Better data
Better lives

Interim review 2017-2019: Three years of Center for Big Data Statistics at Statistics Netherlands

Better data
Better lives

Interim review 2017-2019: Three years of Center for Big Data Statistics at Statistics Netherlands

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information

Introduction

National Statistical Offices (NSOs) have existed for decades, in some cases even more than a century. They were producing and publishing official statistics on a range of topics long before the arrival of the computer, the Internet and the data revolution. These developments have had major transformative effects on production.

The first major shift was brought about by governmental entities such as tax authorities creating digitised, computer-based administrative sources. Government records were being made electronic and stored onto computers, which opened up the possibility of using these data for the production of official statistics. These "administrative sources" have several major advantages but also some disadvantages. Some of the major advantages are that in many cases the records are "complete" i.e. they are not samples but complete records of a large part, or all, of the population. One thus obtains a very rich resource that can be compared to and integrated with other sources using computers. Furthermore, in some cases such as tax data, falsification of data is liable for penalty, which leads to improved data quality. The "disadvantage" was, for example, that the data thus obtained was not originally meant for the production of specific indicators used in official statistics. Definitions and meta-data may not match and often do not. The frequency with which the data is collected may not match either. That is, a "translation" step is required, thus necessitating a serious methodological effort. The production of the required statistical indicators becomes more or less indirect and model based, which for the purists was hard to accept. The period that saw an increase in the use of administrative sources also saw the rapid growth of methodology departments within NSOs. Nevertheless, the use of administrative sources has been a tremendous success. In the final analysis, it allowed for an increase in the quality of official statistics, more detail, lower costs, and a lower administrative burden. One of the most striking illustrations of this, is the replacement of the traditional and extremely expensive (and intrusive) census with an almost completely automated census based mostly on administrative sources, available surveys, and specialised models.

The next major shift for NSOs is the use of automated data-capture through sensors. Even administrative sources rely on manual data-capture somewhere upstream at the start of the production process. And because this is a manual process, it is relatively low-frequent, leading to an often-heard complaint about NSOs that they do not produce current information which can be used in real-time decision-making. Automated sensors can in some cases do thousands of measurements per second, creating huge datasets: big data. Furthermore, the cost of storage capacity has been dropping rapidly, so that most or all of this data can be stored at reasonable costs. And at the same time, hardware, software, and methods have evolved to analyse this dataset and extract useful information from them, often on the fly. For every NSO finding ways to utilise this new data-category and the technologies that comes with them, should be a no-brainer. Of course there are still huge methodological, technological, legal, and publicity challenges that need to be solved before sensitive official statistics can rely predominantly or entirely on sensor data. Nevertheless, it cannot be denied that this development is significant and in time will have a transformative effect on society in general and NSOs in particular. One might even envision a future in which NSOs use proprietary sensor networks in some cases, for example in order to control the quality of the data front-to-end. In some cases combinations

with other datasets, that may be created and owned by others, may be possible. In this respect NSOs are well positioned to consolidate their strong position vis-a-vis other data-intensive organisations. The availability to NSOs of high-quality and high information-dense administrative sources and mandatory surveys, gives them the possibility to combine these sources with sensor data, creating a unique opportunity. NSOs are therefore in a strong position as long as they embrace the opportunities that the aforementioned developments present.

It was with this analysis in mind that Statistics Netherlands (CBS) decided three years ago to launch the Center for Big Data Statistics (CBDS). With limited financial resources, (unfortunately Statistics Netherlands was still living through crippling budget cuts by the Dutch government), but with a seemingly endless supply of enthusiasm and talent in the main actors, CBS has managed to achieve the results described below. Through their commitment and hard work, CBDS has grown rapidly and has produced some striking results. A new website for experimental statistical products was introduced in order to inform the public and invite outsiders to join the journey and provide feedback. External funding from several sources was secured and a strong network of partners was established. Three years after its launch, CBDS is now poised to start a next phase of its development; a phase in which experimentation will still have an important role to play, but will also be characterised by a stronger focus on the integration process of sensor data and cutting-edge techniques in the regular production processes of official statistics. I am looking forward to seeing the CBDS flourish and create some of the statistical methodologies of the future.

Director General Statistics Netherlands
Dr T.B.P.M. Tjin-A-Tsoi

The Hague/Heerlen, March 2020

Content

1.

Introducing CBDS

We mark the third anniversary of the Center for Big Data Statistics (CBDS) by reviewing the first period: 2016 up to and including 2019. In this report, we present the main results that have been achieved in terms of experimental statistics, scientific output, training of data scientists and statistical output. We also present the rewarding collaborations we have had with the many different partners in our data-ecosystem.

1.1 Big Data in society

More and more information is stored digitally by people, sensors and devices. This data contains a wealth of new information and knowledge with economic and societal value. Big data offers opportunities for the community to develop new business models and for the government to make better decisions based on factual, current and the most relevant information. The awareness of the potential value of these data is growing and with it the demand for extraction of the information it contains. In the summer of 2016, Statistics Netherlands (CBS) launched the Center for Big Data Statistics (CBDS), in response to this game changing development.

1.2 Statistics Netherlands (CBS)

CBS supports decision-making by providing the public and private sector with reliable and coherent statistics of undisputed quality. These statistics are also used in scientific research. The information (official statistics) published by CBS covers topics that are relevant to society, including economic growth and consumer prices, but also for instance safety, health and leisure.

Within the Dutch government, CBS is the data expert with 120 years of experience. CBS is the largest data warehouse in the Netherlands and our people have been working with big data for quite some time. Individual big data sources usually only tell part of the story, are often owned by private parties and their quality is unknown. There is a need for an independent party who combines data, sets quality standards and brings in objectivity.

1.3 Center for Big Data Statistics (CBDS)

We see the launch of the CBDS in September 2016 as a logical step in order to support evidence-based policy with new, real-time information. The CBDS offers opportunities for working with data through internships, traineeships and employment in data science. From a government perspective, CBS is the ultimate partner for the academic world and the business community (including Small and Medium-sized Enterprises) in the development of data science techniques and methods. Within the CBDS and jointly with the statistical divisions and the CBS partners, knowledge, infrastructure and data are brought together in order to meet the information needs of society. The CBDS works on socially relevant topics such as economic growth, the energy transition, mobility, the labour market, health, the housing market and safety. Cross-border statistics receive special attention in the work program, which includes product development, research into new methods and techniques and access to new sources (data scouting).

The CBDS team in Heerlen

The strategic goals of the CBDS are:

1. Increase impact on and relevance for society by developing more relevant, more detailed and timely statistics to allow policy making in order to address complex societal questions.
2. Increase ability to innovate through the right statistical methodology and IT infrastructure.
3. Set up and intensify the cooperation with external partners, government and businesses and promote innovation through professional communication.
4. Address legal issues that enable the (re)-use of data in hands of other parties.
5. Increase capacity building among government staff by developing data science skills and expertise allowing implementation of the experimental statistics in the official statistical production process.

The CBDS management team

Astrid Boeijen is a board member at Statistics Netherlands, senior director of the Data Services, Research and Innovation division and director of the Heerlen Branch of Statistics Netherlands. After her chemistry study in Amsterdam, she obtained a PhD in combinatorial chemistry at Utrecht University. She was subsequently employed at Maastricht University, where she worked for fifteen years, holding the positions including director Academic of Affairs and the Student Services Centre. In her current role at Statistics Netherlands, she is responsible for innovation and statistical services to governments, among other areas. She was responsible for the funding of the Statistics Netherlands Center for Big Data Statistics and the establishment of a Provincial Data Center with the Dutch province of Limburg.

Joost Huurman is director Research and Development at CBS and managing director of the Center of Big Data Statistics. Joost holds a Master's degree in Physics from the University of Leiden obtained in 2002. Joost started working at CBS directly after obtaining his Master's degree. Prior to his current position, he was project manager, unit head and program manager among other positions within various CBS department.

Reinoud Stoel is head of the methodology team at CBS in The Hague. Reinoud holds a Master's degree in psychology and a PhD in psychometrics from the University of Amsterdam, the Netherlands. After completing his PhD, he worked at the Department of Psychology of the Vrije Universiteit Amsterdam in the area of Behavioral Genetics for one year. He was Assistant Professor Methods and Statistics from 2004 to 2008 at the University of Amsterdam. From 2008 to 2019 Reinoud worked as a forensic statistician and he was head of the forensic statistics team at the Netherlands Forensic Institute (NFI).

Sofie De Broe is head of the methodology team at CBS in Heerlen and scientific director of the Center for Big Data Statistics. She received a Master's degree in demography from the University of Louvain la Neuve, Belgium, and a PhD in reproductive health from the University of Southampton, UK. Before starting at CBS in 2015, she was a senior researcher at the Office for National Statistics (UK) and lecturer at the University of Duisburg–Essen (Germany).

The CBDS data science team

This multidisciplinary team consists of a data scout, a subsidiologist, data scientists, product developers, methodologists and statisticians such as AI and IT experts, text mining experts, economists, an evolutionary biologist, sampling theorists, health scientists, physicists, R software experts, visualisation experts, sociologists, business analysts and psychologists.

The CBDS product developers and data scouts

The mission of the CBDS is to create new and improved official statistics based on new (big) datasets and new methods. This is not a straightforward process. The product developers have an important role in defining new potential product ideas, providing access to required data and funds to develop the product, figuring out the legal boundaries, developing partnerships with other parties that have a stake in this new product, working together with the statistical divisions within CBS, getting support for new ideas within CBS itself, managing the development process and finally communicating about developed beta products with articles and presentations. As product developers, we act as the glue that holds everything together to enable the data scientists to perform their work in generating new insights from the data we collect and process. This is a diverse role which requires a great deal of flexibility! Our data scout is continuously looking for new data sources. See chapter 2.5 below on all data scouting activities.

