

Web scraping for business statistics

Examples from Statistics Netherlands

Second DSLN sprint

Barteld Braaksma, Innovation Manager, Statistics Netherlands (b.braaksma@cbs.nl)

Dubai, 23 January 2024

Topics to be discussed

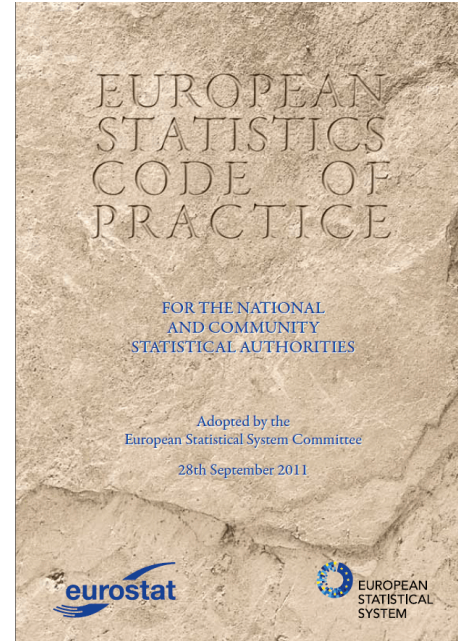
- Setting the Stage
- Detecting Innovative Companies
- Measuring the importance of the Internet
- Statistics in times of Covid

Setting the Stage



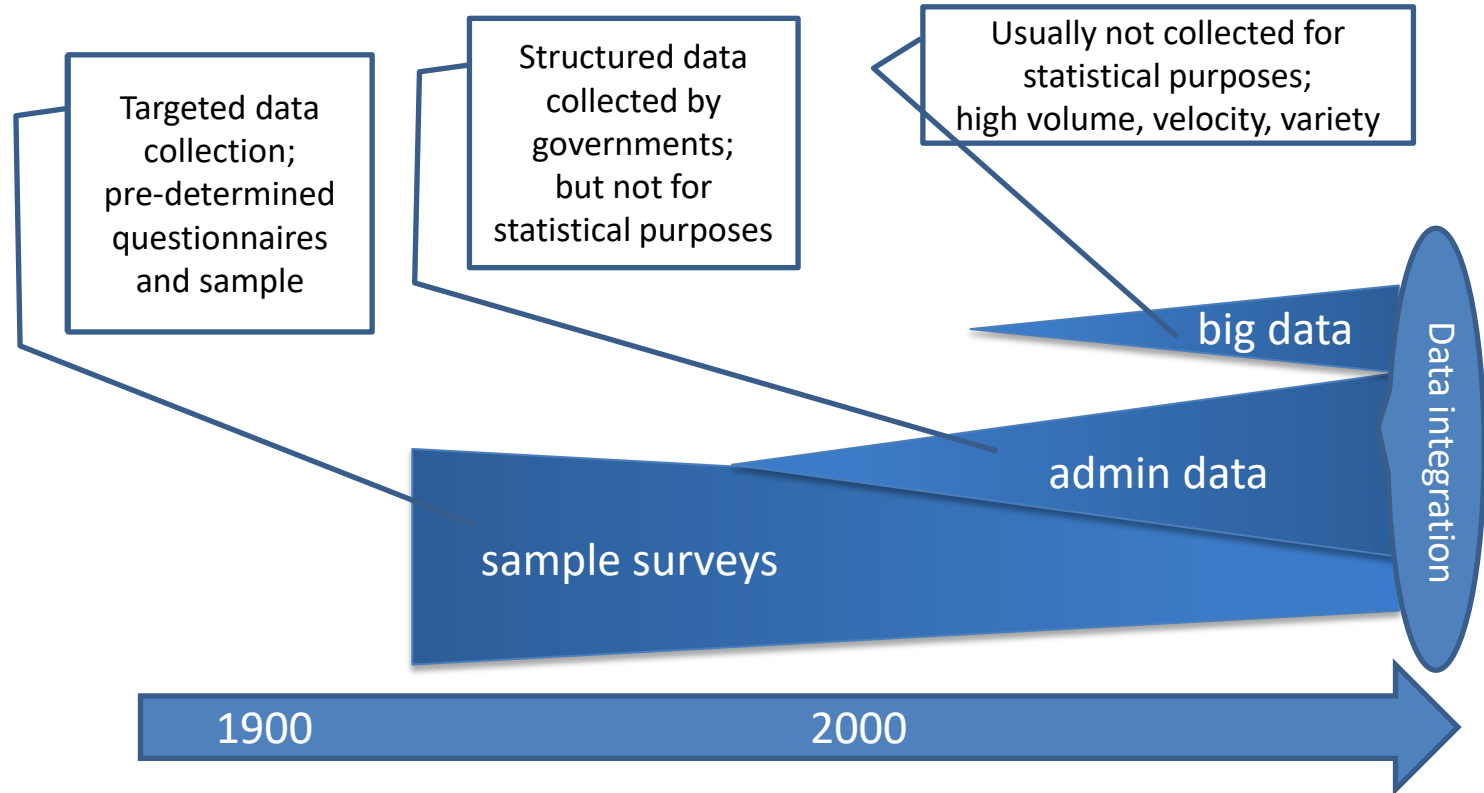
Mission of the European Statistical System

