Using website texts to identify different types of companies

Presentation for ONS-UNECE ML-group, online meeting

Piet Daas (Statistics Netherlands & Eindhoven Univ. of Technology)

15 June 2022

Main focus of this presentation

- Webscraping
- How can the text on websites be used for official statistics production?
  - Can it replace survey data?
  - Can it assist survey data?
  - Examples to illustrate its application

Webscraping

1. Need a list of URLs associated with Dutch companies
   - Currently available on request in our office
   - Still needs to be enriched with additional data

2. Need a program/script to scrape websites

   In principle, 3 options

   a. Directly in Python, e.g. urllib, requests, Scrapy (fast)
   b. Use headless browser, such as Selenium (reasonably fast, but may crash)
   c. Use an actual browser and a shell-script (slowest but most generic, can -in principle- be used to scrape everything)

Example 1:
Detecting Innovative Companies via their website

Center for Big Data Statistics

Detecting innovation

- Web pages of companies provide information
  - The pages can be collected fairly easy
  - The text can be extracted fairly easy

- Here we look at:
  - The potential of *web pages* to provide information on the *innovative* character of a company
  - For both *large* and *small* companies

The Community Innovation Survey

- The Community Innovation Survey (CIS)
  - Focusses on the innovativeness of companies
  - Is a European standardized survey

- The questionnaire is send every other year to about 10,000 companies in the Netherlands
  - Stratified sample of companies
  - With a minimum number of working persons (WP) of 10

  - *So no info on small companies (such as start-ups!)*

Website URLs

- From the CIS response we took all Innovative and a sample of the Non-innovative companies
  - 3340 Innovative and 3002 Non-innovative

- For each company we needed the corresponding website
  - Initially the URLs were searched (via Google)
  - Later on, obtained via a commercial company (DataProvider)
  - All URLs were manually checked

Model building: import considerations

- Web pages were scraped, processed (html-files) and words extracted
  - Only scraped the main-page of each URL
  - Effect of various pre-processing steps
    - Language detection, stopwords removal, stemming, etc.
  - Later on: additional removal of words

- A supervised classification task
  - Tested various algorithms (80/20 training/test set)
  - Compared various metrics (Accuracy is best choice)
  - Started with TF-IDF value in DTM (Log(TF-IDF+1) is a bit better)
  - Effect of including words above a specific number of characters (2, 3)
  - Effect of including Word embeddings (relation between words)

Logistic Regression with L1-norm + WordEmb. worked best

Model details

- A single model including words in both languages trained on 20.000 classified company websites
  - 88% Accuracy over various datasets
  - 580 stemmed words included in the model
  - Revealed U-shape curve on 20% test set

- An English website is a positive indication for innovation (compared to Dutch)
- Depending on the language, there are words that are clearly positive associated with innovation.
  - ***Positive***: **Technology, innovation, software, data**
  - *Negative*: Sale, buy, shopping car, exclusive, service

External validation

- Tested the model on:
  - Web sites of start-ups (~900 in total)
    - 92% Innovative
  - Web sites of small companies (WP < 10) in our Business Register
    - ~ 33% were Innovative
- Manually checked samples to verify these findings