The CBDS team in The Hague

1.4 The CBDS ecosystem

Big data invites us to collaborate with partners and to break through barriers as innovation can be disruptive. The CB(D)S is associated with a large number of renowned national and international partners in the business sector, the academic world (including knowledge institutions such as TNO and colleges and universities), the public sector and citizens, committed to a quadruple helix network at local, national and international level.

CBS brings to the CBDS a wealth of knowledge and experience about data processing and analysis, privacy and IT infrastructure, in addition to the large amount of data that CBS already has in house and that can possibly be linked to new data sources. The CBDS brings critical mass, data, knowledge, skills, infrastructure, capacity, internship opportunities and innovation. This joint effort offers powerful opportunities for all innovation stages: from developing new ideas, looking for funding, collaboration to exchange knowledge and data, joint work on projects, validation of the experimental products, development of new business models and shared use of infrastructures, publication of end results and implementation in the official statistical process.

The CBDS in the region

The Center for Big Data Statistics makes the most of the synergy with the Brightlands Smart Services Campus (BSSC) in Heerlen, the opportunities offered through cooperation with campus partners and the Maastricht University. Brightlands Smart Services Campus offers the latest R&D and knowledge infrastructures, on-campus education, science-oriented business support, and business development services. This offers opportunities for CBS, the Maastricht University and the BSSC. As a national statistical agency, CBS is a unique partner with expertise in the field of statistical and data science methods and techniques, privacy protection and data storage, offering considerable added value for the information needs of all smart services campuses in the area and the Maastricht University.

Local stakeholders and partners about Statistics Netherlands:

Bernd Burger, project manager CGI: "CBS is highly skilled in analysing data using social media and Google and has a great deal of knowledge about migration and its impact on the Netherlands and Europe."

Andre Dekker, Professor Clinical Data Science (Maastricht University): "CBS has some very interesting data, and as a partner it has a lot of drive. Even more interesting is the willingness of CBS to act as a partner in innovation with the CBDS."

Stakeholders and partners about the CBDS:

Martin Paul, president of the board at Maastricht University:
"*The CBDS succeeds in providing knowledge that fits the regional needs. The relationship between the University of Maastricht and the CB(D)S has changed positively.*"

Peter Verkoulen, ex-CEO of Brightlands Smart Services Campus (Heerlen):
"*The added value of CBS and CBDS in the local ecosystem is evident. The knowledge and (open) data of the CBS are hugely relevant for other organisations to use. It increases the possibilities for everyone to get value out of data.*"

Significance of the CB(D)S for the region:

1. 120 years of experience in the field of data and an extensive national and international network.
2. CBDS offers traineeships, jobs and a workplace for interns, PhDs and researchers. It attracts top talent from around the region in the field of data science
3. By combining their networks, the Brightlands Campus and CBS can connect regional companies and start-ups with national parties.

4. The CBS department responsible for taking surveys, located in Heerlen, helps tackling dwindling response rates in the traditional CBS surveys among companies and citizens.
5. Together with the University of Maastricht, CBS does research on privacy preserving data sharing and several PhD positions were funded through the four flagship projects.

The CBDS in Europe

CBS is at the forefront of innovation in statistics, and the CBDS has been in close contact with other National Statistical Institutes (NSIs) for research collaboration, exchanges of staff and experiences. CBS is well represented at the European level through Eurostat and its presence in the Task Force and Steering Committee Big Data as well as in the methodology and IT directors groups.
NSIs collaborating closely with CBS include Statistics Norway and the Office for National Statistics (ONS) in the UK with their Data Science Campus. Below, the Director of the ONS Data Science Campus, Tom Smith, and head of methodology Anders Holmberg of Statistics Norway (currently at Australian Bureau of Statistics) share their views on collaborating with the CBDS:

Tom Smith, Director of the ONS Data Science Campus:
"Statistics Netherlands and the Centre for Big Data Statistics have a well-deserved reputation of ranking as global leaders in the field of big data and data science for official statistics and innovation. The ONS Data Science Campus and the CBDS have been working closely from the very start in terms of sharing knowledge, staff exchanges, project ideas, and strategic goals. We share a common drive to strengthen statistics and make data relevant for society, and I look forward to continuing to work closely with the CBDS and the team."

Opening of the ONS Data Science Campus

Anders Holmberg, chief methodologist Statistics Norway:
"Statistics Norway congratulates the CBDS on its anniversary. We have had some very fruitful staff exchanges which have enhanced our knowledge of incorporating new methods and new data sources in our business. My organisation is very pleased with our collaboration and we look forward to more activities and continue to work together with CBS on innovative topics for statistics. For a relatively small national statistical institute such as Statistics Norway, the CBDS has proven to be a very useful and practical platform to build capabilities, build professional networks and get early experiences about new methodologies."

Marc Spearing, President and (interim) Vice Chancellor of the University of Southampton and the Director-General of Statistics Netherlands Tjark Tjin-A-Tsoi at the signing of the MOU in September 2019 in Southampton.

CBS has also invested substantially in strategic partnerships in order to stimulate research activities in the area of data science and artificial intelligence (AI). In September 2019, CBS signed a Memorandum of Understanding with the University of Southampton, one of the University Partners of the Alan Turing Institute, which is at the forefront of research in that area. This involves a close collaboration in the use of AI on a number of specific topics including social deprivation, determining characteristics of companies and synthetic data.

CBDS participates at several international conferences

In the context of the European Statistical System (ESS), CBS is also heavily involved in research on the possibilities of big data for official statistics. This is organised in the so-called ESSnet Big Data, in which currently 28 partners from 23 countries of the ESS collaborate using EU funding. This ESSnet started in 2016 and will run until the end of 2020 with total funding amounting to approximately 4.5 million euros. CBS is in charge of the overall coordination. At the beginning, the ESSnet consisted of a number of pilot projects exploring the possible use of various big data sources. For some of these data sources, the ESSnet is now carrying out implementation projects. Other pilot projects are being carried out in parallel including one on so-called smart statistics, are related to the Internet of Things for example. In order to make the results readily applicable throughout the ESS, special efforts are undertaken to coordinate the approaches to methods, quality, IT and processes, including their architecture. The ESSnet makes use of a wiki, on which all products and a lot of other material are made public.

One of the ESSnet projects concerns using data from ship tracking systems for official statistics. This project is also led by CBS. Other projects in which CBS is participating focus on the use of webscraping to collect enterprise information; the use of so-called earth observation data for several purposes; the use of data from mobile networks for official statistics; and the integrated use of several data sources for tourism statistics. Furthermore, CBS participates in the smart statistics project and various coordination projects.

2. Achievements of the CBDS since 2016

In a span of three years, the CBDS has grown from a small innovation incubator within CBS into an indispensable unit and a national and regional developer of innovative statistical information for policy making. CBDS now forms part of the larger R&D department within CBS. To date, a total of 26 experimental statistical products (also called beta products) have been published in collaboration with the statistical divisions within CBS and a network of more than 40 partners from the government, the business and the scientific community. In addition to product development, research on new data science methods has been carried out, data science training has been organised and there has been a strong focus on the acquisition of new data sources (data scouting). In this chapter we look back on some of the results achieved so far.

2.1 Experimental statistics

With experimental statistics CBS aims to improve current official statistics by making them more detailed, more real time and/or cheaper to make. Sometimes, our experimental statistics focus on new phenomena and make new statistical output. But not all of the work is focused on new or improved products. We also work on paid assignments for third parties and contribute to deepening methodological knowledge on data sources and data science methods and techniques without a direct application for our statistics.

An experimental statistic usually starts with a proof-of-concept that investigate whether a data science technique, new (big) or combined data sources or contentrelated hypothesis is valuable or not for statistical output. If we are convinced we have an interesting case for official statistics, we further developer the proof-of-concept into a beta product. For every beta product we publish articles in which we describe the data sources and methods that are used, how we deal with privacy issues and what are the outstanding (methodological) challenges if we want to develop these products into official statistics. These articles are published on the innovation website of CBS.

We make a clear distinction between experimental statistics and official statistics and we aim to elicit feedback. Ultimately, it is our goal to make these beta products ready for implementation in the official statistical process through methodological validation.

Examples of the beta products where we worked closely with our partners:

- Together with the Dutch police force, we were able to identify cybercrime cases from police reports using text mining techniques
- We used register data in a study that showed that machine learning is as good at indicating the likelihood of people moving within the next two years as a survey on this subject
- Together with Logius we worked on insights into the use of digital government services
- We developed new visualisations such as the dotmap
- We used transaction data from pharmacies to create a hayfever index
- We showed how social tension and emotions can be distilled from social media data
- Together with T-Mobile, we mapped out the crowds in municipalities with anonymised and aggregated mobile phone data.

These beta products are often developed in co-creation with partners in the ecosystem, sometimes initiated by these partners and sometimes on commission. Beta products have also found their way to the (social) media (see box). Below we present four beta products where two examples illustrate the relevance for the region and two examples illustrate the strength of cross border collaboration and international statistics (Appendix 1 provides a complete list of the published beta products). Most of these experimental statistics were funded through a subsidy from a ministry or a grant.

