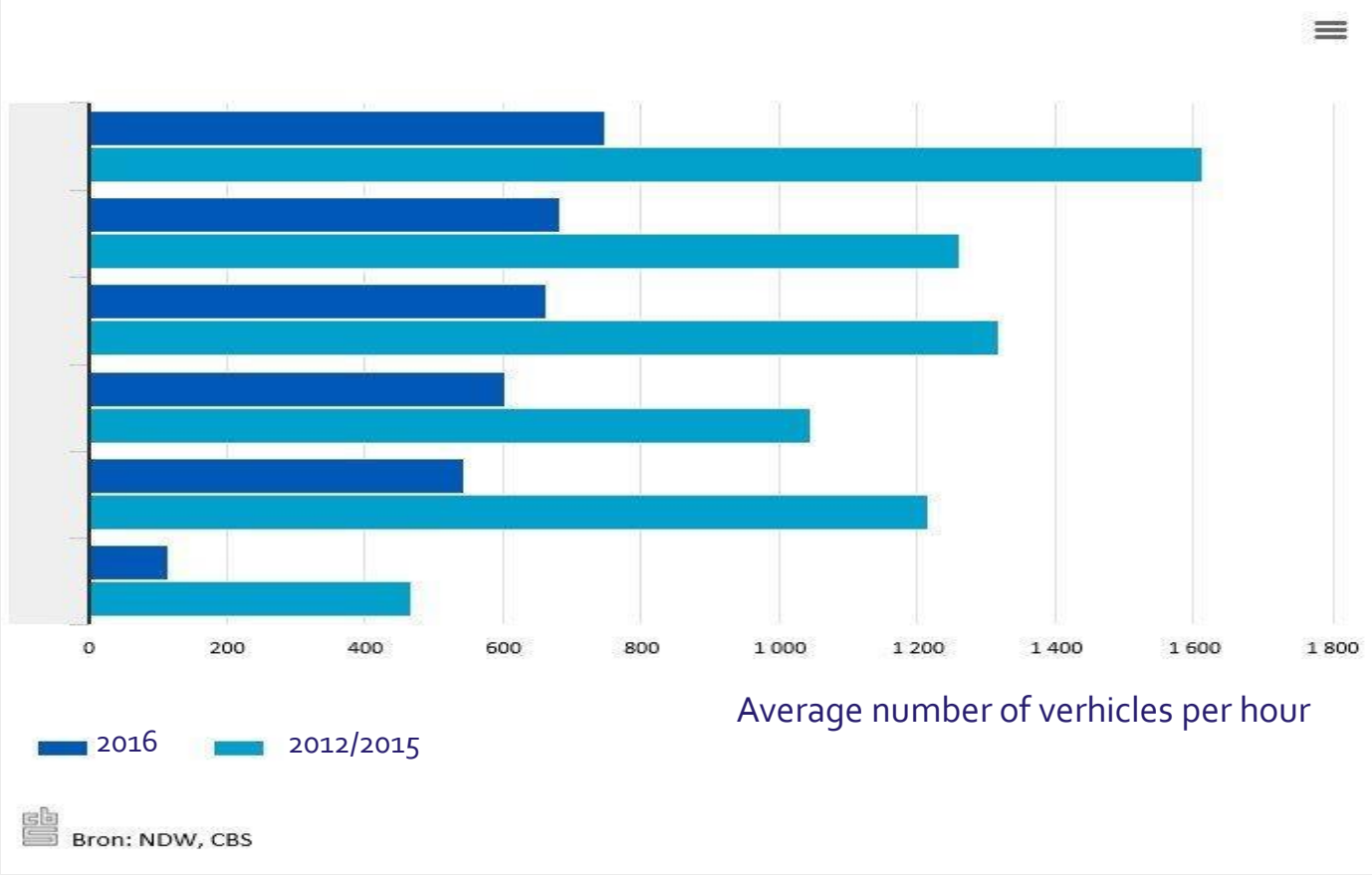


Frosted Roads at the Beginning of January…



… and a Press Release on 8 January!

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



Traffic Index Production Process

Big Data part (80 TB)

a

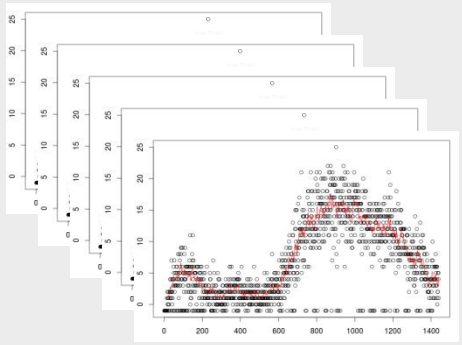
Frame

b

c

Estimation

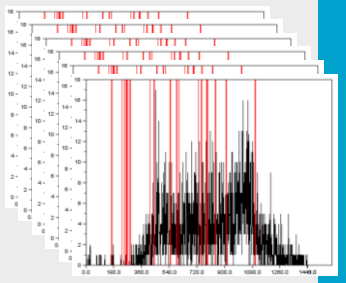
5



4

Editing

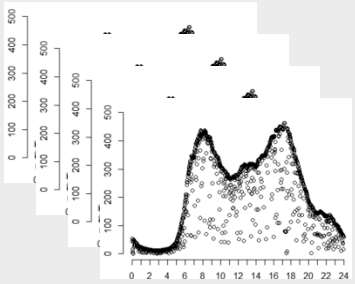
3



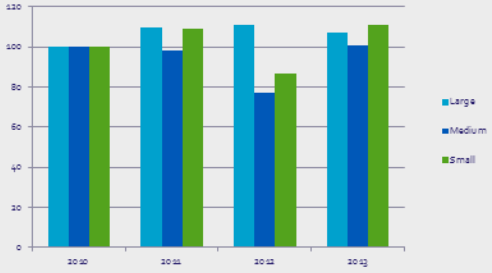
2

Big Data part (70 GB)

