sciendo

Journal of Official Statistics, Vol. 38, No. 2, 2022, pp. 557–598, http://dx.doi.org/10.2478/JOS-2022-0026

Some Thoughts on Official Statistics and its Future (with discussion)

Yves Tillé[1],
Comments by Marc Debusschere[2], *Henri Luomaranta*[3], *Martin Axelson*[4],
Eva Elvers[5], *Anders Holmberg*[6], *and Richard Valliant*[7]

In this article, we share some reflections on the state of statistical science and its evolution in the production systems of official statistics. We first try to make a synthesis of the evolution of statistical thinking. We then examine the evolution of practices in official statistics, which had to face very early on a diversification of sou rces: first with the use of censuses, then sample surveys and finally administrative files. At each stage, a profound revision of methods was necessary. We show that since the middle of the 20th century, one of the major challenges of statistics has been to produce estimates from a variety of sources. To do this, a large number of methods have been proposed which are based on very different f oundations. The term "big data" encompasses a set of sources and new statistical methods. We first examine the potential of valorization of big data in official statistics. Some applications such as image analysis for agricultural prediction are very old and will be further developed. However, we report our skepticism towards web-scrapping methods. Then we examine the use of new deep learning methods. With access to more and more sources, the great challenge will remain the valorization and harmonization of these sources.

Key words: Deduction; foundations; induction; Lasso; *p*-value; registers; sampling; statistical learning.

1. Introduction

Official statistics is a somewhat special field of statistics. The methods used there have been developed to deal with particular problems. In official statistics, the main concern is not decision-making but the quality of the proposed estimates. In this article, we try to glimpse the future of statistical methodology. Currently statistics is confronted with many

[1] University of Neuchâtel, Institut de Statistique, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland, Email: yves.tille@unine.ch
[2] Statistics Belgium, Koning Albert II laan 16, B-1000 Brussels, Belgium. Email: marc.debusschere@economie.fgov.be
[3] Statistics Finland. Työpajankatu 13, 00580 Helsinki, Finland. Email: henri.luomaranta@stat.fi
[4] Statistics Sweden, Klostergatan 23, SE-701 89, Örebro, Sweden. Email: martin.axelson@scb.se
[5] Statistics Sweden, Solna strandväg 86, SE-171 54, Solna, Sweden. Email: eva.elvers@scb.se
[6] Australian Bureau of Statistics, Methodology Division, Locked bag 10, 2617, Belconnen, Australia. Email: anders.holmberg@abs.gov.au
[7] University of Michigan, Institute for Social Research, 4620 North Park Avenue Apt 1406W Chevy Chase, Maryland, 20815, U.S.A. Email: valliant@umich.edu
Acknowledgments: The author, Yves Tillé would like to thank four anonymous referees for their constructive comments and two editors-in-chief for their support. Their feedback has enabled the author to improve this article considerably.
The discussant Richard Valliant, is grateful for comments and ideas from Michael L. Cohen, Frederick Conrad, Michael Elliott, John Eltinge, Barry Graubard, Steven Heeringa, Steven Hipple, Roderick Little, and Tommy Wright.

© Statistics Sweden

epistemological questions, the most emblematic of which is the crisis in the *p*-value. We describe the specificities of official statistics by tracing the his tory of sources and the his tory of controversies around methodology. Then, we analyze the impact of new data sources and new statistical methods. An evolution of the methodology will be necessary and must be supported by quality fundamental research.

2. The Statistical Science

Statistics is a science that aims to study a reality through the processing, analysis, modeling and interpretation of data. Traditionally, there are several statistical approaches.

1. *Descriptive* or *exploratory* statistics, which consists of presenting data in a more condensed way by means of tables and graphs. The aim consists of reducing the complexity of the data by means of dimensionality reduction or cluster analysis. Descriptive statistics is above all linked to an interpretive and exploratory process. John Wilder Tukey was a strong supporter of descriptive statistics. His book *Exploratory Data Analysis* (Tukey 1977) remains a reference of the approach. Similarly, in France, the school of data analysis initiated by Jean-Paul Benzécri promoted descriptive and exploratory statistics (Benzécri 1973a,b; Bastin et al. 1980).
2. *Analytical* or *inferential* statistics which aims to deduce properties on a population from data. This population can be notional and be, for example, a probability distribution or a model. Inferential statistics rely entirely on probability calculations. It includes the theory of statistical hypothesis testing and decision theory. The main founders of this theory are Karl Pearson, William Sealy Gosset, Ronald Fisher and Jerzy Neyman.
3. *Modeling* consists in describing reality by a general model described by one or more equations. A model is necessarily a simpler approximation than reality, as Box and Draper (2007) wrote: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful". Models can be used either to describe relationships between variables or to make predictions.

3. Interpretation of the Statistical Approaches

One usually distinguishes two scientific approaches: The *inductive approach* is based on observations and data and aims to build a general theory. The *deductive approach* (or hypothetico-deductive) is based on a theory. It aims to deduce particular results from a general theory in order to check whether these particular results can be confirmed by observations.

Exploratory statistics can be associated with an inductive approach. In this approach, we start from the data to try to empirically find a global explanation. Inferential statistics and more particularly the theory of statistical hypothesis testing can be seen as a hypothetico-deductive approach. One formulates a hypothesis and then we decide on that hypothesis by confronting it with data. The two approaches are qualified as complementary. The exploratory analysis allows hypotheses to be formulated. Then the inferential statistic can possibly confirm them. A return to exploratory analysis finally makes it possible to

formulate new hypotheses and so on. This is referred to as the induction-deduction cycle that would be the engine of the production of knowledge. Often this cycle has been completely perverted, for example, by using the same data to formulate and to test hypotheses which makes the application of test theory completely lapsed. This approach is ironically called HARKing by Norbert L. Kerr (1998) whose acronym comes from *Hypothesizing After the Results are Known*.

However, this vision of the induction/deduction opposition is far from being unanimously shared. The role of induction has been the subject of controversy between Jerzy Neyman and Ronald Fisher. Fisher (1935) promoted the inductive reasoning. Jerzy Neyman (1957) had a more decision-making approach in the hypothesis testing theory but then he was reluctant to extend this to inductive reasoning. Inferential statistics can be considered as an inductive reasoning because it aims to extrapolate the results of a sample to a population or to a model. It is therefore a generalization process that can be seen as fundamentally inductive (see among others Lehmann 1993; Capel et al. 1996).

Costantini and Galavotti (1986) assert that estimation methods, such as the maximum likelihood method, correspond to an inductive approach and theory of hypothesis testing typically corresponds to a deductive approach. Other interpretations exist. For example, Andrew Gelman (2011) discusses a philosophy of statistics that sees frequentist statistics as deductive and Bayesian statistics as inductive.

Faced with these different interpretations, we can only subscribe to the assertion of Capel et al. (1996) who wrote: "In any event, a real understanding of the role of inductive reasoning in the humanities that is common to the statistician, the researcher and the practitioner clearly does not exist, and in particular, as we have seen, with regard to the use of hypothesis tests" (translated from French).

Moreover, the distinction between induction and deduction is sometimes considered outdated. Karl Popper (2005) rejects the description of the scientific process as an induction/deduction cycle. He outright rejects the interest of the inductive approach in science: "Yet even supposing this were the case – for after all, 'the whole of science' might err – I should still contend that a principle of induction is superfluous, and that it must lead to logical inconsistencies." Karl Popper maintains that the scientific method consists in formulating scientific propositions that must be able to be falsified by an experiment. Indeed, no theory can be proven by an experiment. The fact that a model is compatible with data never proves that the model is true. Indeed, another model could also be compatible with the same data set.

The induction/deduction cycle also completely loses its meaning in certain statistical applications. In official statistics, this distinction can hardly be applied. Indeed, the objective often consists in estimating certain characteristics of a population using scattered sources (surveys, administrative files, censuses). The objective is therefore not to establish a scientific theory but simply to overcome the impossibility of obtaining a complete and correct measurement on all the units of the population of interest.

This induction/deduction cycle also loses its meaning with the advent of new methods known as *statistical learning* (support vector machine, neural network, nearest neighbor, sparse methods, random forests) which make it possible to predict without actually modeling. One could see these methods as typically inductive because they do not imply any a priori formalization of reality (Harman and Kulkarni 2012). The word *statistical*

learning is however a bit misleading because these methods do not really lead to a theorization that would be an automatic modeling of data. The so-called induction/deduction cycle is somehow bypassed. For example, if we do not know the income of an individual in a database, we can assign the income of the statistical unit that most closely resembles her/him. This is a forecast by the nearest neighbor method. Can we however say that this approach consists in modeling the income? At most, we can discuss what it means to "resemble", but we are often forced to use only the variables available in the files to define a distance between the units. However, this approach can be very effective. It is not clear how one could say that this approach would be inductive or deductive. For the statistician, the main question concerning this type of method is to evaluate and estimate its precision. Neural network methods also do not allow us to finely understand the relationship between variables. However, the family of Classification And Regression Tree (CART) allows for post-interpretation. With random forests, it is also possible to have an idea of the importance of the variables in the predictions.

The new methods of *statistical learning* therefore shake up a basic principle of the scientific process that would like all knowledge to be able to be transmitted through discourse. While forecasts by *neural network*, *nearest neighbors* or *random forest* can work very well to make a prediction, they do not allow us to establish a general theory or a principle that would allow us to understand the relationships between variables.

4. Crisis of the *p*-value

In statistical hypothesis testing, the *p*-value is the probability that under a hypothesis (conventionally called the null hypothesis), we obtain the same value or an even more extreme value than that obtained with observed data. If the *p*-value is low, then we reject this hypothesis. The error of type 1 is defined as the probability of rejecting the null hypothesis given that it is true. If we perform the test with an error of type 1 of 5% for example, we reject the hypothesis if the *p*-value is less than 5%.

The *p*-value has become a decisive argument in many sciences: humanities, economics, finance or biology. However, in scientific journals of statistics, one may be surprised at the very small number of *p*-values used in published articles. The *p*-value is often misinterpreted as the probability that the null hypothesis is true, which is obviously not the case.

In many publications several *p*-values (sometimes tens) appear without any reflection on the probability of having at least one *p*-value less than 5% if all the null hypotheses were true. Very often, researchers perform a large number of hypothesis tests and only publish results for *p*-values below 5%. This approach is similar to HARKing. Another methodological error consists in identifying a model by choosing the variables having *p*-values less than 5% among a very large number of variables. This procedure inevitably leads to over-specification of the model. However, new solutions have been developed to select variables using for example the Lasso method (Tibshirani 1996, 2011).

These uses of *p*-values were denounced in a provocative article: *Why Most Published Research Findings Are False* (Ioannidis 2005). The subject has become controversial to such an extent that one can say that there is a crisis in statistics (Gelman and Loken 2014; Fraser and Reid 2016) and that the *American Statistical Association* felt compelled to issue a statement on the *p*-value (Wasserstein and Lazar 2016).

5. What About Official Statistics?

The Swiss Official Statistics Charter is available on the website of the Ethics Council for Official Statistics (Conseil d'ethique de la statistique 2012). In this charter, the mission of official statistics is defined: "The mission of official statistics is to meet the needs for statistical information of general interest of society as well as those relating to the conduct of public policies" (translated from French).

The mission therefore does not consist of interpreting, modeling, establishing knowledge, or deciding. The induction-deduction cycle for the production of knowledge is thus not an appropriate framework in official statistics. In fact, in official statistics, the approach is neither exploratory nor decisional. The statistical methods developed are specific, because it is not a question of carrying out scientific research to establish new knowledge, but of meeting information needs by providing reliable, lasting and high-quality statistics.

There exists however a cycle between civil and political society and official statistics in order to determine the needs. Official statistics must ensure continuity but also renew their statistical production by opening up to new themes such as gender inequalities or environmental concerns. The statistical methods used in official statistics are therefore neither exploratory nor decision-making. Official statisticians have focused on quality. An exemplary document is *The Statistics Canada's Quality Assurance Framework* whose basic principles are relevance, accuracy, timeliness, accessibility, coherence, interpretability (Statistics Canada 2017).

6. Data Sources in Official Statistics and Their Integration

Official statistics have therefore developed somewhat specific methods to meet their objectives. The advantage is that official statistics are immune to the p-value crisis because its use is relatively limited. The history of official statistics methods can be summed up in a few eras (see among others Hansen and Madow 1974; Kruskal and Mosteller 1980; Hansen 1987; Bellhouse 1988; Bethlehem 2009; Tillé 2020). First, there is the era of censuses that covers the entire 19th century. During this period, only comprehensive data compilation was considered scientific. This doctrine is clearly stated by the statistician Adolphe Quetelet (1846). The rupture is initiated by the director of the Norwegian Institute of Statistics: Anders Nicolai Kiær (1896, 1899, 1903, 1905) who proposed to use partial data and therefore a sample. After a long controversy, the idea of using samples was finally accepted by the International Statistical Institute (Jensen 1926). Ken Brewer (2013) interprets this debate as the first controversy in survey sampling.

