Big Data in Official Statistics

MAY 26, 2023

Prof. dr. Piet J.H. Daas

Mathematics and Computer Science, Statistics

Official Statistics

Official statistics are statistics that are produced by governmental agencies or other public bodies, such as Statistics Netherlands (in Dutch: ‘Centraal Bureau voor de Statistiek’).

Official statistics provide an indispensable element in the information system of a democratic society by serving it with data about important themes, such as the economic, demographic, social, and environmental situation of a country.

What data sources are used to produce Official Statistics nowadays?

(Big) Data in Official Statistics

Currently official statistics are predominantly based on: survey data (samples), administrative data (registers), or a combination of both. There is another source of data available, called Big Data, that has great potential.

- Survey data: Data collected by surveys (often questionnaires). Under control of the NSI.
- Administrative data: Data collected by other, often governmental organizations. Delivery arrangements have to be made. Not under direct control.
- Big Data: new data sources, with great *potential*. But what is it exactly?

Big Data, what is it?

Definition of Big Data within an official statistical context:

> *Big data are usually (extremely) large datasets that can contain both structured and unstructured data and that, when analyzed computationally, may reveal patterns, trends, and associations relating to the behavior and interactions of the units included.*

Potential advantages of Big Data are: Speed, More detail, Newness, and Burden reduction

How can Big Data be used in official statistics?

Big Data-based statistics either in production or experimental (1)

No.	Product	Big Data source(s) used	Off/Exp	Countries	Freq
1	Consumer Price Index (inflation index)	Scanner data, web-scraped prices (± surveys)	O	Multiple	M (2W)
2	Biodiversity trends (incl. butterfly index)	Internet observations (+ survey)	O	NL, EU	Y
3	Traffic intensity statistics	Road sensors	O	NL	M
4	Internet economy	Websites (+ admin data)	O	NL	Y
5	Online platform statistics	Websites (+ survey)	O	NL	Y
6	Land use, crop/vegetation detection	Satellite/aerial pictures (± admin data)	O/E	CA,AU/Multiple	Y (Q)
7	Public transport monitor	Public transport smart card (+ survey, admin)	E	NL	Once
8	Mobility patterns (during COVID)	Mobile network operating data	E	Multiple	D
9	Job vacancy/advertisement statistics	Online job ads	E	Multiple	M
10	Enterprise characteristics	Websites	E	Multiple	Y
11	Daytime population/commuting stat.	Mobile phone data, transport data	E	Multiple	M
12	Innovative tourism statistics	Multiple sources (e.g., websites, road sensors)	E	Multiple	M
13	Social media sentiment	Social media messages	E	Multiple	D,W,M
14	SDGs (incl. urbanization)	Satellite/aerial pictures	E	EU,UN	Once
15	Electricity/energy consumption	Smart meter data	E	ES,DK,NO,UK	D
16	Maritime and inland waterway stat.	Automatic identification system data	E	NL,GR,PL,UN	Once
17	Innovative company detection	Websites	E	NL,DE,BE	Once
18	Outbound tourism statistics	Mobile network operating data	E	ES,FI,AT	Once
19	Solar panel detection	Aerial pictures	E	NL,BE,DE	Once
20	Suicide numbers/mental health index	Weblogs, Twitter	E	KR,ID	D
21	Accommodation statistics	Data from most important booking platforms	E	EU	M,Y
22	World Heritage sites popularity	Wikipedia page views	E	EU	Once
23	Social mood on economy Index	Twitter messages	E	IT	D
24	Social unrest indicator	Social media messages	E	NL	D
25	Retail sales index	Debit card transaction data	E	NO	Once
26	Road accidents	OpenStreetMap (+ population, vehicle fleet)	E	IT	Once

Big Data-based statistics either in production or experimental (2)