Data scientists working at CBDS

Beta products in the media

- 14 articles in technical journals on innovation
- 10 feedback items on average per week on beta products
- Most visited pages of the CBS.nl websiteare Dotmaps and migration background: 3,017 unique visitors
- Mapping solar power in a smart way: most often featured in the media
- Website visits since 2018 up until now: 55,000 unique page views

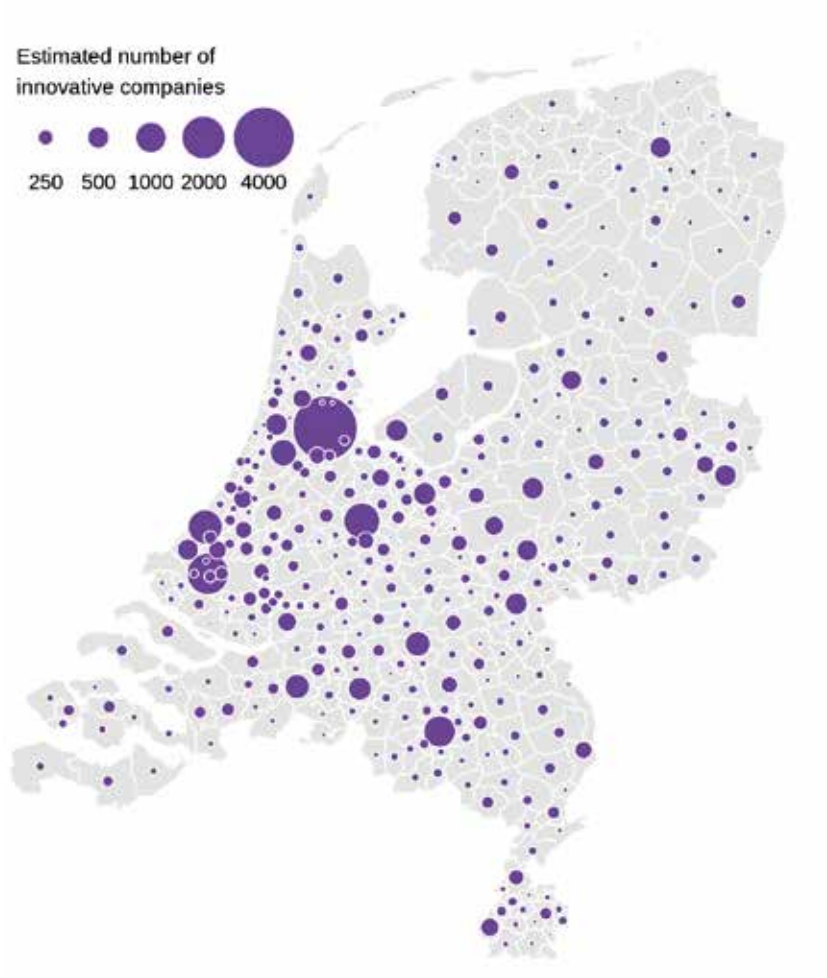
Example 1: Detecting innovative companies

How do you get a good overview of all the innovative companies in the Netherlands? At present, the survey on innovative companies by CBS covers companies with more than 10 employees. However, most start-ups are small companies. In order to include small innovative companies, a method has been developed that makes use of the text on a company's website. Statistics based on this method provide information so that a municipality or province can see whether the number of innovative companies in the region is growing or declining and what their characteristics are. In this project, funded by the Ministry of Economic Affairs, we worked together with Innovation Spotter, a small start-up itself.

The text on the homepage of the website is used to determine whether a company is innovative or not by looking at specific terms. Because we know which companies in the current survey are classified as innovative and which are not, the websites of those large companies are used to train the algorithm. This eventually resulted in a collection of terms

to determine innovation: words like 'technology', 'new product', 'innovation' and 'software'. In 93% of the cases, the final algorithm appears to identify the websites of companies well. Subsequently, 0.5 million small companies were selected from the CBS business register, texts were collected from websites and were classified using the algorithm. It was not known in advance whether these companies were innovative or not. Manual verification of the results confirmed that the algorithm correctly classified a very large number of companies as innovative. This new method ensures a more up-to-date, more complete and detailed picture, and at the same time is less of a burden on companies. The collected information can be combined with all kinds of background characteristics of the companies, such as size, turnover, etc. Applying this technique to the Dutch province of Limburg clearly shows the innovative character of the Brightlands regions (the larger the dots, the higher the number of innovative companies in Figure 1) and allows us to have a better visual image of innovation. The method also enables us to detect other characteristics of companies and will be applied in further projects.

Figure 1. Technological innovative companies at the municipality level in the Netherlands. The larger the dots, the higher the number of innovative companies.



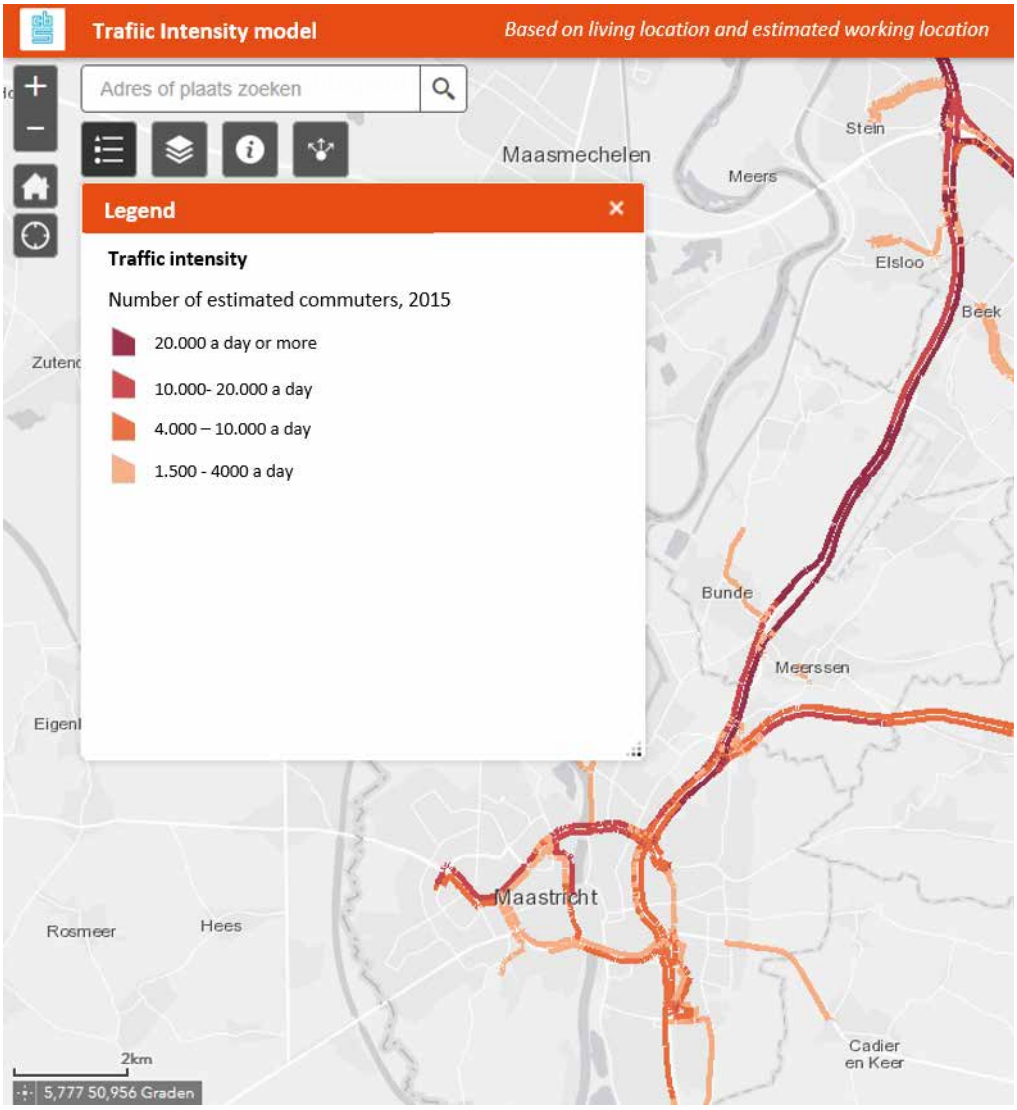
Example 2: On the way to motives for mobility

Policymakers are increasingly confronted with complex issues such as accessibility of public facilities in a city, reducing air pollution, the impact of new mobility solutions and the costs involved. In order to arrive at an effective policy, we map out where people are located and the motives behind moving between these locations. This provides tools for adaptation of mobility flows, or shifts in time or location during rush hours.

This project was a follow-up of a course organised by the Erasmus University and Rotterdam School of Management. The latter project was carried out together with the City of Rotterdam. In this project we linked new data sources, registry data and survey data. Examples of these new data sources are inductive-loop traffic detectors , anonymized data from mobile phone providers, data from navigation systems, public transport data. We use information from registers to identify various travel motives: information where a person lives and works provides insight into commuting, the education register indicates whether and where people study and the file with addresses of public facilities gives an indication of other motives.

We combine these movement motives from register sources with actual counts of movements from new data sources (for example data from the main public transport operator in Rotterdam, RET), thus creating a model. This provides more detailed visualizations of the reasons why people travel from A to B, the choices they make about the time of travel and the mode of transport (see Figure 2 below). These interactive visualizations are now available as beta products for road and public transport use, which can be viewed on the innovation website.

Figure 2. Estimated maximum contribution of daily commute on traffic intensity



Example 3: Automatically detecting solar panels with aerial photos

The current official statistics on solar energy are based on a survey among 350 importers of solar panels. This method only provides national figures on an annual basis, whereas monitoring the energy transition demands real time and detailed statistics at both national and regional levels. After all, it is important to know how much solar energy is generated and used by households and enterprises and whether it is increasing or not. More detailed information would also allow for stabilisation of the grid (matching of offer and demand) and measuring the achievement of sustainability targets.

