Discussion Paper

Big Data and Methodological Challenges in Official Statistics

2016 | 08

Kees Zeelenberg

Contents

Summary

We identify several research areas and topics for methodological research in official statistics. We argue why these are important, and why these are the most important ones for official statistics. We describe the main topics in these research areas and sketch what seems to be the most promising ways to address them. Here we focus on:

- Quality of National accounts, in particular the rate of growth of GNI
- Big data, in particular how to create representative estimates and how to make the most of big data when this is difficult or impossible

We also touch upon

- Increasing timeliness of preliminary and final statistical estimates
- Statistical analysis, in particular of complex and coherent phenomena

These topics are elements in the present Strategic Methodological Research Program that has recently been adopted at Statistics Netherlands

1. Introduction

This paper deals with four questions:

1. What are the major challenges to official statistics?
2. How may methodology help in meeting these challenges?
3. How may big data help in meeting these challenges?
4. What is the best research strategy for methodologists to follow in dealing with big data?

So we will start from rather general points, and then focus on methodology and on big data. The paper tries to present this in some kind of overview, in terms of what seems to be important when it comes to big data and official statistics. The examples as well as the main points come from, or rely on, the methodological research program and the innovation program at Statistics Netherlands. At some points, it is more a personal view, and we are not yet doing research very actively.

2. Challenges to official statistics

I will focus on two of the challenges to official statistics:

- Quality
- User needs

Quality is of course our "niche". This is what distinguishes us from many data producers. Below it will be shown that we do have some quality problems with for example GNP and its growth rate.

Secondly, nowadays, users demand not so much data but information. They want us to tell the story behind the data. They want it more timely, and they want it to be based on better data. The main drive behind this need, is the increasing complexity of society and the increasing pace at which society changes. This leads to three topics for NSIs to focus on:

- Timeliness
- Information and analysis
- Complex relations in society.

Big data are new sources and pose new ways to create information, and so they can be very useful here; this is the so-called analytical use of big data.

For the analysis of complex relations and for making more timely contributions to debates in society, it is first of all necessary that we are able to use link and combine data from various databases. This is the so-called data lake, that we should try to

create; not a single database for all purposes, but more a system of databases that can be linked in a flexible way and from which data may be quickly extracted.

The analysis of complex relations and the emphasis on information, will lead to more emphasis on methods, and so, even apart from big data, methodologists will have to play an important role here. Besides more specific methods for big data (see section 4), multilevel methods seem to be promising here.

We must also learn to create and work with flexible classifications. For enterprises the standard industrial classification (ISIC and its national or regional counterparts) should no longer occupy a central position, since for many complex phenomena, other classifications are much more important, such as small / large enterprises, exporting / domestic enterprises, and IT using / IT producing enterprises. The standard industrial classification is also conceptually outmoded: enterprises are classified according to their main activity, but nowadays enterprises change much more frequently in their activities, and their subsidiary activities may be almost as important as their main activity. Moreover, from the viewpoint of society, it is much more important which products and services are being supplied and demand than in which industry they are being produced. Again, this will lead to more use of statistical methods, such as cluster analysis and other classification methods as well as methods for analysis.

Even if the methods from section 4 may not be usable for official (descriptive) statistics, they may profitably be used for analytical statistics, and thus for the analysis of complex relations in society.

Of course, there are other challenges, such as

- Nonresponse, especially now that we have a web first strategy for data collection, where web surveys are more and more going to take the place of other surveys.
- Linking, in particular for big data, since these do not contain a clear linking variable such as the social-security number. Statistical matching, probabilistic linking and imputation have therefore become much more relevant.
- Privacy is still very important with the general public, and more so now that we are in the era of big data and open data.
- And of course, budget cuts, leading to demands for more efficiency.

Methodologists have a lot to contribute here, and we are working on these topics. But I will not discuss these here, and will focus on quality and user needs (information instead of data).

3. Example: The arrival of spring

Let me start with a nice example that does not use survey data or administrative data, but uses some kind of raw data found on the internet, and let's see what we can learn from that. The example is about the windflower, or wood anemone. This is the first flower that in spring blooms in Western Europe. Dates of first blooming are registered by volunteers, who go out in parks and woods to observe flowers.[1] So you can use these dates of first blooming to see how spring arrives.[2] The analysis has been done by Maaike Hersevoort (Statistics Netherlands) and Hamed Mehdipoor (Twente University).

