Big data to improve policy and decision making: The Experience of Statistics Netherlands

Sofie De Broe, PhD
Scientific Director CBDS, Statistics Netherlands
Magchiel van Meeteren, MSc
Managing Director CBDS, Statistics Netherlands
Piet Daas, PhD
Lead data scientist CBDS, Statistics Netherlands
Ben Laevens, PhD
Data scientist CBDS, Statistics Netherlands

Abstract. Can we improve policy and decision making by using Big Data in official statistics? At the CBS Center for Big Data Statistics (CBDS) we aim to accelerate the introduction of big data sources in official statistics while taking our responsibility in issues of privacy and security. We believe that using Big Data in our statistical products deepens insights in relevant societal questions and improves both timeliness and level of detail. The mission, scope and ecosystem of CBDS will be elaborated, results in terms of beta or experimental statistics and challenges in the transition from experimental to official statistic will be discussed.

1. The CBS Center for Big Data Statistics

Statistics Netherlands (SN) serves to inform policy and decision makers as well as researchers, entrepreneurs and the general public on relevant societal phenomena. Statistics Netherlands has a long tradition in collecting, processing and integrating large amounts of data in order to compile and publish independent, high-quality statistics while observing privacy and security aspects. To continue doing so, SN has to deal with a rapidly changing society in which increasing availability of data means that the speed by which it is constantly needed is increasing. Policy makers have a strong ambition to underpin new policy measures with timely and relevant data which our survey results and administrative data, mainly due to a time-lag, do not always cover. More indicators and hence more timely data are needed to describe societal phenomena to their full extent. We experience a data gap that needs to be addressed before we are able to fully support evidence-based policy making. There are technical possibilities and a societal desire to process, combine and use new, generated data sources leading to innumerable new applications. We see that seizing these

opportunities requires multi-organizational collaboration, data access, strong leadership and a more entrepreneurial approach.

Since 2009 SN has been exploring the use of big data (BD) sources for official statistics. Our initial research soon resulted in new official statistical products paving the way for a formal embedding of BD research within the organization. In September 2016, SN established the Center for Big Data Statistics (CBDS). One of its main goals is to be a pressure cooker for BD related innovation; the center brings data and expertise from different internal and external partners together to allow analysis of combined data and better insights into complex societal phenomena. Hence the CBDS is championing a new role for National Statistical Institutes (NSIs) to provide information as a service for a society where timely policies are necessary to tackle the big questions. The mission statement of the SN Centre for Big Data Statistics reads, as follows:

The CBDS explores and exploits new data sources, applying state-of-the-art methodology in collaboration with partners, in order to provide timely, comprehensive information on social phenomena relevant to users.

The mission of the CBDS is to implement the use of BD sources in official statistics and improve existing statistics (e.g. higher spatial resolution, near real time, higher frequency or additional breakdowns) while lowering the administrative burden on society. We are in a unique position because of our ability to integrate these big data sources into the wealth of survey and administrative data which SN - and in the Netherlands only SN - already has available. This position turns us into the government data hub of the Netherlands. At the same time, we are taking our responsibility for all issues of privacy and security related to big data. On the other hand, tension between the potential benefits of increased data usage versus preserving individual privacy seems to be growing. In the context of high demand, budget cuts, new up and coming players, CBDS has to rise to its task and therefore increase its role in society.

2. The CBDS ecosystem

Statistical offices are challenged by companies claiming to deliver the information needed for policy making. Therefore, it is of crucial importance to join forces with these companies to assure the information conforms to the high quality standards of SN. At the CBDS, we actively work together with over 40 partners, both in the Netherlands and abroad: universities and research organisations, NSIs and government institutions and the private sector. Within these partnerships, new insights and products can be realised in new business models; we share knowledge, privacy preserved data and infrastructure to create value and improve the strength of both organisations.

2.1 Business models

Statistics Netherlands has a lot of experience when it comes to cooperating with universities and knowledge institutions, where the focus lies on research, financed by subsidies. In contrast, collaborating with private companies is new. Such a collaboration is possible and can lead to a variety of business models, for example: a joint development on data from a private party with CBS data, could lead to paid products offered to third parties; a joint tender with a major international business service provider; orders for a ministry in partnership with start-ups; investigating with various companies whether we can achieve paid services (licences for real-time streaming insights); analysing a private source for its quality and usability for statistics whereby the company gains insight into the quality of that data source and the application possibilities for its primary process. CBDS is marketing our products through an annual large seminar, publication of products and promotional material, and organizing sessions for SMEs and start-ups together with field labs where these possibilities are being investigated.