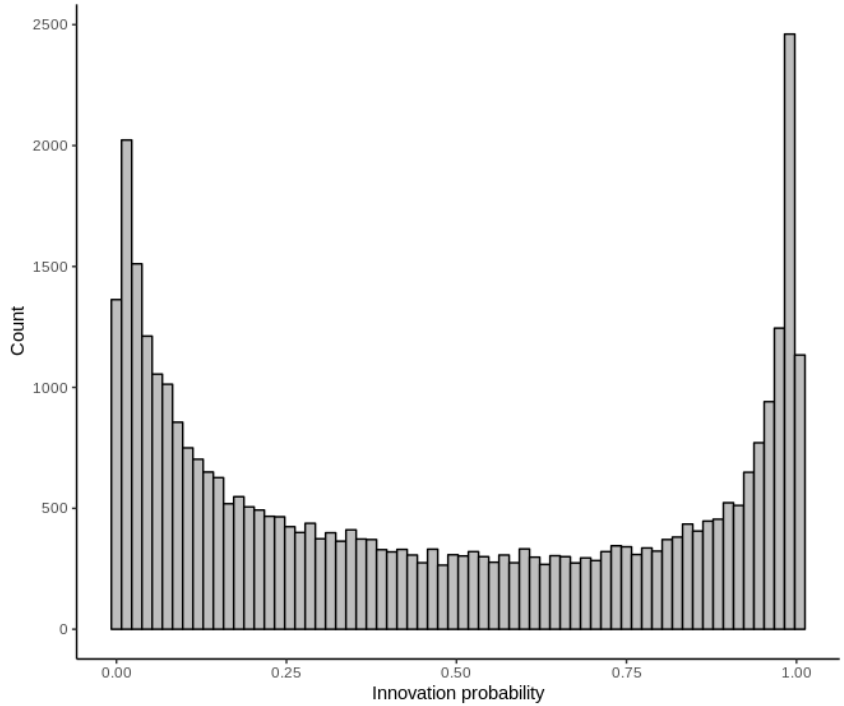
Center for Big Data Statistics

Appling the model

- Applied the model to websites of all companies included in Dutch Business register for which the website was known
  - Scraped ~850.000 web sites
  - Census-based approach
  - Includes web sites of
    - Large companies (WP >= 10)
    - Small companies (WP < 10)

Estimates of number of large innovative companies

- Large companies (WP >= 10) Web-based
  - Total: 37,576, Innovative: 17,783



Probability distribution classified Large comp.

Estimates of number of large innovative companies

- Large companies (WP >= 10) Web-based
  - Detected: 17,783
  - Corrections applied
    - 1. Model bias
      - Imbalance between FP and FN:

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Estimates of number of large innovative companies (2)

- Repeated the whole procedure 1000x

  - Web text based estimate 19,276 ± 190
  - CIS survey based 19,916 ± 680

Center for Big Data Statistics

What about Small innovative companies?

- The Model is developed on large companies
  - All WP >= 10
  - But test on small company data demonstrated that the Model can be used to detect small innovative companies as well
    - Startups & manual checking of BR sample
  - How do the small company results look like?
    - WP 0 through 9

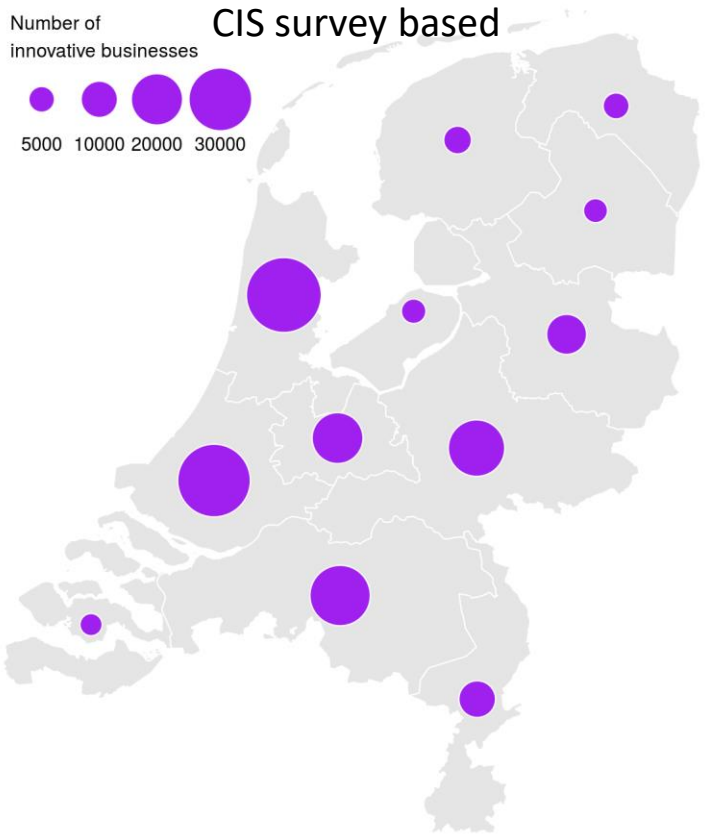
Estimates of number of innovative companies

