Discussion paper

New data sources and inference methods for official statistics

Jan van den Brakel

July 2019

Content

Summary

National statistical institutes are under increasing pressure to reduce administration costs and response burden for the production of official statistics. This could potentially be accomplished by using large data sets - so called big data. However, there are problems that must be addressed when using such data source for the production of official statistics. Two different research lines can be identified on how big data sources can be used in the production of official statistics. The first approach to be presented is to combine big data sources with sample data in a model-based inference approach. This implies that big-data sources are used as covariates in models used for small area estimation and time series models, where cross-sectional and temporalcorrelations are used to improve precision and timeliness of sample statistics. The second approach is to use big data sources as a primary data source for the compilations of official statistics. This can be considered if a big data source covers the intended target population and not suffer to much from under- and over-coverage, e.g. the use of satellite and areal images for deriving statistical information on land use. In most cases, however, adjustments for selection bias are required.

Keywords

Survey sampling, Small area estimation, now casting, non-probability samples

1. Introduction

National statistical institutes are responsible to produce reliable statistical information about economic and social developments of a society. This information is often referred to as official statistics. The required data are obtained via registrations or collected through surveys, usually on the basis of a probability sample. For decades, design-based and model-assisted inference methods have been the preferred methods for national statistical institutes to produce official statistics. The prevailing opinion at national statistical institutes is that official statistics should not be based on explicit statistical model assumptions that are hard to verify. On the other hand there is an increasing pressure on national statistical institutes to reduce administration costs and response burden. In addition, declining response rates compromise the quality of sample estimates and stimulate the search for alternative sources of statistical information. This could be accomplished by using administrative data like tax registers, non-probability samples or other large data sets - so called big data - that are generated as a by-product of processes not directly related to statistical production purposes. Examples of these include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook, sensor data, and internet search behaviour from Google Trends. A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. Other incentives for national statistical institutes to make more use of big data sources in combination with model-based inference methods is an increasing pressure to produce more timely statistical information at a higher frequency and a more detailed level.

The question arises to what extent national statistical institutes in the future can afford to exclusively use traditional probability samples in combination with design-based or model-assisted inference procedures for the production of official statistics. Model-based methods known from small area estimation and nowcasting literature can be used to make more precise and timely predictions for detailed sub populations. New data sources can potentially be used as covariates in these models, since they come at a high frequency and are therefore very timely and also cost effective. The advantage of using big data sources as covariates in models for sample surveys is that problems with selectivity can be circumvented. If, however, big data sources are directly used to produce statistical information, then the potential selection bias of these data sources must be accounted for. In this case statistical modelling also plays a vital role.

The purpose of this paper is to discuss the potentials and risks for national statistical institutes to use these new data sources in combination with model-based inference procedures for the production of official statistics. The chapter is organised as follows. In section 2, the traditional approach of probability sampling in combination with design-based inference methods is reviewed. In section 3 the advantages of model-based inference procedures are described. In Section 4 the pros and cons of new data sources or big data sources are described. In Section 5

the potentials of using these new data sources as covariates in model-based inference procedures is discussed. In Section 6 different methods that account for selection bias of non-probability samples are reviewed. Section 7 concludes with a discussion of the challenges and issues of these new data sources and inference methods for national statistical institutes.

2. The role of probability sampling in official statistics

National statistical institutes gather and publish reliable statistical information about finite populations, generally all people residing in a country or all enterprises registered in a country. This information is often defined as totals, means or proportions. Consider a finite population U of size N. Let $y_i, i = 1, \dots, N$, denote the values of a variable of interest of population unit i. Population totals are typically defined as $Y = \sum_{i=1}^{N} y_i$. Means are simply obtained as $\bar{Y} = Y/N$. This information is not only required at the national level but also for all kind of subpopulations, like municipalities, age classes, gender classes, etc. The population U can be divided in D subpopulations or domains U_d of size N_d. In this case, domain totals are defined as $Y_d = \sum_{i=1}^{N} \delta_{i,d} y_i$, with $\delta_{i,d}$ an indicator taking a value equal to one if element i belongs to domain d, and zero otherwise.

The population values for these variables are generally unknown. Until the beginning of the twentieth century this kind of information was obtained by a complete census of the target population. This is very laborious and expensive. At the beginning of the twentieth century, it gradually became clear that large data sets are not a sufficient condition for valid inference. Despite an impressive 2.3 million respondents, the 1936 Literary Digest poll completely failed to correctly predict the outcomes of the USA presidential elections, because both the sample and the response were selective and not appropriately dealt with (Squire, 1988). This was a strong incentive to embrace the concept of random sampling, which has been developed, mainly on the basis of the work of Bowley (1926) and Neyman (1934), as a method of obtaining valid estimators for finite population parameters based on a relative modest but representative sample, rather than on a complete census. Other important milestone papers are Hansen and Hurwitz (1943), Narain (1951), and Horvitz and Thompson (1952). Under this approach the probability sample s of size n is drawn from the target population U with n<<N. Each element i in the population has a non-zero probability, say π_i, to be included in the sample. An estimator of the unknown population total is obtained as the sum over the observations in the sample, expanded with the so called design weights, i.e. $\hat{Y} = \sum_{i=1}^{n} d_i y_i$, with $d_i = 1/\pi_i$. This estimation procedure is called design-based since inference is completely based on the randomization distribution induced by the sampling design. Statistical modelling of the observations obtained in the survey does not play any role so far.