This opens the era of sampling, which from the start has been the subject of assiduous and fruitful research, (notably on the part of Jerzy Neyman 1934, 1938, 1952). He shows in particular that it is necessary to over-represent in the sample the categories where the dispersion is larger. This result renders the still too often used concept of representativeness obsolete. With the development of information technologies, then came the era of registers and administrative files in the 1970s. Administrative data are often presented as the new source of data. However, it should not be forgotten that some countries like Finland have had a population register for more than 50 years. The technical means for creating a register have been around for a very long time. Obstacles to the use of

registers for statistical purposes are primarily organizational, political and legal. They are often the result of data protection rules or communication difficulties between different administrations. For a long time, the problems have obviously not been technical.

The development of new data sources has not put an end to previous practices. Many National Institutes of Statistics still carry out censuses and almost all of them carry out sample surveys. The integration of administrative files into official statistics is far from obvious. The registers depend on the administrative specificities of the countries, which does not facilitate statistical harmonization. Administrative files often contain many errors, as evidenced by the Serafe crisis in Switzerland.

Serafe is a private company responsible for collecting audiovisual license fees in Switzerland since 2019. During the first invoices sent, a large number of errors were noted. Serafe had sent the invoices on the basis of the "control of inhabitants" registers updated by the municipalities. Many errors appeared to be related to household composition in multi-unit buildings or to obsolescence of data. These population control registers are also used for administrative and statistical purposes, but the impact of these errors is relatively small for these applications. In administrative files, it is often observed that only the variables which are necessary for immediate administrative operation are correctly updated.

We cannot therefore expect that all official statistics will be produced from the same type of source. Each type of source contains specific errors, whether they are data produced by statistical offices or not. Problems to be addressed can be sampling errors, errors due to nonresponse, under-coverage, over-coverage, duplicates, measurement errors, errors due to fraud. Official statistics are a perpetual struggle against these errors.

In each source, one can find a particular reliability. One type of reliability is the correct identification of statistical units, which can often be obtained through an administrative file or register. However, this register may contain poor quality or obsolete variables. A sample survey, even one with nonresponse, may contain more reliable measures for the variables of interest. The problem then comes down to combining what is most reliable in the different sources. The crucial issue in official statistics is thus the integration of data from different sources in order to best enhance the reliability of each source.

From the beginning, research in official statistics has been concerned with the problem of data integration. Almost a century ago, the controversy between, on the one hand, Corrado Gini and Luigi Galvani and, on the other hand, Jerzy Neyman (Gini and Galvani 1929) revolves around balanced sampling and random sampling. Gini and Galvani had selected a sample of 29 districts (*circondari*) out of 214 so as to return the same means as those of the census for several known variables. It is therefore a question of using a source (the census) to improve the collection of a sample. Neyman criticizes this way of doing things, because the selection of the sample is not random and therefore it is not possible to make an inference. We obviously now know that it is possible to select a sample that is both balanced and random (Deville and Tillé 2004). Ken Brewer (2013) interprets this discussion as the second controversy in survey sampling.

Another issue is the adjustment of survey data to census data. On this subject, the article by W. Edwards Deming and Frederick F. Stephan of 1940 is considered as a founding text in official statistics (Deming and Stephan 1940). The problem dealt with is the adjustment of a table obtained by sampling on marginal totals known by a census. The integration of

different sources is therefore present from the start of the use of sampling methods. For the record, the article is however mathematically false since it maintains that the *raking ratio* method is obtained by minimizing the chi-square distance under the constraints given by known marginal totals. This is obviously not true since the Kullback-Leibler divergence must be minimized. The fact remains that this article and all those that follow seek to optimize the integration of survey and census data.

The article of Deville and Särndal (1992) which defines the general method of calibration is the outcome of this research. These authors provide a methodology for calibrating survey data on census data while freeing itself from modeling. The originality is that the processing is carried out in order to obtain a weighting system applicable to any variable, which makes its application extremely practical. Deville and Särndal also provide a method for measuring the accuracy of the estimates obtained. The method has become essential and typical of official statistics. It is now applied by all survey statisticians (see also Särndal 2007; Devaud and Tillé 2019a,b). The use of calibration has become widespread, especially since calibration is currently used not only to correct the sampling error but also to correct the nonresponse and measurement error (see among others Dupont 1994; Fuller et al. 1994; Lundström and Särndal 1999; Deville 2000; Särndal and Lundström 2005; Kott 2006; Brick 2013; Valliant et al. 2013; Haziza and Lesage 2016; Devaud and Tillé 2019a).

A developing problem is the integration not only of two sources but of a multitude of distinct sources. The calibration methods can be generalized to harmonize several sources. These sources can be two samples or several samples and a census (see among others Guandalini and Tillé 2017). Yang and Kim (2020) provide an overview of modern methods for integrating data from different sources.

The use of sources may change quite rapidly. Many European countries have abandoned large population censuses. Statistics based on these censuses are now produced by using registers and new surveys complete these registers. National Institutes of Statistics still carry out a lot of sample surveys. Sampling methods are increasingly used to check or improve the quality of censuses or registers. Thus, sampling may become more of a quality control tool than a direct method of producing data.

The evolution of practices does not go without raising new questions. Most emblematic has been the resignation of Martha Farnsworth Riche from the head of the United States Census Bureau. According to the Wikipedia contributors (2020) site: "Although she cited only personal reasons in her resignation, it was seen as a sign that Congressional Republicans were winning in their fight to prevent the Census Bureau from using sampling techniques to correct for persistent undercounting of minorities and other under-represented groups." This political intervention in the methodology is obviously extremely worrying and goes against a principle of methodological independence of the National Institutes of Statistics.

7. Modeling in Official Statistics

The debate on the place of the model in survey sampling theory is interpreted by Ken Brewer (2013) as the third controversy in survey sampling. The idea of introducing a model to exploit the auxiliary information resulting from a census in a survey was initially

proposed by this same Ken Brewer (1963). However, this idea was mainly developed by Royall (1970, 1971, 1976) (see also Valliant et al. 2000; Chambers and Clark 2012). The model-based approach involves building a model that is estimated using the sample, and then predicting population data that is not in the sample.

The model approach is opposed to the so-called design-based approach which consists in weighting the sample units by the inverse of their probabilities of being selected. The initial survey weights are the inverse of the inclusion probabilities. These weights are then slightly modified by means of the calibration technique of Deville and Särndal in order to return exactly the totals of the variables known by a register or by a census.

Official statisticians are inherently reluctant to model. The reason is this: Modeling is somehow expressing an opinion on observations. Modeling can therefore be interpreted as contrary to the principle of impartiality of official statistics. However, model-based and design-based approaches are not necessarily contradictory. Särndal et al. (1992) advocate a model-assisted and nonetheless valid design-based approach. It is therefore possible to construct the concept of double robustness in the sense of an estimate that would be valid either when the model is correct or when the inclusion and response probabilities are correctly identified.

However, there are issues that cannot be solved without some modeling. Research concerning estimation for small areas has been very active during the last decades. These methods consist in producing estimates at very low levels (districts, municipalities) from a sample survey and a register or census. Once again, the problem is to make the best use of the two sources of information. The sample contains the variable of interest. The register contains the list of all units of the population and auxiliary variables. Composite estimators are often used, which are mixtures of direct estimates computed from the sample and estimates obtained from a model linking the auxiliary variables. These composite estimators can be obtained by means of mixed models for which the domains are random effects (see the very complete book of Rao and Molina 2015).

The National Institutes of Statistics, however, have been relatively cautious and have rarely published estimates obtained by these methods. One of the reasons is that there may be peculiarities well known at the level of local communities that would not be taken into account by a model. For example, Molina and Strzalkowska-Kominiak (2020) proposed labor force estimates in Swiss districts using the Swiss Structural Survey. A Swiss district bordering on Liechtenstein was peculiar because a large number of Swiss border workers work in Liechtenstein. This district has therefore undergone special treatment. However, if it is not possible to identify the singularities before modelling, the estimates may turn out to be very far from reality.

Another area where it is difficult to work without modeling is nonresponse. There are two ways to do this. Either we can predict the missing values (imputation), or we can estimate the probability that these values are missing in order to weight the responding observations. The models can be very simple. One can impute by a closest neighbor. One can also impute by a simple ratio or by a prediction by regression. One can also predict by taking at random an individual belonging to a small homogeneous stratum. In order to avoid a bad specification of the model, one often seeks to treat the nonresponse in a doubly robust way. Two models are used: the first one allows us to predict the missing value and the second one allows us to predict the probability of being missing. Double robustness

means that the estimation is then approximately unbiased if at least one of the two models is well specified (see on this subject Kang and Schafer 2007; Han and Wang 2013; Kim and Haziza 2014; Boistard et al. 2016; Chen and Haziza 2017).

8. Big Data and Official Statistics

Tim Harford (2014) gives a strong opinion on big data: "As with so many buzzwords, "big data" is a vague term, often thrown around by people with something to sell. Some emphasize the sheer scale of the data sets that now exist – the Large Hadron Collider's computers, for example, store 15 petabytes a year of data, equivalent to about 15,000 years' worth of your favorite music." Indeed, we can often attend presentations of "experts" announcing the era of "big data" whose slides only contain lists of words starting with "V", numbers in yottabytes and potatoes linked by arrows without ever giving a real and concrete application. The slides are embellished with pontificating quotes and cartoons. Making lists of words beginning with a letter is the opposite of a serious scientific approach.

The fascination with big data can be irritating in several ways. Talking about "big data" in official statistics is not appropriate. Large administrative files such as population or business registers cannot be qualified as big data. They can contain millions of records but are processable over any desktop computer. These large files have been around in some countries for over fifty years and these countries have always found the computer means to manage them. They are not big data.

I am convinced that the use of web scraping will remain very limited in producing official statistics. For example, inflation cannot be calculated exclusively by automatically fetching prices from the Internet. Ten Bosch et al. (2018) give a list of experiences and lay the foundations of a methodology for web scraping in official statistics. They show that, in most cases, projects rarely end up going into production. The main reasons are that official statistics must be sustainable and that the Internet is not stable over time. Second, because official statistics must have a documented methodology and sources. Because it is not enough to look at a price on the Internet, it must also be possible to verify them. In addition, most National Statistical Institutes do not need to consult prices on the Internet because they use the price files of the main distributors directly. Often the data collected by scraping the web can be more simply obtained by another, more stable means.

The number of applications of web scrapping actually put into practice for the production of official statistics is therefore very limited. One can mention some applications in price indices for products sold almost exclusively via the Internet. For instance the U.S. Bureau of Labor Statistics use the web to track the price changes of airline fares (U.S. Bureau of Labor Statistics 2021). However watching prices on the web cannot be really called "web scraping". Social network analysis can however be interesting for social science analysis, as noted by Connelly et al. (2016). Nevertheless, official statistics needs a stable and reproducible production system that allows temporal comparisons, which is difficult to obtain from the web.

An anonymous referee drew our attention to the billion price project that plans to calculate a price index using web scraping (see Cavallo and Rigobon 2016). This initiative does not come from the world of official statistics but from two professors of MIT Sloan

and Harvard Business School. This approach can be interesting when official statistics are deficient, as was the case in Argentina, but it has not been integrated into official statistics until now. I am also convinced that the methodology is more important than the sample size and that observing more than a billion prices is not necessarily a guarantee of the quality of the estimates.

The huge data streams of social networks are not valuable either. They belong to the giants of the web. This data has immense commercial value for performing advertising profiling. It would also be extremely dangerous to become dependent on these giants. Above all, these data are not reliable in terms of identifying statistical units and potentially usable variables.

However, there are areas where the data is really massive, such as image analysis for territorial statistics. In this case, the data can be produced (aerial or satellite photos, cartography) by the country which uses them. For instance, the Joint Research Centre of the European Union in Ispra has a long tradition of estimation of crop production using satellite images (see for instance Gallego et al. 1993; Taylor et al. 1997; Gallego 2004; Carfagna and Gallego 2005; Kussul et al. 2016). However, this practice was developed before the popularization of the word "big data". Indeed, image analysis is an area of research that emerged with the early days of computer science. Significant progress has been made, notably through the use of neural networks, which allows us to foresee a major development of applications in official statistics.