No.	Product	Big Data source(s) used	Off/Exp	Countries	Freq
27	Caravan home identification	Real estate websites (+ address register)	E	UK	Once
28	Travel flows	Oyster card data (+ census data)	E	UK	Once
29	Determining residence and mobility	Twitter data	E	UK	D
30	Effect of ships on underwater life	Automatic identification system (+ geo) data	E	NL	Once
31	Monthly postal trade statistics	Postal receptacle identifiers	E	UN	M
32	Monthly global trade statistics	Global trade data	E	UN	M
33	Trade volume nowcasts	Automatic identification system data	E	UN	D
34	Air transport index (during COVID)	International civil aviation organization data	E	UN	M
35	Migration indicator	Mobile phone data, social media data	E	UN	Q,Y
36	Displacement and disaster statistics	Mobile phone data	E	UN	Event,Y
37	Information society statistics	Mobile phone data, internet connection speed	E	UN	Y
38	Early economic indicator (spending)	Payment card transaction data	E	US	D
39	Economic sentiment index	Google Trends data	E	CH, IMF	D
40	Poverty/income/demography indicator	Google Street View images	E	Multiple	Once
41	Environmental statistics, water	Geospatial and earth observation data	E	UN	Once
42	Skills of graduates	Aggregated LinkedIn data (+ statistics)	E	NL	Once
43	Environmental health indicator	Emission and noise data (+ geo data)	E	NL	Once
44	Economic activity detection	Websites	E	AT,NL	Once

44 Examples are found of which 6 have been officially produced. All others (38) are experimental and have either been produced once (17) or multiple times (21).

A more detailed version of the table is available online (http://pietdaas.nl/table1).

Working with Big Data: An example

Working with Big Data, differs from working with survey and/or administrative data.

Example: Let's study the question: "when was the term Big Data first used?"

There are a number of ways to find the answer:

1) Via a literature study (a paper by F. Diebold, 2021)
2) Ask a 'search' engine (Google & ChatGPT)
3) *Study 'historical' texts (= data driven)*

Working with Big Data: Ask a ‘search’ engine

Google search: “Some argue that [Big Data] has been around since the early 1990s, crediting American computer scientist John R Mashey, considered the ‘father of big data’, for making it popular. Others believe it was a term coined in 2005 by Roger Mougalas and the O’Reilly Media group. …”

ChatGPT (v3): “The term ‘Big Data’ was first used in the early 2000s, although its exact origin is not clear. Some sources attribute the term to John Mashey, a computer scientist who used it in a presentation in 1997. Others attribute it to Doug Laney, an analyst at Gartner, who used the term in a report in 2001. …”

Seems to indicate ‘Big Data’ originated somewhere between the early 1990s and ~2005

Working with Big Data: a data driven approach

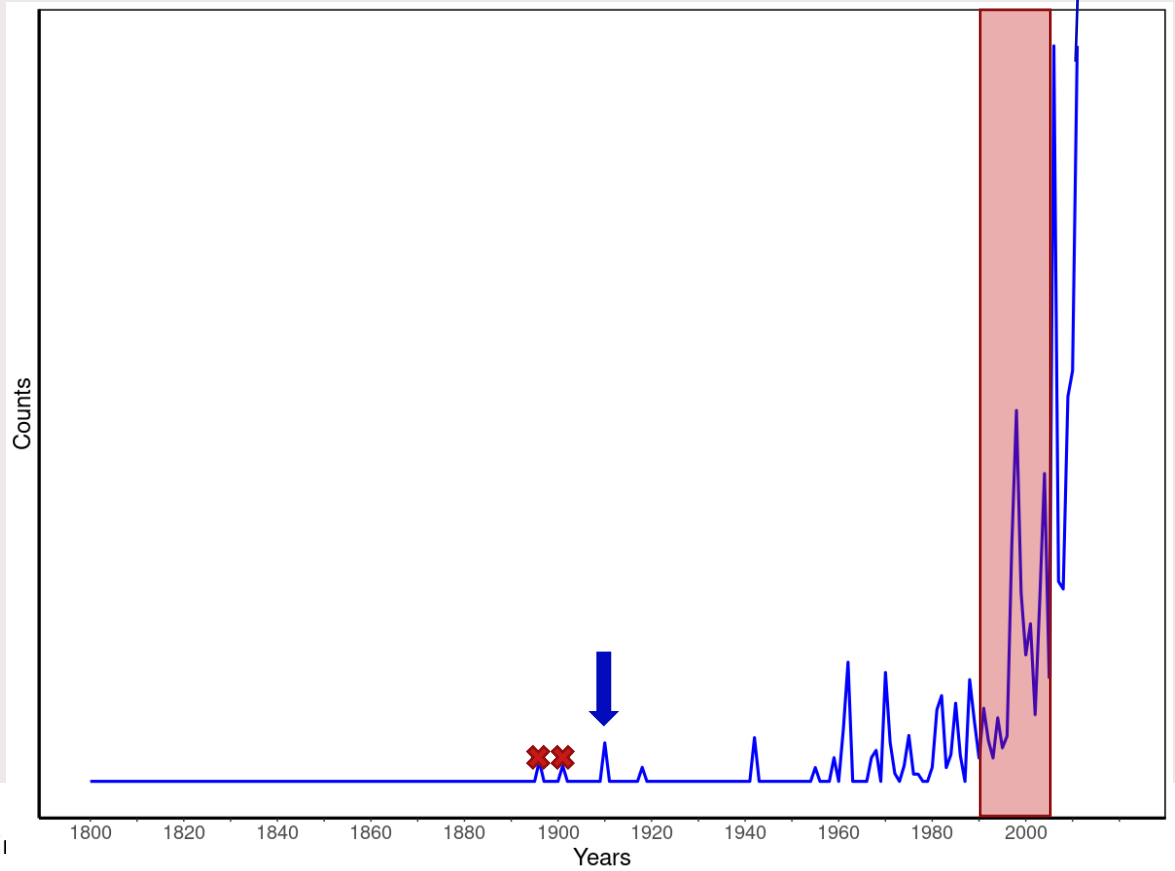
Look in ‘all books’ published and count how often the term Big Data is included per year