National statistical institutes often have auxiliary information about the target population from external sources, e.g. censuses and administrative sources. This information can be used to improve the precision of the sample estimates. One way is to improve the efficiency of the sampling design, e.g. stratified sampling with optimal allocation and sampling designs where selection probabilities are approximately proportional to the target variable. Another way is to use this auxiliary information in the estimation procedure via the so called generalized regression estimator proposed by Särndal et al. (1992). The generalized regression estimator expands the observation in the sample with a regression weight such that the sum over the weighted observations is an approximately design unbiased estimator of the unknown population total. Let $\boldsymbol{x}_i$ denote a vector containing auxiliary variables for which the population totals $\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{x}_i$ are known from a register or census. The design weights d_i are adjusted such that the sum over the weighted auxiliary variables in the sample equates to the known population totals, i.e. $\sum_{i=1}^{n} w_i \boldsymbol{x}_i = \boldsymbol{X}$, where w_i are the regression weights. This results in a correction for groups that are underrepresented in the sample, for example due to selective nonresponse. The regression estimator for the population total is now obtained as $\hat{Y}^R = \sum_{i=1}^{n} w_i y_i$. Generally the purpose of a survey is not limited to estimates at the national level but also to produce statistical information for subpopulations or domains. Direct estimates for domain totals are obtained by $\hat{Y}_d^R = \sum_{i=1}^{n} w_i \delta_{i,d} y_i$.

In the model-assisted approach developed by Särndal et al. (1992) this estimator is derived from a linear regression model that specifies the relationship between the values of a certain target variable and a set of auxiliary variables for which the totals in the finite target population are known, i.e. $y_i = \boldsymbol{\beta}' \boldsymbol{x}_i + e_i$. Most estimators known from sampling theory can be derived as a special case from the generalized regression estimator. Examples are the ratio estimator and poststratification. Generalized regression estimators are members of a larger class of calibration estimators, Deville and Särndal (1992).

The generalized regression estimator has two very attractive properties. Although this estimator is derived from a linear model, it is still approximately design-unbiased. If the underlying linear model explains the variation of the target parameter in the population reasonably well, then the use of this auxiliary information will result in a reduction of the design variance compared to the Horvitz-Thompson estimator and it might also decrease the bias due to selective non-response, Särndal et al. (1992), Särndal and Swenson (1987), Bethlehem (1988), and Särndal and Lundström (2005). Model-misspecification might result in an increase of the design variance but the property that this estimator is approximately design-unbiased remains. From this point of view, the generalized regression estimator is robust against model-misspecification. The linear model is only used to derive an estimator that uses auxiliary information but the resulting estimator is still judged by its design-based properties, such as design expectation and design variance. This is the reason that this approach is called model-assisted.

Design-based and model-assisted inference is a very powerful concept since it is based on a sound mathematical theory that shows how under the right combination of a random sample design and estimator, valid statistical inference

can be made about large finite populations based on relative small samples. In addition, the amount of uncertainty by relying on small samples can be quantified through the variance of the estimators. A strong advantage of probability sampling in combination with a design-based or model-assisted inference is that it has a built-in robustness against model misspecification. This is useful in a production process where there is not much time for extensive model evaluation. For these reasons, design-based and model-assisted inference is still used in modern statistical science and is the standard for most national statistical institutes for producing official statistics.

3. Towards model-based inference in official statistics

Model-based inference refers to estimation procedures that rely on the probability structure of an explicitly assumed statistical model, whereas the probability structure of the sampling design plays a less pronounced role. This is the position taken by authors like Gosh and Meeden (1997), Valliant et al. (2000), and Rao and Molina (2015).

Results published by national statistical institutes must enjoy public confidence. For decades, this has resulted in the prevailing opinion that methods used to produce official statistics, particularly if they are used for planning and implementing policies, must be free from model assumptions and should therefore be based on the above-mentioned design-based and model-assisted approaches. The reason for this is that models depend on assumptions that are hard to verify, which raises concerns about the validity of the results. Design-based and model-assisted approaches, however, have some limitations. In the case of small sample sizes the design-variances of the sample estimates become unacceptably large, which makes the built-in robustness against model misspecification of less use. Furthermore, they do not handle measurement errors effectively. In such situations model-based estimation procedures can be used as an alternative. The rapid rise of large data sets - so called big data - that are generated as a by-product of processes not directly related to statistical production purposes is another incentive for national statistical institutes to move towards model-based inference procedures as will be detailed in Sections 4, 5 and 6.

Important quality aspects of official statistics are accuracy, relevance, timeliness, and comparability with preceding periods. Relevance of statistical information increases with the level of detail and the frequency of the information. For policy making monthly figures at a low regional level are in general more relevant than annual figures at the national level. Figures for reference period t are more relevant if they become available in t+1, instead of a delay of multiple time lags. This results in detailed breakdowns of a target population in domains or subpopulations with respect to regions or socio-demographic classifications in

combination with short reference periods. In such situations domain sample sizes rapidly become too small to produce sufficiently precise domain estimates with design-based or model-assisted procedures. As an alternative model-based procedures, which explicitly use a statistical model can be used to improve the effective sample size of a particular domain with the information from other domains or preceding sampling periods. These methods are in the literature referred to as small area estimation, see e.g. Rao and Molina (2015), Pfeffermann (2002, 2013).

Small area estimation is predominantly based on multilevel models. These methods can be classified in area level models (Fay and Herriot, 1979), and unit level models (Battese et al., 1988). These models are predominantly used to take advantage of cross-sectional sample information that is observed in other domains. In an area level model, the direct estimates of the domains are modelled in a multilevel model, while in a unit level model the sampling units are the input for a multilevel model. They consist of a regression component, where available auxiliary information is used to explain the variation in the survey data, and a random component, which describes the unexplained variation between the domains. Through the regression component, sample information from other domains is used to improve the precision of the estimates for each domain separately. To define an area level model a measurement error model is assumed for the observed domain estimates; $\hat{Y}_d^R = Y_d + e_d$, with e_d the sampling errors which are assumed to be normally and independently distributed; $e_d \sim N(0, \psi_d)$. Subsequently a linear model for the true population parameter is assumed; $Y_d = \boldsymbol{x}_d'\boldsymbol{\beta} + v_d$, with $\boldsymbol{x}_d$ a vector of auxiliary information at the domain level, $\boldsymbol{\beta}$ a vector with regression coefficients and v_d the random domain effects that are assumed to be normally and independently distributed; $v_d \sim N(0, \sigma_v^2)$. Assuming that the design variances ψ_d are known, estimates for $\boldsymbol{\beta}$, v_d and σ_v^2 can be obtained with maximum likelihood methods or Bayesian methods. Finally model based predictions for Y_d including approximations of its uncertainty can be derived. See Rao and Molina (2015) for details. With a unit level model a similar multilevel model is defined but now at the level of the observations of the sampling units. We further focus on the area level model, since most auxiliary information from new data sources are fuzzy and difficult to match at the unit level but are often available at the domain level.