For information at the regional level, it is important to have a complete picture of the locations of solar panels. To that end, the CBDS has combined several existing administrative sources (such as the VAT refund of purchased solar panels) with new sources such as the registration of solar panels in the Netherlands and open source data. In addition, we made use of the Addresses and Buildings (BAG) key registration, so that the breakdown by type of dwelling is possible.

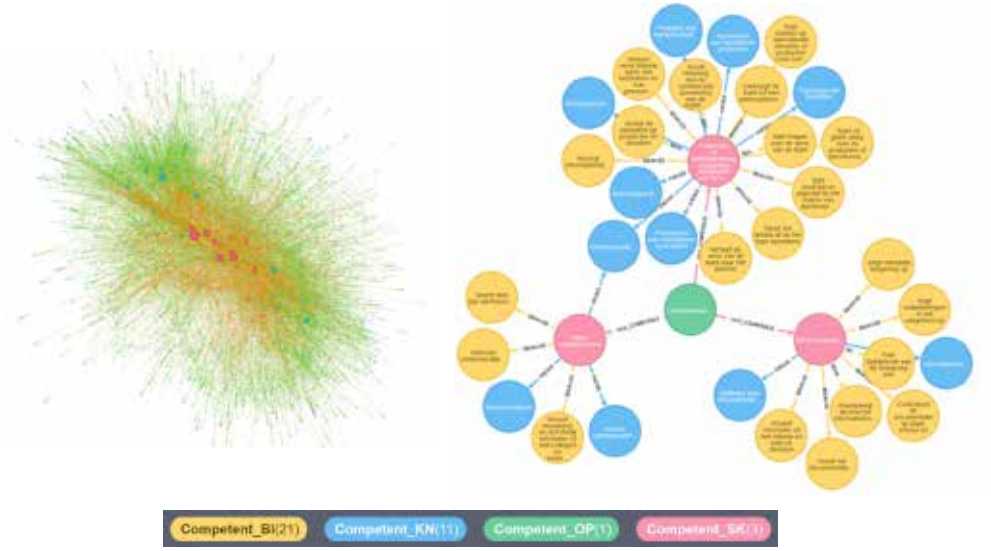
Furthermore, models have been developed using open data with characteristics of solar installations and weather data and on data from the high-voltage network. In doing so, we take into account the inclination of the roof and weather conditions.
We are also working with BISS and the Open University in a European context (North-Rhein Westfalen in Germany and Flanders in Belgium) on the use of aerial photographs to detect solar panels. This project was funded through a Eurostat grant. In this study, machine learning techniques are used to automate the detection of solar panels, which are validated with existing register information and completed with the additional data sources mentioned above. As such CBDS provides insight into where solar panels are located (see Figure 3 below), the potential for installing this type of renewable energy according to type of district and neighbourhoods and the amount of solar power yield that is generated by these panels.

Figure 3. Aerial picture of the City of Heerlen: green dots are roofs with solar panels, red dots without

Example 4: Signalling new labour market competencies

The public employment service (VDAB) of the Belgian region of Flanders has developed a complex network of professional competencies called Competent, where skills are linked explicitly to people and where occupational profiles correspond to sets of competencies. The Dutch Employee Insurance Agency (UWV) wants to extend its own classification of professions with the ones of Competent correlating with the Dutch labour market. In this project, funded by Interreg Europe funding, CBS is developing an algorithm called Big Data Ontology Enrichment (BiDOE) to supplement the Competent ontology (see figure 4 below) on the basis of (actual) big data: large amounts of labour market documents such as Dutch vacancy texts are used to extend the coverage of Competent beyond the Flemish market and to make this complex system responsive to current trends. This data is provided by Textkernel.

Figure 4. A visual representation of the Competent ontology, containing occupational profiles (OP), skills (SK), knowledge items (KI) and behavioural indicators (BI), all linked together



BiDOE identifies Big Data skills (BDS) from the source documents and finds connections to existing Competent concepts by scoring the overlap in skill descriptions. When a BDS has a low score, it can be either a fundamentally new skill or a new (or regional) description of an existing skill. Further, BiDOE can identify a new occupational profile when finding a new and prevalent combination of skills. Output of BiDOE will be presented periodically to domain experts from VDAB and UWV using a front-end tool. They will evaluate the output and decide whether and where to store a new entry in the Competent ontology. They will be offered context information on the BDS sources, some of which is based on the high-scoring connections found in the same source documents. Updates to the ontology and feedback from the user interaction with the front-end tool will be incorporated into BiDOE, so that it will not repeat its suggestions.

The efficiency of a labour market depends on its ability to match people (and their competencies) to job openings. By using the Competent ontology, the similarity between professions can be directly derived from the degree of skill overlap and connectivity. The deployment of BiDOE will make this job market tool more dynamic and applicable to an interregional market, so that job seekers can find appropriate jobs, quickly and across borders. They will also be able to develop necessary skills more efficiently, as the deployment of BiDOE will enable the publication of statistics such as trends in skill demand.
Implementation: from beta products to official statistics

Implementation: from beta products to official statistics

CBS is making steps to increase the number of experimental statistics that are implemented as an official statistic. So far, two experimental statistics have been implemented: the traffic intensity using inductive-loop traffic detectors and the consumer price index using webscraping and scanner data. CBDS is working hard together with the statistical divisions to implement a greater number of these experimental statistics (as stated in the Bucharest memorandum): maritime statistics, detecting innovative companies, cybercrime statistics and more detailed statistics on solar power yield are some of the experimental statistics under investigation. At least two of these are scheduled to go in production in 2020.

2.2 Data science knowledge, expertise and research

CBS supplies microdata to research institutes under certain conditions and continuously makes data available via Statline or the open data app (this can also be done automatically via interfaces). An important part of the activities of the CBDS is the research and development of data science expertise in collaboration with local partners (the Brightlands Smart Services Campus, Maastricht University, the municipality of Heerlen, the Zuyd University of Applied Sciences and the Open University), national partners (among them Universities of Eindhoven, Amsterdam, Utrecht) and international partners such as the other NSIs through for example staff exchanges. In Appendix 2 you will find an overview of the scientific output that has been achieved through collaboration with CB(D)S. By linking a total of 6 PhDs to the CBDS, we offer a lot of space for applied research.

The European Union promotes the international exchange of data scientists from the public, private and academic sector. In light of this, CBDS had attracted a grant from the European Union's Horizon 2020 research and innovation program which allows its data scientists to work temporarily for another type of organisation (private or academic). With this, the project is "in a unique position to deliver cutting edge multi-disciplinary research to advance academic thinking on Data Science in Europe".

The grant under the NeEDS program (www.riseneeds.eu) consists of six academic participants and eight industrial partners from Europe, the United States and Latin America. It aims to advance academic thinking on data science in Europe and beyond through multidisciplinary research and to improve the data science capabilities of industry and the public sector. The project is led and coordinated by the Copenhagen Business School. The exchange of knowledge and experience is embraced and facilitated by the Copenhagen Business School: 3 CBDS employees went on an exchange with a different institution (University of Oxford, Copenhagen Business school and the University of Sevilla) and we received Yu Zhang, a PhD student at the University of Oxford, who has conducted research on Active Learning at CBS.

Michel Dumontier, professor at Maastricht University:
"The CBDS can play a crucial role in the regional ecosystem by facilitating studies with data and knowledge, preferably in combination with other data. NSIs form an important role in this."

Yu Zhang , Ph.D. student in the Computer Science Department of Oxford University:
"My research interest lies in intelligent user interface and data visualization. My previous projects mainly focused on data labeling systems. Many supervised machine-learning techniques require a large amount of human-labeled training data to train a model. In many real-world machine learning applications, the preparation of labeled training set takes the primary time and monetary cost. While much research efforts have been spent on developing machine-learning algorithms, much less attention has been paid to methods for preparing labeled datasets, which are the fuel for machine learning algorithms to function."

For research activities we have further explored funding through the National Research Agenda, Eurostat and Horizon 2020. The CBDS promotes cutting edge research on topics such as text mining, machine learning, deep learning and AI, synthetic data and visualization techniques. In Appendix 3 you will find an overview of the research activities which fit into the CBS research program.

The 2-day seminar on Big Data Methods at the Brightlands Smart Services Campus (BSSC) in Heerlen, October 2018

CBS has invited more than 40 partners from the CBDS ecosystem (www.data-ecosysteem.nl) at the BSSC in Heerlen. The highlights were the two well-attended editions of Big data Matters, both with more than 200 guests, from more than 100 organisations from at home and abroad. The 2-day seminar on Big Data Methods, an international event aimed at universities and other statistical institutes, explored the latest developments in big data methods. We have also organised workshops (such as an Open Data Visualisation workshop with our regional partners), held Master classes (such as Using Existing (big) Data to Advance your Product or Idea in health or life sciences), contributed to hackathons such as Brighthack and organised inspiration sessions around energy transition together with students from the Business Intelligence and Smart Service (BISS) master at the University of Maastricht.

2.3 Education

Training talented people with data science skills is essential to make data an important pillar of the economy. This is in line with both the set-up of the CBDS as an open innovation platform and the ambition of CBS to increase the level of knowledge about data science within the government. The CBDS has also offered training courses for its own staff on text mining, artificial intelligence and visualisation techniques and to external future data scientists and policy makers through the CBS Academy and the European Statistical Training Program organised by Eurostat.
On the work floor of the CBDS we have a mix of data scientists, methodologists, students for our data science activities. CBS has several visiting professor positions at different universities (Maastricht, Utrecht, Amsterdam) and Piet Daas, our senior methodologist recently started a visiting professorship in Big Data for Official Statistics at the University of Technology in Eindhoven. Other authorities (such as the Ministry of the Interior and Kingdom Relations BZK) but also other NSIs (from UK, South Korea, Norway, Finland, Mexico) have seconded people to us on a temporary basis.