- Focus on books written in English and available online

How?

- By using Google Ngram Viewer (https://books.google.com/ngrams/)
  - Type in ‘Big Data’ and search in books published between 1800 and 2019
  - (case insensitive and smoothing = 0)

Working with Big Data: a data driven approach (2)



Import Big Data topics (in Official Statistics)

The most important topics studied are:

1. *Quality of Big Data*
   - Special attention to *Concept*, *Population*, and *Stability over time*

2. *Combining Big Data with other sources*
   - PhD-research of Yvonne Gootzen (CBS & TU/e), increase the potential use of Big Data

3. *Using Data Science Methods (within Official Statistics)*
   - With particular focus on Machine Learning

1. Quality of Big Data

Many Big Data sources are produced by private organizations. Not a lot is known about the data generating mechanism, which is also highly subject to change.

Many quality aspects have been identified, but in a Big Data study, I usually start by focussing on:

- **Concept**: What is the concept included in the source and what concept do we want to measure? Direct vs. Indirect
- **Population**: What part of the population is included in the source and what is the target population of the study? The 'representativity' question
- **Stability over time**: How stable are the findings over time? Reproducibility

1a. Concept

- Some data source *directly* provide the concept of interest.
  - The price of a product
  - Age of a person

- Some don't. Here, the concept could be obtained *indirectly*
  - Online platform detection from website texts
  - Innovative companies from website texts
  - Consumer confidence from social media message sentiment
  - ….

Satellite picture of Antarctica: Penguin colonies and more
DOI: 10.1002/rse2.176

1b. Population

- Which part of the population is included in Big Data?
  - Preferably all (census approach)
  - Representativity (how to determine it?)

- Challenging topic because:
  - Units in source may not represent the target population well
  - Often limited background characteristics are included
  - Potential solutions: census way of working, combining with other sources, use location info, non-prob. based correction methods, …?