9. Statistical Learning in Official Statistics

We often talk about new statistical methods. Most of these methods are actually not so new and were all developed in the 20th century. Several methods (random forests, support vector machine, neural network, nearest neighbors) make it possible to make predictions without having to think about the relationships between the dependent variables and the variable of interest. Can we follow the famous sentence of Deng Xiaoping who said "it doesn't matter whether a cat is black or white, if it catches mice it is a good cat"? Can we use in official statistics methods that make it possible to predict without understanding? Indeed, the Statistical Charter of the Swiss Ethics Council for Official Statistics specifies that "Statistical information is documented in order to facilitate its understanding and allow its correct use" (translated from French). Is it compatible with statistical learning methods? Is it sufficient to specify the method used?

We believe that *statistical learning* methods can be used, but with certain precautions. The generalized regression estimator (see Särndal et al. 1992) allows a prediction to be incorporated into the estimate while remaining approximately unbiased under the sampling design. This estimator thus avoids slippage due to poor specification of the model. So, official statisticians use fairly simple models that can be integrated in a method that remains valid under the design.

One might think that public statistics are not affected by the p-value crisis because they mainly produce descriptive statistics. However, we have seen that underneath this production, there is a very technical machinery. Surveys are imputed, weighted by nonresponse models, and then calibrated. Thus, it is necessary to choose calibration variables, design nonresponse models and imputation models. These operations require

modeling and this modeling is often based on hypothesis testing, that is, p-values. The use of parametric statistics is therefore a hidden task. In this field, the use of new deep learning methods can be very promising.

Official statisticians have also long been familiar with light non-parametric methods for dealing with nonresponse, such as imputation by the nearest neighbor or by an individual selected from a homogeneous stratum. Methods like *support vector machine* or *random forests* ultimately consist in splitting the space of explanatory variables to define one or more neighborhoods of the unit in order to make a forecast.

The generalized regression estimator has enabled Breidt and Opsomer (2000), for example, to construct estimates assisted by a model whose predictions are made by the local polynomial method. By following this same approach, all forecasting methods can be used without introducing disproportionate risks into the estimates. Thus, the recent work of Beaumont and Bocci (2008), Goga and Shehzad (2014), Breidt and Opsomer (2017), McConville et al. (2017), Mayor- Gallego et al. (2019), Chen et al. (2019), Tan (2020), and Dagdoug et al. (2020a,b) which integrates the methods of *shrinkage* and *statistical learning* in calibration and the treatment of nonresponse, probably shows the way forward on future research in official statistics.

The so-called *shrinkage* methods like the Lasso make it possible to choose variables in a modeling or a calibration. Survey statisticians often tend to overcalibrate surveys because more and more variables are available in the registers. However, calibration is above all an estimation technique which aims to reduce the variances of the estimates. In general, both calibration and modeling must be subject to a principle of parsimony. It consists of finding the most efficient model as possible while being as simple as possible. The Lasso method makes it possible to reduce the number of calibration variables. It has already been applied to survey data by McConville et al. (2017).

10. Conclusions

The term big data should maybe be dropped. When one asks what "big data" is, one often gets an answer worthy of Borges' classification of animals in his new "The Analytical Language of John Wilkins" (Borges 2012). The term "big data" can include both certain types of data or a set of methods. For example, image analysis, social network stream analysis, large administrative files, smart meter data, neural network methods, random forest methods, support- vector machine methods, sparse statistical methods (such as Lasso). The word "big data" thus gathers a heterogeneous set of data and methods in which it is necessary to make a distinction. It is important to keep things in perspective. Image analysis is a very old problem which has been used in public statistics for at least 40 years. The "new" statistical methods were almost all developed in the last century before the big data fad. Obviously all the "new" statistical methods are interesting and are increasingly applied in public statistics sometimes to "classical" problems such as the treatment of nonresponse where the data sets are not necessarily large. The valorization of administrative data is a very important challenge. However, I am particularly skeptical about statistical production directly from the web.

The multiplication of sources in official statistics can be misleading because it does not necessarily imply an improvement in quality (see among others Deville 1997).

As Tim Harford (2014) reminds us, abundance of data is not and never will be synonymous with quality. Administrative records often contain a large number of errors because they were not designed to be used for statistical purposes. These new sources must be combined with all other available sources. The integration of sources is an issue that researchers have tackled from the very beginning of official statistics. These methods must be further developed, because more and more sources will be available. It would be useful to have a general theoretical framework for data integration.

New statistical methods must also be integrated or combined with existing methods. We believe that there will certainly not be a clean slate of methodology but an evolution towards practices that are more varied and more focused on the types of data to be processed. More than ever, official statistics must promote and develop methodological research around its issues.

An anonymous arbitrator rightly pointed out to us that an important issue is to maintain and increase the independence of national statistical institutes. A number of autocrats have questioned the legitimacy of official statistical offices, which can bring their output into disrepute. This is obviously a very important issue that we cannot unfortunately develop here. We are convinced that international institutions have a role to play in regulating these practices, but this is well beyond the scope of this article.

11. References

Bastin, C., J.P. Benzécri, C. Bougarit, and P. Cazes. 1980. *Pratique de l'Analyse des Donnees*. Paris: Dunod.

Beaumont, J.-F., and C. Bocci. 2008. "Another look at ridge calibration." *Metron* 66(1): 5–20. Available at: https://www.dss.uniroma1.it/RePec/mtn/articoli/2008-1-1.pdf (accessed February 2022).

Bellhouse, D.R. 1988. "A brief history of random sampling methods." In *Handbook of Statistics*, edited by P.R. Krishnaiah and C.R. Rao, 6: 1–14, New York, Amsterdam. Elsevier/North-Holland. DOI: https://doi.org/10.1016/S0169-7161(88)06003-1.

Benzécri, J.-P. 1973a. *L'analyse des données: tome 1: La taxinomie*. L'analyse des données. Paris: Bordas.

Benzécri, J.-P. 1973b. *L'analyse des données: tome 2: L'analyse des correspondances*. L'analyse des données. Paris: Bordas.

Bethlehem, J.G. 2009. *The rise of survey sampling*. The Hague: Statistics Netherlands. DOI: https://pure.uva.nl/ws/files/24636645/74286_312955.pdf (accessed February 2022).

Boistard, H., G. Chauvet, and D. Haziza. 2016. "Doubly robust inference for the distribution function in the presence of missing survey data." *Scandinavian Journal of Statistics* 43(3): 683–699. DOI: https://doi.org/10.1111/sjos.12198.

Borges, J. 2012. *Inquisiciones — Otras inquisiciones*. Penguin Random House Grupo Editorial España.

Box, G.E.P., and N.R. Draper. 2007. *Response Surfaces, Mixtures, and Ridge Analyses* 649. Hoboken: John Wiley & Sons. Available at: https://www.wiley.com/enus/Response+Surfaces%2C+Mixtures%2C+and+Ridge+Analyses%2C+2nd+Edition-p-9780470053577 (accessed February 2022).

Breidt, F.J., and J.D. Opsomer. 2000. "Local polynomial regresssion estimators in survey sampling." *The Annals of Statistics* 28(4): 1026–1053. DOI: https://doi.org/10.1214/aos/1015956706.

Breidt, F.J., and J.D. Opsomer. 2017. "Model-assisted survey estimation with modern prediction techniques." *Statistical Science* 32(2): 190–205. DOI: https://doi.org/10.1214/16-STS589.

Brewer, K.R.W. 1963. "A model of systematic sampling with unequal probabilities." *Australian Journal of Statistics* 5: 5–13. DOI: https://doi.org/10.1111/j.1467-842X.1963.tb00132.x.

Brewer, K.R.W. 2013. "Three controversies in the history of survey sampling." *Survey Methodology* 39(2): 249–262. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11883-eng.pdf (accessed February 2022).

Brick, M.J. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29(3): 329–353. DOI: https://doi.org/10.2478/jos-2013-0026.

Capel, R., Monod, D., and Müller, J.-P. 1996. "Essai sur le rôle des tests d'hypothèse en sciences humaines." *Actualités Pédagogiques* 1: 1–51. Available at: http://moityca.com/pdfs/Essai.pdf (accessed February 2022).

Carfagna, E., and F.J. Gallego. 2005. "Using remote sensing for agricultural statistics." *International statistical review* 73(3): 389–404. DOI: https://doi.org/10.1111/j.1751-5823.2005.tb00155.x.

Cavallo, A., and R. Rigobon. 2016. "The billion prices project: Using online prices for measurement and research." *Journal of Economic Perspectives* 30(2): 151–178. DOI: https://doi.org/10.1257/jep.30.2.151.

Chambers, R.L., and R.G. Clark. 2012. *An Introduction to Model-Based Survey Sampling with Applications*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198566625.001.0001.

Chen, J.K.T., R.L. Valliant, and M.R. Elliott. 2019. "Calibrating nonprobability surveys to estimated control totals using LASSO, with an application to political polling." *Journal of the Royal Statistical Society* 68(3): 657–681. DOI: https://doi.org/10.1111/rssc.12327.

Chen, S., and D. Haziza. 2017. "Multiply robust imputation procedures for the treatment of item nonresponse in surveys." *Biometrika* 104(2): 439–453. DOI: https://doi.org/10.1093/biomet/asx007.

Connelly, R., C.J. Playford, V. Gayle, and C. Dibben. 2016. "The role of administrative data in the big data revolution in social science research." *Social science research* 59: 1–12. DOI: https://doi.org/10.1016/j.ssresearch.2016.04.015.

Conseil d'éthique de la statistique. 2012. *Charte de la statistique publique de la Suisse*. Office fédéral de la statistique. Neuchâtel. Available at: https://www.conseilethique-stat.ch/fr/assets/File/charte/940-1400.pdf (accessed February 2022).

Costantini, D., and M.C. Galavotti. 1986. "Induction and deduction in statistical analysis." *Erkenntnis* 24: 73–94. DOI: https://doi.org/10.1007/BF00183203.

Dagdoug, M., Goga, C., and D. Haziza. 2020a. *Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison.* Available at: https://arxiv.org/abs/2007.06298 (accessed February 2022).

Dagdoug, M., Goga, C., and D. Haziza. 2020b. *Model-assisted estimation through random forests in finite population sampling*. Available at: https://arxiv.org/abs/2002.09736 (accessed February 2022).

Deming, W.E., and F.F. Stephan. 1940. "On a least square adjustment of sampled frequency table when the expected marginal totals are known." *Annals of Mathematical Statistics* 11: 427–444. DOI: https://doi.org/10.1214/aoms/1177731829.

Devaud, D., and Y. Tillé. 2019a. "Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem." *TEST* 4: 1033–1065. DOI: https://doi.org/10.1007/s11749-019-00681-3.

Devaud, D., and Y. Tillé. 2019b. "Rejoinder on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem." *TEST* 28: 1087–1091. DOI: https://doi.org/10.1007/s11749-019-00681-3.

Deville, J.-C. 1979. *Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement?* Document de travail – Institut national de la statistique et des études économiques. Insee.

Deville, J.-C. 2000. "Generalized calibration and application to weighting for non-response." In *Compstat – Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands* 65–76. New York: Springer. DOI: https://doi.org/10.1007/978-3-642-57678-2_6.

Deville, J.-C., and C.-E. Särndal. 1992. "Calibration estimators in survey sampling." *Journal of the American Statistical Association* 87: 376–382. DOI: https://doi.org/10.1080/01621459.1992.10475217.

Deville, J.-C., and Y. Tillé. 2004. "Efficient balanced sampling: The cube method." *Biometrika* 91: 893–912. DOI: https://doi.org/10.1093/biomet/91.4.893.

Dupont, F. 1994. "Calibration used as a nonresponse adjustment, studies in classification, data analysis, and knowledge organization." In *New Approaches in Classification and Data Analysis*, edited by E. Diday: 539–548. Springer-Verlag.

Fisher, R.A. 1935. "The logic of inductive inference." *Journal of the royal statistical society* 98(1): 39–82. Available at https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2397-2335.1935.tb04208.x (accessed February 2022).

Fraser, D.A.S., and N. Reid. 2016. "Crisis in science? or crisis in statistics! mixed messages in statistics with impact on science." *Journal of Statistical Research* 48(1): 1–9. Available at: http://www.utstat.utoronto.ca/reid/research/273-v8-nr.pdf (accessed February 2022).

Fuller, W.A., M.M. Loughin, and H.D. Baker. 1994. "Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide food consumption survey." *Survey Methodology* 20: 75–85. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994001/article/14429-eng.pdf?st = uKDnj6Zz (accessed February 2022).