Maria Yli-Heikkilä (Statistics Finland):
"Visiting CBDS showed me how state-of-the-art data science is deployed in statistical production. CBDS has built a dream team of people with various skills across the data science. I learned a lot about practices in data analytics, and had fruitful conversations, for example about remote sensing, machine learning, and developing new indicators."

Younghee Nam (Kostat South Korea)*:*
"I am seconded as a trainee from South Korea Kostat. I have worked with the Labour Force Survey, E-commerce Survey and PR at Statistics Korea. It has been 9 months since I arrived at the Center for Big Data Statistics. I'm interested in job vacancy statistics using big data sources and matching them with survey data. My research will be on how to apply what I have learned about statistics in Korea such as economic register data or labour survey data. It is a great pleasure and honor for me to learn, experience, and be shared other different techniques and knowledge from CBDS."

Students sometimes progress from an internship at the CBDS to an internal PhD position and sometimes newly graduated students from the region find a job at CBDS as a data scientist. At the moment we have an average of 12 to 15 internships per year at professional and academic educational level. Three students follow the trainee program Knowledge Engineering at Work (KE@work) from the Maastricht University and are working at CBDS for a period of 2 years on the impact of air pollution on society, Big Data and diluted signals and deep learning techniques in official statistics.

*Aaron Slots (*internship student):
"I am in the third year of my Bachelor's degree in computer sciences at Avans University of Applied Sciences in 's-Hertogenbosch.
I am working on public transportation data with Marko Roos as supervisor.
I am passionate about doing a Master's at the Jheronimus Academy of Data Science after my Bachelor's degree. Next to my studies I do Krav Maga. I like to play board or card games as well."

Dewi Peerlings (started as a trainee in 2016 and is now working as a PhD student at the CBDS):
"I am very grateful that CBS gives me the opportunity to continue the research I have started during my internship with them before with the help of Marco Puts. Combining research at Maastricht University together with the environment that the CBDS creates is exciting and very interesting. Being able to combine this in Limburg is even more special."

Jonas Klingwort (PhD student since 2017 at CBDS):
"I am close to finishing my PhD, which is on the combination of sensor, survey, and administrative Data. My project benefited in particular from the collaboration between the CBDS and the Methodology department. In particular, the different views of the professions from both departments enabled me to ask the right questions to derive the most insights out of the data. I am already looking forward to the projects I will be working on after my doctoral thesis at the CBDS."

The Scientific Director of the CBDS, Sofie De Broe, actively participates in the education program. She sits on the board of the Maastricht University Master Business Intelligence and Smart Services program and on the Advisory Board for setting up the Bachelor Business Analytics with Prof. Dr. Alex Gregoriev.

Prof. Dr. Alex Gregoriev (Maastricht University):
"We undertook two student projects for the BISS master at the UM on GIS systems, this is new for the students and was fully supported by the CBDS and Urban Data Centres."

2.4 Ecosystem allowing data scouting and sharing

The results above show that the CBDS aimed to create impact for the local, national and international ecosystem through the realisation of innovative (beta) products as well as research output and data science education. Access to new data sources is crucial to create new relevant output. In addition to the large amount of data that CBS already has at its disposal, we are proud of the access to additional (big) data sources that we have realised in recent years, including mobile phone data, navigation system data, data on company networks, public transport data, online job vacancy data, flight freight data, smart meter data, housing market data and information on websites. In addition, all kinds of open data sources were used, such as social media, satellite data, KNMI data, Google APIs, etc (see Appendix 4 for an overview). CBS appointed a datascout who is continuously looking for data, see below for a summary of what datascouting entails.

Datascouting: objectives and definitions
Datascouting at CBS has the objective to organise and facilitate the exploration and acquisition of new (big) data sources for the purpose of statistical innovation (e.g. improve existing statistics, or to make new statistics). For many years, CBS has been acquiring new data sources for individual statistics, but the professionalisation of the data scouting function and its structural positioning in the organisation is a more recent development. This reflects our recognition that secondary (big) data sources are becoming increasingly important for official statistics due to the fast developments around datafication in today's information society.

Given this background, today's ambition at CBS is to structurally anchor data scouting in the organisation, with an integral approach. An integral approach means that new data sources are evaluated with a holistic view across individual thematic domains, rather than solely assessing them based on their relevance to single statistics. This requires establishing and formalising them as a central support function in the organisational structure. At CBS, this is done by creating the datascout function and developing an organisation-wide Datascouting Community, to coordinate and streamline all datascouting activities throughout the various sectors of the organisation.

In essence, datascouting is a bridge-building function between the internal stakeholders at CBS and external stakeholders and partners. Datascouting is based on building relations

with other organisations, and liaising with business, legal, technical and domain experts on what is possible (and what is not). It is thus by definition a process that involves a diverse range of stakeholders, including (but not limited to): statistical domain experts, product developers, data scientists, management, account managers, legal experts, subsidy experts, IT specialists and communication staff.

It is important to recognize that the datascouting process is not a uniform process. Each project is different, depending on its specific mix of data needs, stakeholder interests, the nature of the data etc. However, there are a number of process steps and aspects that play a role in every datascouting project, which we briefly summarise below.

At the start of the process is the definition of the data needs, together with both the internal and external users. To this end, the datascout maintains an inventory of data needs, our so-called wishlist. This wishlist contains ideas and suggestions for potentially relevant sources or types of data that can come from within the organization as well as from contacts with external parties. Our wishlist is a broad pool of ideas that gives a good overview of which data sources might be relevant and which sources might be more closely inspected for potential use.

Such closer inspection is a crucial next step in the datascouting process, and aims to establish whether a potential source is suitable and usable for our statistical purposes. In this step, we first perform a quick-check on various dimensions, and if this results in a promising initial evaluation, we perform a much more in-depth evaluation of the data source, typically together with the data owner.

When this inspection of the data source leads to a positive evaluation and agreement is reached with the data owner to share the data, there are several options for accessing the data. CBS offers multiple ways for partners to deliver data, from secure connections for data transport to methods where the data is not transported but CBS is granted access to the data (either electronically or on-site). CBS is currently also carrying out several pilot studies on innovative ways to securely share data, such as multi-party computation. Finally, once the data is accessible to the CBDS, it can be used for further testing in pilots and beta-products, eventually leading to new or improved official statistics.

The key dimensions of in-depth evaluation of the new data sources are:

- Needs assessment: who are the users of the data (internal/external), and what are their goals? How would it interfere with the landscape of existing processes and partners? Etc.
- Data characteristics and quality: how does the data look in terms of aspects such as representativeness, validity and bias, periodicity, structure and data model, potential for linking with other data? Etc.
- Technical aspects: what are the characteristics in aspects such as accessibility (including options for remote or on-site access), transport/download options and limitations, storage implications, security, format, interoperability? Etc.
- Legal and ethical aspects: is CBS allowed to use the data by means of a legal mandate and its work program? Is the data source compatible with requirements from the national legal framework (e.g. CBS legislation) and international requirements (e.g. GDPR)? Etc.
- Business case: is there a positive business case for acquiring and using the data (both for CBS and the data owner)? How can CBS and data owner arrange a sustainable partnership with structural access for CBS to the data in the long term? Etc.

Sensor data

One of the CB(D)S' objectives is to investigate the potential of sensor data. This would include the use of e.g. inductive-loop traffic detectors' data and satellite images. At the Department of Methodology, CBS also looks at sensor data at micro level, i.e. at the chain of data from sensor data and acquisition to an official statistical product at enterprise level. For this theme, smart farming was chosen as a first use case (but other smart industry sectors could be explored as well). In this case, contact was sought with an innovative farmer and the Eindhoven University of Technology to gain access to the data. Very soon, farmers asked: "What is in it for me?". Relation management was very important to establish trust in order to get access to the data. The ultimate goal of using sensor data is threefold: can we use sensor data instead of surveys as a source for existing official statistical purposes, thus lowering the response burden? The second goal is to derive information from these sensor data that is addressing some of the questions policymakers and farmers have, resulting in new, additional statistical products? A third goal is "to close the data cycle", and produce timely indicators that are of use to farmers themselves. As to the first goal: the first results indicate that these sensor data are promising as a source for statistics, but we still have a long way to go. CBS however has the ambition to be involved right from the very start. In addition to the use of existing sensor data there is a possibility of actively installing proprietary sensors for statistical purposes (i.e. a network of sensors installed by CBS), and in this way acquire relevant data instead of using surveys. The ambition is to have answers to these questions within the time frame of the 5-year research plan. In the field of sensor data, the following themes among others are elaborated upon: potato harvests, livestock farming, horticulture and water. We work together with the Eindhoven University of Technology, HAS High School Den Bosch, Fontys High School Venlo, the water boards and the water company Limburg.

Sharing data or working in joint collaboration

What we notice when accessing new sources is that, in addition to technical issues, ethical and legal issues often arise as well. What are we legally allowed to do with the data? What is legally accepted? What needs to be arranged? We have therefore started several projects to develop not only technical solutions to facilitate data sharing, but also an associated ethical and legal framework. We have done this through Techruption and also together with Maastricht University in the Responsible Value Creation Big Data project (VWData), a part of the National Science Agenda (NWA) programme. This is one of the five flagship

projects we are carrying out together with Maastricht University. On this topic, the CBDS is also active internationally. CBS has a lot of expertise in the methodology groups on this theme and organised a Privacy Preserved Data Sharing Seminar on 1 November 2019 for its international partners.