“We provide the European Union, the world and the public with **independent high quality information** on the economy and society on European, national **and regional** levels and make the information **available to everyone** for decision-making purposes, research and debate.”



➤ Output-oriented, no restriction on source data

From surveys via admin data to big data



Is the time of statistics over?



In a post-truth world, statistics could provide an essential public service
John Pullinger (National Statistician UK)

By combining the best of both worlds

Statistics Netherlands

<https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy>

Data, data everywhere

Information has gone from scarce to superabundant.

The Economist

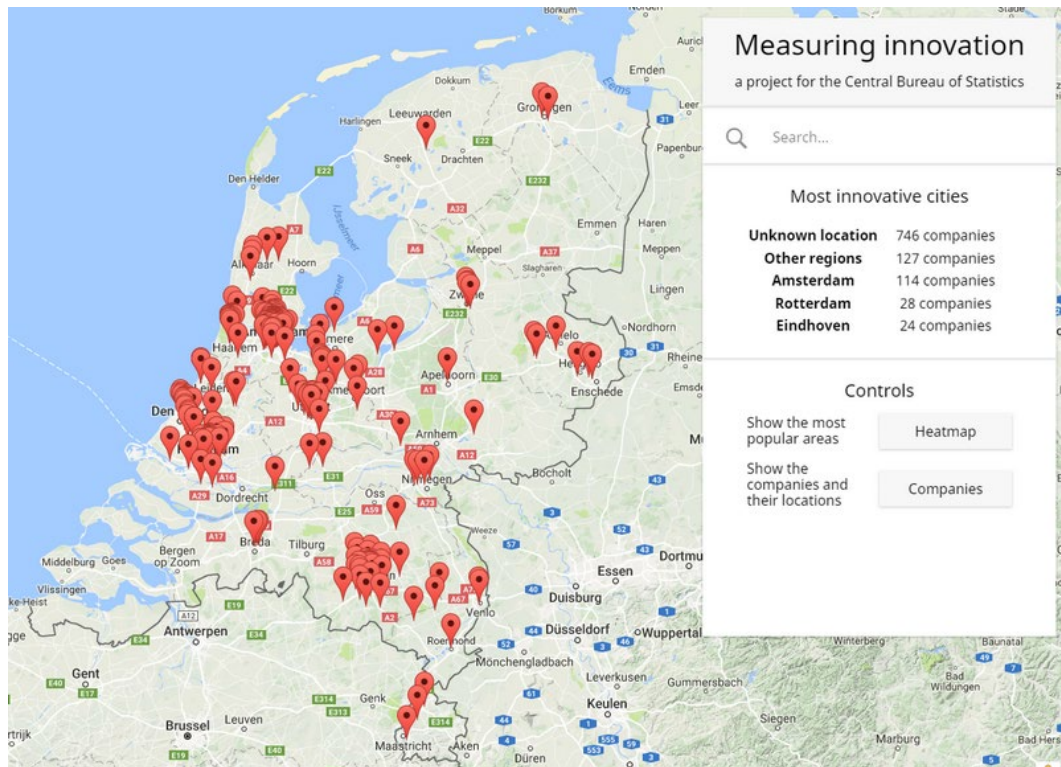


Detecting Innovative Companies

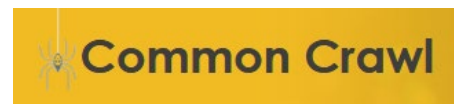
using webscraping, NLP and AI



First a student project

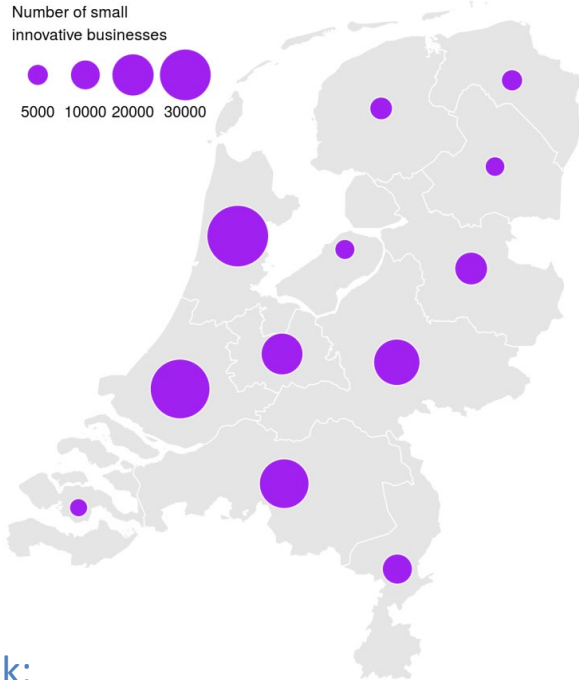
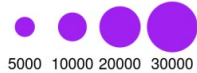


- Promising results
- Presented in Brussels (DG GROW, DG RTD)