1



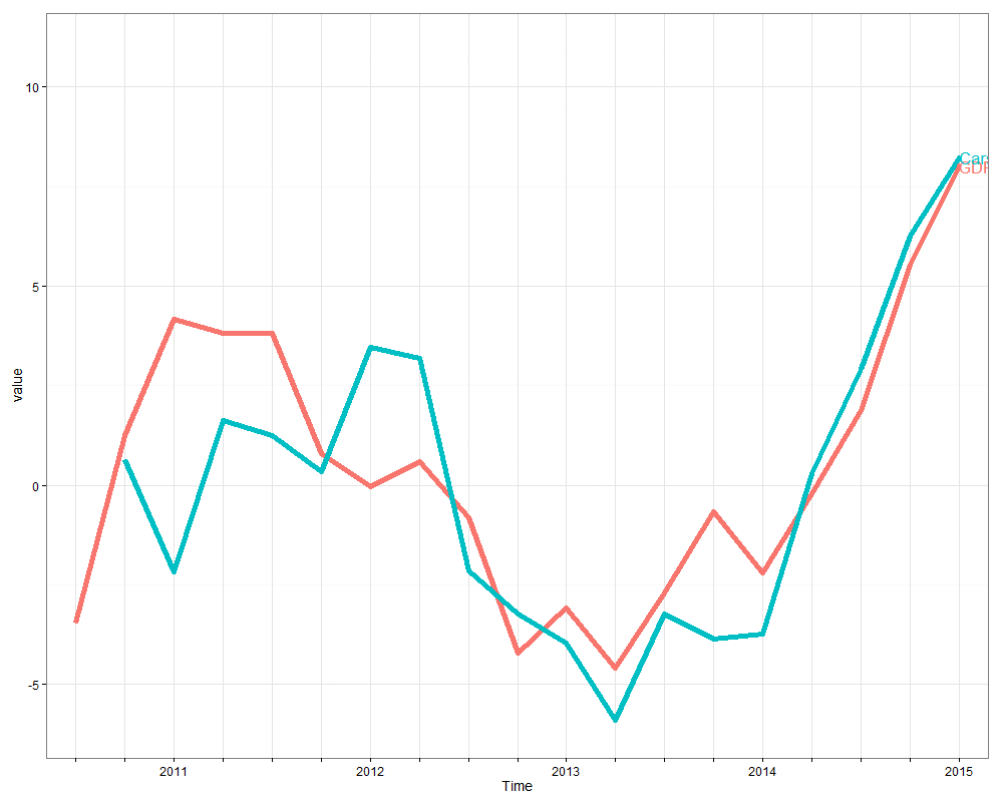
Select
+
Transform



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

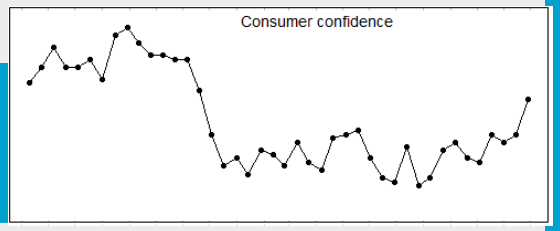


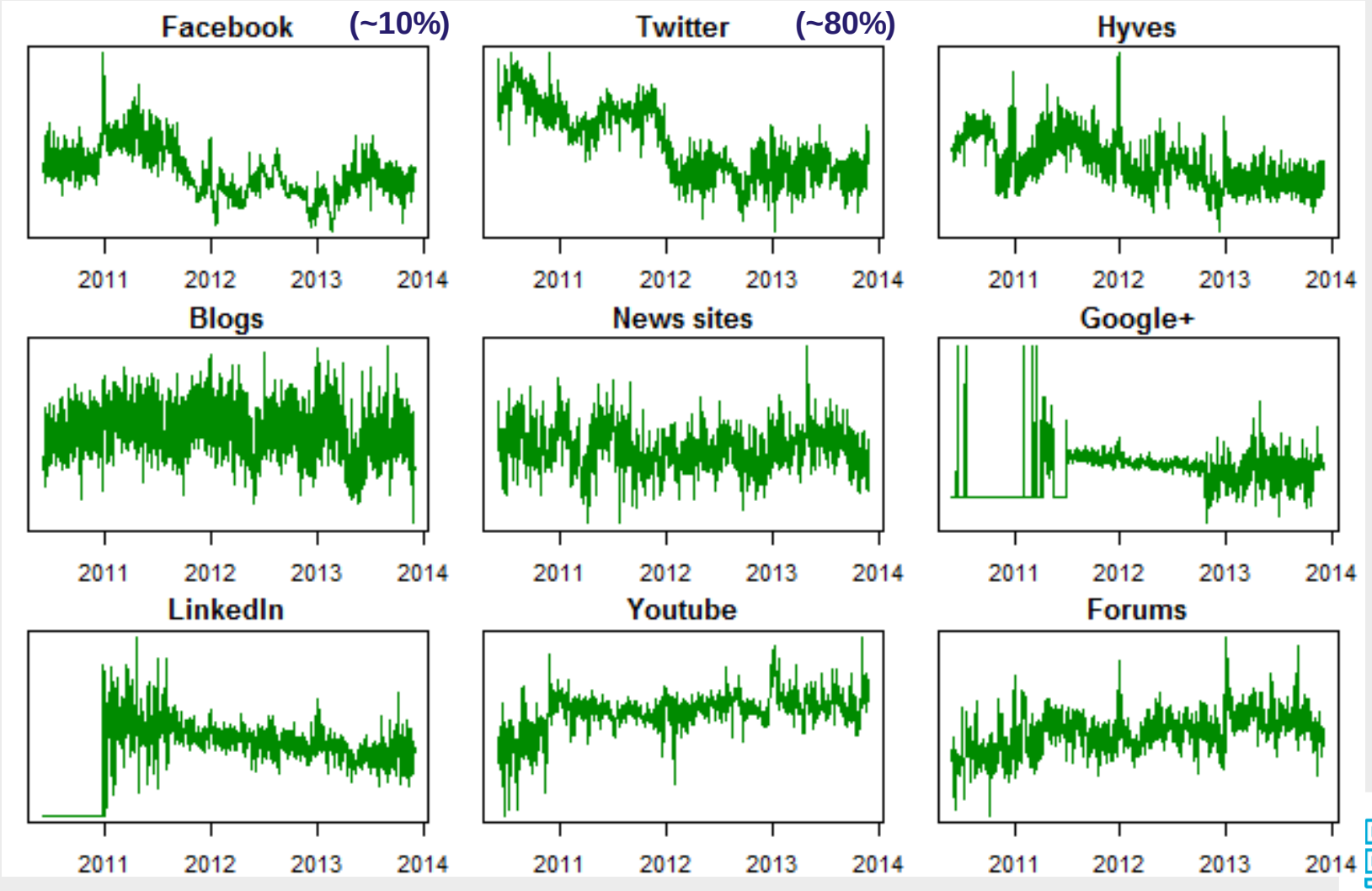
- GDP
- Traffic

2. 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