Michel Dumontier (professor at Maastricht University):
"With the VW data, we managed to identify a topic in which both organisations could work on their strategic goals and we also found funding for it. These combinations are the most powerful and should be the model to strive for."

The CBDS together with the statistical divisions seek to work closely with statistical organisations across the border so we can join forces in making cross-border statistics. A very fruitful relationship exists between CBS and the statistical offices of Belgium and Germany. On 17 December 2019, CBS and Stats Flanders, with Roeland Beerten as its National Statistician, jointly organised a conference called "Data science for better decisions" in Brussels.

Roeland Beerten: (Statistics Flanders)
"As a relatively young setup, Statistics Flanders has already created links with the team at the CBDS. As close neighbours there are lots of interesting data topics we will be working on – for example on cross-border statistics. We are also sharing experiences and skills through a joint conference on Data Science for Better Decisions, and we will be looking at organising joint workshops and seminars. Having this strong cooperation with the experienced CBDS team will really help us in maximising the opportunities from big data and data science applications for official statistics to improve decision-making."

Many foreign delegations of statistical offices visit CBS to learn about innovation and the approach taken. To facilitate this, a space has also been set up for groups in which the use of big data is made visual. The CBDS was also an invited speaker at conferences such as the APG Investor Day, the Spring Meetings of the International Monetary Fund (IMF) in Washington, the seminar on big data organised by INEGO in Mexico City and the ISI international conference in Kuala Lumpur.

Internally, we are also working on making data science an important part of our own statistical process. To this end, we have set up a Data Science Community, which creates a strong network of data scientists from the CBDS with the statistical divisions in order to disseminate and reuse knowledge.

2.5 Conclusion and outlook

As the CBDS management team, we can look back on three successful years. We will do everything we can with our team to continue to maintain these results in the coming years. In order to achieve even better results, we are continuously aiming to improve our output to achieve our strategic goals and to increase our impact through the production of innovative statistics that address complex policy questions. After all, the success of the CBDS lies in permanent change.

We are explicitly aiming at the implementation of beta products into an official statistical product. A strong focus on this point has led to four projects currently underway that will lead to official statistics in 2020. This is anticipated in areas such as solar energy production, cybercrime, innovative companies and maritime statistics.

We will continue to develop data science skills of our own staff but also to be a centre that stimulates data science expertise within and beyond the Dutch borders. The CB(D)S needs to invest further in its IT infrastructure to allow large computation exercises.

We will continue to look for new data sources and seek legal support to access the data that allow us to make official statistics. Sensor data offer great potential to replace surveys on specific topics and are a source of new statistical information as well, while reducing the response burden. These projects are ideally developed in a consortium consisting of all the partners in the quadruple helix: statistical offices, universities or knowledge institutes, private partners, citizens. As such, we are making the most of our ecosystem whilst increasing our relevance and impact for national and local policy through the provincial and urban data centres.

Appendices

Appendix 1 Overview of beta products

- Measuring price developments of financial real estate December 2019
- More than 1 billion euros spent at European webshops December 2019
- Regional differences between owner-occupied homes in the picture December 2019
- Pupils to school by bike or car December 2019
- The value of combining care data December 2019
- Linked-In graduate skills December 2019
- Social tensions and emotions in society 2019
- Determining forms of innovation using text analysis October 2019
- Insight in the users of the website MyGovernment July 2019
- How many visitors on Kingsday in Amsterdam April 2019
- Update hayfever index April 2019
- Estimation of current moving behaviour in the Netherlands February 2019
- Detecting cybercrime in declarations February 2019
- Modelling chance to moving with machine learning January 2019
- Automatically detecting solar panels using areal pictures December 2018
- Measuring movements using mobile phone data December 2018
- Social tension-indicator: measurement point in society December 2018
- Classify occupations with data in online vacancies November 2018
- Deriving a wish to move from social media October 2018
- Innovation in small businesses July 2018
- Better flash estimation on inland waterways thanks to sensor data June 2018
- Hay fever index June 2018
- Deepening of maritime statistics: trans-shipment June 2018
- Predictions of pregnancies economic recessions? June 2018
- Smart solar mapping May 2018
- Towards motives for mobility April 2018

Appendix 2 Overview of scientific output

Papers and bookchapters

2017:

- Brakel, J.A. van den, E. Söhler, P. Daas and B. Buelens (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Survey Methodology, vol 43 pp. 183-210

2018:

- Brakel, J.A. van den (2018). Combining time series obtained with repeated sample surveys and auxiliary series from Big Data sources for official statistics. Report for the methodology advisory committee January 2018.
- Buelens B., J. Burger, J. van den Brakel (2018). Comparing inference methods for non-probability samples. International Statistical Review. Vol 86, pp. 322-343.
- Delden van, A., Daas, P., ten Bosch, O., Windmeijer, D. (2018). Tekstanalysemethoden: Toepassingen in de officiële statistiek. STAtOR 2 (juni), pp 8-12.
- Hürriyetoğlu, A., Oostdijk, N., & van den Bosch, A. (2018). Estimating Time to Event based on Linguistic Cues on Twitter. In K. Shaalan, A. E. Hassanien, & F. Tolba (Eds.), Intelligent Natural Language Processing: Trends and Applications (Vol. 740). Springer International Publishing. http://www.springer.com/cn/book/9783319670553.
- Tennekes, M. (2018) tmap: Thematic Maps in R. Journal of Statistical Software 84(6), 1-39.

2019:

- Braaksma B., Zeelenberg, K. , De Broe S. (forthcoming in 2020). Big Data in official statistics: a perspective from Statistics Netherlands. Forthcoming in: Lilli Japec, Lars Lyberg et al (eds): Big Data Meets Survey Science, Wiley: New York.
- Klingwort, J., Buelens, B. & Schnell, R. (2019) Capture-recapture techniques for transport survey estimate adjustment using road sensor data. In: Social Science Computer Review. Special issue on big data 39 (4).
- Klingwort, J., Buelens, B., Burger, J. & Schnell, R. Correcting survey measurement error using road sensor data. Submitted to: International Total Survey Error Workshop (ITSEW2019).

2020

- Curier, R.L., Ziemons, H., de Jong T, Iren D., Bomuri S. Semi-automated analysis of aerial images for the detecting of photovoltaic solar panels. Proceedings of Statistic Canada Symposium 2018.
- Daas P., van der Doef S. Detecting innovative companies via their website. Paper for the methodology advisory committee at Statistics Netherlands 17 September 2019.
- De Broe S., Struijs P., Daas P., van Delden A., Burger J., van den Brakel J., ten Bosch O., Zeelenberg K., Ypma W., Updating the Paradigm of Official Statistics: new quality criteria for integrating new data and methods in official statistics. Paper for the methodology advisory committee at Statistics Netherlands 17 September 2019.
- De Broe S., Meijers R., ten Bosch O., Buelens B., Laevens, B. Priem A., de Jong T. and Puts M. (2019). From experimental to official statistics: The case of solar energy. IAOS 35 (3):371-385.
- Delden A., Towards implementing a text mining model to detect cybercrime in police reports . Paper for the methodology advisory committee at Statistics Netherlands 17 September 2019.

2020

- Klingwort, J., Buelens, B. & Schnell, R. Transport survey estimate adjustment by permanently installed highway-sensors using capture-recapture techniques.Proceedings of Statistics Canada Symposium 2018: Growth in Statistical Information: Challenges and Benefits.
- Laevens B. and ten Bosch O., Towards an observational daily and regional solar energy statistic for the Netherland, Methodology advisory board, 2019.
- Puts, M., Tennekes, M., Daas, P.J.H., de Blois, C. (2019) Using huge amounts of road sensor data for official statistics, AIMS Mathematics, 4(1), 12-25.
- Schiavoni, C., F. Palm, S. Smeekes and J.A. van den Brakel (2019). A dynamic factor model approach to incorporate big data in state space models for official statistics. Discussion paper January, 2019, Statistics Netherlands, Heerlen.
- Waal, T, de, Delden, A. van en Scholtus S. (2019). Multi-source Statistics: Basic Situations and Methods. International Statistical Review.

Abstracts

2017

- Tennekes, M., Offermans, M. Heerschap, N. (2017) Determining an optimal time window for roaming data for tourism statistics. Paper presented at the NetMob 2017 conference, Milan, Italy.
- Tennekes, M., Jonge, E. de (2017) The compositional dot map: a visualization of spatial data. Paper presented at the 2017 New Techniques and Technologies for Statistics (NTTS) conference, Brussels, Belgium.

2018

- Buelens B., De Broe S., Meijers R., Bosch ten O. and Puts M. (2018). From Beta to Offstat: the case study of energy statistics. BigSurv conference, Barcelona October 26-27, 2018.
- Bosch, O. ten, Windmeijer, D., Delden, A. van and Heuvel, G. van den (2018). Web scraping meets survey design: combining forces. Bigsurv conference, www.bigsurv18.org , October 25-27, 2018, Barcelona, Spain.
- Klingwort J., Buelens B., Schnell R. (2018). Capture-recapture Techniques to Validate Survey with Sensor Data, BigSurv conference, Barcelona October 26-27 2018.
- Salgado, D. et al. (2018) Estimation of population counts combining official and mobile phone data. Paper for the European Conference on Quality in Official Statistics 2018, Krakow, Poland.
- Snijkers G., De Broe S. (2018). Smart business statistics: how to integrate S2S technology and official statistics? European conference on Quality in official Statistics. Krakow June 26-29 2018.
- Struijs P., De Broe S. (2018). Big data strategies for official statistics. European conference on Quality in official Statistics. Krakow June 26-29 2018.
- van der Doef S., Daas P. Identifying innovative companies on-line. BigSurv conference, Barcelona October 26-27 2018.
- Tinto, A., F. Bacchini, B. Baldazzi, A. Ferruzza; J.A. Van den Brakel, R.M.A. Willems; N. Rosenski, T. Zimmermann, Z. Andrási, M. Farkas, Z. Fábián (2018). International and national experiences and main insights for policy use of well-being and sustainability frameworks. Paper for the 16th IAOS conference.
- Zeelenberg K., Braaksma B., De Broe S. (2018). Big data in official statistics: a perspective from Statistics Netherlands. BigSurv conference, Barcelona October 26-27 2018.