2.2 Sharing data

Statistics Netherlands has already acquired scanner data and traffic loop data in the past. CBDS has taken a proactive stance in getting access to all types of new data sources. For instance, CBDS has acquired mobile phone data, data on housing market, navigation system data, datasets on the energy transition, sensor data of smart farms, audience figures and increasingly uses open data such as satellite data, weather and geo-spatial data and available on-line. CBDS is also revisiting existing data sources it has in house for policy relevant topics (such as aerial pictures of roofs to detect solar panels). All of these combined data sources offer great potential for mapping offer and demand (for example in the labour market or elderly care), answer complex policy questions such as energy transition and making real time and more detailed regional statistics of increased relevance for local policy makers.

CBS is reluctant to pay for data, so other ways of adding value to a data source owner are sought. Because more data will be shared between organisations, much attention is being paid to developing privacy-preserved data sharing (PPDS) techniques and the ethical and legal framework that must accompany this. We also believe that awareness must increase among Dutch citizens on the possible societal benefits and how their privacy is being safeguarded.

2.3 Funding

In order to fund our different activities we make use of both commissioned work, grants and own funding. The commissioned work serves to develop products that demonstrate a new form of visualisation and statistics within a theme. Grants are used to develop both new statistics, as well as new methodology. For this purpose specifically, a grant development office was initiated. The grant development office connects strategic research goals to suitable grants. In this process the office offers support and guidance for the principle investigator. The office tries to see which consortia would be the best match in order to achieve the grant conditions and research goals, and acts as a sparring partner while the proposal is written.

3. Results

3.1 Beta products

Results are published frequently on www.cbs.nl/innovation, where we also solicit active feedback from the public. These so-called beta products are not official statistics. As a proofs of concept new sources are investigated for potential usage, their quality and their applicability to specific social issues are demonstrated. In addition, new visualisations and new insights are presented. The beta products should then be transferred to internal production by other SN statistical divisions, provided the quality is sufficient and there is permanent demand. Focus topics include energy transition, air quality, innovation, mobility, health, sentiment using social media [1], economic indicators, housing market, safety, labour market, Sustainable Development Goals and smart cities. Furthermore, CBDS directs efforts towards methodological issues [2,3]. New methods such as machine or deep learning, text mining or Artificial Intelligence are being developed to help extract relevant statistical information. New methods are needed for the integration of heterogeneous and volatile data sources, since the value of the CBDS lies particularly in this combination [4]. Other topics we are looking at include data access, data integrity, ethics and privacy. In a picture:

Figure 1: Activities and product development topics at CBDS

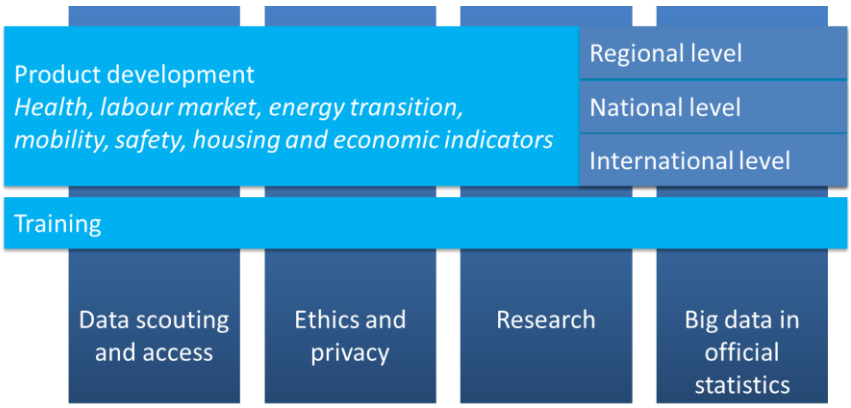
---

[1] Daas, P.J.H., Puts, M.J.H. (2014), Social Media Sentiment and Consumer Confidence. *Proceedings of the Workshop on using Big Data for Forecasting and Statistics*, Frankfurt, Germany

[2] Daas, P.J.H., Puts, M.J., Buelens, B. and van den Hurk, P.A.M. (2015) Big Data as a Source for Official Statistics. *Journal of Official Statistics* Vol. 31(2), 249-262.

[3] Daas, P.J.H., Puts, M., Tennekes, M., Priem, A. (2014), Big Data as a Data Source for Official Statistics: experiences at Statistics Netherlands. *Proceedings of Statistics Canada International Methodology Symposium 2014*, Gatineau, Canada.