Sentiment per Platform





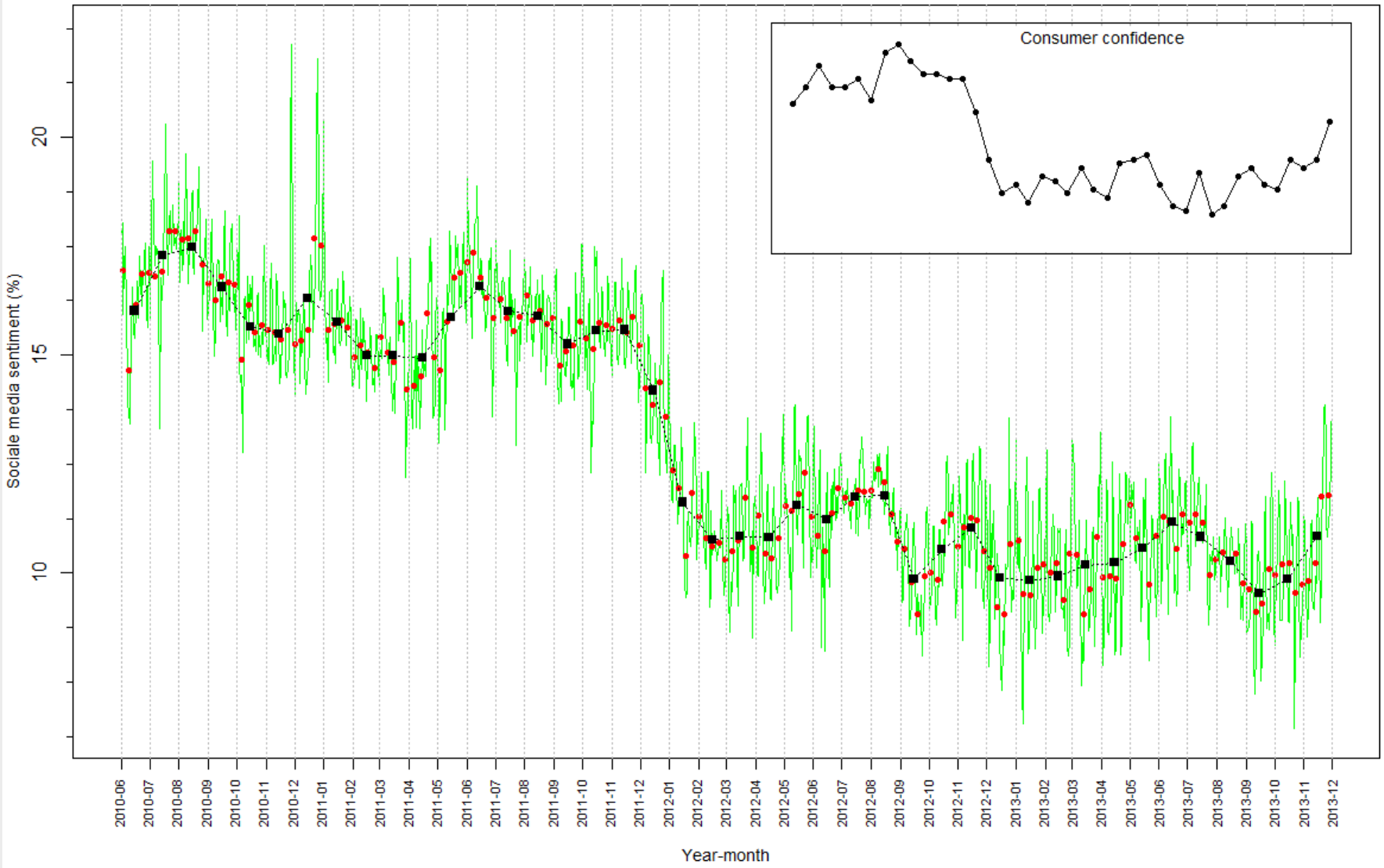
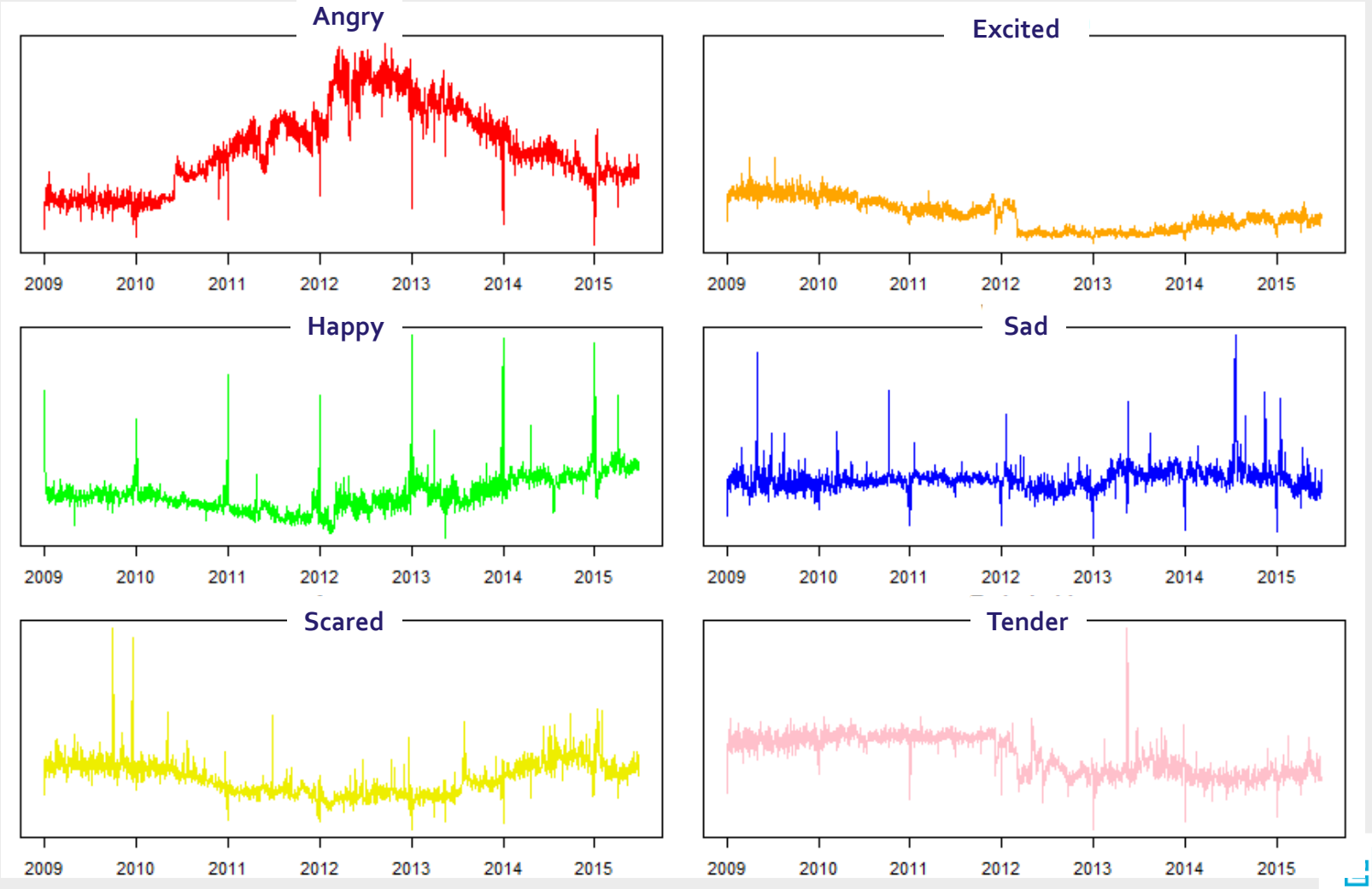


Figure 1. Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. In the insert the development of consumer confidence is shown for the identical period.