Most surveys conducted by national statistical institutes are conducted repeatedly over time. A natural approach for small area prediction as well as nowcasting is to extend the Fay-Herriot model with related information from previous editions of the survey. Rao and Yu (1994) extended the area level model by modelling random domain effects with an AR(1) model. Other accounts of regional small area estimation of unemployment, where strength is borrowed over both time and space, include Tiller (1992), Datta et al. (1999), You (2008), Pfeffermann and Tiller (2006).

Temporal information can be included in the area level by assuming a structural time series (STS) model for the unknown domain parameters. Similarly to the area level model, a time series model for survey estimates observed with a periodic survey starts with a measurement error model, $\hat{Y}_{t,d}^R = Y_{t,d} + e_{t,d}$, where subscript t

refers to the time periods of the survey, $t = 1, \dots, T$. Subsequently a structural time series model is assumed for the domain parameters. For simplicity we assume a basic structural time series model, which assumes that a series can be decomposed in a stochastic trend model, say $L_{t,d}$, for modelling the low frequency variation, a stochastic seasonal component, say $S_{t,d}$, to model a cycle pattern with a period of one year, and a white noise, say $v_{t,d}$, for the remaining unexplained variation. This leads to $Y_{t,d} = L_{t,d} + S_{t,d} + v_{t,d}$. This model can be extended with other cycles, regression components and AR or MA components. See Durbin and Koopman (2012) for an introduction to STS modelling. For the components stochastic models are assumed, which makes them time dependent. A frequently applied trend model is the local linear trend model, which is defined as

$$L_{t,d} = L_{t-1,d} + R_{t-1,d} + \xi_{t,d}, \quad \xi_{t,d} \sim N(0, \sigma_\xi^2)$$
$$R_{t,d} = R_{t-1,d} + \eta_{t,d}, \qquad \eta_{t,d} \sim N(0, \sigma_\eta^2)$$

For the seasonal component the dummy or trigonometric seasonal component can be used, see Durbin and Koopman (2012) for an expression. The white noise terms are independently normally distributed; $v_{t,d} \sim N(0, \sigma_v^2)$. Inserting the STS model into the measurement error model gives $\hat{Y}_{t,d}^R = L_{t,d} + S_{t,d} + \varphi_{t,d}$, with $\varphi_{t,d} = v_{t,d} + e_{t,d}$ and assuming that $\varphi_{t,d} \sim N(0, \psi_d \sigma_\varphi^2)$ with ψ_d assumed to be known. See Van den Brakel and Krieg (2015) for details. STS models can be fitted using the Kalman filter after writing them in state-space form, see Durbin and Koopman (2012) for details.

The univariate STS model can be seen as a form of small area estimation, where sample information from preceding periods is used to improve the effective sample size for the last period. This model can be extended in several ways. A first generalization is to combine the time series of all D domains in one multivariate STS model. In this case the D domain estimates for one period are stacked in one vector $\hat{\boldsymbol{Y}}_t^R = (\hat{Y}_{t,1}^R, \dots, \hat{Y}_{t,D}^R)'$. Each series has its own trend and seasonal component. By modelling the correlations between the level disturbances of the domains $\xi_{t,d}$ cross-sectional information from other domains can be used. This assumes a $D \times D$ full covariance matrix for the vector $\boldsymbol{\xi}_t = (\xi_{t,1}, \dots, \xi_{t,D})'$. In a similar way the correlation between the slope disturbances $\eta_{t,d}$ can be modelled as well as the disturbance terms of the seasonal components. This results in a multivariate STS model that uses temporal and cross-sectional information to improve the effective sample size for the different domains. This approach is followed by Pfeffermann and Burck (1990), Pfeffermann and Bleur (1993), Van den Brakel and Krieg (2016), Boonstra and Van den Brakel (2019).

Another useful application of STS models is to account for non-sampling errors. As long as the survey design of a repeated cross-sectional survey is not changed, non-sampling errors like measurement bias and selection bias remain rather invisible. In some situations the effects of non-sampling errors become visible. The first example are rotating panel designs, which are frequently used by national statistical institutes for labour force surveys. In a rotating panel on each survey occasion a new panel is added to the sample, and followed for a number of periods according to a predetermined pattern, after which the panel is (normally) dropped and replaced by a new one. Generally there are systematic differences

between the responses of the subsequent waves, which is referred to in the literature as rotation group bias (RGB), see Bailar (1975). Pfeffermann (1991) proposed a multivariate STS model where time series of direct estimates of the different waves of the rotating panel serve as the input and the RGB is explicitly modelled. This model can be used as a form of small area estimation and accounts for RGB induced by the rotating panel design. Another occasion where non-sampling errors become visible are major redesigns of the survey process of a repeated survey. When methods are necessarily updated it generally causes a change in the series. Such systematic differences are distinct from the sampling error and are known as a discontinuities. One way to avoid confounding real period-to-period change from discontinuities is to model the effect of a redesign with an STS model. In this case the above proposed model is extended with an intervention variable which changes from zero to one at the moment of implementing the new survey design. The corresponding regression coefficient can be interpreted as the discontinuity, see e.g. Van den Brakel and Roels (2010).

Finally the STS model can be augmented with related auxiliary series. This can be done by extending the univariate STS model with a regression component or by defining a bivariate STS model where the input vector contains the survey estimate and the auxiliary series, say $\left(\hat{Y}_{t,d}^{R}, x_{t,d}\right)'$. Both series have their own trend and seasonal component. The correlation between level disturbance terms of both series can be modelled in a similar way as explained for the multivariate STS model for all domain estimates. Also the correlation between the disturbance terms of other model components can be modelled. In this way the additional information from related auxiliary series is used to improve the survey estimates, see e.g. Harvey and Chung (2000) and Van den Brakel and Krieg (2016).