Gallego, F.J. 2004. "Remote sensing and land cover area estimation." *International Journal of Remote Sensing* 25(15): 3019–3047. DOI: https://doi.org/10.1080/01431160310001619607.

Gallego, F.J., J. Delincé, and C. Rueda. 1993. "Crop area estimates through remote sensing: stability of the regression correction." *International Journal of Remote Sensing* 14(18): 3433–3445. DOI: https://doi.org/10.1080/01431169308904456.

Gelman, A. 2011. "Induction and deduction in Bayesian data analysis." *Rationality, Markets and Morals* 2(67–78): 1999. Available at: https://link.springer.com/article/10.1007/BF00183203 (accessed February 2022).

Gelman, A., and E. Loken. 2014. "The statistical crisis in science: data-dependent analysis-a "garden of forking paths"-explains why many statistically significant comparisons don't hold up." *American scientist* 102(6): 460–466. DOI: https://doi.org/10.1511/2014.111.460.

Gini, C., and L. Galvani. 1929. "Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921)." *Annali di Statistica, Series* 4: 1–107.

Goga, C., and M.A. Shehzad. 2014. "A note on partially penalized calibration." *Pakistan Journal of Statistics* 30(4): 429–438.

Guandalini, A., and Y. Tillé. 2017. "Design-based estimators calibrated on estimated totals from multiple surveys." *International Statistical Review* 85: 250–269. DOI: https://doi.org/10.1111/insr.12160.

Han, P., and L. Wang. 2013. "Estimation with missing data: beyond double robustness." *Biometrika* 100(2): 417–430. DOI: https://doi.org/10.1093/biomet/ass087.

Hansen, M.H. 1987. "Some history and reminiscences on survey sampling." *Statistical Science* 2: 180–190. DOI: https://doi.org/10.1214/ss/1177013352.

Hansen, M.H., and W.G. Madow. 1974. "Some important events in the historical development of sample survey." In *On the History of Statistics and Probability*, edited by D.B. Owen: 75–102. New York: Marcel Dekker.

Harford, T. 2014. "Big data: A big mistake" *Significance* 11(5): 14–19. DOI: https://doi.org/10.1111/j.1740-9713.2014.00778.x.

Harman, G., and S. Kulkarni. 2012. *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge, Massachusetts: MIT Press. Available at: https://mitpress.mit.edu/books/reliable-reasoning (accessed February 2022).

Haziza, D., and É. Lesage. 2016. "A discussion of weighting procedures for Unit Nonresponse." *Journal of Official Statistics* 32(1): 129–145. DOI: https://doi.org/10.1515/jos-2016-0006.

Ioannidis, J.P. 2005. "Why most published research findings are false." *PLoS medicine* 2(8): e124. DOI: https://doi.org/10.1371/journal.pmed.0020124.

Jensen, A. 1926. "Report on the representative method in statistics." *Bulletin of the International Statistical Institute* 22: 359–380.

Kang, J.D.Y., and J.L. Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical Science* 22(4): 523–539. DOI: https://doi.org/10.1214/07-STS227.

Kerr, N.L. 1998. "Harking: Hypothesizing after the results are known." *Personality and Social Psychology Review* 2(3): 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4.

Kiær, A.N. 1896. "Observations et expériences concernant des dénombrements représentatifs." *Bulletin de l'Institut International de Statistique* 9: 176–183.

Kiær, A.N. 1899. "Sur les méthodes représentatives ou typologiques appliquées à la statistique." *Bulletin de l'Institut International de Statistique* 11: 180–185.

Kiær, A.N. 1903. "Sur les méthodes représentatives ou typologiques." *Bulletin de l'Institut International de Statistique* 13: 66–78.

Kiær, A.N. 1905. "Discours sans intitulé sur la méthode représentative." *Bulletin de l'Institut International de Statistique* 14: 119–134.

Kim, J.K., and D. Haziza. 2014. "Doubly robust inference with missing data in survey sampling." *Statistica Sinica* 24(1): 375–394. DOI: https://doi.org/10.5705/ss.2012.005.

Kott, P.S. 2006. "Using calibration weighting to adjust for nonresponse and coverage errors." *Survey Methodology* 32: 133–142. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9547-eng.pdf?st = 63fu-3Bc (accessed February 2022).

Kruskal, W., and F. Mosteller. 1980. "Representative sampling, IV: The history of the concept in statistics, 1895–1939." *International Statistical Review* 48: 169–195. DOI: https://doi.org/10.2307/140315.

Kussul, N., Lemoine, G., F.J. Gallego, S.V. Skakun, M. Lavreniuk, and A.Y. Shelestov. 2016. "Parcel-based crop classification in ukraine using landsat-8 data and sentinel-1a data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9(6): 2500–2508. DOI: https://doi.org/10.1109/JSTARS.2016.2560141.

Lehmann, E.L. 1993. "The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?" *Journal of the American Statistical Association* 88(424): 1242–1249. DOI: https://doi.org/10.1080/01621459.1993.10476404.

Lundström, S., and C.-E. Särndal. 1999. "Calibration as a Standard Method for Treatment of Nonresponse." *Journal of Official Statistics* 15: 305–327. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/calibration-as-a-standard-method-for-treatment-of-nonresponse.pdf (accessed March 2022).

Mayor-Gallego, J., J. Moreno-Rebollo, and M. Jimenez-Gamero. 2019. "Estimation of the finite population distribution function using a global penalized calibration method." *AStA Advances in Statistical Analysis* 103(1): 1–35. DOI: https://doi.org/10.1007/s10182-018-0321-z.

McConville, K.S., F.J. Breidt, T.C.M. Lee, and G.G. Moisen. 2017. "Model-assisted survey regression estimation with the lasso." *Journal of Survey Statistics and Methodology* 5(2): 131–158. DOI: https://doi.org/10.1093/jssam/smw041.

Molina, I., and E. Strzalkowska-Kominiak. 2020. "Estimation of proportions in small areas: application to the labour force using the swiss census structural survey." *Journal of the Royal Statistical Society* A183(1): 281–310. DOI: https://doi.org/10.1111/rssa.12498.

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: 558–606. DOI: https://doi.org/10.2307/2342192.

Neyman, J. 1938. "Contribution to the theory of sampling human population." *Journal of the American Statistical Association* 33: 101–116. Available at: https://doi.org/10.1080/01621459.1938.10503378.

Neyman, J. 1952. "*Lectures and Conferences on Mathematical Statistics and Probability*." Graduate School, U.S. Department of Agriculture, Washington. DOI: https://doi.org/10.1038/142274a0.

Neyman, J. 1957. "Inductive Behavior" as a basic concept of philosophy of science." *Revue de l'Institut International de Statistique* 25: 7–22. DOI: https://doi.org/10.2307/1401671.

Popper, K. 2005. *The logic of scientific discovery*. London: Routledge.

Quetelet, A. 1846. *Lettres à S.A.R. le Duc régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités appliquées aux sciences morales et politiques*. Bruxelles: M. Hayez.

Rao, J.N.K., and I. Molina. 2015. *Small Area Estimation*. New York: Wiley. Available at: https://www.wiley.com/en-us/Small+Area+Estimation%2C+2nd+Edition-p-9781118735787 (accessed February 2022).

Royall, R.M. 1970. "On finite population sampling theory under certain linear regression models." *Biometrika* 57: 377–387. DOI: https://doi.org/10.1093/biomet/57.2.377.

Royall, R.M. 1971. "Linear regression models in finite population sampling theory." In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott: 259–279. Toronto, Montreal: Holt, Rinehart et Winston.

Royall, R.M. 1976. "The linear least squares prediction approach to two-stage sampling." *Journal of the American Statistical Association* 71: 657–664. DOI: https://doi.org/10.1080/01621459.1976.10481542.

Särndal, C.-E. 2007. "The calibration approach un survey theory and practice." *Survey Methodology* 33: 99–119. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X200700210488 (accessed February 2022).

Särndal, C.-E., and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: Wiley. Available at: https://onlinelibrary.wiley.com/doi/book/10.1002/0470011351 (accessed February 2022).

Särndal, C.-E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer. Available at: https://link.springer.com/book/9780387406206 (accessed February 2022).

Statistics Canada. 2017. *Statistics Canada's Quality Assurance Framework. Documentation of the internet site of Statistics Canada.* Statistics Canada, Ottawa. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-586-X (accessed February 2022).

Tan, Z. 2020. "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data." *Biometrika* 107(1): 137–158. Available at: https://doi.org/10.1093/biomet/asz059.

Taylor, J., Sannier, C., J. Delincé, and F.J. Gallego. 1997. *Regional crop inventories in Europe assisted by remote sensing. Synthesis Report, Office for Publications of the European Commission.* Available at: https://op.europa.eu/en/publication-detail/-/publication/dbec0d28-e918-452d-a70b-6706946ae935 (accessed February 2022).

Ten Bosch, O., D. Windmeijer, A. van Delden, G. van den Heuvel G. 2018. "Web scraping meets survey design: Combining forces." In Big Data Meets Survey Science Conference, October 25–27, Barcelona, Spain. Available at: https://www.europeansurveyresearch.org/ bigsurv18/uploads/73/61/20180820_BigSurv_WebscrapingMeetsSurveyDesign.pdf (accessed February 2022).

Tibshirani, R.J. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society* 58(1): 267–288. Available at: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R.J. 2011. "Regression shrinkage and selection via the lasso: a retrospective." *Journal of the Royal Statistical Society* B73(3): 273–282. DOI: https://doi.org/10.1111/j.1467-9868.2011.00771.x.

Tillé, Y. 2020. *Sampling and Estimation From Finite Populations*. Hoboken: Wiley. Available at: https://www.wiley.com/en-us/Sampling + and + Estimation + from + Finite + Populations-p-9780470682050 (accessed February 2022).

Tukey, J.W. 1977. *Exploratory Data Analysis*, volume 2. Reading, MA: Addison-Wesley.

U.S. Bureau of Labor Statistics. 2021. Consumer price index. Available at: https://www.bls.gov/cpi/factsheets/airline-fares.htm (accessed February 2022).

Valliant, R., Dever, J.A., and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer. Available at: https://link.springer.com/book/10.1007/978-1-4614-6449-5 (accessed February 2022).

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley. Available at: https://www.wiley.com/enus/Finite+Population+Sampling+and+Inference%3A+A+Prediction+Approach-p-9780471293415 (accessed February 2022).

Wasserstein, R.L., and N.A. Lazar. 2016. "The ASA statement on *p*-values: Context, process, and purpose." *The American Statistician* 70(2): 129–133. DOI: https://doi.org/10.1080/00031305.2016.1154108.

Wikipedia contributors. 2020. "Martha Farnsworth Riche – Wikipedia, the free encyclopedia." Available at: https://en.wikipedia.org/wiki/Martha_Farnsworth_Riche (accessed February 2022).

Yang, S., and J.K. Kim. 2020. "Statistical data integration in survey sampling: A review." *Japanese Journal of Statistics and Data Science* 3: 625–650. DOI: https://doi.org/10.1007/s42081-020-00093-w.

Comments by Marc Debusschere

Let's cut to the chase. This article sounds like a discussion about the relative merits of methods, but underlying it is a seemingly semantic discussion on the meaning of 'big data' and more particularly 'webscraping' (essentially saying that 'good' big data/webscraping is NOT big data/webscraping at all...). And underlying that, at an even deeper level, is a 'cultural' issue.

As in many areas of human endeavour, official statistics has its conservatives and its progressives. Overstating somewhat, the conservative statisticians are very wary of new sources and methods, they want to stick to the old and trusted and familiar, and they think the focus should be on ever more refining the current way of doing things, leading to increasingly specialised and intricate methodologies. Consequently, they are extremely critical of the weak points, real or contrived, of alternatives.

Equally overstating, we have in the other corner the progressives, the free thinkers, and statistical revolutionaries. Which one could also call the fad- followers, not into detail or complexity, moving from buzzword to buzzword (of which 'big data' is the almost-latest one, the very latest being 'smart statistics'). Progressives enthusiastically embrace innovation, tend to be less than critical when assessing costs and benefits, and are rather quick in dismissing all that went before to the dustbins of history.

I plead for an intermediate pragmatic attitude. The discussion about what to call big data (or smart statistics) is not terribly important. Let's just label anything of potential use in official statistics which is not a survey or administrative database a 'new data source', to be investigated and exploited if at all possible. Without a priori exclusives or taboos, but also avoiding uncritical acceptance. There is a risk, certainly, that official statistics will invest in sources and methods which in the end turn out to be unproductive. But there is also a huge potential benefit. The current methods based on surveys and administrative data cannot provide the timeliness and level of detail we need, and they are unable to answer many important questions. Some new data sources have already proven they can fill the gap. Official statistics must integrate them in its regular production, and keep on the look-out for additional ones. If official statistics doesn't use them, other data handlers with other agendas will.