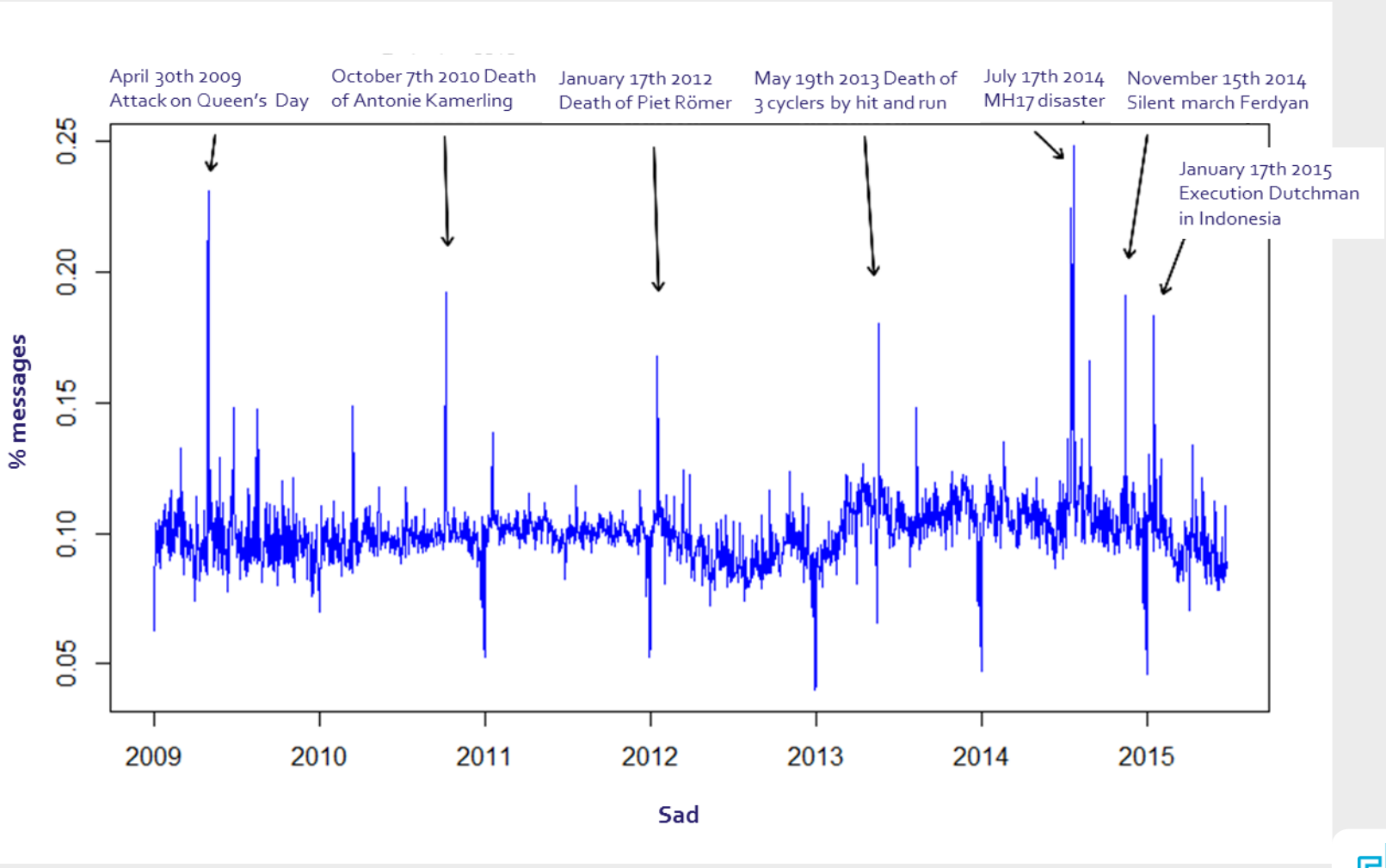
Basic Emotions in Social Media

Some basic emotions

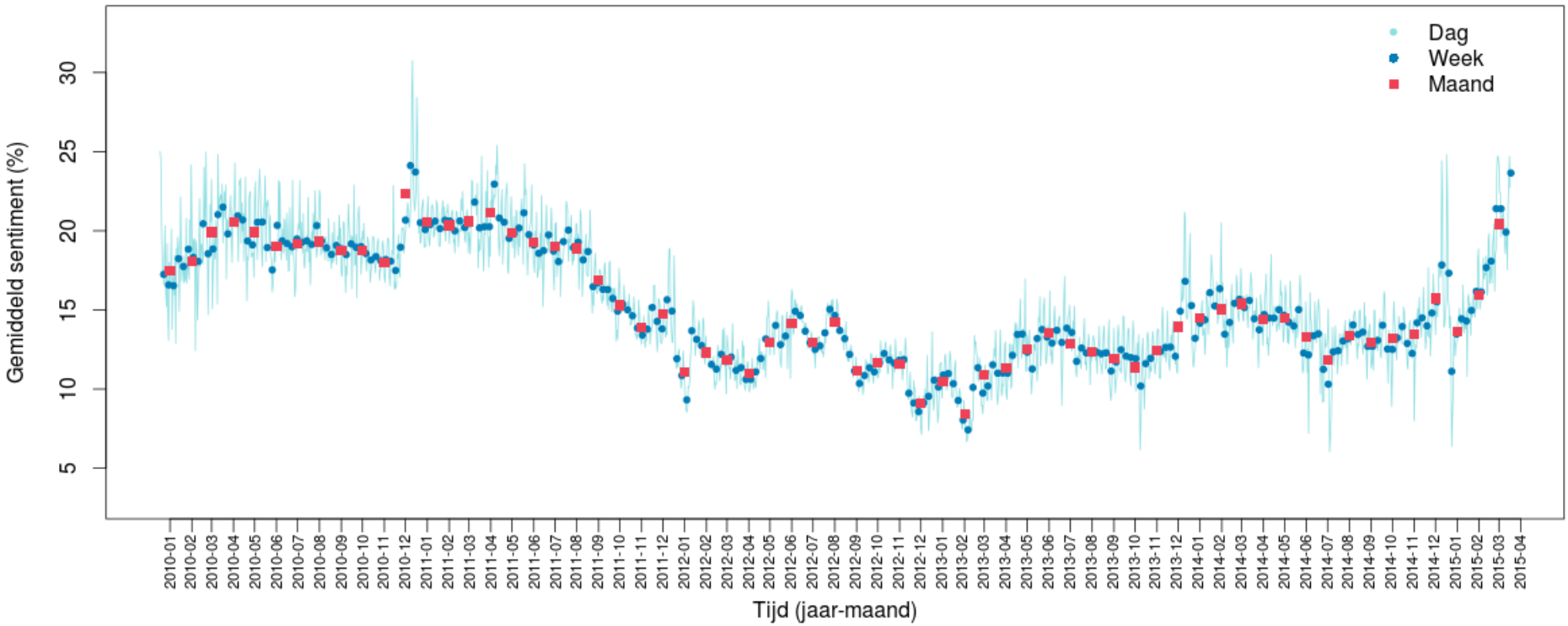
First Results



Sad


April 30th 2009
Attack on Queen's Day
October 7th 2010 Death
of Antonie Kamerling
January 17th 2012
Death of Piet Römer
May 19th 2013 Death of
3 cyclers by hit and run
July 17th 2014
MH17 disaster
November 15th 2014
Silent march Ferdyan
January 17th 2015
Execution Dutchman
in Indonesia
% messages
0.05
0.10
0.15
0.20
0.25
2009
2010
2011
2012
2013
2014
2015
Sad

Social Media: 'Sentiment' Indicator for NL



Based on the average sentiment of *public* Dutch Facebook and Twitter messages

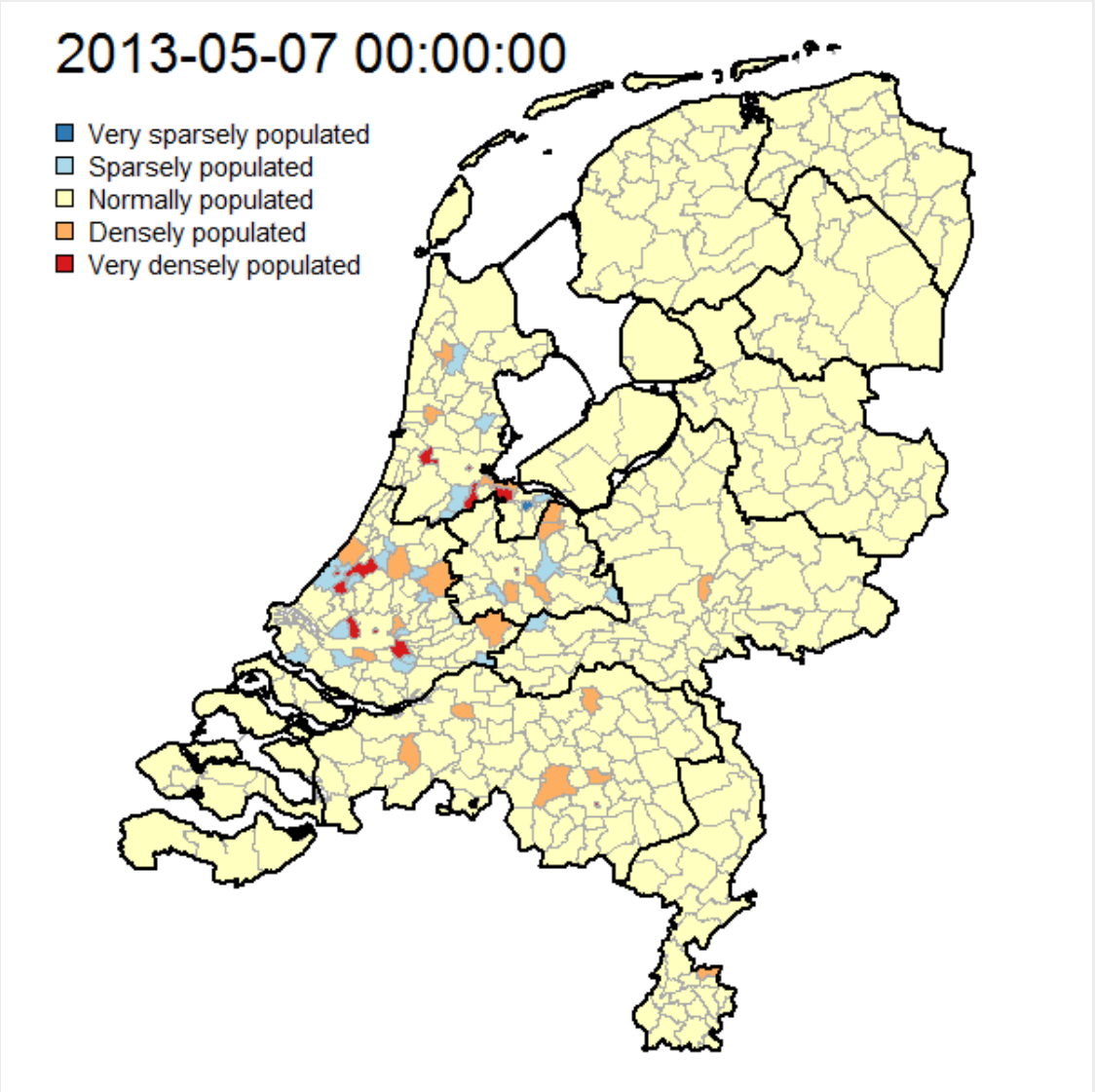
3. Mobile Phone Location Data

- 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
  - But the grid is continously changing

- Data of a single mobile company was used
  - Hourly aggregates per area
  - Confidentiality ensured (threshold of 15 events)
  - Algorithm supplied by Statistics Netherlands

Daytime Population Based on Mobile Phone Data



Scanner Data

- Use by Statistics Netherlands from 2002
- Currently: data from 6 supermarket chains
- Weekly > 120.000 product codes