- EC commissioned NESTA and Deloitte to develop web scraping methods

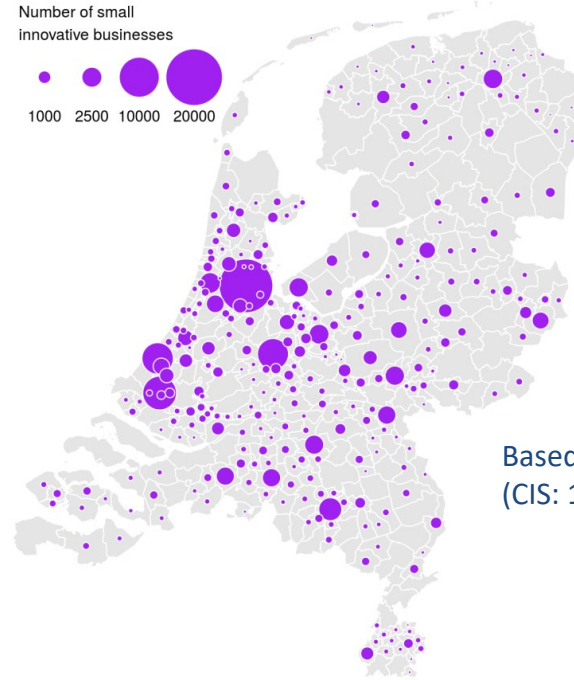
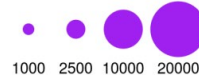


Then a beta product (experimental statistics)

Number of small innovative businesses



Number of small innovative businesses



Based on data from 500,000 SMEs
(CIS: 10,000 non-SMEs)

Link:

[Innovation in small businesses](#)



Method description and feedback form

(for all beta products)

Innovation in small businesses

How can we get a good idea of all the innovative businesses in the Netherlands? Statistics Netherlands (CBS) has researched this question at the Center for Big Data Statistics. CBS is currently only surveying companies that have more than ten employees about their level of innovation, an approach which by definition excludes a great many Dutch companies. To make it possible to collect information about the other businesses too, a big data method has been developed that analyses the text on a company's website. This 'web scraping' method is mainly useful for identifying small innovative businesses, such as start-ups.