[4] Van den Brakel, J., Söhler, E., Daas, P., Buelens, B. (2016) Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Discussion paper 201601, Statistics Netherlands, The Hague/Heerlen, The Netherlands

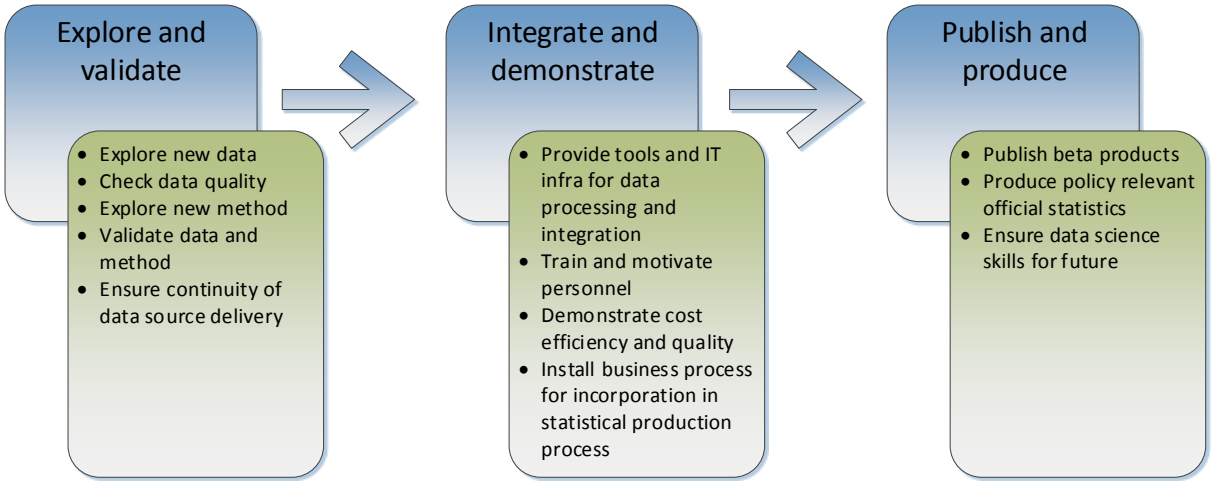


3.2 From Beta products to official statistic

For experimental statistics to be relevant for policy makers they need to be embedded in the statistical process and production. Our initial research soon resulted in new official statistical products paving the way for a formal embedding of BD research within the organisation. The CBDS' ethos is to produce experimental statistics first and turn them into official statistics at a later stage when they prove successful. Two examples have shown that the transition from experimental to official statistics can be successful: the consumer price index based on scanner data (data collected from supermarkets) and more recently internet robots [5,6] and the traffic intensity indicator based on road sensor data [7]. Both benefitted from feedback of internal and external users during the experimental stage. These are known examples at SN and will not be further discussed. Defining the process where an experimental statistic (a useful definition of an experimental statistic can be found on the Office for National Statistics website [8]) transitions to an official statistical product requires a stepwise validation and quality check for each of the process steps and methodological issues. This is mostly an iterative process where one goes through the several steps more than once (first for the beta publication and thereafter for the official statistic). An attempt to define a generic process for future products is presented below.

Figure 2: Issues and steps in the transition from Beta to Offstat

[5] Bosch ten O., Windmeijer D., (2014), On the Use of internet robots for official statistics. *Proceedings of the UNECE MSIS Conference*, Dublin, Ireland.
[6] Hoekstra R., ten Bosch O., Harteveld F. (2012), Automated data collection from web sources for official statistics: First experiences, *Statistical Journal of the IAOS*, Vol. 28 (3-4), 99-111
[7] Puts, M., Tennekes, M., Daas, P.J.H., de Blois, C. (2016), Using huge amounts of road sensor data for official statistics. *Proceedings of the European Conference on Quality in Official Statistics 2016*, Madrid, Spain.
[8] https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/guidetoexperimentalstatistics



Data:

- Often data sources are used which have not been used before. Only after performing explorative data analysis on multiple versions of these data sources one can develop a feeling for the nature and quality characteristics of these data; for implementation from beta to official statistics a continuous quality monitoring system of the data is needed. Quality indicators tailored for use on large amounts of data play an important role here.
- Stability of the source: successful research on a new data source does not guarantee it is also fit for production. One must make sure that the required information can be extracted from the source, even if it changes over time, such that future output is guaranteed;
- Variables: interpretation of the definition of used variables should be clear and consistent and differences in variable definitions between different sources made explicit;
- The relationship between the population comprised in the data source and the target population needs to be made explicit if possible;
- Measurement error needs to be assessed for every source being used in the analysis;
- Meta data should be (made) available.