Improving precision of direct estimates is an argument for national statistical institutes to move towards model based estimation procedures in the production of official statistics. Statistics Netherlands made some steps in this direction. Boonstra et al., (2008), summarizes the first research results in small area estimation at Statistics Netherlands. Based on this work Statistics Netherlands currently uses a multivariate STS model in the production of monthly Labour Force figures to handle problems with small sample sizes, rotation group bias and discontinuities since 2010, Van den Brakel and Krieg (2015). A similar model has been implemented in 2017 for producing official figures for the Consumer Confidence Index. A Battese-Harter-Fuller unit level model is in use since 2015 to produce annual municipal unemployment figures, Boonstra et al. (2011). A multi-level time series modelling approach, based on an extension of the model proposed by Bollineni-Balabay et al. (2016) will be implemented in 2019 to estimate official trend figures in time series of the Dutch National Travel Survey.

4. New data sources

The accuracy of statistics is measured by its variance and bias. The variance is inversely related to the sample size and will generally be a major uncertainty component for survey sample statistics, because sample surveys usually have limited sample sizes. A strong point of sample surveys is that a national statistical agency has control over the quality of the survey outcomes through the design of the sample survey. The precision of the sample estimates can be controlled in advance via variance and sample size calculations and the choice of an optimal sampling strategy, i.e. the combination of a sample design and estimator. In addition the national statistical institute is in control of the availability of the data source as well as its frequency. Repeated sample surveys are therefore a stable data source for measuring the evolution of social-economic phenomena over time.

Concerning bias, we distinguish between selection bias and measurement bias. The selection bias of sample survey statistics is approximately zero under complete response. In practice however, selection bias arises due to selective nonresponse, under-coverage of the sample frame and to what extent the field work strategy successfully reached the target population. Particularly non-response can be informative and result in biased estimates if not appropriately accounted for (Pfeffermann and Sverchkov, 2003, 2009). The measurement bias in sample statistics typically depends on the extent to which the conceptual variables to be measured are correctly implemented in the questionnaire, on the mode of data collection and on the quality and skills of the interviewers in the case of telephone and face-to-face surveys. Problems with measurement bias in surveys arises, since measurements of the variables of interest are indirect in that respondents are asked to report about their behavior, introducing all kind of measurement errors.

Drawbacks of sample surveys are that data collection is costly, its quality is compromised by non-response and measurement bias, and they are generally not very timely. In addition survey samples induce response burden, which is particularly an issue in business surveys. For national statistical institutes this is an argument to make more use of administrative data like tax registers, or other large data sets - so called big data - that are generated as a by-product of processes not directly related to statistical production purposes. Examples of these include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook, internet search behavior from Google Trends, information found on the internet, web scraping, scanner data and sensor data like e.g. satellite images, aerial images and road sensor data. A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. A challenging problem in this context is to use this data for the production of official statistics that are representative of the target population. There is no randomized sampling design that facilitates the generalization of conclusions and results obtained with the available data to an intended larger target population. Hence, extracting statistically relevant information from these sources is a challenging task.

A strong point of administrative data sources and some big data sources is that they contain direct measurements of people's behavior, and are therefore unaffected by measurement bias induced by questionnaires. Examples include smart meters to measure electricity consumption, GPS trackers in mobile phones to measure mobility and travel of populations, search and purchase behavior on the internet. If similar information has been collected via questionnaires, substantial measurement bias might occur. This only holds, however, for specific examples.

A problem with registers and big data sources is that a national statistical institute has no control over the quality, availability and stability of this data source. Major changes in the behavior of the public on social media and internet have a disturbing effect on the comparability of series over time. Also the use of these media might fluctuate over time. For example a Google-trend series on search related to vacancies might track an official series on unemployment. It does not measure unemployment, however. Search behavior before the start of the financial crisis in 2009 might be completely different compared to the period directly after the financial crisis, invalidating measurements of the intendent concept. Another example is the frequency with which administrative data become available. For the short term business statistics, published on a monthly frequency, Statistics Netherlands changed from survey data to administrative data of value added tax in a period that businesses were required by law to declare value added tax on a monthly frequency. Later on this legislation changed and businesses were allowed to choose whether they declared tax on a monthly, quarterly or even annual frequency.

Particularly in the case of big data with immense volumes, the variance will often be a minor uncertainty component. The bias, however, might be substantial. The size of particularly the selection bias depends on the extent to which the non-probability data source represents or covers the intended target population. Data obtained from smart meters, GPS trackers and internet behavior are currently considered for production of official statistics, because they measure the individual behavior very precisely in a cost-effective way. Here the question is how to account for selection bias.

The rise of the big data era is somewhat reminiscent of the developments of probability sampling in the early 20th century due to problems with the use of large non-probability samples like the 1936 Literary Digest poll. The volume of big data might lure some into the same trap of narrowing accuracy to precision, ignoring selection bias. This paradox has been mathematically formalized by Meng (2018) who derived an expression for the error of estimates derived from non-probability samples. The error contains three components; 1) a data quality measure or data defect index which measures the level of departure from simple random sampling, 2) a data quantity measure which measures the fraction of the target population covered by the big data sample, and 3) a problem difficulty measure, which is the standard deviation of the target variables. This measure shows that selection-bias in non-probability samples becomes an issue if the data defect index becomes substantial even if the sample size is voluminous.

The non-probability nature of the data therefore requires dedicated methods of inference to produce statistics about the intended, finite target population. Broadly spoken, there are two ways to use non-probability data sources in the production of official statistics. The first approach is to use them as covariates in model-based prediction methods for survey data. The second approach is to use them directly as a data source for official statistics and correct for possible selection bias.