2019

- Beuningen, van J., Meiers R., Jong de T., Burger J., Buelens B. (2019). Replacing a survey question by predictive modeling using register data. Paper that will be presented at the 62nd ISI World Statistics Congress, Kuala Lumpur
- Buelens, B. and J.A. van den Brakel (2019). Estimating unmetered photovoltaic power consumption using causal models. Invited paper presented at the NTTS, 12-14 March 2019, Brussels
- Bosch ten O. et al. (2019). ClairCity: official statistics as an enabler in a citizen-led European air quality project. Paper presented at the NTTS 2019 Conference, Brussels, 12 – 14 March 2019.
- Burger, J., Buelens, B., de Jong, T. & Gootzen, Y. (2019). Replacing a survey question by predictive modeling using register data. Paper presented at the 62nd ISI World Statistics Congress, Kuala Lumpur
- Daas P. and Puts M. (2019) IT infrastructure for big data and Data science: Challenges at Statistics Netherlands. Paper presented at the NTTS 2019 Conference, Brussels, 12 – 14 March 2019.
- Daas P. (2019) Using big data in official statistics. Paper presented at the DagStat Conference, München, 18-22 March.
- Daas, P., Gootzen, Y., van der Doef, S. (2019) Detecting Innovative Companies via Their Website. Abstract for the 2nd annual Symposium on Data Science and Statistics, May 29-June 1, Bellevue, WA, USA.
- De Broe, S. (2019). How to create innovative environments to harness big data in official statistics. Paper presented at the 62nd ISI World Statistics Congress, Kuala Lumpur
- De Broe S., Schouten B., Snijkers G. (2019). Sensor data at the heart of innovation in official statistics. Paper presented at the 62nd ISI World Statistics Congress, Kuala Lumpur
- De Broe S. (2019) Big data to improve policy and decision making: The Experience of Statistics Netherlands, Mexico City, 16-18 May
- Gootzen Y. and Puts M. (2019), Mobile device tracking and, transportation mode detection. Paper presented at the NTTS 2019 Conference, Brussels, 12 – 14 March 2019.
- Klingwort, J., Buelens, B., Burger, J. & Schnell, R. (2019). Implementing big data in official statistics: Capture-recapture techniques to adjust for underreporting in transport surveys using sensor data. New Techniques and Technologies for Statistics NTTS: Brussels, Belgium. 12.03.2019.
- Klingwort, Jonas, Bart Buelens, Joep Burger and Rainer Schnell *(2019*)*: 'Graph-based Inference from Non-Probability Road Sensor Data'. Ed. by Hocine Cherifi, José Fernando Mendes, Luis Mateus Rocha, Sabrina Gaito, Esteban Moro, Joana Gonçalves-Sá and Francisco Santos. Complex Networks 2019. The 8th International Conference on Complex Networks and their Applications. Book of Abstracts. Lisbon: 599–601.
- Ponsen, M., Gootzen, Y., Puts, M., Tennekes, M., Jonge, E. de, Shah, S., Offermans, M. Towards Official Tourism Statistics – Machine Learning for Processing Signalling Data. Paper to be presented at NetMob 2019, 8-10 July, Oxford
- Ricciato, F., Wirthmann, A., Tennekes, M., Sakarovitch, B., Radini, R., Salgado, D. Towards a Reference Methodological Framework for processing of Mobile Network Operator data for Official Statistics. Paper to be presented at NetMob 2019, 8-10 July, Oxford
- Scannapieco M., Stateva G. and Struijs P. (2019). Estimating Enterprise Characteristics from Web Data: Achievements and Future Developments. Paper presented at the NTTS 2019 Conference, Brussels, 12 – 14 March 2019.
- Tennekes, M. (2019) Reproducible maps for everyone. Paper presented at the NTTS 2019 Conference, Brussels, 12 – 14 March 2019.
- Van den Brakel, J.A. (2019) Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources – Makswell. Paper presentation at the NTTS, 12-14 March 2019 Brussels

- Van den Brakel, J.A., C. Schiavoni, S. Smeets, and F. Palm (2019). State-space dynamic factor model for now casting unemployment. Invited paper for the 62nd ISI World Statistics conference Kuala Lumpur.
- Van den Brakel, J.A. (2019). Big data in official statistics. Invited paper for the Conference on Current trends in Survey Sampling, 13-16 August 2019, Singapore.

Reports

2018

- Consten, A., Chavdarov, V., Daas, P.J.H., Horvat, V., Maslankowski, J., Quaresma, S., Scannapieco, M., Six, M., Tuoto, T. (2018) Report describing the IT-infrastructure used and the accompanying processes developed and skills needed to study or produce big data based official statistics. Deliverable 8.3, Workpackage 8, ESSnet big data, 5 March 2018.
- Consten, A., Puts, M., de Wit, T., Bisioti, E., Pierrakou, C., Bilska, A., Bis, M., Langsrud, Ø. (2018) Determining emissions using AIS. Deliverable 4.6, Workpackage 4 ESSnet big data, 31 March 2018.
- Consten, A., Puts, M., de Wit, T., Bisioti, E., Pierrakou, C., Bilska, A., Bis, M., Langsrud, Ø. (2018) Report about possible new statistical output based on (European) AIS data. Deliverable 4.7, Workpackage 4 ESSnet big data, 31 March 2018.
- Luomaranta, H., Puts, M., Grygiel, G., Righi, A., Campos, P., Grahonja, Č., Špeh, T. (2018) Report about the impact of one (or more) big data source (and other) sources on economic indicators. Deliverable 6.6, Workpackage 6, ESSnet Big Data, 16 March 2018 (Draft).
- Luomaranta, H., Puts, M., Grygiel, G., Righi, A., Campos, P., Grahonja, Č., Špeh, T. (2018) Report and recommendations about the methodology and process of calculating estimates for at least one early economic indicator. Deliverable 6.7, Workpackage 6, ESSnet big data, 16 March 2018 (Draft).
- Luomaranta, H., Puts, M., Grygiel, G., Righi, A., Campos, P., Grahonja, Č., Špeh, T. (2018) Example of calculated concrete estimates for one of the economic indicators with quality assessment of the input, throughput and output phase of the process. Deliverable 6.8, Workpackage 6, et: ESSnet big data, 16 March 2018 (Draft).
- Luomaranta, H., Puts, M., Grygiel, G., Righi, A., Campos, P., Grahonja, Č., Špeh, T. (2018) Example Report and recommendation about IT infrastructure needed for the storage, analysing, combining data sources and the process of calculating early economic indicators. Deliverable 6.8, Workpackage 6, ESSnet Big Data, 16 March 2018 (Draft).

2019

- Daas, P., Harmsen, C., Offermans, M. (2019). Kwaliteitstoets Mezuro, CBS rapport, March.
- Brakel, J.A. van den et al. (2019). Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources; Aspects of existing databases, traditional and non-traditional data sources and collection of good practices. Report of Makswell Deliverable 2.1.
- Brakel, J.A. van den et al. (2019). Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources; Methodological aspects of using big data. Report Makswell Deliverable 2.2.

– Consten, A., de Wit, T., van der Spoel, M., Dahlmans, D., Meijers, W., Schenau, S., Bisioti, E., Pierrakou, C., Bilska, A., Bis, Ø. (2019) Defining products for AIS implementation. Deliverable E.1, Workpackage E ESSnet Big Data II, 19 May 2019.
– Laevens B. and ten Bosch O., Towards an observational daily and regional solar energy statistic for the Netherland, Methodology advisory board, 2019.
– Van der Valk, J., Souren, M., Tennekes, M., Shah, S., Offermans, M., De Jonge, E., Van der Laan, J., Gootzen, Y., Scholtus, S., & Mitriaieva, A. (2019). Experiences of using anonymized aggregated mobile phone data in The Netherlands. In: City data from LFS and big data. Report for the European Commission. https://ec.europa.eu/regional_policy/en/information/publications/studies/2019/city-data-from-lfs-and-big-data.

Appendix 3 Research focus @ CBDS

The section on big data is described in the CBS research plan Big Data 2020-2025 and is coordinated by Piet Daas. This document provides an overview of the research and activities that are partly carried out by the CBDS.

As part of this research programme, CBS will in the coming years invest in strategic relationships with Universities such as of Southampton, Maastricht (Artificial Intelligence and Data Science), Eindhoven, VU, JADS, MIT and research institutions such as VITO and TNO. Training staff through the Academy and setting up a 'cutting edge' infrastructure is also essential.