Procedure

The text on the website's homepage is used to decide whether a business is or is not innovative. Punctuation marks and common general terms are removed from the text on each site, and the remaining words form the initial dataset for the development of an algorithm that can distinguish between innovative and non-innovative businesses. Because we know which of the businesses in the CBS innovation survey are innovative and which are not, we use the websites of these larger companies to train the algorithm. Ultimately, this produces a list of words that are important when classifying innovation, such as 'technology', 'new product', 'innovation' and 'software'. The language in which the website is presented is another important indicator. A company whose website is in English is statistically more likely to be innovative than a business with a Dutch website. Some words actually indicate that a business is not innovative; these include 'shop', 'transport', 'restaurant' and 'service'. Of course this does not necessarily mean that a shop can never be innovative; the combination of the other words on the website is also relevant. The latest version of the algorithm has been shown to be able to identify the innovativeness of large companies' websites with 93% accuracy.

Findings

The next step was to select half a million companies with fewer than ten employees from CBS' business register. The text of these companies' websites was then collected and classified using the algorithm. We did not know in advance whether these businesses were innovative or not, but a prediction was made based on the algorithm's results. A manual check of a large section of the results confirmed that the algorithm also works well on

Feedback

We look forward to hearing your views about this innovation and about its potential applications, and we are always open to ideas to help us further refine this web scraping method. Please send us your feedback using the form below.

Name

Organisation

E-mail

Feedback

Submit



...which caught the attention of a Ministry

Ministry of Economic Affairs asked for further research

- Together with dedicated company  Innovatiespotter
- Report presented early 2020, describing combined results and comparing approaches
- After some other priorities kicked in (Covid, AI)
- But autumn 2022: Ministry asks CBS to investigate feasibility of regular statistics ('innovation register')

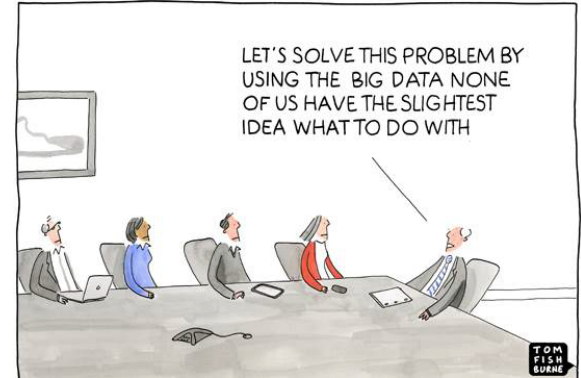
Why is the Ministry of Economic Affairs interested?

- Innovation stimulation is among their core business
 - Performance of industries, allocation of subsidies, impact of policies
 - Also relevant at EU level (DG GROW, DG RTD)
 - ... and at regional/local level: attract new business, employment
- Traditional CIS survey: too little, too late 😞
- Need for more statistics
 - Regional, more timely, by industry, startups, ...
- Also need for additional information beyond statistics (!)
 - Lists of innovative companies



Current situation

- Further research on several aspects
 - Concept drift
 - Use of LLMs
 - Ensemble methods
 - ...
- **Key outstanding issue: concept drift**
 - Important obstacle to robust results, comparable over time
- Preparations for structural implementation by production department, now responsible for CIS survey
 - Knowledge transfer not easy
 - Expectations of Ministry not well articulated
 - Limited resources



Detecting Innovative Companies using Web Mining and Ensemble Learning

BY

JAN STEPNIEWSKI

SUPERVISED BY:

PIETER KLEER (TILBURG UNIVERSITY)

PIET DAAS (CBS AND TU/E)

BARTELD BRAAKSMA (CBS)