© Statistics Sweden

Comments by Henri Luomaranta

In "Some Thoughts on Official Statistics and its Future", the author provides nalysis of the evolution of official statistics and its epochs, lets us know where we are now and assesses the future that lies ahead. The author takes us back to the days of censuses, to sampling design and related methodologies, and to the register-based approaches typically found in the Nordic countries. In fact, it becomes clear to the reader that these epochs are the building blocks for the key capability of official statistics, the integration of many, possibly highly complex, data sets (be they big or not) to produce relevant information that serves as the backbone of democratic processes. In what follows, I offer considerations focusing on the new topics pursued by statistical offices, where there is perhaps more ongoing discussions among practitioners.

As we all know, the latest, perhaps dramatic, evolution of official statistics is the adoption of big data and statistical learning. The author points out, and I agree, that these new methods and data sources can and should be integrated to the many hidden processes found in official statistics, for instance on editing and imputing data. The comparison of current practices to an AI system is interesting. In statistical systems, a set of methodological practices translate several input data into statistics, adding tremendous value. Certainly, statistical learning and "big data" can make the statistical systems and these rules-based processes more efficient.

There are roadblocks before such evolutions can take place, more in some offices than in others. The author points a finger at some statisticians who fear sacrificing impartiality if modelling, and especially statistical learning, is widely adopted, as modelling requires taking a stance on the relationship between variables. Personally I do find that Deng Xiaoping's assessment rings true: "if it catches mice, it is a good cat". We should adopt those methodologies that work best for the problem at hand. Testing is key, obviously. Engineering worldview promotes testing as the process that makes anything trustworthy. Methodology, data sources and procedures can be described by scientific standards even if modelling is used, as required by the 3rd UN principle of official statistics (United Nations 2014). But sure enough: for building trust, we have to understand the processes and be able to stand by them.

The buzzword "big data" irritates the author semantically, and the article suggests dropping the use of the term as it encompasses multitude of both methods and data sources, with no commonly agreed definition. The author expresses concerns over the use of Internet as a data source, as we cannot control or verify whether it is true what is posted online. Another important risk raised is the dependency from private firms, if we should try to obtain data from search engines and Internet platforms. At the same time, it is recognized that having some data is better than no data, and having no data is a real concern as nonresponse is a widely acknowledged problem.

I would like to point out, that modern statistical offices are already highly dependent from other actors for data. For a while now, data has been increasingly obtained from

© Statistics Sweden

many private and public organisations and the data generating process is not controlled by the National Statistical Institutes (NSI). This is a risk that we must manage, as reversing the trend is unlikely. Perhaps it is better to go all in, learn how to manage that risk optimally by developing trusted partnerships, and then pay back by producing trusted, impartial statistics. Strujs et al. (2014) bring this point home arguing that impartiality and respect for privacy of NSIs, if communicated successfully, will put the NSIs as the uniquely positioned trusted party also in the new world of big data, which may change how NSIs operate altogether in their view.

While recognizing that the doubts of using web scraping as a source are well grounded, I would like to maintain a more optimistic view of Internet as a source, as we have barely begun looking at the possibilities. Many aspects of modern lives are online. The Internet is evolving rapidly, as is the way we use and regulate it. Learning how to harness the data on the Internet-economy by accessing websites seems valuable and not doing so might present a risk of missing out. And, as by the 5th UN principle of official statistics, we need to consider the cost effectiveness and response burden in choosing data sources, it would be a mistake to discard web scraping at this stage.

In another piece on the future of official statistics MacFeely (2020), explains how we have undergone series of data revolutions, all building the importance of data in today's economies. These revolutions must have consequences for official statistics if we are to remain relevant and is probably an argument why we should embrace some of the hype around big data. Businesses, that have adopted big data and data driven strategies, have produced remarkable innovations. When compared, the official statistics could easily be seen as dated, or useless. As a consequence, an official statistician who is desperately clinging on to a too strict interpretation of quality with the expense of relevance, may describe dated statistics with an impressive scientific precision, but provides no real value to users.

This is why modernising the official statistics production systems is important. The main focus should be on the ability of producing relevant information, flexibly and in a more timely fashion, for example by exploiting new data sources and technologies. But, this presents an additional dilemma, as we must somehow strike a balance between relevance and impartiality (as stressed in Sæbø and Holmberg 2019). In the European Code of Practice of official statistics, impartiality and professional independence are high on the list and places importance on "how" statistics are produced, but emphasize also the relevance and pursuit of innovation in statistical outputs (Eurostat 2017). If "how" remains at the discretion of the statistical offices, close cooperation with users, and even actively obtaining outside financing for producing innovative new statistical products should not be a problem. Of course, care is needed to communicate the impartiality aspects to the public and the results should be made available in an impartial manner.

Production of new statistics is one thing, another is disseminating them to users – how do we do that? An example of this is the recent trend of publishing experimental statistics, which has risen precisely from the viewpoint of relevance and impartiality. In Finland, these new statistics have allowed more flexibility and have made the innovation process more visible to users, creating an environment where the statistical office could be seen as able to respond to emerging data needs, harnessing the coveted relevance in the eyes of the public. We have been able to explore some aspects of the "new" statistical learning

technologies and have created new statistical products, such as nowcasting (see, e.g., Fornaro and Luomaranta 2020), which has improved the timeliness dimension of economic statistics. Otherwise, big data is slowly but surely being integrated into the statistics production systems; for example the adoption of scanner data in price index compilation has improved the quality and efficiency of price index production (see, e.g., Suoperä et al. 2017). There are also cases where the adoption of web scraping allowed describing new phenomena altogether, as in some cases, prices are indeed only available online – and this trend is expected to strengthen.

It is easy to agree with the author that we should continue improving on the competitive advantage of data integration capacity in official statistics. No one produces comparable, high-quality statistics better than official statisticians, and this should be clearly communicated to the public as well. This emphasis of valorising the skills in data integration allows the opening of statistical service to new areas too, as the existing data already collected by statistical offices allows for far more combinations, breakdowns and data products than what are currently published. This way we can do more with what we have cost effectively, and remain relevant as the data revolutions are playing out.

As many of us younger methodologists are excited about the prospects of new shiny things, the pragmatism and grounding offered in this article are truly valuable. Mission first.

References

Fornaro, P., and H. Luomaranta. 2020. "Nowcasting Finnish real economic activity: a machine learning approach." *Empirical Economics* 58(1): 55–71. DOI: https://doi.org/10.1007/s00181-019-01809-y.

Suoperä, A., Nieminen, K., and S. Montonen. 2017. *Analysis of Change and Constructing Index Series for Complete Micro Data*. Available at: http://www.stat.fi/static/media/uploads/meta_en/menetelmakehitystyo/analysis_of_chain_driftfor_complete_micro_data_final.pdf (accessed February 2022).

Struijs, P., Braaksma, B., and P.J.H. Daas. 2014. "Official statistics and Big Data." *Big Data & Society* 2014: 1–6. DOI: https://doi.org/10.1177/2053951714538417.

MacFeely, S. 2020. "In search of the data revolution: Has the official statistics paradigm shifted?" *Statistical Journal of the IAOS* 36(4). DOI: https://doi.org/10.3233/SJI-200662.

Sæbø, H.V., and A. Holmberg. 2019. "Beyond code of practice: New quality challenges in official statistics." *Statistical Journal of the IAOS* 35(2): 171–178. DOI: https://doi.org/10.3233/SJI-180463.

Eurostat. 2017. *European Statistics Code of Practice*. Available at: http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice (accessed February 2022).

United Nations. 2014. *UN Fundamental Principles of official Statistics.* Available at: https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed February 2022).

Comments by Martin Axelson and Eva Elvers

1. Introduction

First, we would like to thank the Editors for inviting us and to thank the Author for an interesting and thought-provoking article about the field of official statistics and its future. The article provides an important contribution to the literature on official statistics and the use of statistical theory in their production. Official statistics are more specific than statistics in general, but still open to some controversies and debates. We will in our discussion of official statistics now and in the future not cover all aspects that the Author raises but rather focus on some of them, in order to present a somewhat complementary view of the topic at hand.

2. Statistical Theory

2.1. Survey – In a Broad Sense

A *statistical survey* is here considered in a broad sense; it is not limited to a sample survey with direct data collection. There may be already available administrative data to use, solely or partly, and there may be other data sources. The design of the survey considers if, and how, to draw a sample; observe data with appropriate collection methods; use administrative data; edit and integrate data; etc. The data sets to be integrated into one set, or otherwise to be used together, are often micro data, but there may be aggregates.

When the Author uses the term survey, it mostly refers to a sample survey. Integration of data is added for several situations.

Official statistics are based on statistical surveys. There is an ongoing development of surveys, with expansion in data sources, collection and integration methods, and estimation methods.

2.2. Two Main Parts in Statistical Theory

Statistical theory has two main parts:

1. probability theory, and
2. statistical inference (with estimation and testing).

These two parts can be considered as statistical computations going in two directions. One direction goes from general to specific, through assessment of possible outcomes beforehand. The other direction goes from specific to general, through use of realised outcomes to draw conclusions, that is, to infer something about the general.

© Statistics Sweden

We will describe and discuss the two main parts in the setting of official statistics and two approaches to statistics in Subsections 2.3–2.5 below. Our views deviate somewhat from those of the Author.

2.3. *Observed Data and Induction*

A Swedish encyclopaedia says about statistical inference: an inductive science, where conclusions are drawn from empirical data with regard to uncertainty caused by randomness influencing data.

Statistics make use of observed data together with notions about untainted data and possibly also scientific models. Observed data are rarely perfect and complete. On the contrary, the data are typically associated with uncertainty, due to, for instance, measurement errors, incomplete coverage, and missing data (nonresponse). The observed data may refer to a sampled set of units rather than a full population; a set that is further reduced due to nonresponse. The observed data may consist of several separate data sets, possibly of different coverage and with different content in terms of variables.

A step is taken from the observed data, associated with uncertainty: the observed data are aggregated to statistics. This step may, with reference to philosophy, be called an induction; it is a reasoning with evidence but without certainty about the truth. The step is built on statistical theory, and it is called statistical inference. The inferential step aims to:

- reduce consequences of data deficiencies such as missing data and
- describe the uncertainty that remains in the statistics.

2.4. *Deduction and an Illustration*

How to observe data is an essential issue – data that are appropriate for the overall goal of the statistics and where quality, costs, and respondent burden (when direct data collection is used) are balanced.

First, there is an ex-ante reasoning about which data to observe and how. This reasoning may, again with reference to philosophy, be called a deduction, where the data behaviour is described in advance based on the statistical models. This reasoning depends largely on probability theory. Using previous experience, different possibilities can be studied and compared, with respect to expected quality, costs, and respondent burden. The initial, designing, stage – where both deduction and induction are considered and an outline is made – is indeed important.

Figure 1 is a summary that illustrates deduction and induction in one or more successive sub-steps. At the same time, Figure 1 shows how to use probability theory and statistical inference for official statistics. Ideally, the two sets of statistical assumptions are made in such a way that agreement between them is achieved. If so, the statistical inference will be in accordance with the probability theory. Typically, alternative possibilities to obtain the intended set of statistics are studied, thus implying that both sides of Figure 1 are used iteratively.

