Big Data (and official statistics)

Piet Daas and Mark van der Loo*

Statistics Netherlands

* With contributions of: Edwin de Jonge and Paul van den Hurk

Overview

- What's Big Data?
  - *Definition and the 3 V's*
- Can Big Data be used for official statistics?
  - *Examples from Statistics Netherlands*
- Future challenges
  - *What has to change?*

What is Big Data?

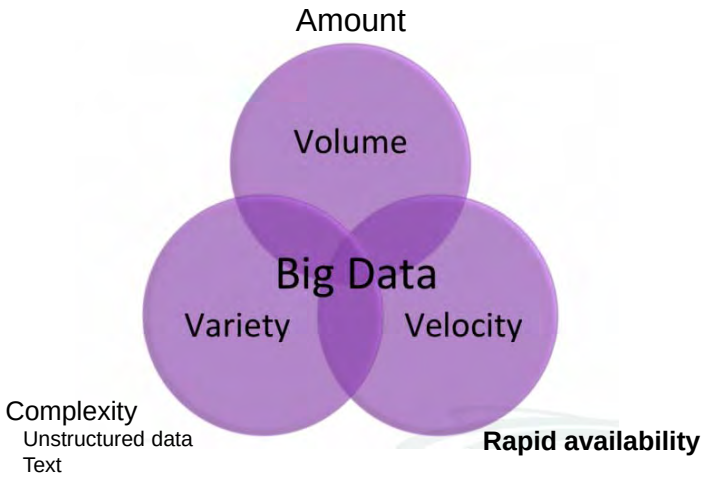
- According to a group of experts

  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."

- According to a user

  "*Data so big that it becomes awkward to work with*"

The most 3 important characteristics of Big Data



3 Big Data case studies

Can Big Data be used for official statistics?

Examples from Statistics Netherlands

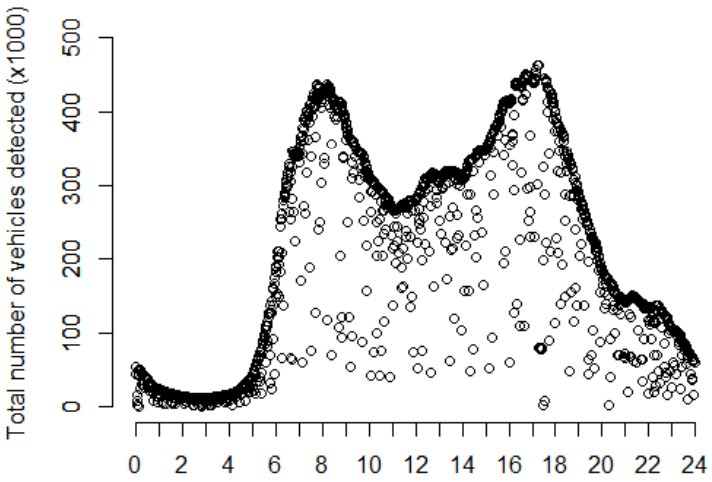
1. Traffic loop detection data (100 million records/day)
   - Traffic & transport statistics
2. Mobile phone data (35 million records/day)
   - Day time population, tourism
3. Dutch social media messages (1~2 million messages/day)
   - Topics and sentiment

1. Traffic loop detection data

- Traffic ‘loops’
  - Every minute (24/7) the number of passing vehicles is counted by >10,000 road sensors & camera’s in the Netherlands
    - Total vehicles and in different length classes
  - Interesting source to produce traffic and transport statistics (and more)
    - Huge amounts of data, about 100 million records a day

Locations

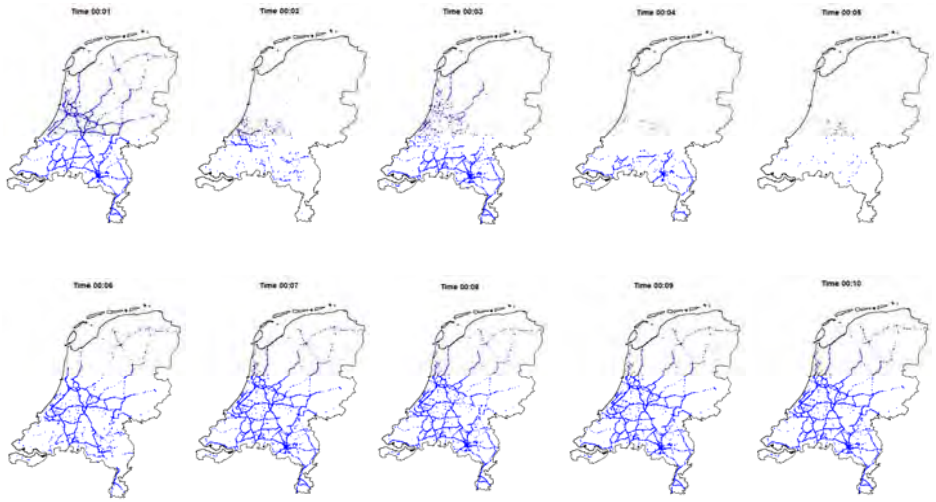
Number of detected vehicles on a single day