5. Big data as auxiliary variables

Problems with selection bias of non-probability data sources can be circumvented, at least partially, if they are used as covariates in prediction models for sample survey data. One potential application are small area estimation models. Most big data sources are fuzzy and volatile and the records typically do not coincide with the units of an intended target population or the sampling units of a probability sample. Therefore linking units in big data sources with sampling units in a probability sample will often be a heroic task. These complications can be avoided, at least partially, by using area level models instead of unit level models for small area estimation. The area level model was briefly introduced in Section 4. Covariates traditionally used in small area prediction models are available from registers and censuses. The value of new data sources is multiple. First of all in developing countries and combat areas, the availability of registers, frequently updated censuses and survey data are generally scarce. Satellite images and mobile phone data can have valuable information for making detailed regional predictions. Also in developed countries, new timely data sources offer valuable additional information, e.g. once a census, which is typically conducted with a frequency of 10 years, becomes outdated. The high frequency with which new data sources become available allow for more frequent updates of official statistics (Powell, et al, 2017, Hand, 2018).

Parallel to the development of the small area estimation literature, several authors proposed methods to combine survey data with non-probability data sources available from, e.g. sensor data and mobile phone data with the purpose to make detailed regional predictions for well-being and poverty. Many applications apply machine learning algorithms to establish the relation between survey data and sensor or mobile phone data and use the latter data set in a second step to make detailed regional predictions. Noor et al. (2008) analyzed the correlation between night-time light intensity from satellite images and survey sample data on household income in Africa. They report a high correlation and used this empirical finding as a motivation to use night-time light intensity as an alternative measure for poverty. Although one can question whether night-time light intensity is an efficient construct to measure poverty, their empirical findings illustrate the potential of using remote sensor information as covariates in small area prediction

models. Engstrom et al. (2017) used day time satellite images to predict well-being. In a first step they applied deep learning to extract features from satellite images that are potentially related to well-being, like number of cars, building type, roof type, etc.. In a next step they applied a Lasso to construct a linear model that relates the relevant images features with survey data on well-being. This relation is used to predict well-being on a fine regional detail in Sri Lanka. Blumenstock et al. (2015) applied machine learning methods to combine mobile phone data with survey data on poverty and used this to predict poverty and well-being on small regional level in Rwanda. Steele et al. (2017) combine survey data and mobile phone and satellite data in a generalized linear model to predict poverty in Bangladesh. This literature illustrates the potential value of these new forms of data for official statistics.

Some caution, however, for making fine regional predictions with the use of machine learning algorithms for over-reliance on a model is required. One step into this direction is made by Marchetti et al. (2015) who used mobility patterns of cars tracked with GPS as a covariate in a Fay Herriot model for predicting poverty for small regions in Italy. This class of small area estimation predictions are specified as a composite estimator of a model-based prediction and a design-based estimate where the weights are based on their measure of uncertainty and provide mean squared error approximations for the uncertainty of the small domain predictions. Similarly Schmid et al. (2017) use mobile phone data as a covariate in a Fay Herriot model to predict literacy in Senegal.

In Section 3 it was emphasized that STS models are particular appropriate as a form of small area estimation, since official statistics are based on repeated surveys. Multivariate STS models are therefore appropriate to borrow strength over both time and space. Multivariate STS models can be used in a similar way to combine time series obtained with repeated sample surveys with auxiliary series derived from registers or big data sources.

This serves two purposes. Extending the time series model with an auxiliary series allows modelling the correlation between the unobserved components of the structural time series models, e.g. trend and seasonal components. If the model detects a strong correlation, then the accuracy of domain predictions will be further increased. Harvey and Chung (2000) propose a time series model for the Labor Force Survey in the UK extended with a series of claimant counts. Information derived from non-traditional data sources like Google trends or social media platforms are generally available at a higher frequency than series obtained with repeated surveys. This allows to use this time series modelling approach to make predictions for the survey outcomes in real time at the moment that the outcomes for the big data series are available, but the survey data not yet. In this case the auxiliary series are used as a form of nowcasting. Van den Brakel et al. (2017) applied a bivariate STS model to estimate the Consumer Confidence Index, based on a monthly cross-sectional sample, in real time using an auxiliary series derived from messages left on social media platforms. Google Trends in particular has been used in the economic forecasting literature for this purpose, see e.g. Vosen and Schmidt (2011) and the references therein.

To exploit the timeliness of the auxiliary series obtained with big data sources, the multivariate STS model can be expressed at the high frequency of the auxiliary series. This requires a disaggregation of the unobserved time series components of the target series observed with a repeated survey at a low frequency to this higher frequency. After fitting the model, estimates for the survey parameters are obtained by aggregating the underlying components to a monthly frequency. Details of mixed frequency state-space models are described in Harvey (1989), Ch. 6.3, Durbin and Quenneville (1997), and Moauro and Savio (2005).

With data sources like Google trends, a large number of potential auxiliary series might be obtained easily. Combining them in a full multivariate STS model as outlined before, limits the degrees of freedom for model fitting. Due to the so-called "curse of dimensionality", prediction power of such models will be low. From this perspective, factor models are developed to formulate parsimonious models, despite the fact that a large number of auxiliary series is considered. Factor models are developed and widely applied by central banks to nowcast GDP on a quarterly frequency using a large amount of related series observed on a monthly frequency, Boivin and Ng (2005), Stock and Watson (2002a, 2002b), and Marcellino et al. (2003). More recently, Giannone et al. (2008) and Doz et al. (2011) proposed a state-space dynamic factor model. They propose a two-step estimator. In a first step a small amount of common factors are extracted from a large set of series using principal component analysis. In a second step, the common factors are combined with the target series in a state-space model and are fitted using the Kalman filter. This approach is applied by Schiavoni et al. (2019) to estimate monthly unemployment figures in real time with claimant count series and Google trend series.

6. Big data as direct data source for official statistics

If non-probability data sources are considered as a primary data source for compiling official statistics, then the question arises to what extent results obtained with a non-probability data source can be generalized to an intended, larger target population. Contrary to probability samples, the data generating process of these data sources is generally unknown. As a result, statistical information derived from non-probability samples can suffer from large selection bias if it is used for these purposes.