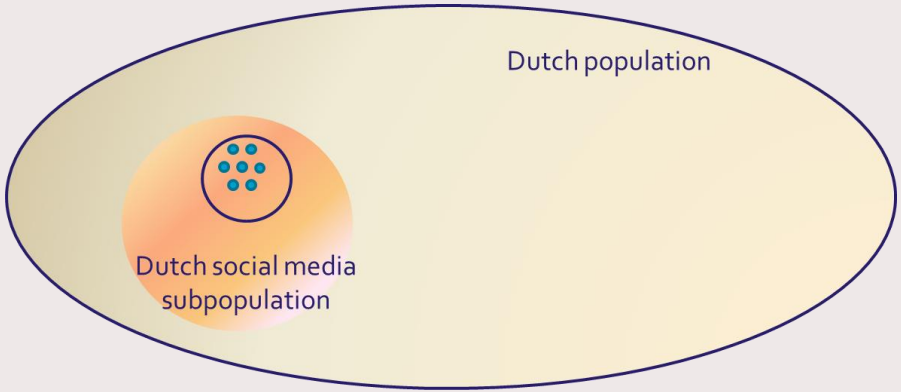
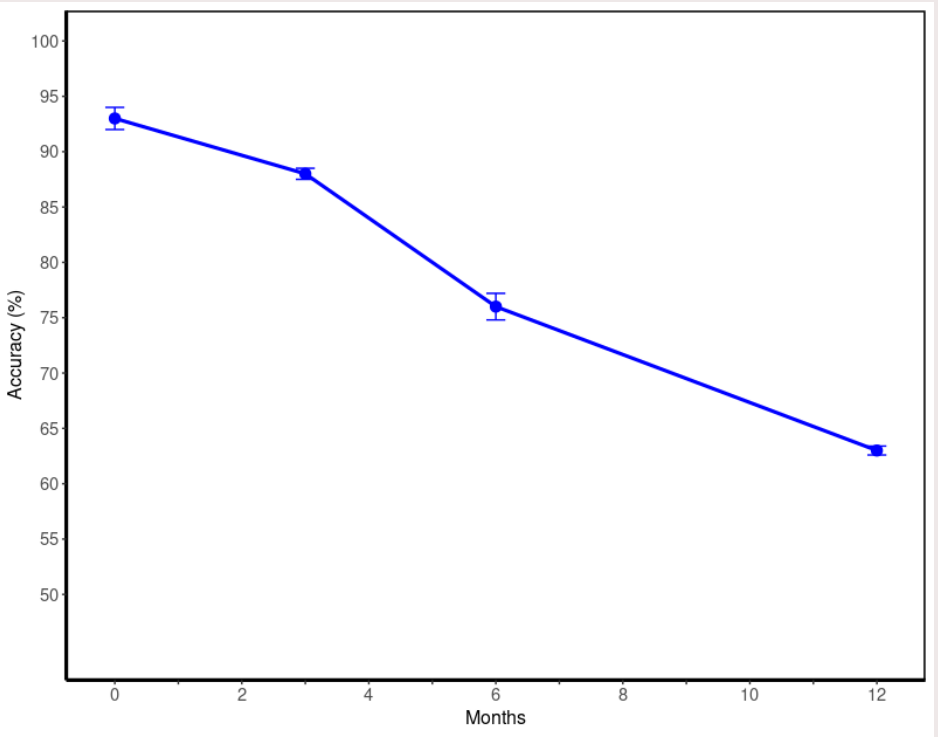


Illustration of selectivity of population included in social media messages

1c. Stability over time

- It is important to check how stable the findings are over time (especially when indirectly measuring concepts)
  - Reproducibility
  - Reduce chance of spurious correlations

- An example is shown of earlier work on detecting innovative companies from website texts
  - Accuracy of model reduced from 93% to 63% in 1 year
  - Recent results suggest a solution!



Accuracy of innovation model (trained at t = 0) over 1 year

2. Combining Big Data with other sources

Many of the official statistics based on Big Data that are in production combine it with other sources. This is advantages for a number of reasons, one of which is dealing with the non-representativeness of (some) Big Data sources.

- Topic studied by PhD-researcher Yvonne Gootzen (CBS & TU/e)
- Study in a systematic and logical way, how the metadata of various data sources can be used to find the best possible way(s) to combine them
- Ultimate goal is to develop a (software-implemented) strategy to enhance the use of Big Data within official statistics

2a. Combining sources: Example

- Study the *mobility by car*, by combining 4 data sources
  1. Big Data: Road sensor data
  2. Big Data: OpenStreetMap data
  3. Survey: National Dutch Travel Survey
  4. Admin data: Combination of registers, location work-home at neighbourhood level

- Combining the data sources requires
  A. Relate car modality and person characteristics (3)
  B. Create Origin-Destination (OD) matrix at neigh-bourhood level for traveling from home to work (A,4)
  C. Use route planning software, StreetMap data (2) and OD-matrix (B) to obtain routes for all car trips.
  D. Compare *combined* findings (C) with road sensor data (1) and calibrate