3.1. The arrival of spring in the Netherlands, 2013 and 2014

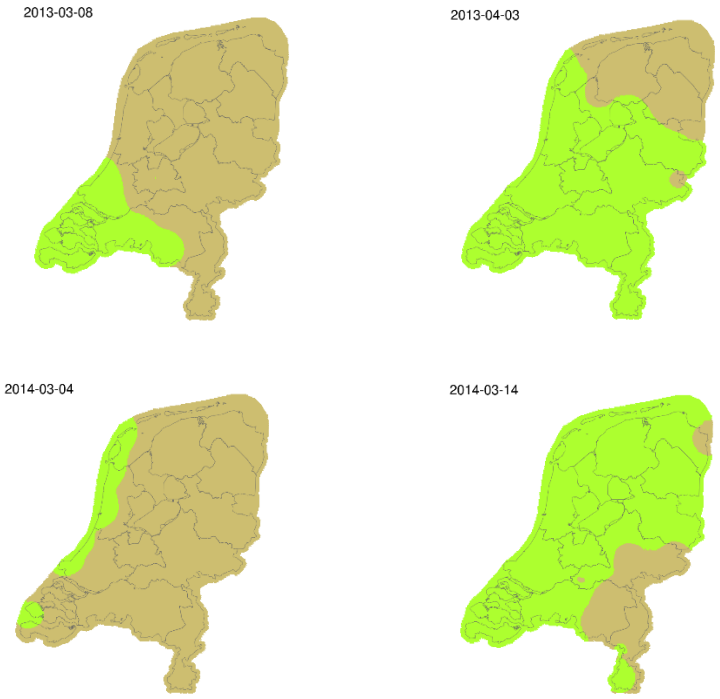


Figure 3.1 shows the arrival of spring, measured by the blooming of windflower. In 2013, spring arrives from the south, in 2014 from the west. Also, spring in 2013 starts 1½ week later than 2014, and finishes about 3 weeks later; so in 2013 spring arrives later and spreads at first somewhat quicker and thereafter slower than in 2014.

What can we learn from this example?

[1] The data are at www.natuurkalender.nl
[2] In fact, the arrival of spring is the topic of one of the most famous poems in Dutch language: May by Herman Gorter; see Gorter (2015).

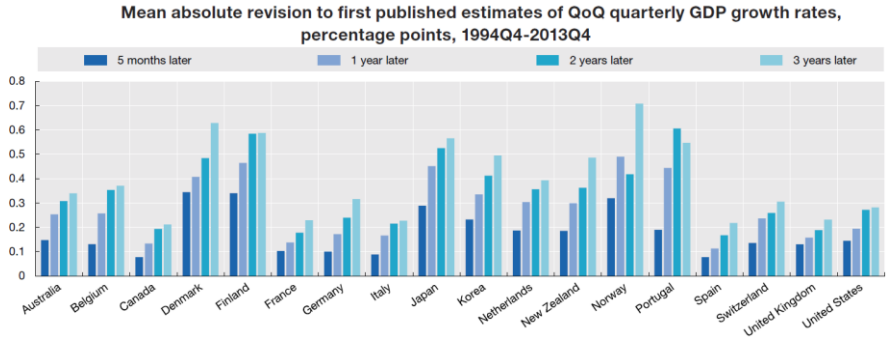
- The data are collected by volunteers. It's not the same observers every year, they do not go out on every day or at the same time of the day. So, these data are not evenly spread in time and space. In more familiar words: we have nonresponse and selectivity.
- That's why the authors of this example had to rely on modelling to improve quality.
- But they also found that modeling is labor intensive: so, building a production process is not always simple or cheap.
- And what they have done, is they have created information, about the arrival of spring, and not just collected or cleaned some data set.

These are some of the problems that I will discuss.

4. Quality of GNP

Figure 4.1 (Figure 3 from Zwijnenburg, 2015) shows, for nearly twenty OECD member states, mean absolute revision (MAR) to the first estimate of GNP growth, for several time lags. Let us focus on the final revisions after 3 years, the light blue bar, most to the right for each country. First, Canada, France, Germany, Italy, Spain, Switzerland, the United Kingdom and the United States show the lowest revisions, with an average MAR below 0.25%-point. A second group, composed of Australia, Belgium, the Netherlands and New Zealand, presents medium size revisions, with an average MAR between 0.25%-point and 0.35%-point). And a third group with Denmark, Finland, Japan, Korea, Norway and Portugal records the highest revisions, with an average MAR above 0.35%-point. So, for several countries the revisions are quite large, and the Netherlands are somewhere in the middle, both in the figure and in the size of the revision.