Different methods are proposed in the literature to account for selection bias in non-probability samples. Some authors apply standard weighting and calibration methods known from classical probability sampling to non-probability samples, which are referred to as pseudo-design-based inference methods (Baker et al., 2013). Several authors apply propensity scoring, proposed by Rosenbaum and

Rubin (1983), to construct weights that correct for selection bias. Estimating response probabilities and using them in Horvitz-Thompson type estimators to account for unequal selection probabilities is sometimes called pseudo-randomization. Valliant and Dever (2011) propose different models to estimate response probabilities in opt-in Web panels and discuss under which conditions they correct for selection bias. Deville (1991) proposed models for quota samples, which can be used to construct post-stratification estimators or linear weighting type estimators (Dever et al., 2008). There are many references in the literature where propensity scores are used to correct for selection bias in non-probability samples (see, e.g. Lee, 2006, Lee & Valliant, 2009, Schonlau et al., 2007, 2009). Auxiliary information typically available for weighting and calibration are demographic variables like age class, gender, regional classifications. Buelens et al. (2018), compared pseudo-design based, model-based and algorithmic methods and concluded that such demographic auxiliary variables do not sufficiently explain the data generating process of a non-probability sample to correct successfully for selection bias.

Another class of methods to correct for selection bias apply a statistical model to predict the units not in the sample (Royall, 1970, Valliant et al., 2000). This approach is based on the specification of an appropriate super-population model that captures the variation of the target variables instead of adjusting selection probabilities.

Some methods combine a non-probability sample that contains the target variable of interest and auxiliary variables with a reference sample that is based on a probability sample and only contains auxiliary variables. The reference sample is used to assess the selectivity of the non-probability sample. One approach, quasi randomization, is to construct propensity models to estimate selection probabilities for the non-probability sample (Isaksson and Forsman, 2003, Elliot and Vaillant, 2017, Vaillant et al. 2013). Sample matching is also applied as an attempt to reduce selection bias in opt-in Web panels using covariates obtained in a small reference sample to construct propensity weights without collecting observations for the target variables (Vavreck and Rivers, 2008, Rivers and Bailey, 2009, Terhanian and Bremer, 2012). These ideas are related to approaches that are also used in microsimulation to match probability samples with population or census data (Tanton and Edwards, 2013). Kim and Wang (2018) proposed inverse sampling. In a first step, important weights are derived for the units in the non-probability sample, using the auxiliary variables in the reference sample and the non-probability sample. In a second step, a sample using unequal probability sampling proportional to the important weights is drawn from the non-probability sample, such that it can be interpreted as a simple random sample from the target population. As an alternative Kim and Wang (2018) proposed data integration which implies that a parametric model is assumed to construct weights for the units in the non-probability sample, which are subsequently used in standard weighting methods. Rivers (2007) proposed imputation of the target variables observed in the non-probability sample in the reference sample using nearest neighbor imputation and subsequently applying standard weighting methods.

A consequence of combining a large non-probability sample with a high-quality smaller reference sample is that the precision of the large non-probability sample reduces to the standard error of the smaller reference sample. These methods nevertheless might improve the accuracy, in terms of mean squared error, of estimates derived from non-probability samples. The methods summarized above are based on strong ignorability assumptions and can lead to serious bias if these assumptions are not met.

In the case that the non-probability sample and the probability-based reference sample both contain the target variable and some auxiliary variables, Kim and Tam (2018) propose a design-based inference method that can be regarded as a post-stratification estimator where one stratum is the subpopulation that is completely observed with the non-probability sample. Model-based approaches for informative sampling (Pfeffermann and Sverchkov, 2003, 2009), where the selection probabilities are related to the target variables, might potentially be fruitful to correct for selection bias in non-probability samples for situations where no reference sample is available.

7. Discussion

National statistical institutes face multiple challenges. There is an increasing pressure to reduce administration costs and response burden. Non response is a gradually increasing problem which compromises the quality of traditional sample surveys. In order to remain relevant for data users, the level of detail, frequency and timeliness of statistical information must increase. This raises the question whether national statistical institutes can continue to base official statistics on probability samples in combination with design-based or model-assisted inference methods solely. The advantage of this approach is its low risk level. With sample surveys, a national statistical institute has full control over the availability of the data, as well as the quality and frequency of the statistical output. Model assisted inference methods have a built-in robustness against model miss-specification, which make these methods attractive for multipurpose surveys in the production of official statistics where there is usually very limited time for model building and evaluation. Repeated sample surveys therefore provide a safe method to produce consistent time series that measure period-to-period change in a reliable way.

In order to improve the level of detail, frequency and timeliness of statistical information, without increasing sample sizes and thus data collection costs, model based inference procedures known from the literature of small area estimation, time series analysis and nowcasting can be considered. This, however, increases the risk level for a national statistical institute, since model miss-specification can result in biased statistical information. The output, however, is primarily based on sample survey data, collected by the national statistical institute. This implies that the risks concerning availability, frequency and quality of the data are still

managed by the national statistical institute. In this context new data sources can provide useful additional information as covariates in small area prediction models, particular for countries without registers or timely census data. Many big data sources are available at a high frequency which makes them potentially useful to make more precise predictions of sample statistics in real time with nowcasting models.

Replacing sample surveys for registers or other types of non-probability data sources, implies a substantially increased risk level, since in this situation a national statistical institute has no control over the availability, comparability, and quality of the data source over time. Another issue with using big data as a primary data source to compile statistical information is to account for selectivity. Big data are used successfully in many different disciplines. The use of these data sources in the context of official statistics is, however, different. The problem, which is unique for official statistics, is the question to what extent statistical results can be generalized to larger intended target populations (Pfeffermann et al. 2015, Pfeffermann, 2019).

As highlighted in Section 6, there is a substantial amount of literature for correcting for selection bias in non-probability samples. There are nevertheless a lot of issues with the application of these methods in the daily practice of official statistics. One issue is that all methods are based on strong ignorability assumptions conditional on the available covariates, which are difficult to verify. A more practical issue is that all methods assume that the records in a big data source contains besides the target variable a set of auxiliary variables which correspond to the units in a target population or a reference sample. Unfortunately, these conditions are seldom met. Most big data sets are fuzzy, records do not correspond with units in the target population or a reference sample and auxiliary information is generally not available since owners of the big data source are reluctant to provide them due to privacy issues. Mobile phone data, e.g., are mostly a file of call detail records that contain time and location information generated by devices. Mobile phone companies generally do not provide the demographic information of the owners of the devices. As a result, methods summarized in Section 6 to correct for selection bias cannot be applied in a straightforward manner in these situations. Attempts to use these data to produce for example day time population statistics are based on machine learning methods which attempt to derive demographic information from the observed mobility patterns of the devices, followed by rather naïve post-stratification corrections. It is not likely that this sufficiently corrects for selectivity.