Method

- The method should be justified. In case of several methods, a clear decision should be taken with respect to which methodology provides the best results and stable output;
- New methods: a validation assessment of the new method needs to be undertaken through peer-review;
- New tooling or IT infrastructure available to implement the beta product into the statistical process could be necessary;
- Standard, proven methods should be applied as much as possible. However new data sources might require new methods. This should be explained and if useful, the new methods should be added to the standardised methodology handbooks for further re-use. Algorithm design and efficient computing is often required for implementation in a production process.

Output

- Reproducibility: results should be reproducible based on data obtained at different time intervals. This may be a problem for streaming data or some BD sources which are too big to store and archive.

- Stability: assumptions should be made explicit and justified based on the literature wherever possible. These should be regularly re-tested as part of the implementation process.
- Validation (and cross-validation) of the results should be undertaken through comparative analysis of other data and disclaimers of the initial beta product addressed.
- Output should not be hampered by legal and ethical issues. Different legal boundaries might apply for experiments and a continuous production process.

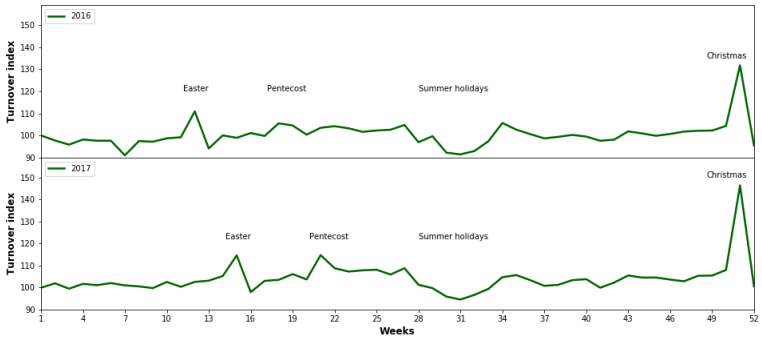
Process

- Involving domain specialists from the start in the development of a new Beta product in order to maximally exploit the knowledge already available on the subject and to allow a smooth acceptance and incorporation into the existing statistical production process;
- Publication of the Beta product in order to get users' feedback on relevance and usability of the product as well as getting users familiar with different forms of statistical output, such as visual products, data portals or dashboards;
- Ideally, an efficient standardised migration process should be in place within the organisation to allow a harmonised and transparent transition from beta to offstat. In practice, however, the transition process will depend on the type of new statistic: whether it's a completely new statistic, a supplement to or extension of an existing statistic or if it will replace an existing statistic.

Below we discuss two examples of experimental statistics that are being assessed for embedding in the statistical production: Gross Domestic Product (GDP) flash estimates and detecting innovative companies. Statistics Netherlands is currently investigating the possibility of using alternative data sources to provide a faster and more accurate first flash estimate of the GDP. One of the biggest challenges of the flash estimate is computing the effect of the contribution of the last month on the overall estimate. A possible good data source is the transaction data of Dutch supermarkets, which could be used as a proxy for household consumption in the Netherlands. Statistics Netherlands receives the sales of each item, aggregated on a weekly basis, from all major supermarkets. These data account for a market share of around 90%. The main idea is to assess whether a correlation can be found between the transaction data and the final GDP number for each quarter. Fortunately a lot of historical transaction data, dating back to 2010, is available. Preliminary investigation has revealed certain challenges for this research. First of all it is important to know what the market share is of each supermarket on a high temporal resolution (ideally monthly or weekly). If supermarkets A, B and C have a market share of 75% in week 37 of 2015, this is not necessarily the case in week 3 of 2012. A second important factor is the purchasing behaviour of consumers. The transaction data need to be converted from weeks to quarters, which means assumptions need to be made on how the total turnover of a supermarket is spread over two quarters (when a week straddles those two quarters). This becomes especially important for festive seasons such as Easter and Christmas/New Year. Figure 3 displays the total turnover, on a weekly basis, for 2016 and 2017.

The turnover is expressed in terms of an index, which is set at 100 for the first week of 2016. While this figure does not display the turnover for all supermarket chains, the same supermarket chains are used for the two years concerned. This figure reveals a spike around festive periods such as Easter, Pentecost and Christmas and New Year. A dearth is equally visible around the Summer holidays. Although giving an indication of consumption behaviour, please note that the turnover indices ought to be normalised for the combined market shares of the supermarket chains in question over both years to get a true sense of consumption.