Data Source and preprocessing

- 2016 Community Innovation Survey as the basis
- The dataset consists of 20 000 scraped websites of companies included in the Business Register
- The following pre-processing steps are applied:
 - Language detection
 - Stop words removal
 - Stemming
 - Words shorter than 3 characters are removed
 - This is one of the examples from website.nl → one example website
- Term Frequency-Inverse Document Frequency (tf-idf) is used to convert text into vectors

Machine Learning Analysis

- Get random set from CIS
 - ✓ 3000 innovative companies
 - ✓ 3000 non-innovative companies
- Split in 66% training set, 33% test set and find websites
- Try various classification approaches
 - ✓ Naive Bayes (accuracy: 81%)
 - ✓ Logistic Regression with L1 penalty (accuracy: **91%**)
 - ✓ Random forest (accuracy: 89%)
 - ✓ Neural Network (accuracy: 90%)
- Check model findings
 - ✓ Websites of 900 Dutch startups (< 1% non-innov.)
 - ✓ Websites in SME-top 100, 7 years (20% non-innov.)
 - ✓ ...
- Beware!
 - ✓ Depends on CIS companies (with 10 or more people employed)
 - ✓ Targeted on *technological* innovation only



Versatile approach

- ✓ Other countries with same/similar different languages
 - BE, DE, (PL), SE
- ✓ Local analysis (city/municipality level) possible
 - Creative industry around Eindhoven
- ✓ Define sample frame for a new survey
 - Application: detect online platforms
- ✓ Narrow down search space for manual inspection
 - Companies active in circular economy
- ✓ ...
- ✓ Not always successful
 - detecting AI producing companies did not work too well
 - ❖ conceptually unclear, no good training set, websites not always honest
 - ...but Ministry did not dare to submit it for the Brilliant Failure Award ☹️



"IT FIGURES. IF THERE'S ARTIFICIAL INTELLIGENCE, THERE'S BOUND TO BE SOME ARTIFICIAL STUPIDITY."

Measuring the importance of the Internet *for the Dutch economy*

Internet economy project, published 7 October 2016

-“What is the importance of the internet for the Dutch economy?”

✓ **Google** commissioned project

✓ **dataprovier.com**: website information

✓  : statistical analysis

- Report presented to Minister of Economic Affairs
- Study repeated in 2020, published [here](#)



General approach to measure internet economy

CBS data

CBS Business Register:

- Size class
- NACE code
- Age
- Region

Statistical microdata:

- Turnover
- Employees

Dataprovider data

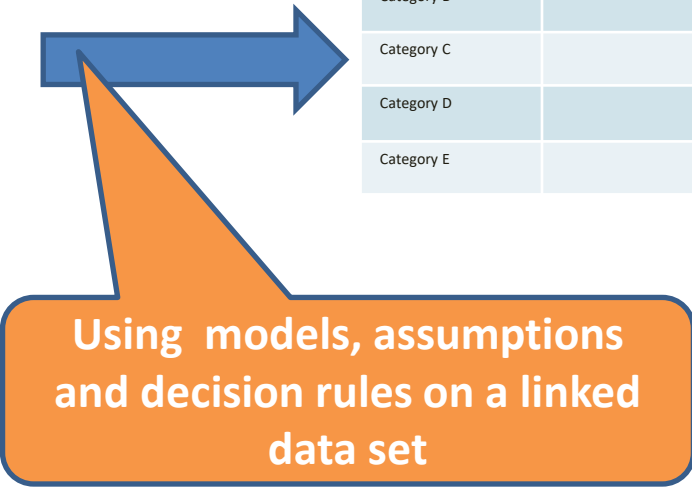
Website attributes:

- Keywords
- Economic footprint
- Payment systems

2.5 million Dutch websites
linked to CBS business register

Results

	Companies (number)	Turnover (mln euro)	Employees (number)
Total			
Category A			
Category B			
Category C			
Category D			
Category E			