Spring in the Netherlands

Flowering of the wood anemone

2013-02-24

2014-02-24

2013 2,5 mean 8 days below zero
2014 8,3 mean 0 days below zero

GWG Big Data Survey

- Global Working Group on Big Data for Official Statistics (UN) sent out a survey: “Global assessment of the use of Big Data for Official Statistics”

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

Projects as Reported to 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)

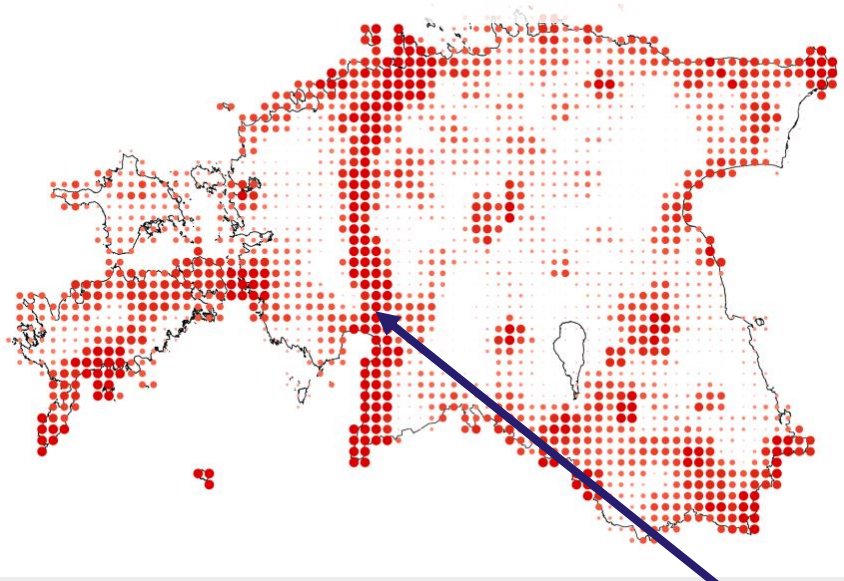
   Land use, crops, 'poverty' …

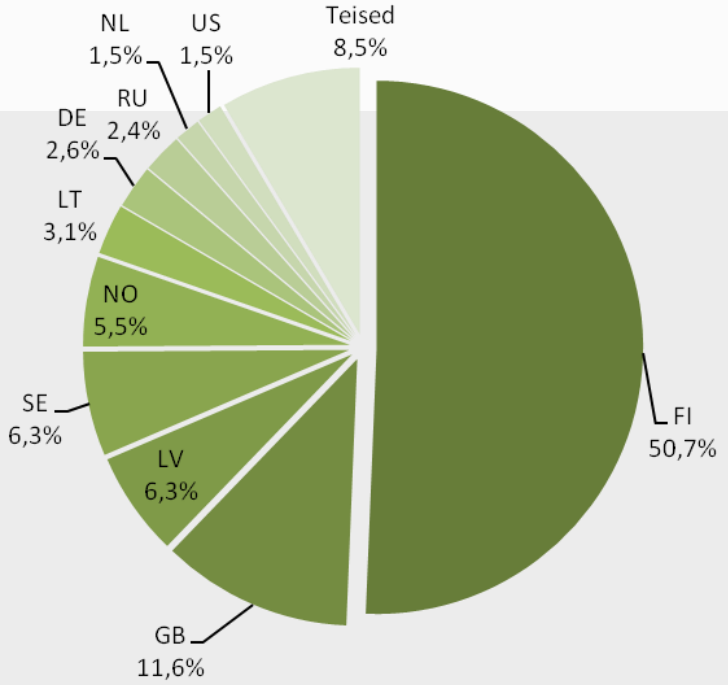
6. Other (46)

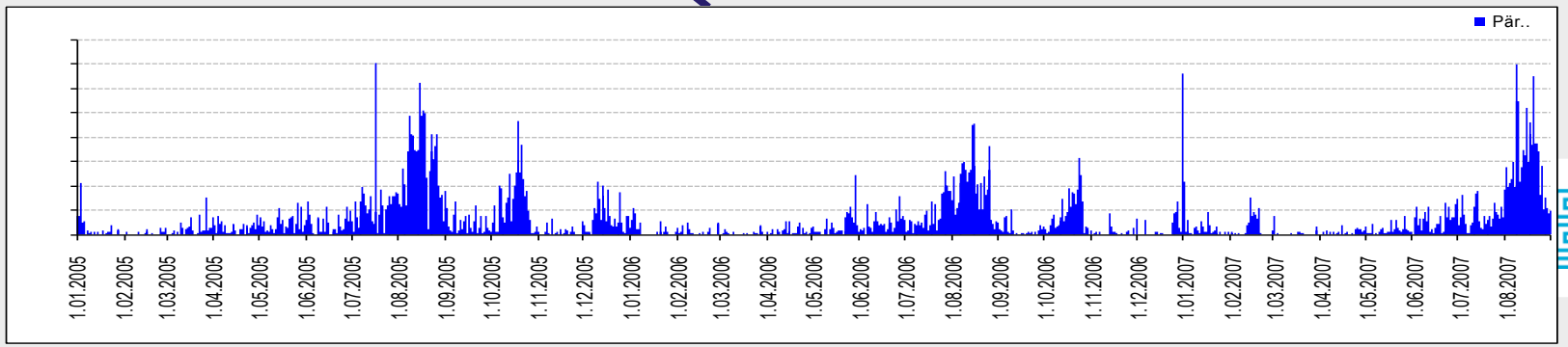
   Smart meters, transport (land, water), health, credit-card, patents …. Incl. admin data

Mobile Phone Data: Tourism

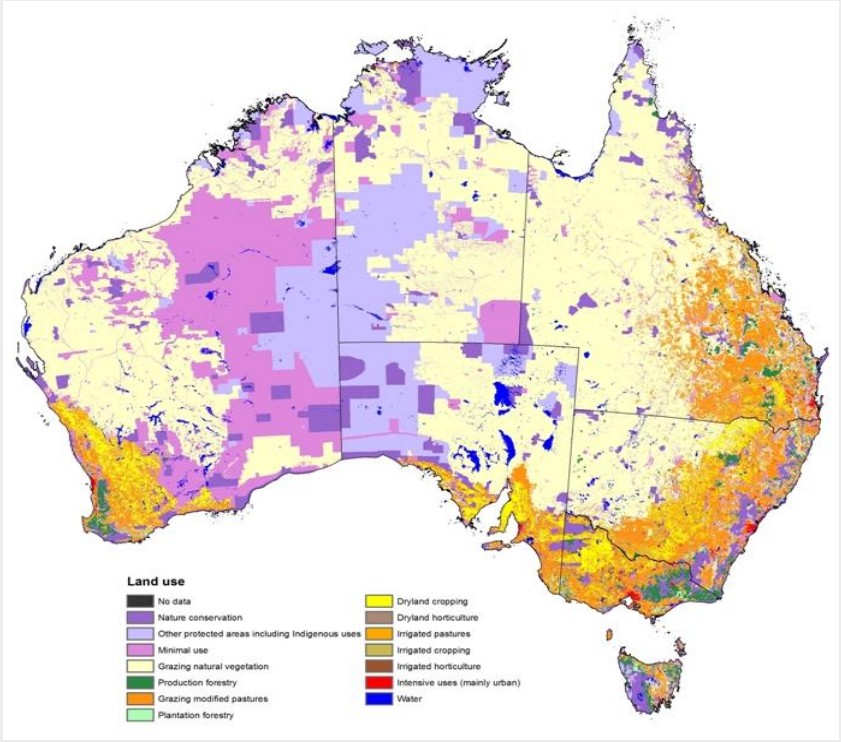
Quantitative tourism statistics Foreigners in Estonia