Figure 3: the total turnover, on a weekly basis, for 2016 and 2017

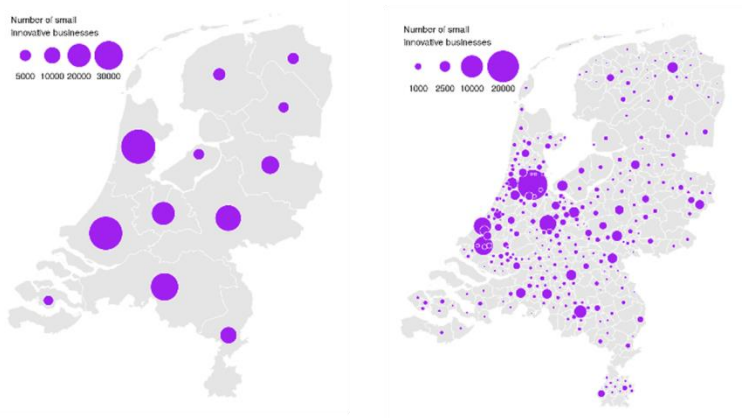


Another example is the identification of small innovative companies. Getting an overview of the innovative companies in a country is a challenging task. One of the ways of doing so is by using a survey to contact a sample of companies; for instance by phone or via a questionnaire. The response can be used to derive how many innovative companies there are in a country or area. This approach, however, puts a burden on companies and may result in a considerable non-response. Another downside is the fact that the focus of the survey is on large companies and not on small companies, such as start-ups. We therefore looked for an alternative approach and came up with the idea of determining if a company is innovative by studying the text on the main page of their website. To enable this the following steps were applied, namely: first, selecting a set of known innovative and non-innovative companies; second, making sure that for each company the corresponding URL of their website is available; third, scrape the main page of each web site and pre-process the text displayed; forth, develop a model to determine if a company is innovate or not based on the pre-processed texts.

We started with a sample of 3000 innovative and 3000 non-innovative companies according to the Community Innovation Survey of Statistics Netherlands. The first thing observed was that two-third of the URL's of the companies selected were absent in the business register. These URL's were added via the URL finding approach developed in WP2 of the ESSnet Big Data (Deliverable 2.2). Next, the text displayed on the main web page of each company was scraped with PhantomJS. After

language detection (usually Dutch or English) of the inner HTML text, punctuation marks and stop words were removed and the remaining words were stemmed. This was used as input for the model. Here, it was found that logistic regression with L1-norm performed well. With a 70%-30% training and test set, the trained model was able to determine if a company was innovative or not with 92% accuracy. Next the main web page from all companies with less than 10 employed persons with a known URL in the business register, about 500.000, were scraped, processed and classified. The results of these companies are shown at the municipality level in Figure 4 and reveal the potential and granularity of this approach. The lowest level of detail that can be obtained by the Community Innovation Survey is the COROP (NUTS 3) level. The web scraping approach easily enables one to produce zip-code 4 maps which are particularly interesting for large cities.

Figure 4: Number of small innovative businesses at the province and municipality level



Both examples show an intermediate level of maturity in the transition from beta to official statistic. Both the quality, methodological validation and organisational support for embedding needs to be further addressed for these beta products to become official statistics of relevance for policy makers.

4. Conclusions

In order to embrace the possibilities of BD and stay relevant in the future, we need genuine and rapid change. From past experiences, we have drawn several lessons and conclusions. Ecosystems and (temporal) consortia are the future. Due to the fast pace of technological changes and growing complexity of society increased access to data sources, co-creation is necessary. Collaboration often requires a (tailor-made) approach in which common goals of organizations are addressed. For collaborations to be set up in policy relevant areas, effort needs to be put into fostering a close dialogue with policy makers to have a better understanding of questions they have and the possibilities innovations at SN could offer. Mainly for this reasons, Urban Data Centers (UDCs) were set up by SN in 2016. Access to data remains a difficult issue. Last year, we gained access to

several data sources such as open data including social media content, sources on cybersecurity and earth observation data. New legislation is underway for this at the national level and we are participating in new EU legislation as well.

Inside SN, CBDS acts like an entrepreneurial startup and has the freedom to explore its role with regards to policy makers, agile processes and IT. The CBDS is connected with the rest of SN and with third parties via flexible, multi-disciplinary project teams. In order to be successful, we need to manage expectations and explain to our surroundings that exploring, cleaning and ascertaining the applicability of a new (big) data source takes time and that retrieving insights out of generated data sources is not trivial. Additionally, we have undertaken research on BD methods, looking at supporting the development of real time, more detailed or new statistics, turning beta into official statistical product. Strong commitment from top management as well as the commitment from statistical divisions' business owners, as early as possible in the process are an absolute prerequisite for success. As a statistical agency with access to a wealth of administrative data and clear ambitions, we are an attractive organization to collaborate with. Finally, in experimental projects small steps are best. Partnerships will only be kicked off if a tangible (small) project can be identified and policy relevant questions can be addressed.