The Author states that the induction/deduction distinction hardly can be applied in official statistics – with reference to scientific theory. We find a reasoning with deduction and induction useful for survey design and the inferential step. This is a perspective shared

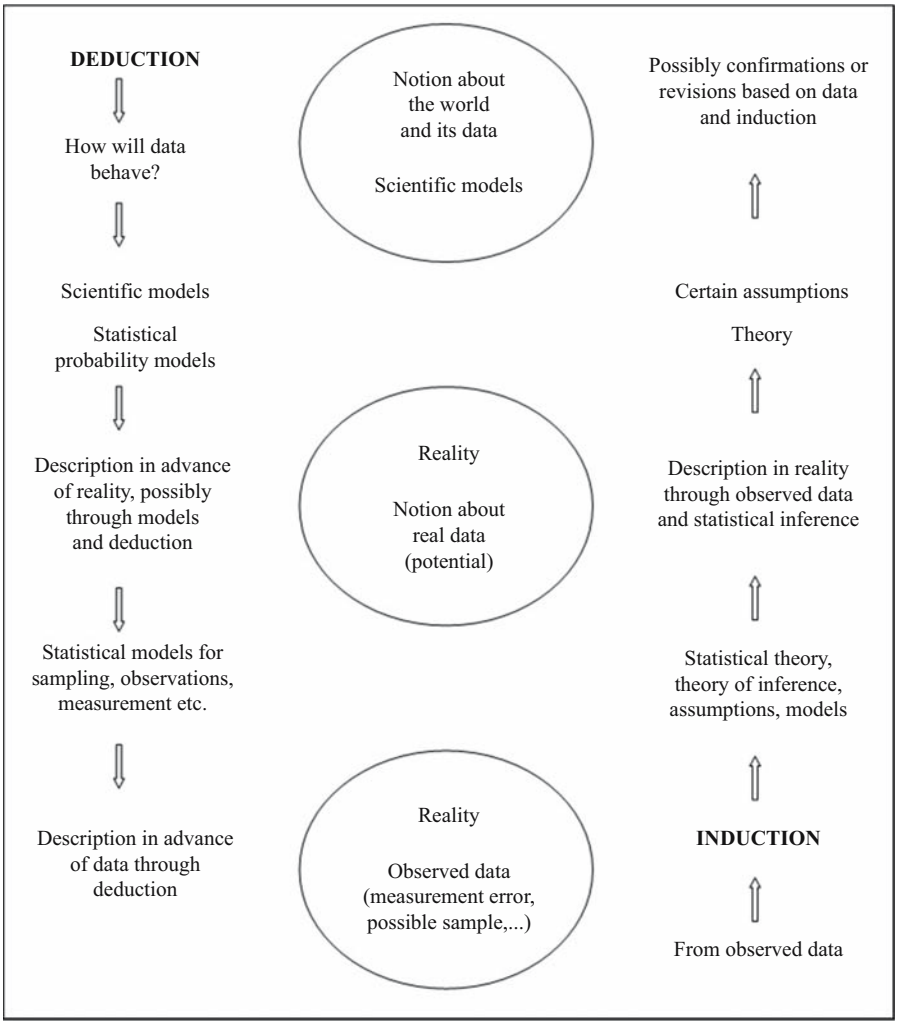


Fig. 1. Induction and deduction

by others. For example, Alleva et al. (2021) say that in particular for official statistics, an iteration and learning loop between the deductive and the inductive approach is characteristic.

When working with the deduction side of Figure 1 the purpose of the statistics should be stated clearly enough to signal the corresponding quality requirements, for example regarding the content of the statistics, accuracy, timeliness, and coherence. Priorities are made in the design in order for the statistics to be *fit for purpose*.

2.5. *Descriptive and Analytic Approaches*

The Author distinguishes between descriptive or exploratory statistics on the one hand and analytical or inferential statistics on the other hand with a statistical approach that is general and not tied to official statistics.

When defining and producing official statistics, the statistics may be characterised as either descriptive or analytic; a how-many type and a why type, respectively.

Many short-term time series are seasonally adjusted in order to make comparisons between consecutive time periods easier – it may thus be easier to see and interpret trends and short-term movements. A suitable model is used for the seasonal adjustment; a time series model with parameters that are estimated from the data. The unadjusted and the adjusted time series provide examples of the two approaches descriptive and analytic statistics, respectively.

Another way to clarify the two approaches is that the unadjusted time series describes the real world, as it is, whereas the seasonally adjusted time series describes something else. It refers to a theoretical concept, aimed at describing a world similar to the real world but without seasonal effects; thus comparisons over time are easier to make. This description builds partly on model assumptions. The distinction between the two types of statistics is not always clear-cut. Model assumptions are often used also for estimates regarding descriptive statistics, as discussed further below.

The inferential step as illustrated in Figure 1 may be to the real world or to some relevant notion of the world, such as a world without seasonal variations.

2.6. Comments

The description in Subsections 2.3–2.5 serves two purposes. Firstly, the intention is to show why we find a reasoning with deduction and induction useful. Secondly, it underlines the importance of being transparent about the inferential step: what world, what parameters and so on. We agree essentially with the Author that official statistics are not produced as scientific research. However, researchers may find much of interest in official statistics to investigate further, for example in long time series and through combining several sets of statistics.

3. Design, Data Integration, and Estimation

3.1. Survey Design

The design of a survey (in a broad sense) considers the inferential step from the data to the real world and, possibly, further steps to descriptions that are assisted by, or even based on, scientific models. Estimation procedures are constructed aiming at statistics with no or little bias and with a small variance. Randomness is part of the picture, but what is fixed and what is random? There are several possibilities and schools of thought.

A sample survey with probability sampling and an estimator based on the sampling design has for a long time been considered the standard approach for official statistics. The population is finite and the observations are considered fixed, whereas the sample is random. Further approaches exist, for instance to consider observations as outcomes of random variables. Even if the inference is still to the finite population, this is a model-based approach, sometimes called a prediction approach. Yet another approach is to consider the finite population as a sample from an infinite population, a so-called super-population. The inference may be to some parameter of the infinite population or still of the finite population. This corresponds to different possible worlds in Figure 1.

3.2. Nonresponse and Calibration

Nonresponse rates have increased considerably for a number of years. The statistical inference is affected, and there is concern about possible bias in the statistics. The Author discusses nonresponse, models, approaches to estimation and desirable properties at several places. With nonresponse, some use of assumptions and models is necessary.

When it comes to nonresponse, the auxiliary variables are, simply expressed, powerful if they capture variations in the target variable and in the propensity to respond. We do not dispute the concept of double robustness described by the Author per se, but we would like to point out that it is not necessarily easy from a practitioner's point of view to prove that model assumptions hold. In our experience, the building of imputation and nonresponse models is based on reasoning about similarities within and between groups rather than on formal hypothesis testing and p-values. The rather small data sets from respondents only is the main explanation.

Calibration is an elegant method that was introduced to decrease variance through use of the auxiliary variables that are correlated with the target variables. The Author provides references. The use of this tool has expanded to adjust for nonresponse and sometimes also for measurement errors and coverage deficiencies, as the Author notes. We are less optimistic than the Author about the overall usefulness of calibration. In our opinion, the tool is too often used without enough reflection on its properties and the conditions under which it indeed adjusts for nonresponse, measurement errors, and coverage deficiencies. We are convinced that official statistics would benefit from the theory on estimation under nonresponse being further developed, preferably in a way that explicitly takes existing subject matter theory into account. For this to take place, statisticians and subject matter experts must work together.

3.3. Data Integration

Calibration may be considered as an example of data integration, where information on the auxiliary variables and information on the target variables are combined. The Author mentions this.

In Sweden, there are official, unique identification numbers for individuals and businesses, to mention just two such unit types. This makes it possible to integrate data from different data sources on the micro level more easily than in many other countries. However, Sweden is far from alone in combining data sources. Overall, there is an increasing interest in using more data sources for official statistics, if possible.

The Author states that it is a crucial issue in official statistics to integrate data from different data sources "in order to best enhance the reliability of each source". It is, of course, a matter of more than reliability.

3.4. Administrative Data and Costs

Sample surveys with probability samples and direct data collection has been a "gold standard" for official statistics for a rather long time, as the Author notes. However, there has been an increased use of administrative data for some time now, with a variation between countries.

Administrative data have been collected for a specific administrative purpose. When they are to be used for statistical purposes the deduction and induction steps of Figure 1 have to be considered.

How can the data reasonably be described on the generating side, and what assumptions and models are reasonable on the inferential side? Maybe some adjustments and models are motivated with respect to coverage and measurement deficiencies. It may be desirable and motivated to use a model to "derive" somewhat new variables for the statistics based on the observed ones.

The usefulness of administrative data for statistical purposes depends highly on the original administrative use when it comes to timeliness, accuracy of the micro data and so on. There may be agreements between agencies about data handling. Our experience of using administrative data for statistical purposes is more positive than that of the Author. Still, there certainly has to be care when designing the use.

3.5. New Data Sources and Improvements

Official statistics are important for decisions in the society, and for general information, investigation, and research. They are to be objective and they are a public good. The money spent on the production of official statistics has to be used with caution and in an efficient way. Hence, the possibilities to use existing data rather than collecting new data have to be considered. The purpose of the statistics, the quality, the costs, and the respondent burden related to direct data collection have to be considered together. Producers of official statistics should not collect data from respondents if these data are already available. If available data are close to the intentions, it may be possible to use these data together with other available data and statistical models. It is a matter of adding value in relation to costs.

Lothian et al. (2019) describe how to use available data, and they call them "it-is-what-it-is". They discuss how to integrate these data with other data. The so-called base registers, which many statistical agencies have on individuals, businesses, and dwellings, are useful.

Alleva et al. (2021) and De Broe et al. (2021) are two recent examples that show the quick development in the production of official statistics. They discuss methodological, technological, and design challenges in the new multisource statistics ecosystem and the paradigm of official statistics, respectively. De Broe et al. (2021) describe cultural changes, too. There are many challenges involved for many employees.

When official statistics are produced as long time series, requirements on stability over time is one of the consequences. The Author states this together with comments on the usefulness of web scraped data, in particular regarding prices, in a way that we find unnecessarily negative, even if we overall find big data to be a bit of a buzzword. New data sources may be used as long as the statistics produced are in line with the quality requirements. It is often a considerable challenge to measure or assess the accuracy.

The Covid-19 pandemic has put pressure on timely statistics and to some extent on new statistics. This has implied new and modified methods for some statistics in an unusually quick way. The pandemic has also demonstrated deficiencies in comparability between countries due to different definitions. More work is needed in the near future.

4. A Few Concluding Remarks

To conclude, we would like to thank the author once more for an interesting and thought-provoking article on an important topic. Even if official statistics are specific as such, they are not essentially different from other statistics when it comes to the use of probability theory and statistical inference. A producer of (official) statistics has an important role when developing and communicating the statistics. The scientific model(s) that the statistics depend on should be clearly stated, and the target of the statistical inference should be explicit – both in discussions with stakeholders and in the general, open, presentation.

Data sources and methods should be chosen carefully and in a way that combines and integrates the data sources and the methods used in an appropriate way. Objectivity, impartiality, a scientific basis, and transparency are key. Quality should be in line with the purpose of the statistics and not be expressed as vaguely as "high-quality" statistics. Here our view deviates somewhat from what the Author advocates.

The Author states in his/her concluding remarks that it would be useful to have a general theoretical framework for data integration. We agree, and we would like to go further; we would like theory for statistical inference from registers and integrated data.

5. References

Alleva, G., E. Baldacci, and P.D. Falorsi. 2021. "Methodological, technological and design challenges in the new multisource statistics ecosystem." *Statistical Journal of the IAOS* 37: 245–254. DOI: https://doi.org/10.3233/SJI-200757.

De Broe, S., P. Struijs, P. Daas, A. van Delden, J. Burger, J. van den Brakel, O. ten Bosch, K. Zeelenberg, W. Ypma, "Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics." *Statistical Journal of the IAOS* 37: 343–360. DOI: https://doi.org/ 10.3233/SJI-200711

Lothian, J., A. Holmberg, and A. Seyb. 2019. "An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics." *Journal of Official Statistics* 35(1): 137–165. DOI: https://doi.org/10.2478/jos-2019-0007.

Comments by Anders Holmberg

I would like to thank the editors for the opportunity to comment on this article. The title contains Official Statistics and it is that lens I am taking. The views expressed are my personal and not those of my employer.

The author tries to connect thoughts on the future of official statistics with the evolution of (parts of) statistical science and when doing so turning against the concept of "Big Data". It is easy to have sympathy for the frustration of how the Big Data terminology sometimes appear. I agree with most of the points made, but the author is preaching to the choir. And I do not see any danger pursuing use of alternate data irrespective of their size or structure.

It is my opinion that official statisticians know that alternate data acquired for statistics without a purpose-built questionnaire will never give you exact what you are looking for. Does that make these data useless for official statistics? No of course not. Does it make such data worse than survey data? Sometimes! But that can be evaluated by a systematic approach using different quality frameworks that has been developed in recent times (see e.g., Zhang 2012; Reid et al. 2015). It is through a quality-cost lens we can find the most appropriate mix of data sources for recurrent Official Statistics.

The article meanders between issues and opportunities without clearly highlighting any main issues for Official Statistics. The "*p*-value crisis" (which has been around for a while and sometimes turn up as mass significance issues) and the synopsis of the statistical science development does not add any new insights or a research synthesis to its future. Although many author statements are agreeable, I miss the connections to basic factors such as the UN fundamental principles (UN 2014), technology drivers, and the changing societal and stakeholder's requirements that often emphasise other quality dimensions than accuracy. Almost all methodological references in the article address the accuracy dimension. This is ok but not sufficient to future proof official statistics to retain public trust and stay relevant. The demands on and the responses from National Statistical Institutes worldwide during the COVID-19 pandemic show that sometime accuracy must be traded. In Australia for example, the Australian Bureau of Statistics used a new data source on payroll and wages to rapidly produce fortnightly indexes to help policy makers.