- Small companies (WP < 10)

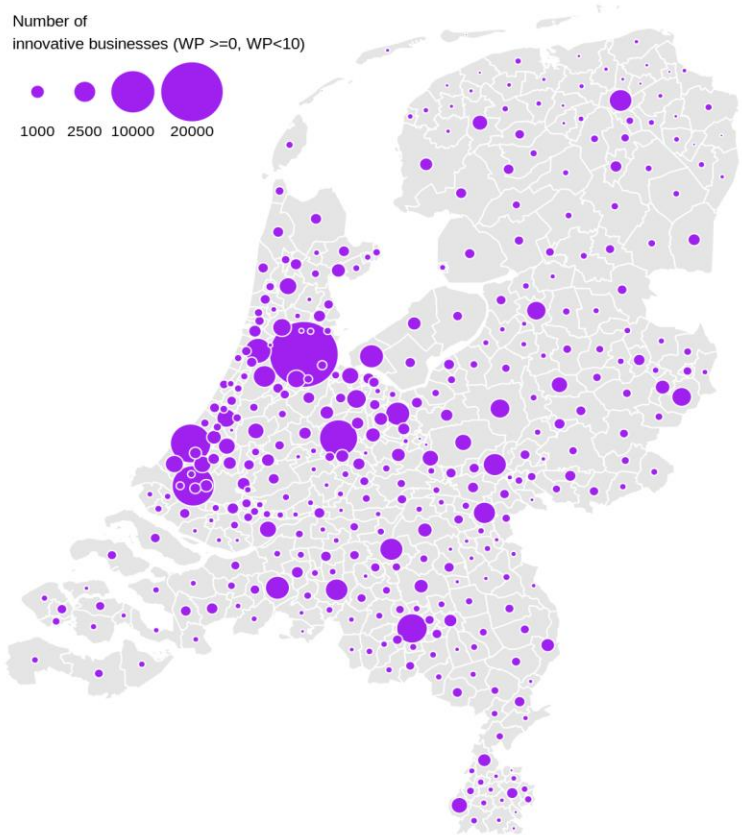
WP range	Websites scraped	Innov. comp.	Conf. inter.	Perc. innov (%)	Description
2-2.9	43,857	15,544	260	35.4	Small, 2 working persons
3-3.9	16,400	5,657	108	34.5	Small, 3 working persons
4-4.9	10,329	3,573	96	34.6	Small, 4 working persons
5-5.9	6,964	2,500	77	35.9	Small, 5 working persons
6-6.9	5,209	1,963	65	37.7	Small, 6 working persons
7-7.9	4,110	1,613	59	39.2	Small, 7 working persons
8-8.9	3,435	1,366	56	39.8	Small, 8 working persons
9-9.9	3,475	1,383	52	39.8	Small, 9 working persons