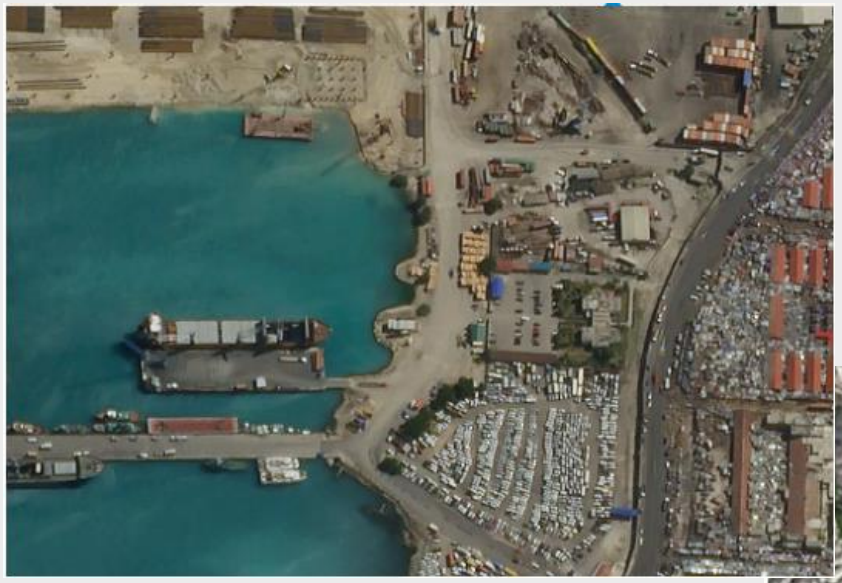


Satellite Data, Land Use/Crops



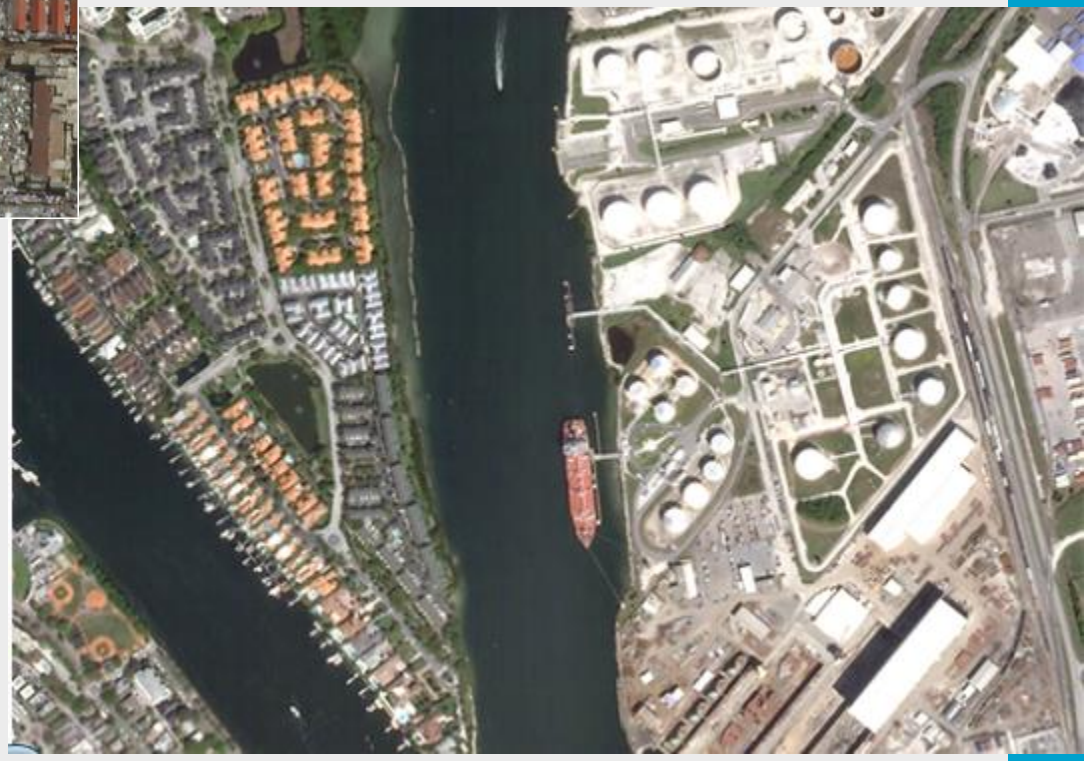
Use of Satellite data for official statistics on land use and crop estimation. Statistics Australia is one of the countries working on this topic.
(photo's are not of ABS and found on-line).

Satellite Data, Economy



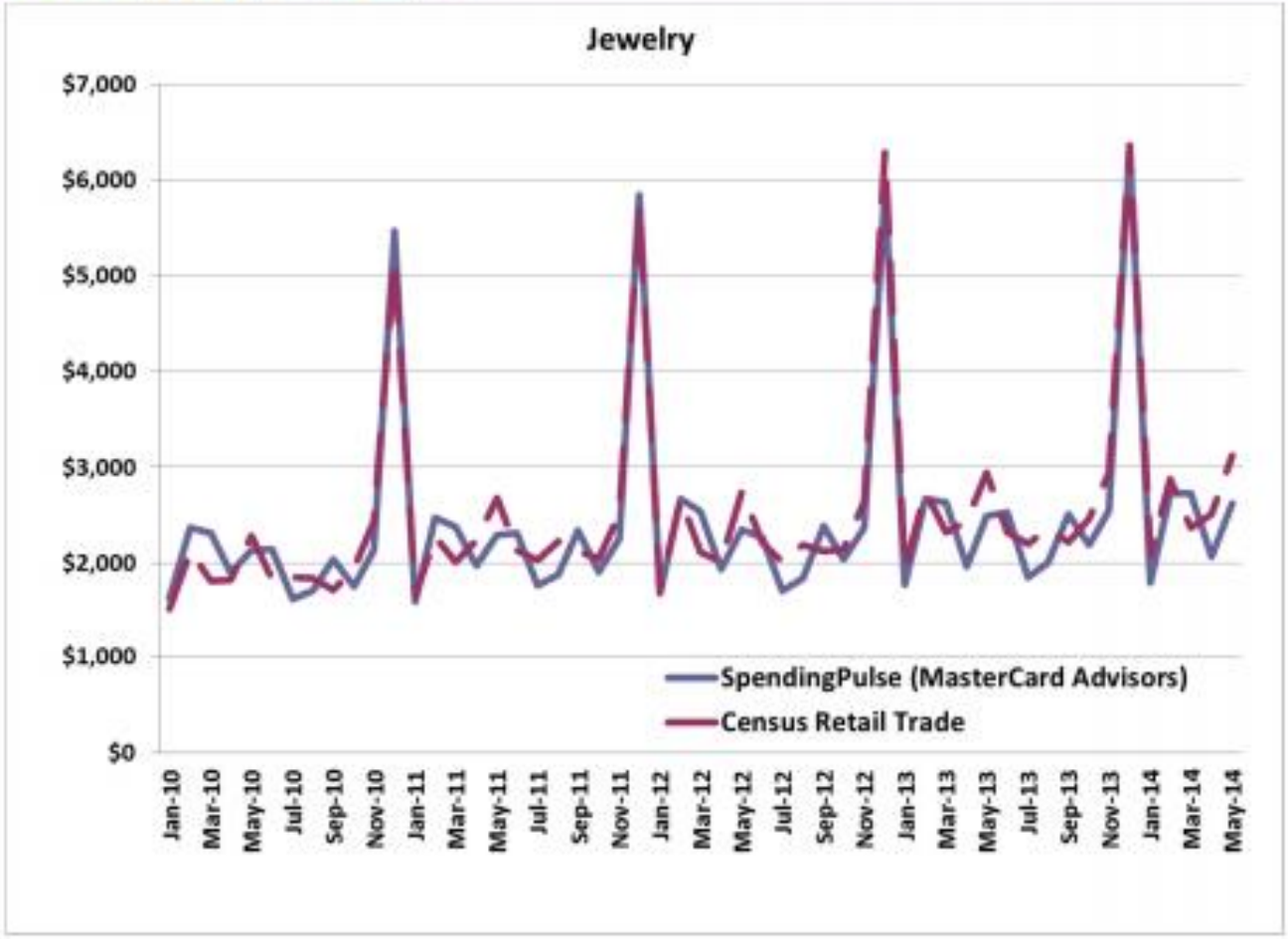
Watch containers and ships move

Watch filling of oil storage tanks change
First estimates were 70% accurate



Credit Card Data: BEA US

Estimates using Monthly Credit Card Data are similar to Retail Trade aggregates



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: beginning of 2017 – May 2018	0.7 M€