At this moment it is not at all clear how big data can be used in the production of official statistics (Pfeffermann et al. 2015, Pfeffermann 2019). National statistical institutes, nevertheless, have to investigate to what extent these new data sources in combination with new inference methods can be used to improve the level of detail, frequency and timeliness of their publications on the one hand and to reduce data collection costs at the other hand. The literature that uses satellite images and mobile phone data to make small area predictions for poverty and well-being on a fine regional level, clearly illustrate the potential of big data sources. Using these new data sources in the production of official statistics

requires more research and insight into the quality of these data sources and an extension of the methodological tools to extract the right information from these new data sources. This is not only an extension from design-based to model-based inference, but also to machine learning methods and artificial intelligence algorithms to extract information from satellite and aerial images or sensor data. The advantage of all these developments is that it makes life of an official statistician more exciting.

References

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau (2013). Summary report of the AAPOR task force on non-probability sampling. Journal of Survey Statistics and Methodology, 1, pp. 90-143.

Battese, G.E., R.M. Harter and W.A. Fuller (1988). An error components model for prediction of county crop areas using satellite data. Journal of the American Statistical Association, 83, pp. 28-36.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. Journal of Official Statistics, 4, pp. 251-260.

Blumenstock, J., G. Cadamuro and R. On (2015). Predicitng poverty and wealth from mobile phone metadata. Science, 350, pp. 1073-1076.

Boivin, J. and S. Ng (2005). Understanding and comparing factor-based forecasts. International Journal of Central Banking. 3, pp. 117-151.

Bollineni-Balabay, O. Brakel, J.A. van den and Palm, F. (2016). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns. Journal of the Royal Statistical Society, A series, vol 179, pp. 377-402.

Boonstra, H.J. and J.A. van den Brakel (2019). Estimation of level and change for unemployment using structural time series models. Survey Methodology, forth coming.

Boonstra, H. J., B. Buelens, K. Leufkens, and M. Smeets (2011). Small area estimates of labour status in dutch municipalities. Technical Report 201102, https://www.cbs.nl/nl-nl/achtergrond/ 2011/02/small-area-estimates-of-labour-status-in-dutch-municipalities, Statistics Netherlands.

Boonstra, H.J., J.A. van den Brakel, B. Buelens, S. Krieg and M. Smeets (2008). Towards small area estimation at Statistics Netherlands. Metron International Journal of Statistics, LXVI, pp. 21-50.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. Bulletin de l' Institute International de Statistique 22, Supplement to Book 1: 6-62.

Buelens, B., J. Burger and J.A. van den Brakel (2018). Comparing inference methods for non-probability samples. International Statistical Review, 86, pp. 322-343.

Datta, G., P. Lahiri, T. Maiti, and K. Lu (1999). Hierarchical Bayes estimation of unemployment rates for states of the US. Journal of the American Statistical Association, 94, pp. 1074-1082.

Deville, J., and C.-E. Särnal (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, 87, pp. 376-382.
Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. Journal of Econometrics, 164, pp. 188-205.

Dever, J. A., A. Rafferty, and R. Valliant (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? Survey Research Methods, 2, pp. 47-60.
Deville, J.-C. (1991). A theory of quota surveys. Survey Methodology, 17, pp. 163-181.

Durbin, J. and S.J. Koopman (2012). Time Series Analysis by State Space Methods. Oxford: Oxford University Press.

Durbin, J. and B. Quenneville (1997). Benchmarking by state space models. International Statistical Review, 65, pp. 23-48.

Elliot, M. R. and R. Vailliant (2017). Inference for non-probability samples. Statistical Science, 32, pp. 249-264.

Engstrom, R., J. Hersh and D. Newhouse (2017). Poverty from Space: Using high resolution satellite imagery for estimating economic well-being. Technical report.
Fay, R.E. and R.A. Herriot (1979). Estimation of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Society, 74, pp. 268-277.

Giannone, D.L., L. Reichlin and D. Small (2008). Nowcasting: The real-time information content of macroeconomic data. Journal of Monetary Economics, 55, pp. 665-676.

Gosh, M., and G. Meeden (1997). Bayesian Methods for Finite Population Sampling. London: Chapman & Hall.

Hand, D.J. (2018). Statistical challenges of administrative and transaction data. Journal of the Royal Statistical Society, A series, Vol. 181, pp. 555-605.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics.14, pp. 333-362.

Harvey, A.C. (1989). Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

Harvey, A.C. and C. Chung (2000). Estimating the underlying change in unemployment in the UK. Journal of the Royal Statistical Society, A Series, 163, pp. 303-339.

Horvitz, D.G., and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, pp. 663-685.

Isaksson, A. & Forsman, G. (2003). A comparison between using the web and using the telephone to survey political opinions. In Annual Meeting of the American Association for Public Opinion Research, Nashville, TN, pp. 100-106.

Kim, K. and Z. Wang (2018). Sampling techniques for big data analysis in finite population inference. International Statistical Review, 87, pp. 177-191.

Kim, K. and S.M. Tam (2018). Data integration by combining big data and survey sample data for finite population inference. Working paper.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. Journal of Official Statistics, 22, pp. 329-349.

Lee, S. and R. Valliant (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods and Research, 37, pp. 319-343.

Marcellino, M., J. Stock and M. Watson (2003). Macroeconomic forecasting in the euro area; country specific versus area wide information. European economic review, 47, pp. 1-18.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Perdreschi, Rinzivillo, L. Pappalardo and L. Gabrielli (2015). Small area model-based estimators using Big data sources. Journal of Official Statistics. 31, pp. 263-281.

Moauro, F. and G. Savio (2005).Temporal disaggregation using multivariate structural time series models. Econometrics Journal, 8, pp. 214-234.