Final estimate ‘small’ companies (WP 2 and more) 33,599 ± 773

Innovative companies: Web sites

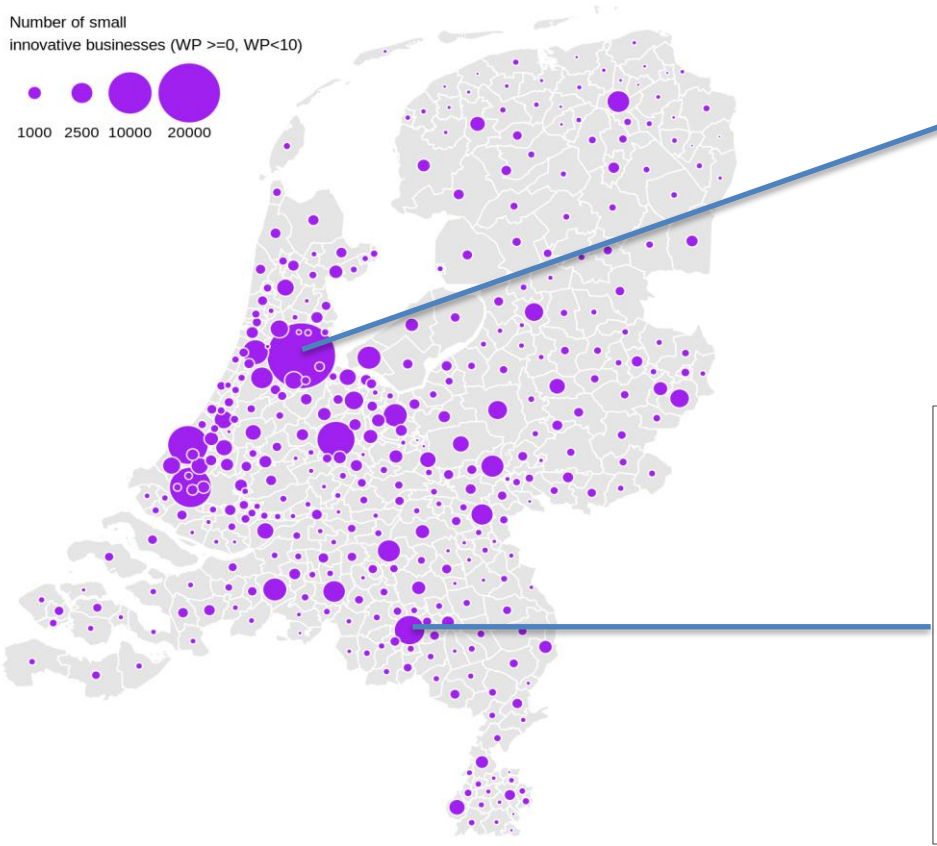


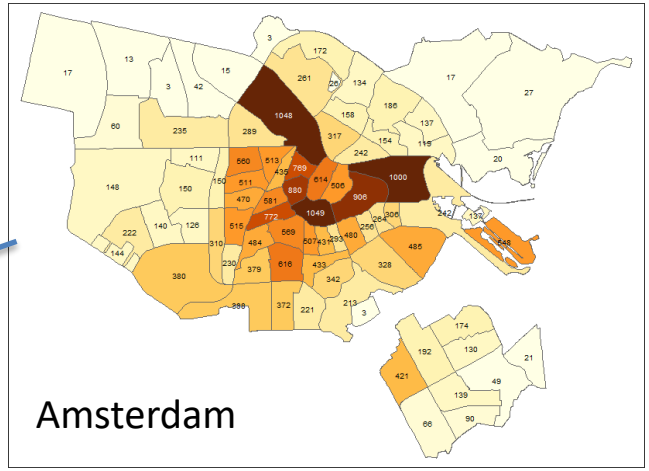
CIS survey based

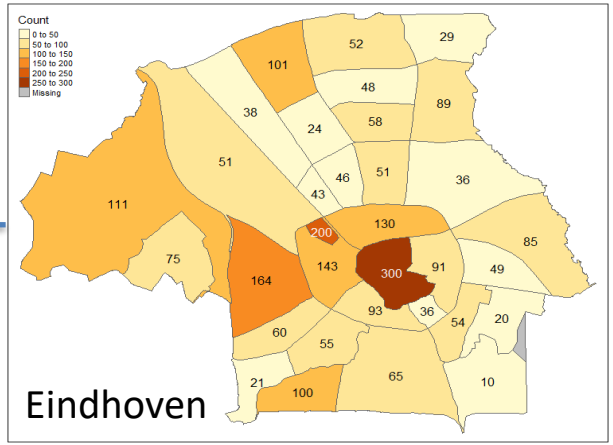


Based on URL's included in Business Register of CBS (~850.000)

Small innovative companies







Additional remarks

- Approach also works:
  - In Germany (work of Kinne and Lenz, 2019)
  - In Flanders (Dutch speaking part of Belgium)
    - Cooperation with Statistics Flanders
  - Did *not* work in: Sweden (Acc. only 70%)
    - Why?
      - Is it a language specific issue (context related)?
      - Effect of type of word embeddings?