By all loops

Total = ~ 295 million

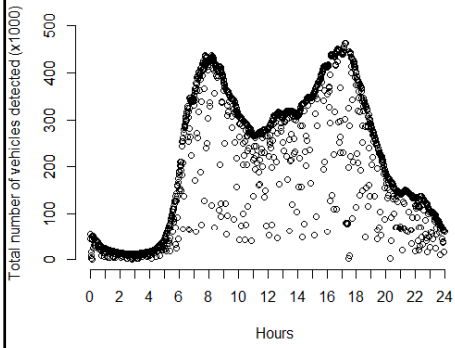
Traffic loop detection activity (only first 10 min.)



Correct for missing data

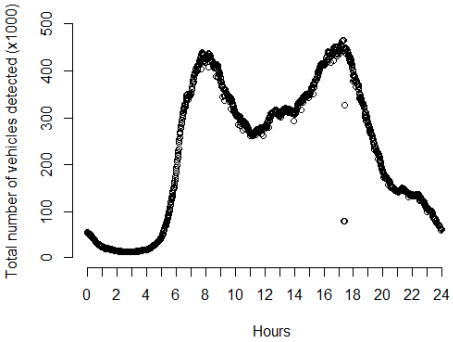
- ‘Corrected’ data (for blocks of 5 min)

Before



Total = ~ 295 million

After



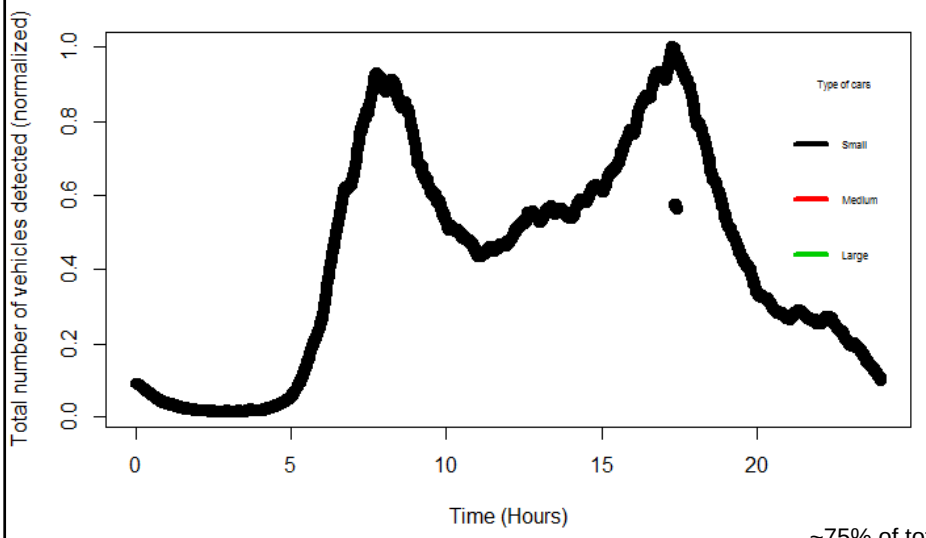
Total = ~ 330 million (+ 12%)

For different vehicle lengths

1 categorie	3 categoriën	5 categoriën
Totaal	Totaal	Totaal
	<= 5.6m	> 1.85 & <= 2.4m
	> 5.6 & <= 12.2m	> 2.4 & <= 5.6m
	> 12.2m	> 5.6 & <= 11.5m
		> 11.5 & <= 12.2m
		> 12.2m