4.1 Mean absolute revision of quarter on quarter GDP growth rates, 1994-2013



However, for the Netherlands, things are much worse when it come to the mean revision, so without taking absolute values (Zwijnenburg, 2015, Figure 2), then the Netherlands is one of the worst performing countries, with a mean revision of 0.1 %-point. Also, nearly one third of the uncertainty in forecasts of GNP by the national forecasting agency (CPB), is due to revisions of earlier estimates of GNP by Statistics

Netherlands (Elbourne et al, 2015). With a an average Q-Q growth of GNP of 0.2 percent (2006-2015), it is clear that we dò have a problem here. And this is reflected in comments by our main users.

The major cause of the revisions in GNP appears to be the quality of the basic data:

For the third month of the last quarter, we have hardly any data.
For the second month, we have for some components of GNP only very preliminary data.
For some industries there are even no monthly data at all.

So, to improve GNP and reduce revisions, we need better preliminary and first estimates!

What can methodology do when it comes to the quality of GNP? First, what we could do, and should do in my opinion, is to look at Bayesian methods to assess the quality of GNP. There have been in the past various studies about integration methods in several countries, for example the UK, Italy, Netherlands, and maybe also other countries. These have led to a successful implementation of benchmarking methods for GNP, for example to benchmark quarterly data to yearly data, component data to totals, et cetera. But we may elaborate this, and starting with a small-scale model of integration, use Bayesian methods to simulate the integration process. There have been some very small examples of this, for example Van Tongeren, Magnus and De Vos (2001). In a second stage, we may expand the model to lower levels and to more components and details. Such a model of the integration process will give us more insight into the quality of GNP.[3]

But also the use of big data may contribute to the improvement of the quality of GNP estimates. There are four possible big-data sources, here:

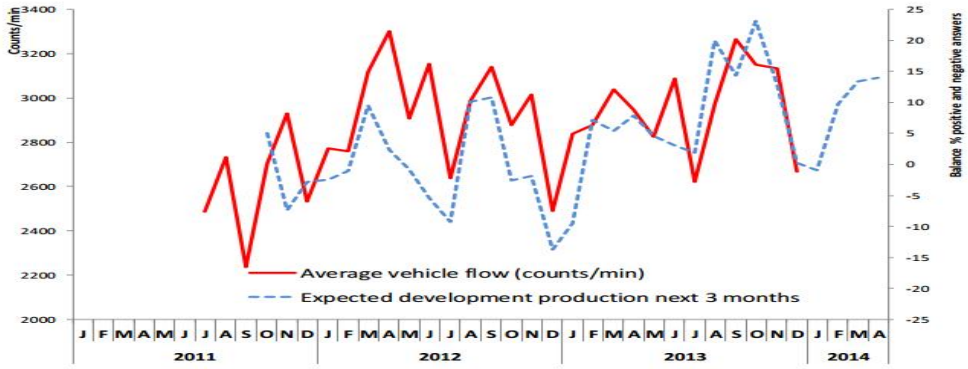
- Tax databases: These are mainly for taxes on wages and on sales. These are not yet rapid enough. They lag about two months: one month for reporting taxes and one month for filling the databases.
- Company accounts: It will become possible within the next few years to have direct access to company accounts. Direct access in the sense that reporting modules for taxation and statistics will have been built in the accounting software. But this is not yet possible, and even when it will have been implemented, then only for annual accounts.
- Banking transactions of enterprises and households: This is clearly the most promising. There is only one clearing-house for banks in the Netherlands, and it is very rapid; also because banks have to report daily, weekly and monthly to the central bank. But there are of course very strong privacy concerns here. So it will be very, very difficult, and will take a long time before even the first steps will be taken and even longer before it will be implemented.

[3] These Bayesian methods are examples of model-based estimates. Some other examples are structural time-series models and integration techniques, for example for micro-data sets. I will come back to this below.

– Model-based estimates based on big data; not a big-data source as such, of course, but a way to use big data. So at the moment this has the best options.

An example of what model-based estimates from big data may have to offer, is given by Van Ruth (2015). Figure 4.2 shows the relation between traffic intensity and production (value added) in an important region of the Netherlands, around the city of Eindhoven. Traffic intensity is measured from traffic sensors in the road surfaces, and production is taken from the Monthly Business Survey. Statistically, the series appear to be coincident, and possibly seasonal adjustment and a trend-cycle decomposition may remove some noise and further improve the model. Now, traffic intensity has a very short time lag, maybe one or two days. So, this and similar relations might be useful in making better first and preliminary estimates of output.