Example 2:
Detecting Platform Economy websites

Center for Big Data Statistics

Detecting Platform economy websites

- Statistics Netherlands wants to produce statistics on platform economy companies
  - "An online platform is a website or application that mediates or supports the exchange of goods, services or information between individuals, companies or organizations. The platform contains either offers from other parties, or, in addition to own offers, also offers from other parties"
  - Examples are: AirBnB, Uber, Amazon, …

- Here we look at:
  - The potential of using the *text on web pages* to detect platform economy websites
  - Particularly as a way to *pre-screen* the population of Dutch companies

Model development

- Statistics Netherlands experts provided a set of examples of platform economy websites (680 positive examples)
- Only a few negative examples were given
  - Added a random sample of non-platform econ. from Business Register
  - In the end 50% positive, 50% negative (remember this!)

- Is there a difference between the text of platform and non-platform economy websites?
  - Developed a ML-model to detect platform economy websites based on the text
  - Standard preprocessing, combined multiple webpages per website (up to 200)
  - Best result: Support Vector Machine Acc. of 82%

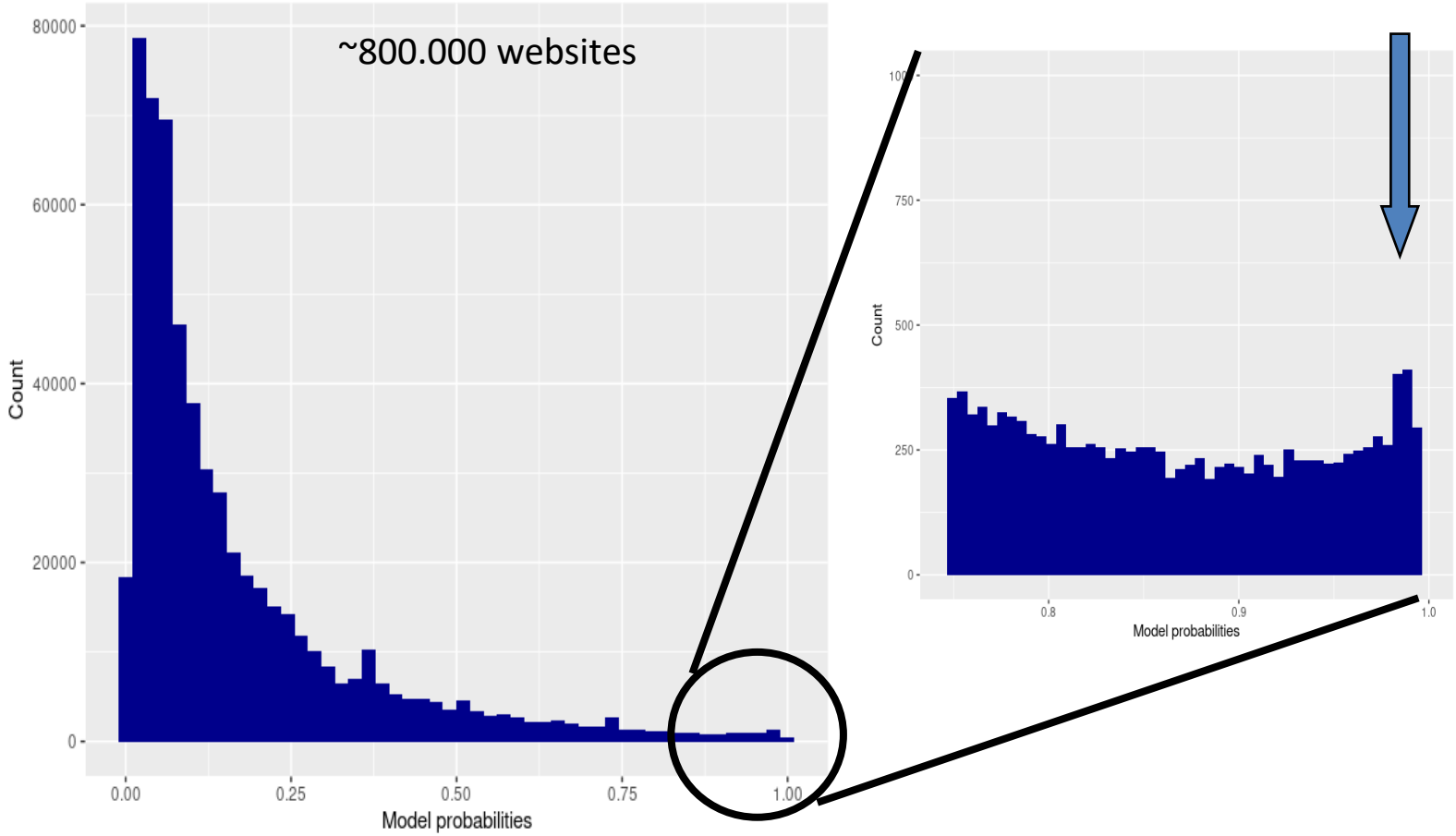
Model evaluation

- Model provided the *chance* of being a platform economy web site
  - Value between 0 and 1
  - U-shape distribution of test set (rel. small set)

- Words positively associated with platform economy:
  - **Register, com, login, platform, invest, sign up, …**
  - Negative associated words are indicative for other type of websites

Center for Big Data Statistics

Applying model to all websites linked to BR