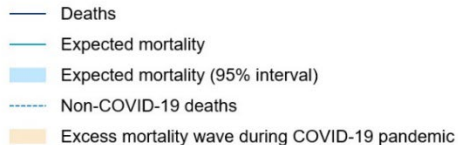
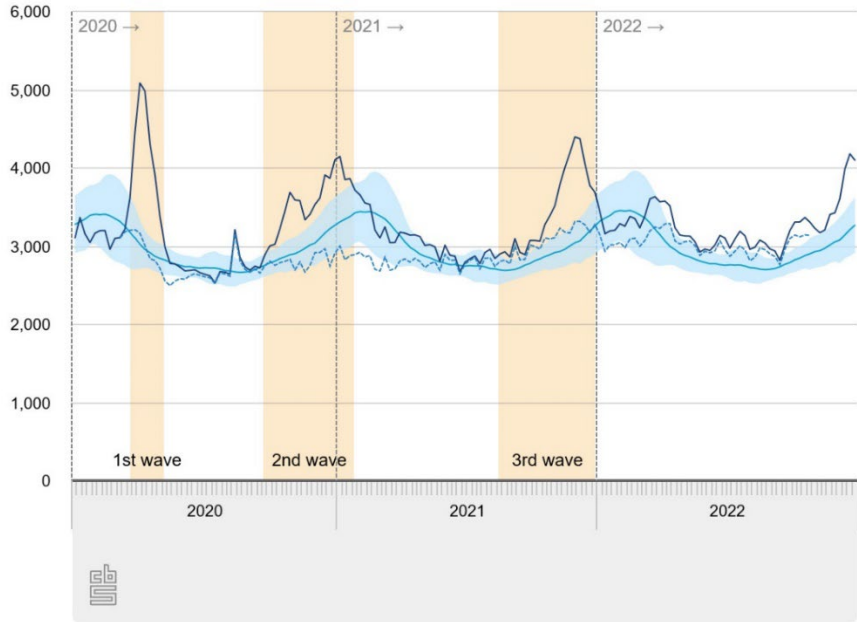
Using models, assumptions
and decision rules on a linked
data set

Statistics in times of Covid

Do we really need big data and data science?



Weekly (excess) mortality



* 2022 provisional figures. COVID-19 mortality known up to October 2022 inclusive.

- “Regular” statistics, based on administrative data
- Process simplified so timeliness and frequency could be improved
- Initiative from (senior) specialists
- Lots of attention in media and elsewhere

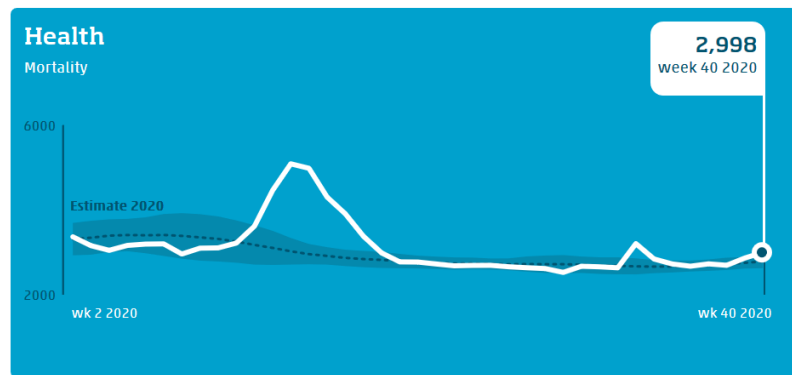


Covid dashboards: reuse of existing data and statistics

Well-being in times of coronavirus

21/09/2020 09:36

Presented here is the most up-to-date overview of well-being in the Netherlands, divided over nine themes. The development in relation to these themes is based on multiple indicators for each theme. A brief explanation [can be found here](#).