The four separate sections discussing data sources and their integration, modelling, big data, and statistical learning have a mixed bag of author reflections. The review of calibration methodology, the development lines of probability sampling and use of modelling all come from the sample survey theory tradition. While that is important it also misses, or underplays important work done on alternate data sources. For example Journal of Official Statistics had a special issue about coverage in administrative data in 2015, the Dutch virtual census came 2001 (Schulte-Nordholt et al. 2004), register-based statistics (UNECE 2007; Wallgren and Wallgren 2014 and many more) and integration of data (Kim and Tam 2021).

© Statistics Sweden

The author notes however, that sampling increasingly is used to check and improve other data sources. This trend of more sampling to observe as opposed to sampling to do surveys is growing in importance. In my view, this is the way forward to get more value from alternate data sources and make the best of integrated data. For official statistics this may not need so much new statistical theory as it needs rethinking statistical designs where probability-based data is adapted to add value to not randomly selected "it-is-what-it-is" data. Nothing stops a statistical system from using both surveys and alternate sources interchangeably and in a complementary way (e.g., see Lothian et al. 2019). The strength of the sample survey approach (That it can be tailored to a problem and better control biases.) speaks in favour for it being a key tool that filling the gaps from first using an alternative data source. I am certain we will soon see a lot more applications of this.

It is therefore easy to agree with the author that data integration is a necessary pathway ahead. I also share the skepticism that data scraped from the web never will pay a dominant role. Not so much because of the data quality, this can be good (e.g., timely and accurate price information) or poor with concept deviations, selection biases etc. I am more concerned about fit for purpose, transparency, ethics and production circumstances to official statistics. Web scraped data can be unsustainable and not fulfilling initial promises of automation and efficiency with one IT-solution per application needing adjustment as soon as sources are updated. A better solution would be to investigate the providers' ability to directly transfer their data.

Finally, a comment on the use of statistical learning and doing predictions when making official statistics. This is not new, and statistical offices will continue to use and develop sound methods for this when appropriate. Again cost versus benefit will be a major driver.

References

Kim, J.K., and S.M. Tam. 2021. "Data integration by combining big data and survey sample data for finite population inference." *International Statistical Review*. 89(2): 382–401. DOI: https://doi.org/10.1111/insr.12434.

Lothian, J., A. Holmberg, and A. Seyb. 2019. "An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics." *Journal of Official Statistics* 35(1): 137–165. DOI: http://doi.org/10.2478/JOS-2019-0007.

Reid, G. Zabala, F., and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics* 33(2): 477–511. DOI: https://doi.org/10.1515/jos-2017-0023.

Schulte-Nordholt, E., M. Hartgers, and R. Gircour. 2004. *The Dutch Virtual Census of 2001, Analysis and methodology*. Voorburg/Heerlen. Available at: http://www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/publicaties/archief/2005/2001-b57-e-pub.htm (accessed February 2022).

UNECE (United Nations Economic Commission for Europe). 2007. *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics*. United Nations Publication. Available at: https://unece.org/DAM/-stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed February 2022).

United Nations. 2014. *Fundamental Principles of Official Statistics.* 68/261 Resolution adopted by the General Assembly on 29 January 2014. Available at: https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed February 2022).

Wallgren, A., and B. Wallgren. 2014. *Register-based statistics administrative data for statistical purposes*. Chichester, England: John Wiley & Sons. Available at: https://www.wiley.com/en-us/Register + based + Statistics°/03A + Administrative + Data + for + Statistical + Purposes-p-978047002778 (accessed February 2022).

Zhang, L.C. 2012. "Topics of statistical theory for register-based statistics and data integration." *Statistica Neerlandica* 66(1): 41–63. DOI: https://doi.org/10.1111/j.1467-9574.2011.00508.x.

Comments by Richard Valliant

This article is a thoughtful discussion of the goals of official statistics and the issues it must confront in the coming years. The author covers some history of official statistics and connects the current debate about inference to the production of official statistics. The author also gives a useful overview of the problems in data integration that agencies are sorting through. These comments expand on some of the authors points about official statistics.

Official statistics is now at a critical juncture because of rapid, technological changes in most societies. The call by Statistics Canada for submissions to the 2021 International Methodology Symposium, titled "Adopting Data Science in Official Statistics to Meet Society's Emerging Needs" is a reflection of this (www.statcan.gc.ca/eng/conferences/-symposium2021/index). As the conference organizers ask, "Advancements in the field of data science provide solutions to some of these new challenges, but what does it all mean for official statistics and what is the impact on how data are collected, compiled, analyzed and presented?"

Perhaps even more daunting than the technical challenges is rising distrust of official statistics in some countries. According to Davies (2017), prior to the 2016 US presidential election, a study in the US found that 68% of Trump supporters distrusted the economic data published by the federal government with nearly half saying they do not trust government economic data "at all". He further notes that in the UK, a research project by Cambridge University and YouGov looking at conspiracy theories discovered that 55% of the population believes that the government "is hiding the truth about the number of immigrants living here". As observed by social psychologist Arie Kruglanski in a Washington Post article on extremism "... trust is low in government, in Congress, in science, in medicine, the church..." (Fisher 2021). Convincing the public that official statistics are still to be trusted will require persuasive public relations in addition to methods that can be justified technically.

1. Growing Nonresponse to Surveys and Censuses

A major problem is getting data in regular probability surveys in the first place. Nonresponse rates have been steadily increasing over the years with no abatement in sight. The general decline in public trust affects governmental data collection in many types of surveys. Williams and Brick (2018) chart the decline in response rates in five major US in-person household surveys over the period 1990–2014. The National Health Interview Survey response rate (RR) declined from 95% to 74%. The RR in the Current Population Survey-Annual Social and Economic Supplement (CPS-ASEC) dropped from about 89% to 76%. The RR in the National Survey of Drug Use and Health went from 84% to about 61%. Since 2014, response rates have declined even more with shocks like the global COVID-19 pandemic causing sudden, additional decreases (e.g., see https://www.cen-

© Statistics Sweden

sus.gov/newsroom/blogs/research-matters/2020/09/pandemic-affect-survey-response.html.

Response rates to other modes like telephone are far worse than for in-person surveys. Kennedy and Hartig (2019) report that RRs to Pew Research telephone surveys declined from 36% in 1997 to 6% in 2018. Although the relationship between bias and RR is not necessarily high (Groves and Peytcheva 2008), such low RRs are not a source of comfort. Using hand-held devices like mobile phones to collect data is one way to facilitate response and increase response rates. Use of these devices is bound to expand in the future when in-person collection is not required.

2. Reacting Quickly to Changing Demands for Data

A health crisis is often the impetus to quickly generate data on new topics – the global COVID-19 pandemic being the latest with the HIV/AIDS epidemic of the 1980s being another. Governmental agencies need to agile enough to mount new collections of data in months or even weeks to serve policy needs. Past development cycles that can stretch into years are not acceptable when needs for data are pressing. An example of a current, quick response US survey is the Household Pulse Survey (Fields et al. 2020) sponsored by the Census Bureau and five other agencies. The Pulse Survey is designed to measure the impact of the COVID-19 pandemic on employment status, consumer spending, food security, housing, and other topics and collects data on-line. Although the survey provides timely estimates, these come with a price. The survey had less than a 4% response rate in April 2020. Although the sampling frame used is probably the best available in the US, assessing the quality of data from surveys with such low RRs may be difficult or impossible.

As response rates progressively diminish in probability samples, the repeated sampling mathematics traditionally used to derive finite population estimators becomes more tenuous. Consequently, model-based justifications for estimating inclusion propensities and for calibrating estimators become more important. The need to validate the quality of estimates applies to all surveys but is especially germane to quick response, low response rate surveys and to nonprobability samples mentioned below.

3. Adapting to New Methods of Data Collection

Collection of biospecimens and other physical measurements has been commonplace in various official surveys for years and is likely to expand in the future. For example, the U.S. National Health and Nutrition Examination Survey has collected blood serum and urine from a national probability sample since 1988. The U.S. Health and Retirement Study (HRS) has genotyped almost 20,000 respondents who provided DNA samples and signed consent forms in 2006–2012 (https://hrs.isr.umich.edu/data-products/genetic-data). There are a number of HRS sister studies of the aging population that have similar goals: English Longitudinal Study of Aging (ELSA), Survey of Health, Ageing, and Retirement in Europe (SHARE), The Irish Longitudinal Study on Ageing (TILDA), and others (https://hrs.isr.umich.edu/about/international-sister-studies).

Use of wearables to collect survey data opens the possibility for these and other studies to build rich time series of data on health measures from representative samples that have

not been available before (see, e.g., De Arriba-Pérez et al. 2016). Data from wearables can be collected for closely spaced time intervals on sleep patterns, exercise habits, and other characteristics that have not been easily accessible before on large samples. The ability to analyze such data for race-ethnicity and other subgroups could be a boon to public health and health policy analyses. Naturally, this technology raises concerns about confidentiality and willingness of demographic groups to cooperate in these data collection efforts (see below).

In the US, the National Institute for Health's All of US (AoU) voluntary national research program whose goal is to collect biometric data on one million people through surveys, electronic health records, and biosamples including blood and saliva (https://allofus.nih.gov/about). In 2019 the Fitbit Bring-Your-Own-Device (BYOD) project was added to AoU where participants can choose to share biometric data from their Fitbit® accounts. Since AoU is a nonprobability sample, how to use these data for population estimates is a problem to be solved, as noted below.

4. Data Integration, Nonprobability Samples, and Calibration

Administrative databases can cover some or all of a population for which statistics are desired. Combining administrative data with survey data seems to be an avenue toward improving survey data by using the administrative data to impute for missing values, to avoid asking survey respondents for certain items if they are available in an administrative database, and to compensate for incomplete sampling frames (Auerbach et al. 2019). Tax records and postal service databases are examples of administrative databases that are useful. As the author notes, administrative data may not be collected with the same care as high-quality survey data. Use of these data in combination with survey data for estimation raises several questions, including: (1) Are the data of the same quality as survey data? (2) Do the data cover the correct universe (or part of it)? (3) Are units in the administrative database defined in the same way as those in the survey database?

Although the founders of probability sampling (Neyman, Hansen, Hurwitz, Madow, etc.) would probably recoil at the notion, use of nonprobability samples is expanding and may inevitably become a part of official data collections at least in part as a way of collecting data cheaper and quicker (Japec and Lyberg 2020). Nonprobability samples may also be the best way of finding units with special characteristics that would be rarely located with a probability sample from a full population. For example, Elliott et al. (2010) and Rafei et al. (2020) discuss combining a nonprobability sample of motor vehicle crashes that lead to occupants being severely injured with a probability sample of all crashes. As this example implies, such a combination may be of most interest in official statistics when estimates are needed for rare subpopulations.

The literature is rapidly growing on how to use nonprobability samples – either alone or combined with a probability sample. "Big data" are not obtained by designed data collection methods and must be treated as nonprobability samples (if they are used at all). The main dangers in using big data are under- and over-coverage, and self-selection. Attempts must be made to statistically adjust for each of these problems to create approximately unbiased estimators. Having a gold standard probability sample that collects some of the same items as a nonprobability sample is one approach to adjustment.

For the sake of efficiency, federal statistical agencies should cooperate in identifying a set of probability surveys to jointly fund for this purpose and modify them to use uniform definitions of terms and items so that the purposes of different agencies are met. Such uniformity may also facilitate linking different data sets. Coordination among data collection systems, of course, should be easier in countries with centralized statistical systems.

As the author notes, calibration of probability samples to control counts from administrative databases is a key part of survey estimation. The uses of calibration in nonprobability samples are the same – adjustment for coverage errors, correction of nonresponse errors, and reduction of standard errors. Calibrating to estimates from well-executed, probability surveys is another option that allows control quantities to be used that are different and potentially more effective than ones from administrative data sets. When nonprobability samples are used, thorough modeling will become more important to estimate propensities of being observed in a sample and to model the relationship between analytic variables (y's) and predictors (auxiliaries).

Whether a particular nonprobability sample should be used at all for official statistics is, of course, the critical question. The author expresses some justifiable doubt about the use of web-scraped data and other sources that are considered nonprobability data sources. Diagnostics for nonignorable selection in such samples (e.g., see Andridge et al. 2019; Little et al. 2020) should be an important step in deciding whether nonprobability samples are "fit for use".