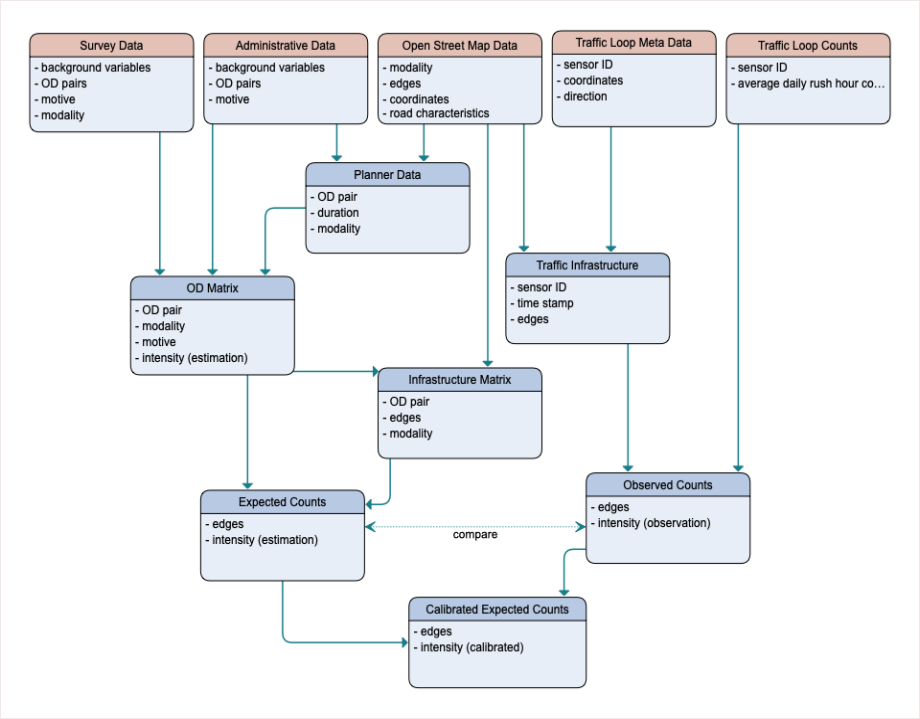
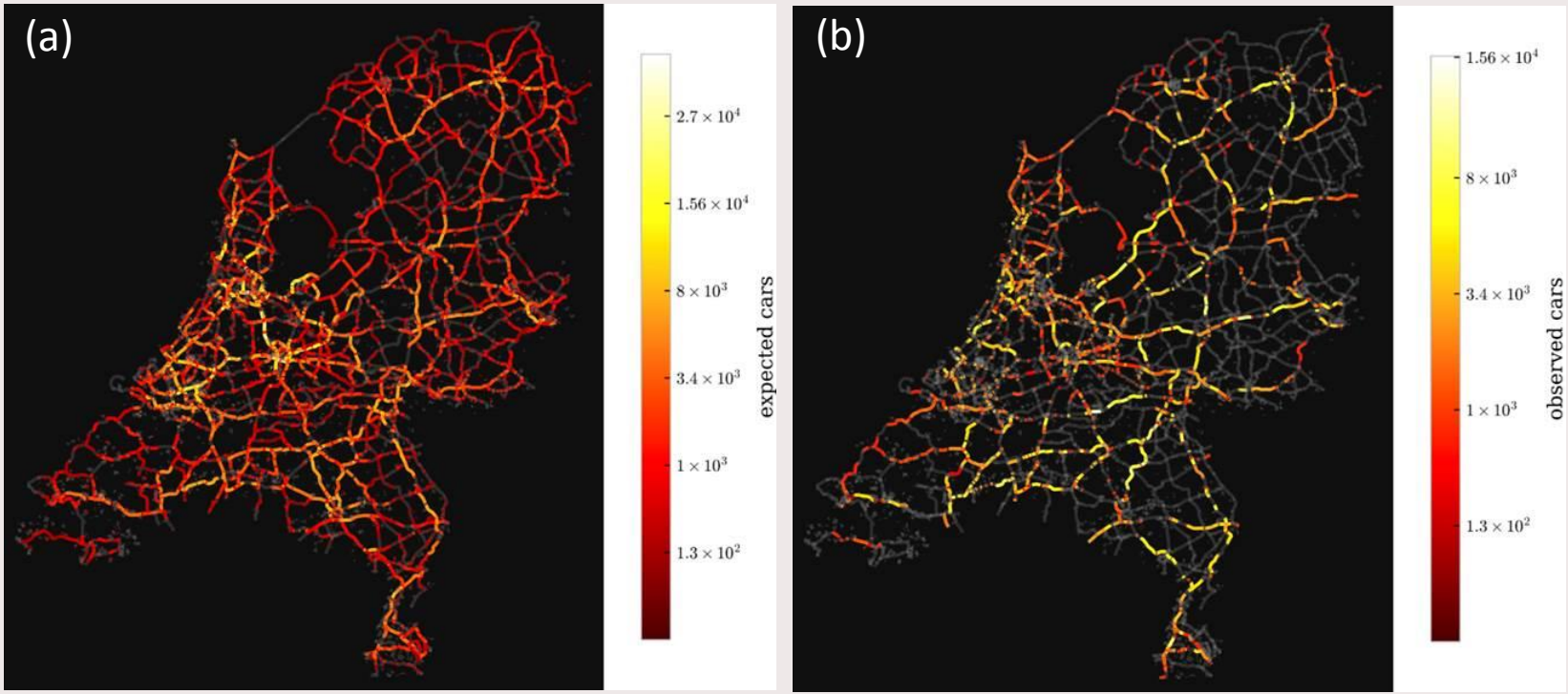