Coronavirus crisis FAQs



Feedback

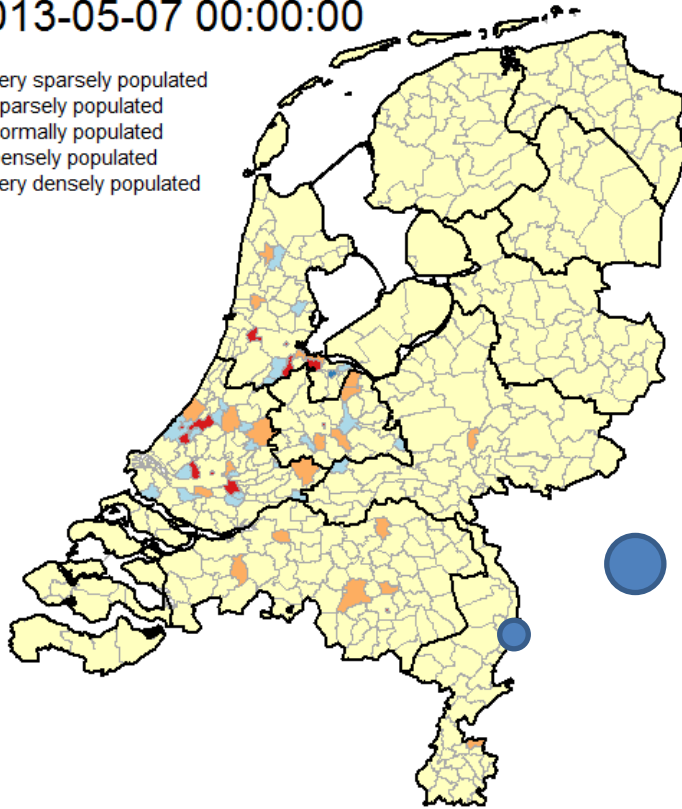
Nine themes:

- Health
- Society
- Safety
- Economy
- Labour and leisure
- Prosperity
- Housing market
- Mobility
- Environment

Whereabouts of people

2013-05-07 00:00:00

- Very sparsely populated
- Sparsely populated
- Normally populated
- Densely populated
- Very densely populated



- ❖ Municipality level
- ❖ Red is increasing
- ❖ Blue is decreasing



First collaboration
(mostly research)

Second collaboration
(production ready june 2020)



T Mobile

Mobile phone use - Ethnographic data analysis

DRIVACY FIRST



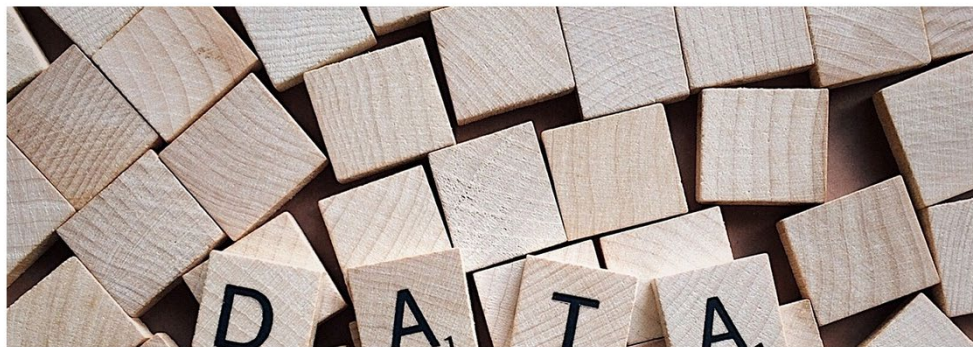
VACATURE PLA...

Nieuws Vernieuwing Het vak Loopbaan Agenda

Vacatures Stages Opdrachten I

Trudy Brandenburg - van de Ven — vrijdag 3 juli 2020, 14:48 | 1 reactie, [praat mee](#)

Autoriteit Persoonsgegevens benadrukt dat wijziging noodwet over delen telecomdata nodig is



Net binnen Uitg

Phishers doen zich v
redactie LINDA.