4.2 Average vehicle flow and production, Eindhoven region, 2011-4



5. Methodological strategy for big data in official statistics

5.1 The representativity problem of big data

As we all know now, it is difficult to turn big data into statistics and into information, because representativity of big data is a problem. The representativity problem of big data is the most important methodological barrier to using big data in official statistics! They are difficult to link and match with other data sets that contain background information. For example, with traffic sensors you can observe that a car is passing, but you don't know who is in the car, who owns the car, and why he or she is driving at that spot. And these big data are selective. For example, on Twitter, males around 30 years old, are overrepresented.

Precisely because linking to background variables is often not possible, the usual methods from survey methodology cannot be directly used. However, we may use methods for non-probability samples. You might think here in terms of a propensity

to be observed in a big data set, just as in a survey, we might analyze the propensity to respond. This looks a lot like what in marketing is called profiling. As an example, we might estimate the probability that a person in a twitter data set, is a man or a woman. And so you might be able to correct for an unbalanced gender distribution. More generally, you try to model the relations between the variables in the big data set and, if possible, other data sets.

Basically, there are three classes of methods for modelling big data sets in official statistics:

- Pseudo design-based methods
- Machine-learning techniques
- Integration methods where simply put, big data are used as auxiliary variables

5.2 Pseudo design-based methods

Here we act as if somehow a probability mechanism has generated the data. We may then proceed in the usual way, for example Use features of the events in the data to make strata, and then use post stratification estimator to correct for selectivity, but this does not always makes sense and it is hard to test.

5.3 Machine-learning techniques

These are much more recent, but there is already a lot of literature about these methods. Simply said, with the help of these techniques, you let the data decide what the strata are. Several well-known techniques are:

- Regression trees
- K-nearest neighbor
- Artificial neural networks
- Support-vector machines

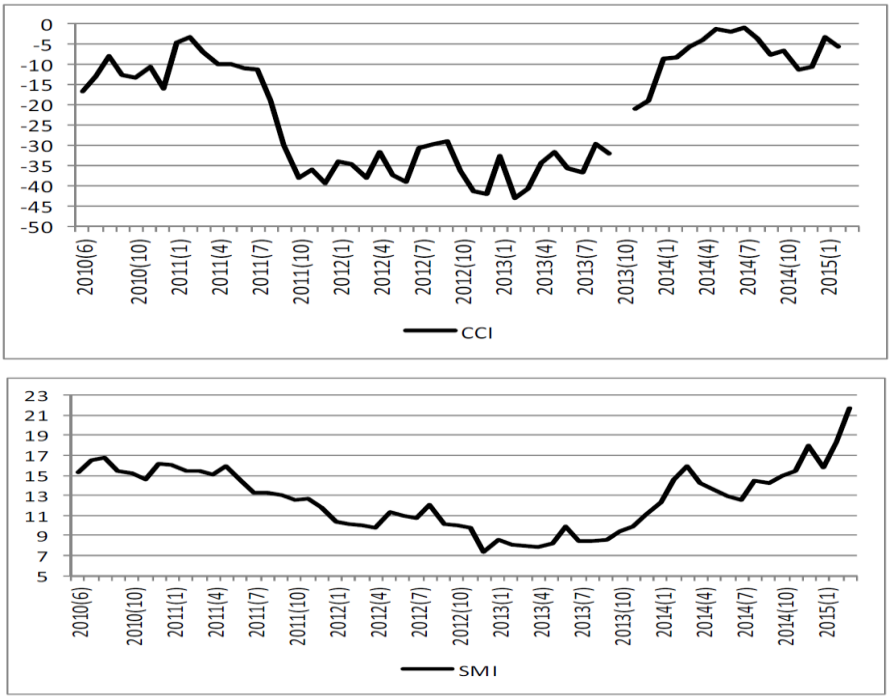
These techniques are also very useful for analytical use of big data, that earlier we mentioned briefly.

With these machine-learning techniques, we have done a simulation study on the Motor Vehicle Database (Van den Brakel, Buelens and Burger, 2015). This is a very large database, with over 7 million records, all the motor vehicles in the Netherlands. The simulation results show that machine-learning techniques are promising. In particular, regression trees are promising, because they are somewhat of a middle way between design-based techniques and the other more form-free machine-learning techniques. And they are also easier to understand what happens “behind the screen”, and so more acceptable to NSIs who always have to be transparent about their methodology.

5.4 Integration techniques

Here we try to improve the ordinary estimate from surveys or from registers with the help of big data. As example, we have looked at the relation between the Consumer Confidence Index and a Social Media sentiment index (Van den Brakel et al, 2016). The Consumer Confidence Index comes from a monthly survey with around 1000 respondents. The Social Media sentiment index is based on big data from Twitter and FaceBook.