NB. still to be drafted and signed

Work Packages

WP0 Co-ordination

WP1 Webscraping / Job Vacancies
WP2 Webscraping / Enterprise Characteristics
WP3 Smart Meters
WP4 AIS Data
WP5 Mobile Phone Data

WP6 Early Estimates
WP7 Multi Domains

WP8 Methodology (foreseen for SGA-2)

WP9 Dissemination

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

WP1 Webscraping / Job Vacancies

WP leader: UK

Partners: Germany, Greece, Italy, Sweden, Slovenia

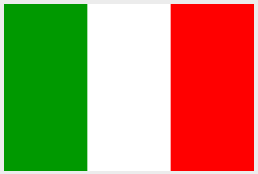
SGA-1:

- Data access: job portals
- Data handling: legal and technical aspects, test webscraping
- Methodology for output production: from semi-structured to structured data

SGA-2:

- Future perspectives: webscraping enterprise websites (?), methodology for future production, explore new products

WP2 Webscraping / Enterprise Characteristics

WP leader: Italy 

Partners: Bulgaria, Netherlands, Poland, Sweden, UK

SGA-1:

- Data access: inventory of target enterprises, URLs; legal and privacy aspects
- Data handling: use cases; actual webscraping
- Testing of methods and techniques (start): proof of concept for selected use cases

SGA-2:

- Testing of methods and techniques (continued): all use cases; build and apply predictor for estimates of enterprise characteristics

WP3 Smart Meters

WP leader: Estonia 

Partners: Austria, Denmark, Sweden

SGA-1:

- Data access: availability of smart meters, legal aspects
- Data handling: coverage assessment, production of cleaned datasets
- Methodology and techniques: linkage with administrative data; methodology for electricity consumption businesses and households; also seasonally vacant living spaces

SGA-2:

- Future perspectives: potential new products, feasibility of using aggregated data

WP4 AIS Data

WP leader: Netherlands 

Partners: Denmark, Greece, Norway, Poland

SGA-1:

- Data access: data availability (in particular EMSA)
- Data handling: processing and storage, aimed at linking with data from port authorities, traffic analyses, journeys
- Methodology and techniques (start): for linking with data from port authorities and traffic analyses

SGA-2:

- Methodology and techniques (continued): estimate emissions
- Future perspectives: qualitative cost-benefit analysis

WP5 Mobile Phone Data

WP leader: Spain 

Partners: Finland, France, Italy, Romania

SGA-1:

– Data access: data availability (workshop with MNOs)

SGA-2:

– Data handling: investigation of IT tools and aggregation level needed
– Statistical outputs: describe a statistical output to be presented to MNO to carry out a pilot

WP6 Early Estimates

WP leader: Slovenia 

Partners: Finland, Netherlands, Poland

SGA-1:

- Data access: sources for consumer confidence index, nowcasts of turnover and early estimates
- Data handling: technical requirements; deployment of collection system
- Methodology and techniques: includes feability of linking administrative and other existing sources

SGA-2:

- Future perspectives: calculation of the consumer confidence index and nowcasts of turnover; pilots for combining sources for early estimates

WP7 Multi Domains

WP leader: Poland

Partners: Netherlands, UK

SGA-1:

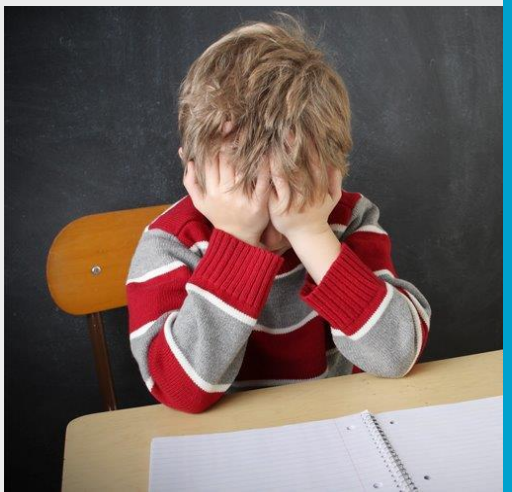
- Data access: data availability (inventory, based on questionnaire), aimed at three domains (populations, tourism / border crossings, agriculture)
- Data feasibility: exploration of combining sources for these domains

SGA-2:

- Data combination: experiments
- Future perspectives: suggest pilots for 2018

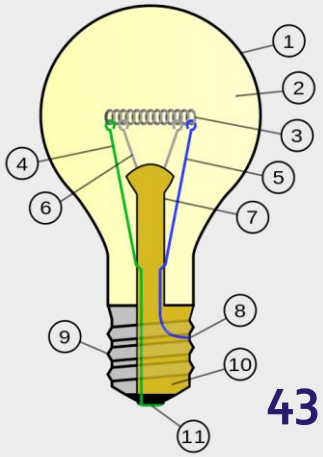
Overview of Issues

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



Possible Responses to the Issues

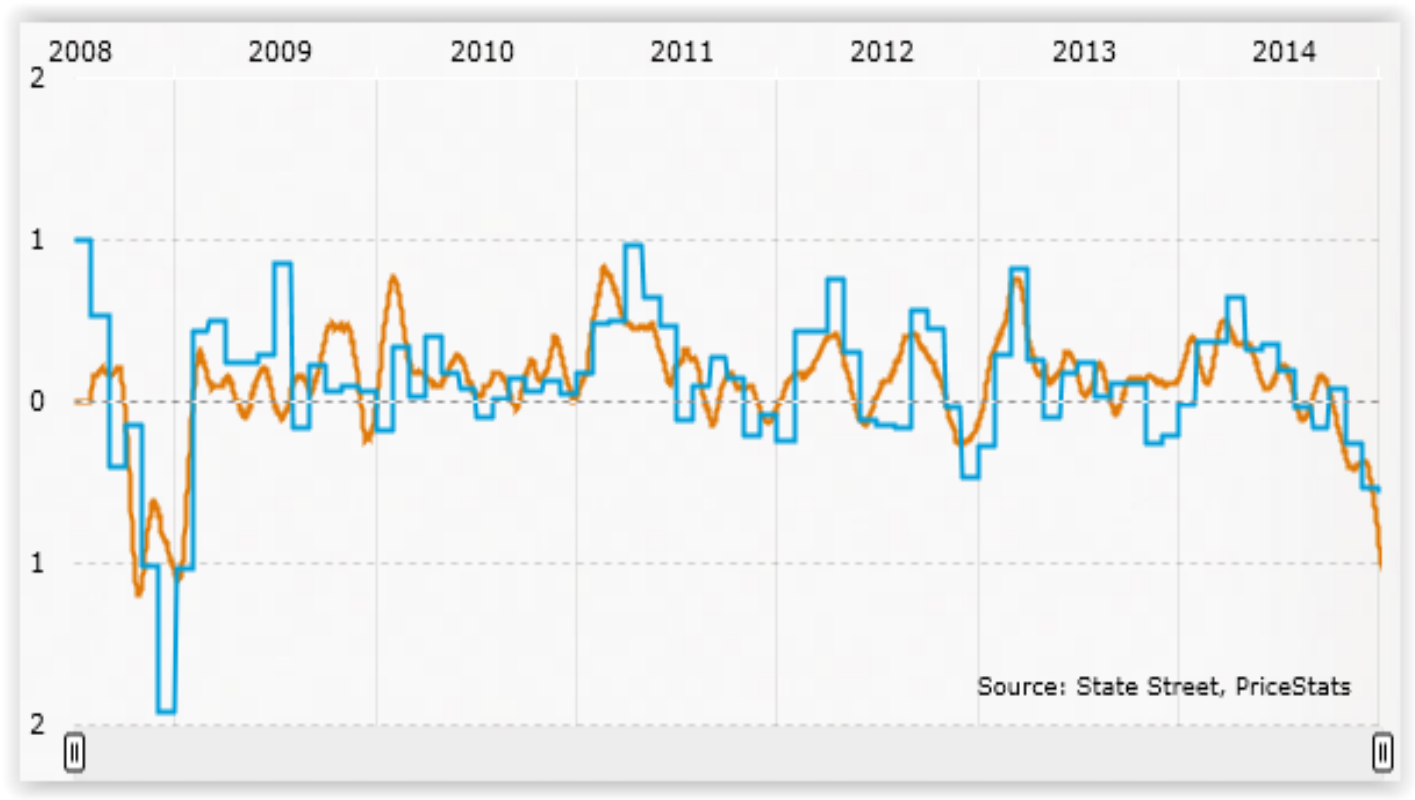
- Invest in good relations with the data provider
- Invest in methodological research and play with the data to get a grip on quality
- Use only aggregate data if possible
- Explore alternatives to population-based estimation methods
- Keep an open mindset
- Take the strategic challenges seriously