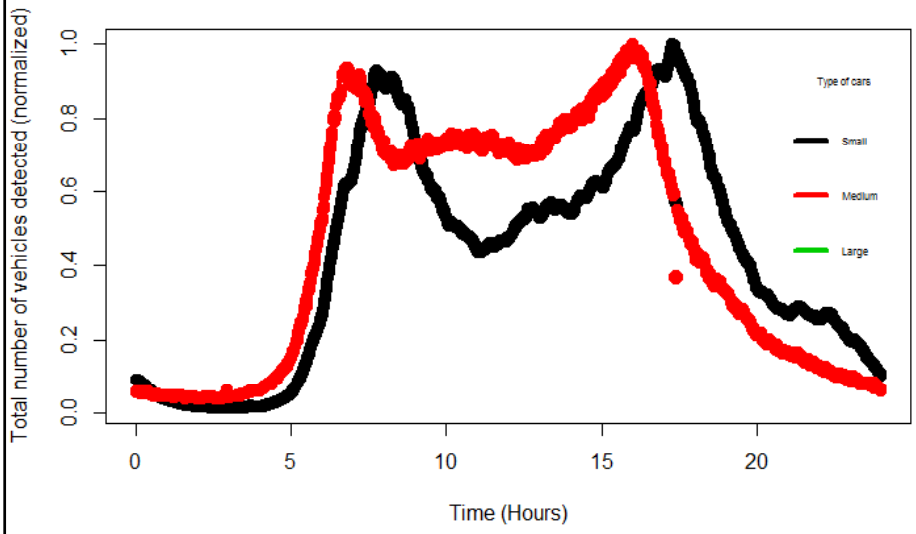
Small vehicles <= 5.6 m
Medium sized vehicles > 5.6 m & <= 12.2 m
Large vehicles > 12.2 m

Small vehicles

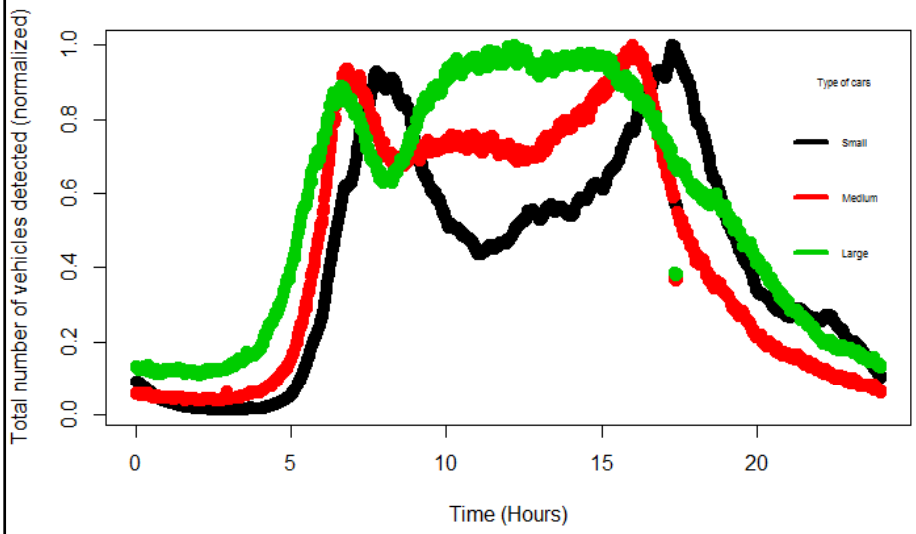


~75% of total

Small & medium vehicles



Small, medium & large vehicles



2. Mobile phone data

- Nearly every person in the Netherlands has a mobile phone
  - On them and almost always switched on!
    - An increasing number of people has a smart phone
- Ideal source of information to:
  - Use mobile phone data of mobile phone companies:
    - Travel behaviour (‘Day time’-population)
    - Tourism (new phones that register to network)
    - Crowd info (for example during events)

Travel behaviour of mobile phones

Mobility of very active
active mobile phone users
- during a 14-day period
- data of a single mob. company

Based on:
- Call- and text-*activity*
  multiples times a day
- Location based on phone masts

Clearly selective:
- Includes major cities
- But the North and South-east of the country much less