Model evaluation (2)

First finding:

- The model indicated 9.802 websites as potential platform economy (all with p > 0.5)
- After manual checking websites with p > 0.8 were found the most interesting. Adult sites, non-linked and non-text were removed.
  - A total of 5.734 websites/4.170 companies remained.
- A total of 3.522 companies received a survey. 2.232 companies responded (63%) of which 537 were identified as platform economy companies.
  - Based on this it was found that platform economy companies all had websites with p >= 0.95 !
- Model has been used 3x already to *preselect* companies
  - - Model was checked each year and found stable

Center for Big Data Statistics

Other applications:
Detecting AI company websites
Detecting Drone companies

Other applications

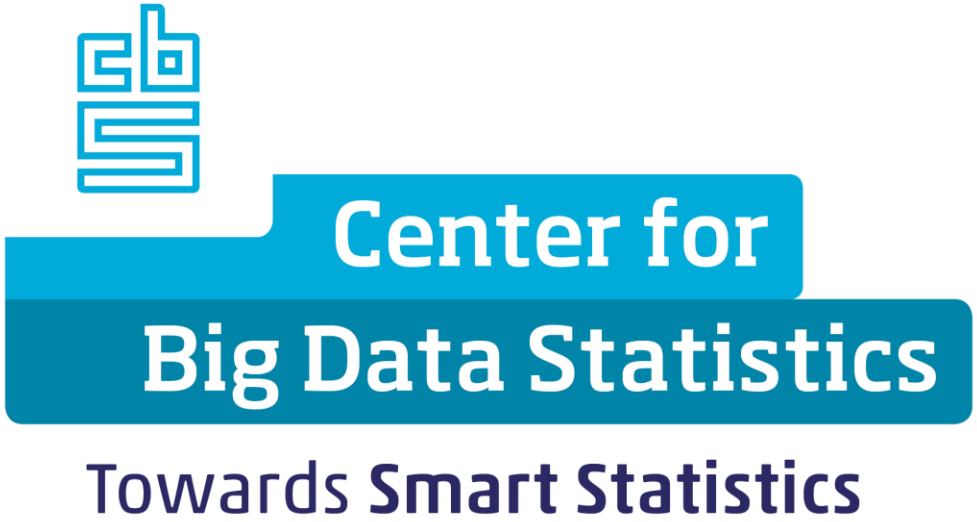
- Detecting AI companies
  - Garbage in-garbage out (bad training set)
  - May be 1500 in NL?
- Detecting Drone companies
  - One model used for Spain, Ireland and Italy
  - Acc. 82-86%, contains English words and translation table

Round up

- Web site texts are a very interesting source of information
  - I have shown some examples that demonstrate the advantages and new possibilities this data offers

- But:
  - Be aware that an association is measured, an indirect relation
  - There is a need to increase the knowledge on using Machine Learning within the real of official statistics

Thank you for your attention !!



Prof. Dr. Piet J.H. Daas

- Statistics Netherlands (CBS)
  - Senior-methodologist (20+ years)
  - Big Data research leader (since 2011)
- Eindhoven Univ. of Technology
  - Part-time Professor “Big Data in official statistics” (since 2019)
- I focus is on the methods needed to (re-)use ‘already existing’ data for official statistics