Digitale journalistiel
conserveren kun je l

Milieu-expert gedag
optreden in uitzendi

Toeziethouder CTIV
inlichtingendienst
verzamelen te veel e
te lang

Rechtszaak bedreigi
Gargard: boetes en

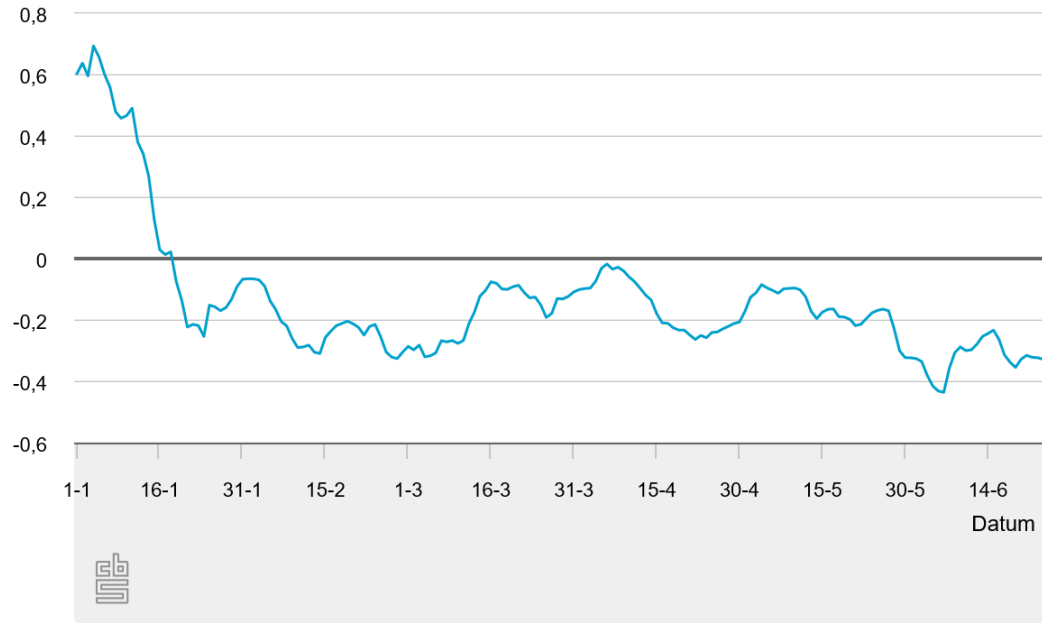


Covid sentiment indicator from Dutch social media

experimental

Figuur 1. Coronasentiment op Nederlandse social media, 2020

[Link](#) to CBS innovation site (Dutch only)



Monitor for public transport use (OV-chipcard)

under development



Cijfers Arbeid en inkomen Economie Maatschappij Regio Corporate

OV-Monitor

Redenen om te reizen

Hoe vaak en ver reizen we

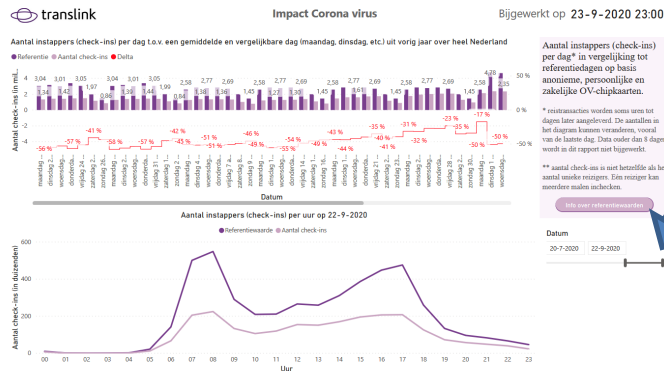
Wanneer is het druk

Hoe bereikbaar is het OV

Veel minder druk in het openbaar vervoer door Corona-virus

De in- en uitchecktransacties met de OV chipkaart worden centraal geregistreerd en vastgelegd. Met die registratie kan bijna real-time vastgelegd worden wat de veranderingen zijn in het gebruik van het openbaar vervoer. Translink gebruikt deze data nu om een actueel inzicht te geven in het gebruik van het openbaar vervoer. In onderstaande visualisatie wordt de drukte van de afgelopen dagen afgezet tegen de drukte op een normale vergelijkbare dag in maart (de referentiedag).

Interactive dashboard,
[link](#) to CBS-site



Microsoft Power BI





Facts that matter