5.1 Consumer Confidence Index and Social Media sentiment index, 2010-2015



The upper panel of Figure 5.1 shows the Consumer Confidence Index and the lower panel shows the Social Media sentiment index; note that the vertical scales are different. The time period is 2009 till 2014 and the data are monthly data. We see that both series move more or less together. This is confirmed in a formal statistical analysis with a structural time-series model, which shows that both series are co-integrated and that is possible to improve the Consumer Confidence Index with the help of the model. The advantages are

- Improved precision
- Higher frequency
- More detailed subpopulations

5.5 Models in official statistics

This section and the previous one have argued that model-based analysis will become much more important than before, in particular when big data are going to used more and more in official statistics. And, somewhat behind the lines, this paper

has also argued that the advantages of big data may be large enough that NSIs may not be able to afford to neglect big data.

NSIs have always, rightly, been careful in using models in official statistics. When it comes to the use of big data, we should always take care to remain within the framework of official statistics. On the other hand, in several areas, we are already using models, such as in seasonal adjustment and in correcting for non-response.[4] To remain within the framework of official statistics, we should therefore not simply refuse to use models, but should look for methods that are acceptable. Guidelines and restrictions for the use of models in official statistics have been developed by Buelens, De Wolff and Zeelenberg (2015); for example:

- No forecasting
- Nowcasting only if based on relevant data, not if based on behavioral models
- Extensive model checking
- Transparency about methodology

5.6 Generic production processes for big data

We must also try to create more generic production processes for big data. This has not bene discussed in this paper, because it is more a task for development than for research. But it ìs important. If we do not have such generic production processes, it will always be very expensive to use big data. And again, solving the representativity problem is one big step towards such more generic production processes.

6. Summary

A methodological strategy for big data in official statistics might be:

Focus on model-based estimates. This useful for

- Improving quality
- Analysis of complex relations
- Solving the representativity problem of big data

[4] Other examples are given in Bethlehem and Van den Brakel (2008) and Buelens, De Wolff and Zeelenberg (2015).

References

van den Brakel, J., and Bethlehem, J. (2008) “Model-Based Estimation for Official Statistics,” Discussion Paper 2008-02, Statistics Netherlands. https://www.cbs.nl/nl-nl/achtergrond/2008/10/model-based-estimation-for-official-statistics

van den Brakel, J., E. Söhler, P. Daas, and B. Buelens (2016), “Social media as a data source for official statistics; the Dutch Consumer Confidence Index, ” Discussion Paper 2016-01, Statistics Netherlands. https://www.cbs.nl/nl-nl/achtergrond/2016/07/social-media-as-a-data-source-for-official-statistics-the-dutch-consumer-confidence-index

Braaksma, B., and K. Zeelenberg (2015), “ ‘Re-make/Re-Model’: Should Big Data Change the Modelling Paradigm in Official Statistics?” Statistical Journal of the IAOS 31 (2), pp. 193–202. doi: 10.3233/sji-150892.

Buelens, B., P.-P. de Wolff, and K. Zeelenberg (2015), “Model based estimation at Statistics Netherlands,” paper presented to the Advisory Council on Methodology and Quality, 2 April 2015, Statistics Netherlands.

Elbourne, A., K. Grabska, H. Kranendonk, and J. Rhuggenaath (2015), “The effects of CBS revisions on CPB forecasts,” CPB Background Document, CPB Netherlands Bureau for Economic Policy Analysis. http://www.cpb.nl/en/publication/the-effects-of-cbs-revisions-on-cpb-forecasts

van Ruth, F. (2014), “Traffic intensity as indicator of regional economic activity,” Discussion Paper 2014-21, Statistics Netherlands. https://www.cbs.nl/nl-nl/achtergrond/2014/34/traffic-intensity-as-indicator-of-regional-economic-activity

Smekens, M. , and K. Zeelenberg (2015), “Lean Six Sigma at Statistics Netherlands,” Statistical Journal of the IAOS 31 (4): 583–86. doi: 10.3233/SJI-150930.

Vincent, P. (2015), Herman Gorter - Poems of 1890: A Selection. UCL Press, London 2015. Open access at http://www.ucl.ac.uk/ucl-press/browse-books/poems-1890/index

Zeelenberg, K. (2015), “Product and service innovation at Statistics Netherlands,” Paper presented to the 63rd plenary session of the Conference of European Statisticians, Geneva, 16 June 2014. http://www.unece.org/index.php?id=38920#jfmulticontent_c48917-3

Zwijnenburg, J. (2015), “Revisions of quarterly GDP in selected OECD countries,” OECD Statistics Brief 22. https://www.oecd.org/std/thestatisticsbrief.htm

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

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

Prepress
Statistics Netherlands, CCN Creatie

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information