1. CBS will position itself as a data hub and as a leading data-science institute:
   - CBS as a data hub: CBS can load external big data sources and link them to internal data in a safe, efficient and consistent way. Important points of attention in this respect are:
     a). A consistent approach to data scouting where aspects such as data quality, source stability, data generation process, possibilities for output for official statistics emerge. Examples of new data sources and collaborations are:
       - Sensor data quality and their applications (TU Eindhoven);
       - Aerial photographs (Open University)
       - Satellite data (Stats Canada)
       - AIS data (ONS)
       - Telephony data (T-Mobile and Vodafone)
       - Internet data (Innovation Spotter)

     b). A CBS data lake that provides access to external open data (which was found useful during data scouting).
     c). Privacy Preserved Data Sharing (PPDS) with non-open data from external parties (which was found useful during data scouting).
     d). Edge Computing: working with non-open external data (that was found useful during data scouting) and sending aggregated data to Statistics Netherlands.
5. CBS as a leading data-science institute: Using state-of-the-art techniques, CBS can process big data into official statistics, for which the CBS quality mark is essential. Important points of attention in this respect are:
   a). State-of-the-art IT capabilities, both in the field of software and hardware. These include the use of GPUs, computer centres and the application of big data processing techniques.
   b). Measuring concepts in big data sources.
   c). Reliable combination of big data with other sources.
   d). Correcting selectivity and bias in big data sources.
   e). Determining the correlation, causality and doing of 'large scale inference' in big data sources.
   f). Investigating the applicability of state-of-the-art techniques in the statistical process, such as Machine Learning / Deep Learning (moving via WvD BZK; collaboration with Stats canada, DeepSolaris), data mining, Natural Language Processing (for example, to detect sentiment or tension).
   g). Data visualisation (e.g. visualisation techniques for air pollution, Clair City).

Interpretability of AI

Machine learning (ML) methods offer new opportunities for the production of official statistics, especially when fed with vast amounts of data and supported by powerful hardware and software. Complex data such as images (e.g. satellite images or profile images on social media) and texts (e.g. scraped from the internet or answers to open questions) can be mined with ML. Algorithmic models can complement mathematical models. The combination of ML, IT infrastructure and big data allows us to develop new statistics and to produce current statistics more precise, more detailed, quicker or more efficient.

Generalisability of ML: Supervised machine learning is used to learn the relationship between input and output from examples and to predict the output of missing units. Predictive performance can be assessed from cross validation within the examples. As long as the examples represent the missing units, the predictive performance can be taken as a quality measure for predictive performance on missing units. However, the examples often may not represent the missing units, since not all units in the target population had a non-zero and known probability of being included in the examples.

Transfer learning and cross-border statistics. A convolutional neural network (CNN) is designed to classify images. Since it requires a lot of examples to estimate many parameters, often a pre-trained network is used as a starting point. The early convolutional layers are frozen because they are assumed to recognize universal, low-level features such as edges. The pre-trained network was developed on more images, but the images were likely taken from another source and the learning task was likely not exactly the same. When will the benefits of transfer learning outweigh the costs? This is related to the bias-variance trade-off in statistics. A potential application could be cross-border statistics on satellite imagery, when a CNN developed in one country is transferred to another country.

Using text in the statistical domain

Text is becoming an increasingly important input source for statistics. Think of descriptions of products and answers to open questions, but also texts on web pages, vacancy texts and the content of social media messages. In order to become suitable for use by an NSI, it is important that the texts are converted into a representation that can be used for statistical purposes. In order to make this possible, the following research themes are proposed:

Building reliable and reproducible text-based models. Because of the distribution and variability of the words included in texts, it is challenging to build reliable and reproducible models. This becomes even more important when reliable estimates have to be produced over longer periods of time. In all these cases, model stability plays an important part. Three important subtopics can be discerned.

The first one is dealing with word frequencies. The second one is finding optimal ways to convert word frequencies to numbers. Traditionally, Term Frequency-Inversed Document Frequency (TF-IDF) is used for this as it is able to reduce the effect of words that occur in many documents and hence are not discriminative. The third and final subtopic is dealing with changes in word use, their meaning and the rise of new topic specific words in our modern world. These will likely result in a reduction of the accuracy of the models developed also called 'concept drift'. Methods to identify and deal with these changes need to be developed to enable the production of reliable statistical text-based time series. Specific text analysis within the statistical domain. The use of texts for official statistics

is very specific in a number of cases. This makes it important to carry out research into the optimal applications of text analysis techniques within this domain. Examples of this are the classification of the economic activity of companies, the detection of innovative companies and the extraction of skills from vacancy texts. This results in the creation of tailor-made code and lists of synonyms as well as the development of statistically specific text processing methods. This is needed as insufficient information is available on businesses from survey and administrative sources alone. This is especially important for small and medium sized companies. More data on these companies enable answering the question how innovative they are and how their production growths? These are important policy issues because small and medium sized companies are considered an important driving force for the economy. This research will initially focus on webscraping characteristics of companies to exploit the full potential of the web as a source of information.

Synthetic data, a.k.a. dealing with the security / innovation paradox

Working responsibly with sensitive data is an important, better yet, vital issue at an NSI. All measures are employed to prevent data leakage of privacy sensitive data. The consequence, typically, is that this data is stored in secured, locked up environments (typically, fully disconnected from the internet). The burden on society (via surveys or collecting administrative data), and as such the costs, needs to be lowered. And there is a demand for more detailed and timely statistics. Artificial Intelligence techniques have great potential to cope with these challenges and NSIs should be able to work with state-of-the-art techniques in the field. However, in practice the state-of-the-art is not available in the secured, locked-up environments that store the sensitive data. Researchers also need to clearly state why they need access to sensitive data and it is not always the case that they are granted access based on (preliminary) research ideas. Rightfully so, because sometimes researchers simply want to 'try out things' without having a strong business case. Facilitating all these (initial) experiments can be costly. And even if access is granted, 'doing all the paperwork' may take multiple months which can be demotivating and slows down (or even kills) innovation.

This in short is what we call the security / innovation paradox. One way of dealing with this paradox is to create synthetic data based on the real sensitive data. The synthetic data should be free of privacy issues and may thus be used in the 'outside' world. Researchers can then freely apply whatever software they see fit to work out their research ideas: innovation. The results can then be evaluated. Good results on the synthetic dataset yields a better business case for the facilitation of used soft- and hardware within the secured environment of the NSI: security.

Visualisation

Data visualisation plays an important role in statistics and data science. As a result of recent developments, this role seems to be becoming increasingly important. This role is addressed in the following three research themes:

Which visualisation method can be used for which type of data sources? Today, there is an abundance of data. There are more and more data sources that can possibly be used for official statistics. In addition to the well-known administrative sources, there are many automatically generated sources, such as smart meters, traffic sensors, camera images, satellite images, mobile telephony network data, public transport data, and social media. To be able to explore such sources, data visualisation methods are needed. Research

into visualisation methods has already been done here (e.g. the tableplot). Visualisation methods are also available for other types of sources, but general guidelines on which method to use for which type of data source are still lacking.

Visualisation of spatial data. Almost all big data sources have a spatial component. Think of coordinates of the sensors, geolocation tag of social media messages, public transport routes, and of course aerial photographs and satellite images. Regular statistics also usually have a spatial component; the desire for low-regional estimates is growing. A number of sub-topics can be distinguished here: visualisation of spatial data at 1 time or 1 reference period, visualisation of flows / networks, visualisation of spatial data over time.

Visualisation methods for Active learning / Explainable AI. Active learning is becoming increasingly important; users want more insight into 'black-box' methods, such as neural networks. For interaction with the user, it is necessary to use data visualisation methods that provide insight into the input and output data and into the AI/ML method, with the aim of explaining the results as much as possible and fine-tuning the method if necessary.

Correlation, causation and large scale inference in big data

The data in new data sources are generated in a way that is not comparable to that of register and survey data. In many big data sources the data generation mechanism is different, this process is often non-transparent and little is known about the units. All this could lead to bias in the figures based on these data. This makes it challenging to use these sources for official statistics. In order to do this in the best possible way, the following research themes are foreseen:

Combination with other sources. By combining big data with sources of which the inclusion rates of the units are known, it will be possible to correct for the selectivity of the data. Detection of subpopulations. In the case of big data sources containing units that lack an identifier that could be used by CBS, the data in the source itself can be used to distinguish different groups. The data of certain variables can be used for this purpose, such as the classification of words, but also by combining variables and making use of so-called unsupervised classification methods.

Development of a big data selectivity correction model. Due to the unknown inclusion rates of the statistically relevant units in big data, it is not immediately possible to apply standard sample theory to correct for the bias. Stochastic models and the results of simulation studies are used to determine which solution offer the best correction.

Causal relations and big data. In most statistical methods, the direction of relationships between variables is usually not taken into account. Cause and effect can therefore not always be easily determined. This is a pity, because, especially in the case of big data sources, the use of causal models offer possibilities to: i) better understand what information a big data source can contribute, ii) distinguish actual (co-) relationships from non-physical (co-) relationships and iii) add extra (external) knowledge. The fact that the data generating mechanism of many big data sources is unknown makes this a very challenging and interesting topic.

Appendix 4 Overview of data sources acquired by CBDS

Source	Data	Application
Cargonaut	Air cargo in NL	Economy and trade
Coosto	Social media posts	Desire to move, social tension
Enelogic /Liander	Smart meter data (test data)	Energy consumption
Jaap	Housing market sales	Housing
KNMI	Sun radiation data	Solar energy production
MijnOverheid.nl (Logius)	Log data from the Dutch citizen platform	Digital skills of NL population
National Police	Police reports	Cybercrime
PV Output	Energy production from solar panels	Energy production
Rijkswaterstaat	Ship movements	Mobility, economy
Stichting KijkOnderzoek	Watch and listening data (TV/radio)	Time use of NL population
Tennet	Amount of used electricity	Electricity consumption
Textkernel, Dataprovider	Online vacancies	Job classification, vacancy statistics
T-Mobile	Aggregated signalling data	Mobility
Tomtom	Floating car data (test data)	Traffic intensity