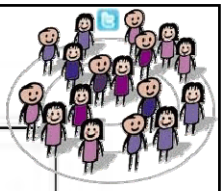
3. Social media messages

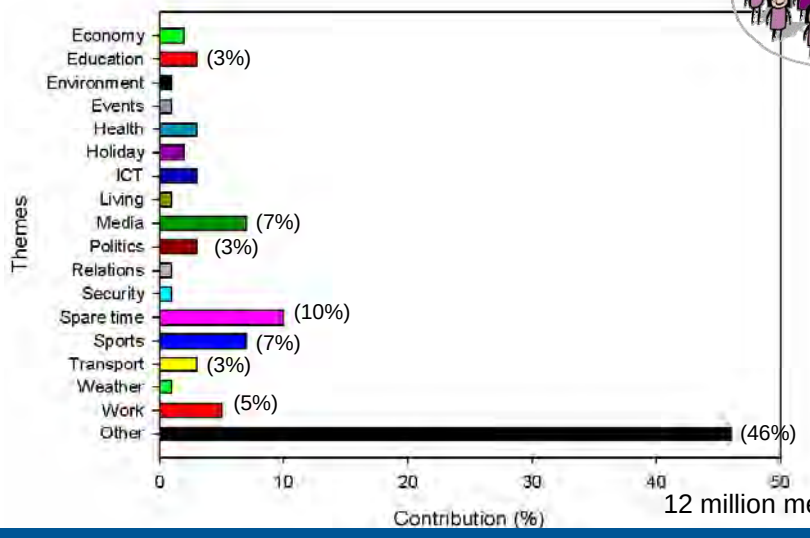
- Dutch are very active on social media platforms
  - Potential information source for:
    - Topics discussed and sentiment over these topics (quickly available!) and probably more?
    - Investigate it to obtain an answer on its potential use

3a. Content:
- Collected Dutch Twitter messages for study: ‘selection’ of 12 million

3b. Sentiment
- Sentiment in Dutch social media messages: ‘all’ ~2 billion

Social media: Dutch Twitter topics





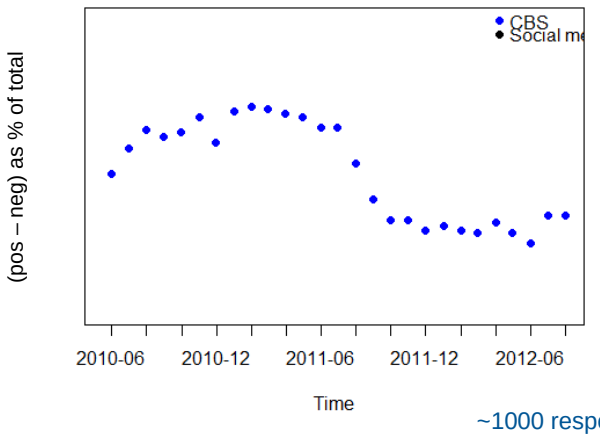
12 million messages

Sentiment in Social media

- Access to Coosto database
  - ~ 2 billion publicly available messages
    - Twitter, Facebook, Hyves, Webfora, Blogs etc.
  - Sentiment of each message
    - Positive, negative or **neutral**
  - Interesting finding
    - Looked at so-called ‘Mood of the nation’ compared to Consumer confidence of Statistics Netherlands

Consumer confidence, survey data

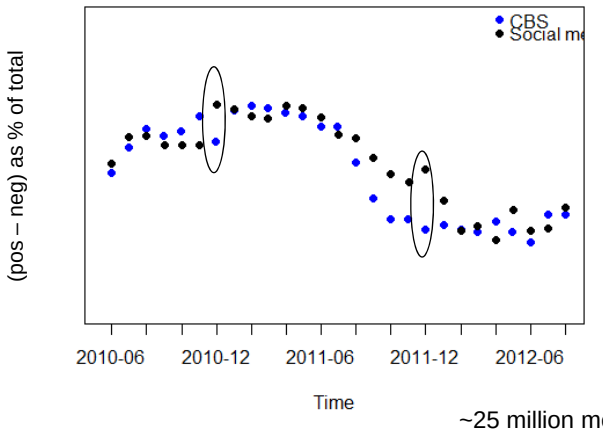
Sentiment towards the economic climate



~1000 respondents/month

Sentiment in social media messages

Sentiment towards the economic climate &
Social media message sentiment



Corr: 0.88

~25 million messages/month

Challenges: Big Data and statistics

- Legal
  - Is access routinely allowed (not only for research)?
- Privacy
  - With more and more data, privacy demands increase
  - We have to be careful here!
- Costs
  - In the Netherlands we don't pay for admin data.
  - Should we pay for Big Data?
- Manage
  - Who owns the data? Stability of delivery/source
  - Because of its volume, run queries in database of data source holder

Challenges: Big Data and statistics (2)

- Methodological
  - Big data sources register events, not units, and they are selective!
  - Methods & models specific for large dataset (fast and 'robust')
  - Try to 'make big data small' ASAP (noise reduction)
- Technological
  - Learn from 'computational statistical' research areas
  - High Performance Computing needs, parallel processing
- People
  - Need 'data scientists' (statistical minded people with programming skills that are curious)
  - That are able to think outside the traditional sample survey based paradigm!


The data deluge