Strategic Aspects

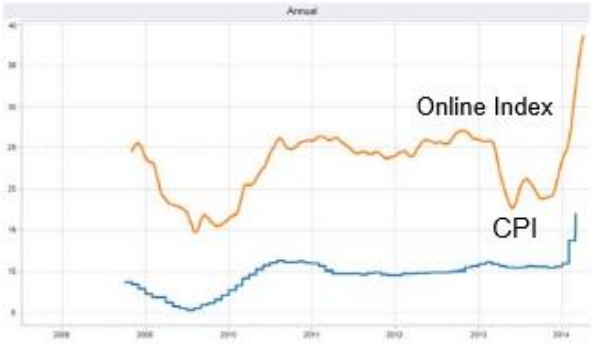
- Others start producing statistics
  - there may be quality issues
  - but they are extremely rapid
  - and there is obviously demand
- Need for good, impartial information (benchmark information) will remain
  - without a monopoly for NSIs
- There is a need for validation of information produced by others

Billion Prices Project MIT: USA

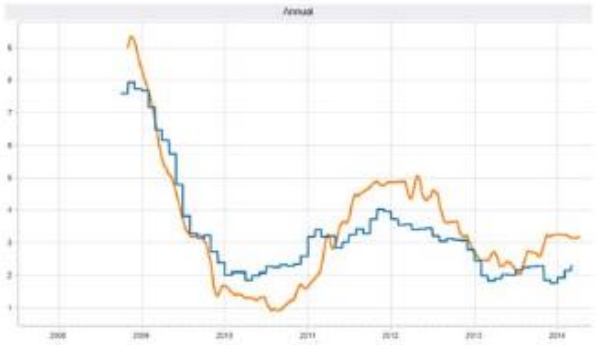


Annual Inflation Rates in Other Countries

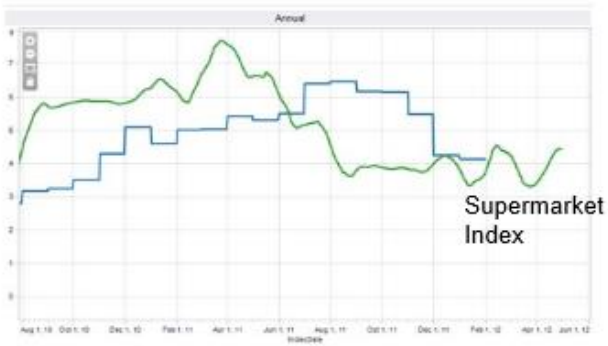
Argentina



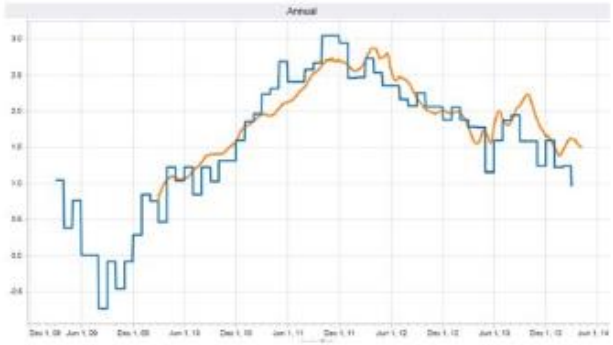
Colombia



China

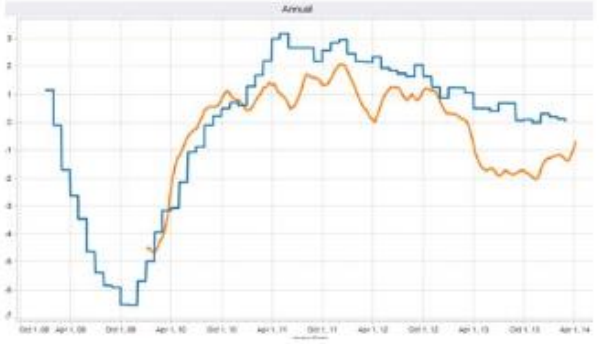


Germany

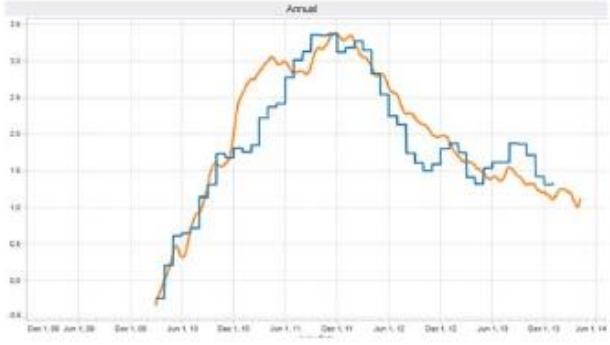


France

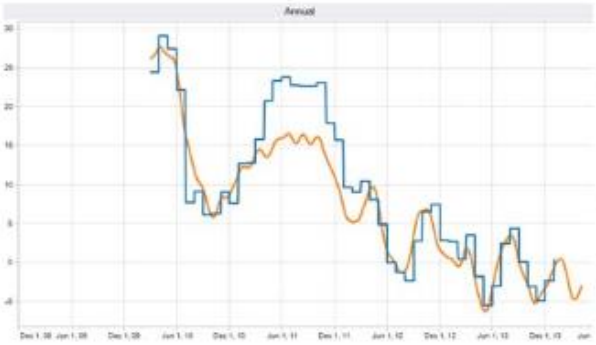
Ireland



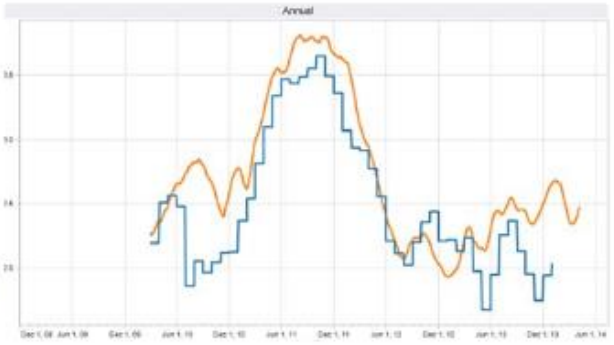
Developed Food



Developed Fuel



Global - Aggregate



Source: PriceStats – StateStreet Inflation Series


CHALLENGES
AHEAD


What
policy?

Roadmap Approach (Statistics Netherlands)

- 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

Supporting Programmes

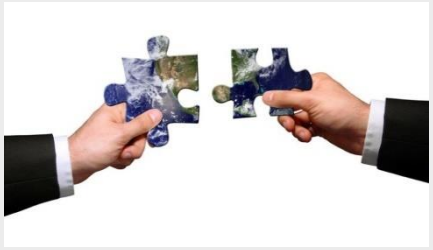
Big Data features in:

- Innovation programme
- Methodological research programme

Cooperation and Collaboration on Big Data

Statistics Netherlands works together with:

- Other NSIs
- UN, UNECE, EU, WorldBank
- ESSnet on Big Data
- Government organisations
- Universities and research organisations
- Data providers
- IT providers
- Big Data collecting firms
- Research consortia (e.g. H2020)



Conclusion: The Way Forward

- Get to know Big Data
- Use Big Data for:
  - new products / demands
  - early indicators
  - efficiency and response burden reduction
- Important to have showcases
- Use new professional methods where needed
- Create the right environment
- Don’t do it alone!

The Future

Questions?

Thank you for your attention!

p.struijs@cbs.nl