Noor, A., V. Angela, P. Gething, A. Tatem, and R. Snow (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. Population and Health Metrics, 6:5, doi 10.1186/1478-7954-6-5.

Pfeffermann, D.A. (2019). Challenges in the production of official statistics with different methods of data collection. Paper presented at the Annual Workshop on Survey Methodology, Brazilian Network Information Centre (NIC.br). Sao Paulo, 20 May, 2019.

Pfeffermann, D.A. (2013), New Important Developments in Small Area Estimation. Statistical Science, 28, pp. 40-68.

Pfeffermann, D.A. (2002), Small Area Estimation – New Developments and Directions. International Statistical Review, 70, pp. 125-143.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. Journal of Business & Economic Statistics, 9, pp. 163-175.

Pfeffermann, D. and S.R. Bleuer (1993). Robust Joint Modelling of Labour Force Series of Small Areas. Survey Methodology, 19, pp. 149-163.

Pfeffermann, D. and L. Burck (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. Survey Methodology, 16, pp. 217-237.

Pfeffermann, D., Eltinge, J. L. & Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics. Journal of Survey Statistics and Methodology, 3, pp. 425-483.

Pfeffermann, D. and M.Y. Sverchkov (2003). Fitting generalized linear models under informative sampling. In Analysis of Survey Data, Eds. Chambers, R. L. & Skinner, C. J., pp. 175-195. Chichester: Wiley.

Pfeffermann, D. and M.Y. Sverchkov (2009). Inference under informative sampling. In Handbook of Statistics, Vol. 29, Ed. Rao, C., pp. 455–487. Amsterdam: Elsevier.

Pfeffermann, D. and R. Tiller (2006). Small Area Estimation with State Space Models Subject to Benchmark Constraints. Journal of the American Statistical Association, 101, pp. 1387-1397.

Powell, B., G. Nason, D. Elliot, M. Mayhew, J.J. Davies and J. Winton (2017). Journal of the Royal Statistical Society A series, Vol. 181, pp. 737-756.

Narain, R. (1951). On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 3, pp. 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97, pp. 558-625.

Meng, X.L. (2018). Statistical paradises and paradoxes in big data. The Annals of Applied Statistics, 12, pp. 685-726.

Rao, J.N.K. and I. Molina (2015). Small Area Estimation, 2nd edition. New York: Wiley.

Rao, J.N.K. and M. Yu (1994). Small area estimation by combining time series and cross-sectional data. The Canadian Journal of Statistics, 22, pp. 511-528.

Rivers, D. (2007). Sampling for web surveys. In 2007 JSM Proceedings, ASA Section on Survey Research Methods, American Statistical Association.

Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 US national elections. In Proceedings of the Joint Statistical Meetings, Washington, DC, pp. 627-639.

Rosenbaum, P.R. and D.B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, pp. 41-55.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. Biometrika, 57, pp. 377-387.

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.

Schonlau, M., A. van Soest, and A. Kapteyn (2007). Are 'Webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? Survey Research Methods, 1, pp. 155-163.

Schonlau, M., A. van Soest, A. Kapteyn, and M. Couper (2009). Selection bias in web surveys and the use of propensity scores. Sociological Methods and Research, 37, pp. 291-318.

Schmid, T., F. Bruckschen, N. Salvati and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society, Series A, 178, pp. 239-257.

Särndal, C.-E., and S. Lundström (2005). Estimation in Surveys with Nonresponse. New-York: Wiley.

Särndal, C.E., and B. Swensson (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. International Statistical Review, 55, pp. 279-294.

Särndal, C.E., B. Swensson, and J. Wretman (1992). Model Assisted Survey Sampling. New York: . Springer-Verlag.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. Public Opinion Quarterly, 52, pp. 125-133.

Steele, J., P.R. Sundsøy, C. Pezzulo, V.A. Alegana, T.J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.A. de Montjoye, A.M. Iqbal, K.N. Haddiuzzaman, X. Lu, E. Wetter, A.J. Tatum and L. Bengtsson (2017). Mapping poverty using mobile phone and satellite data. Journal of the Royal Statistical Society Interface, 14, 127.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Society, 97, pp. 1167-1179.

Stock, J. and M. Watson (2002b). Macroeconomic forecasting using diffuse indexes. Journal of Business and Economic Statistics, 20, pp. 147-162.

Valliant, R. and J.A. Dever (2011). Estimating propensity adjustments for volunteer web surveys. Sociological Methods and Research, 40, pp.105-137.

Valliant, R. and J.A. Dever and F. Kreuter (2013). Practical tools for designing and weighting survey samples. New York: Springer verlag.

Valliant, R., A.H. Dorfman, and R.M. Royall (2000). Finite Population Sampling and Inference, A Prediction Approach. New York: Wiley.

Van den Brakel, J.A. and S. Krieg (2016). Small area estimation with state-space common factor models for rotating panels. Journal of the Royal Statistical Society A series, 179, pp. 763-791

Van den Brakel, J.A. and S. Krieg, (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. Survey Methodology, 41, pp. 267-296.

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

Van den Brakel, J.A. and J. Roels, (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. Annals of Applied Statistics, 4, pp. 1105-1138.

Vavreck, L. and D. Rivers (2008). The 2006 cooperative congressional election study. Journal of Elections, Public Opinion and Parties, 18, pp. 355-366.

Vosen, M. and T. Schmidt (2011). Forecasting private consumption: Survey-based indicators versus Google trends. Journal of Forecasting, 30, pp. 565-578.

You, Y. (2008). An integrated modelling approach to unemployment rate estimation for sub-provincial ares of Canada. Survey Methodology, 34, pp. 19-27.

Tanton, R. and K. Edwards (eds). (2013). Spatial Microsimulation: A Reference Guide for Users, Dordrecht: Springer.

Terhanian, G. and J. Bremer (2012). A smarter way to select respondents for surveys. International Journal of Marketing Research, 54, pp. 751-780.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey, Journal of Official Statistics, 8, pp. 149-166.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/'18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/'14–2017/'18	Crop year, financial year, etc., 2015/'16 to 2017/'18 inclusive

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

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

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