Illustration of the way the 4 data sources are combined

Result of combined (a) and road sensor (b) data



Combined (expected, a) versus road sensor (observed , b) traffic counts on the Dutch road network. Note that the legends do not use the same value range. Since not all road segments contain sensors, such as local roads, (a) contains much more detailed results compared to (b).

3. Data Science Methods

Data Science techniques are able to ‘learn from data’, by detecting patterns in large amounts of data.

Topics especially important in this area, in the context of official statistics production, are:

- Interpretability/transparency of the approaches developed
- *Validation* of the findings, both internal and external
- Causes of *Bias* and how to correct for it
- Creating a good training and test set (improve generalization of findings)
- ….

Let’s focus on 2 topics: *Validation* and *Bias* (in supervised classification)

3a. Internal and External validation

For supervised learning a dataset with known outcomes is used. Often a 50% positive and 50% negative dataset is used for classifications.

Usually an 80% random sample of this dataset is used to train the model. The performance of the model is determined on the (remaining, unseen) 20% test set. This is what we call the *Internal validation*.

For official statistics the performance of the model on the target population is of interest. This is what we call the *External validation*.

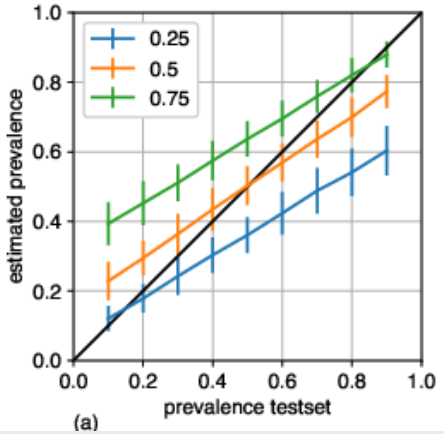
- What is the best way to obtain a model with a high external validity?
- Currently looking at: Construct a representative training/test set with random samples, Iterative model development, New metrics, …

3b. Bias correction

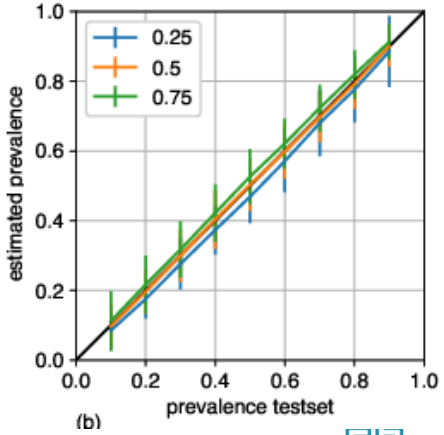
- Using a particular percentage of positive and negative examples in a training set, negatively affects the performance of the model in general
  - It introduces a *bias*

- A Bayesian correction method has been developed to correct for this
  - Code available on: https://github.com/mputs/BayesCCal
  - Works really well!
  - Requires classification models that can produce ‘probabilities’

Before



After Bias correction



Observations and lessons learned

- Using Big Data for official statistics is a relatively new topic. As a result, it is sometimes challenging to convince official statisticians to use it for statistics production
- More experience is needed to deal with the peculiarities of Big Data to ensure it can be reliably used in official statistics production
- The recent COVID crisis revealed that ‘Speed’ is not the predominant reason to use Big Data; these are ‘Newness’ and ‘More detail’
- Combining Big Data and ‘how to deal with its representativity’ are major challenges for which more research is needed
- Data Science techniques need to become Data Science Methods before they can be used in official statistics production.

Acknowledgements

Thank you!!