The technical, statistical problems of data integration need additional research. Lohr and Raghunathan (2017) discuss methods of integration including imputation, use of multiple frames, and small area estimation techniques for integration and how studies might be designed to check on the accuracy of different sources. Kim and Tam (2021) consider techniques for integration that attempt to handle selection bias.

5. Confidentiality and Disclosure Protections that Society will Accept

As distrust of government builds among some segments of society, agencies must present a convincing case that they will safeguard participants' data. This especially applies to personal, health-related data and financial data. Linking different administrative databases, like tax records or medical histories, to survey data expands the kinds of analyses that can be done but requires that people give informed consent to do the linkage (Sakshaug et al. 2012).

Sophisticated confidentiality protection measures like differential privacy have been developed by cryptographers and are being applied to surveys and censuses. For publicly released US 2020 census data, the plan is to use the technique to avoid disclosures of individual information (U.S. Census Bureau 2021). However, objections to its use have already been raised – not for its inadequate protection but because it is claimed to suppress too much data. In a 2020 lawsuit (since withdrawn), the state of Alabama sued the U.S. Census Bureau in 2020 on the grounds that the planned disclosure methods interfered with the data needed to redraw congressional districts (Riccardi and Schneider 2021). Another option for maintaining confidentiality is to create entirely synthetic data for applications

where real data are too sensitive to publish; the United Kingdom is actively investigating this (Kaloskampis et al. 2020).

Some US agencies have established Research Data Centers (RDCs) to allow analysts access to confidential data while protecting privacy of persons or businesses. An RDC is a location where researchers can go to analyze restricted-use microdata that has not been subjected to data suppression or other disclosure protection techniques. Analysts must receive clearance to use an RDC and are restricted in the sort of material that can be taken away from the RDC. Many researchers feel that these strictures are extremely inconvenient and limit the kinds of analysis that can be done or whether a researcher can use the data at all. Making official data more accessible to more researchers while still protecting confidentiality should be an ongoing concern.

6. Recruiting Staff and Competing with the Private Sector

The statistical analyses and methodology required for surveys are growing more and more technologically complex. New and existing staff in government agencies must be up to the new challenges. This will require agencies to invest in continuing education and to hire qualified staff with strong educations in statistics, data science, and methodology. Because national governments have made huge investments in trying to recover from the public health and economic effects of the pandemic, finding money to hire and educate staff will be a major test in coming years. Government must also compete with private sector technology companies that typically pay higher salaries than the government.

In summary, the challenges faced by official statistics are immense. A key goal should be to remain a trusted source for truth in the face of so much misinformation and disinformation that is swirling everywhere. To quote Rademacher (2020), "...facts can play an enormously important part if they are produced with a quality that makes them credible and purpose-specific. In order for this to actually happen, official statistics must continue to actively pursue the modernisation of their working methods."

7. References

Andridge, R.R., B.T. West, R.J.A. Little, P.S. Boonstra, and F. Alvarado-Leiton. 2019. "Indices of non-ignorable selection bias for proportions estimated from non-probability samples." *Applied Statistics* 68(5): 1465–1483. DOI: https://doi.org/10.1111/rssc.12371.

Auerbach, J., Q. Brummet, J. Czajka, G.C. Hough, Jr., E. Hunsinger, and J. Salvo. 2019. "Will administrative data save government surveys?" *Significance Magazine* 16(5): 35–39. DOI: https://doi.Org/10.llll/j.1740-9713.2019.01319.x.

Davies, W. 2017. "How statistics lost their power – and why we should fear what comes next." *The Guardian.* Available at: https://www.theguardian.com/politics/2017/jan/19/-crisis-of-statistics-big-data-democracy (accessed February 2022).

De Arriba-Pérez, F., M. Caeiro-Rodríguez, and J.M. Santos-Gago. 2016. "Collection and Processing of Data from Wrist Wearable Devices in Heterogeneous and Multiple- User Scenarios." *Sensor* 16: 1538. DOI: https://doi.org/10.3390/sl6091538.

Elliott, M., A. Resler, C. Flannagan, and J. Rupp. 2010. "Combining data from probability and non-probability samples using pseudo-weights." *Accident Analysis and Prevention* 42: 530–539. DOI: https://doi.org/10.29115/SP-2009-0025.

Fields, J.F., J. Hunter-Childs, A. Tersine, J. Sisson, E. Parker, V. Velkoff, C. Logan, and H. Shin. 2020. *Design and Operation of the 2020 Household Pulse Survey.* U.S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020_HPS_Background.pdf (accessed February 2022).

Fisher, M. 2021. "Eroding trust, spreading fear: The historical ties between pandemics and extremism." *The Washington Post.* Available at: https: //www.washingtonpost.com/-politics/pandemics-spawn-extremism/2021/02/14/d4f7195c-6bIf-1leb-ba56-d7e2c8de-fa31_story.html (accessed February 2022).

Groves, R.M., and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias a Meta-analysis." *Public Opinion Quarterly* 72: 167–189. DOI: https://doi.org/10.1093/poq/nfn011.

Japec, L., and L. Lyberg. 2020. "Big data initiatives in official statistics." In *Big Data Meets Survey Science: A Collection of Innovative Methods*, edited by C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, and L.E., Lyberg: 273–302. John Wiley & Sons. https://doi.org/10.1002/9781118976357.ch9.

Kaloskampis, I., C. Joshi, C. Cheung, D. Pugh, and L. Nolan. 2020. "Synthetic data in the civil service." *Significance Magazine* 17(6): 18–23. DOI: https://doi.org/10.1111/1740-9713.01466.

Kennedy, C., and H. Hartig. 2019. "Response rates in telephone surveys have resumed their decline." *Pew Research Center.* Available at: https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/ (accessed February 2022).

Kim, J.-K., and S.-M. Tam. 2021. "Data Integration by combining big data and survey sample data for finite population inference." *International Statistical Review* 89: 382–401. DOI: https://doi.org/10.1111/insr.12434.

Little, R.J.A., B.T. West, P.S., Boonstra, and J. Hu. 2020. "Measures of the Degree of Departure from Ignorable Sample Selection." *Journal of Survey Statistics and Methodology* 8(5): 932–964. DOI: https://doi.org/10.1093/jssam/smz023.

Lohr, S., and T. Raghunathan. 2017. "Combining survey data with other data sources." *Statistical Science* 32: 293–312. DOI: https://doi.org/10.1214/16-STS584.

Radermacher, W.J. 2020. *Official Statistics 4.0: Verified Facts for People in the 21st Century*. Berlin: Springer. Available at: https://link.springer.com/book/10.1007/978-3-030-31492-7 (accessed February 2022).

Rafei, A., C.A.C. Flannagan, and M.R. Elliott. 2020. "Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees." *Journal of Survey Statistics and Methodology* 8(1): 148–180. DOI: https://doi.org/10.1093/jssam/smz060.

Riccardi, N., and M. Schneider. 2021. "Alabama sues to stop redistricting delay, privacy initiative." *AP News.* Available at: https://apnews.com/article/alabama-redistricting-lawsuits-census-2020-courts-9627c66d67629942d2f961bf781f4309 (accessed February 2022).

Sakshaug, J.W., M.P. Couper, M.B. Ofstedal, R. David, and D.R. Weir. 2012. "Linking Survey and Administrative Records: Mechanisms of Consent." *Sociological Methods Research* 41(4): 535–569. DOI: https://doi.org/10.1177/0049124112460381.

U.S. Census Bureau. 2021. *2020 Census Data Products: Disclosure Avoidance Modernization.* Available at: https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html (accessed February 2022).

Williams, D., and J.M. Brick. 2018. "Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort." *Journal of Survey Statistics and Methodology* 6(2): 186–211. DOI: https://doi.org/10.1093/jssam/smx019.

Rejoinder by Yves Tillé

I would first like to thank the editors of the *Journal of Official Statistics*, who were kind enough to consider my article and who prompted this discussion with quality people. My objective was to share my point of view on the future of official statistics, on the challenges and to address the questions posed by new data sources and new methods. I would also like to thank all the discussants who took the time to read my article and who added extremely relevant comments.

Dr. Marc Debusschere develops the idea that there are conservatives and progressives in official statistics. I certainly do not think I am a conservative. I believe in scientific innovation and the need for further research in the field of official statistics. Statistics has changed a lot in the last few years to take into account new problems. It is in the best interest of official statistics to embrace these new methods. However, I completely disagree with the statement that archaic statistics would "lead to increasingly specialized and intricate methodologies". Let me explain. We know more or less what to do when processing a sample. There are standard procedures: treatment of questionnaire nonresponse by reweighting, imputation of item nonresponse, calibration and estimation. It is quite illusory to think that the introduction of new statistical methods will simplify the work of the statisticians. It will make it both more complex and richer. For these reasons, research and training will be the key issues of the future.

Concerning the use of new sources, I am however much more reserved. Administrative files should be used, but it is often difficult because they were not designed for statistical purposes. Data coming directly from the Internet suffer for the most part from several problems: a poor definition of statistical units, the impossibility of carrying out a production control procedure, and instability over time, which makes it very difficult to use them for official statistics. The purpose of my article was to warn against the illusion that new methods and new data will magically simplify the work of statisticians. On the contrary, more and more technical skills will be needed to meet these challenges.

Dr. Henri Luomaranta brings very interesting clarifications to my article. The references are also very important. Finland has been a forerunner in the use of administrative files. It is essential to continue in this way because the response rates are falling almost everywhere. For a citizen or a company, it is incomprehensible to have to give information that the State has already collected in one way or another. The reflections of Dr. Henri Luomaranta on the right compromise between speed and accuracy are also very relevant. I think that statistical institutes should not hesitate to publish statistics quickly, even if it means correcting them later.

Dr. Martin Axelson and Dr. Eva Elvers added very constructive comments. My point was to say that the interpretation of the induction/deduction cycle is far from being unanimous in statistics. Test theory can be considered as a deductive approach. The hypothesis is derived from a general model. We then reason about the hypothesis to

© Statistics Sweden

deduce what we should get if the hypothesis were true. We then confront this reasoning with the data to make a decision on the hypothesis. By following this view, Jerzy Neyman was therefore relatively reserved about the idea of inductive reasoning.

As Dr. Martin Axelson and Dr. Eva Elvers explain, inferential statistics can also be seen as inductive because it consists in extrapolating partial data. So one generalizes observations to a larger entity: a population or a model. This is another interpretation of the induction/deduction cycle. In their article, Alleva et al. (2021) also speak of an induction/deduction cycle. It seems that they consider inferential statistics as an inductive approach, but they do not explain clearly what they mean by deductive approach.

In sampling theory, we sometimes refer to the anticipated variance, which is the variance calculated both under the design and under a model. This theoretical variance makes it possible to evaluate the precision that would be obtained under a particular model with a given sampling design. This allows us to choose the best design for a model on the data. This is the reasoning developed in the article of Tillé and Wilhelm (2017) and would then be typically deductive. Then, data collection and analysis allows new hypotheses about the population to be formulated in order to continue optimizing the sampling procedure in the next wave. In this case, the induction- deduction cycle is not applied to build knowledge but only to establish an efficient sampling strategy according to the postulated model. My point was to say that the induction/deduction cycle is subject to many different interpretations in statistics. One cannot really say that there is a consensus about it.

Official statistics is in my opinion specific, not because of its methods, but because of its objectives, which are to produce quality reference statistics. The objective is therefore not to produce knowledge directly. I still think that official statistics is not intended to model all the data it produces. Nevertheless, there is no reason why statistical methods should be fundamentally different from other fields of statistics. For example, modeling can be extremely useful for imputing missing values. However, the goal is not really to theorize the data set but to fill in the holes as best as possible. To do this, one can obviously use very sophisticated methods to reconstruct the missing part of the data. In my opinion, statistical learning methods are not inductive, because they often predict well without really understanding and therefore do not allow theorizing. The example is particularly striking in the field of translation where current translators are very efficient but are not based on any theory of language. The understanding of "why it works" for statistical learning methods is a major issue in current research in the field.

I thank Dr. Anders Holmberg for his important comments and references. It is true that I have mainly talked about methods and data sources because this is my field of research. There are many other issues concerning official statistics: political, administrative, organizational, ethical. I am very much in favor of exploring all possibilities of using new data sources. However, my main point was to say that there will be nothing magical and nothing automatic. The new data sources all contain gaps. The integration of the various sources will require more and more technical skills. A general methodology for integration with multiple sources has yet to be built.

I would like to thank Prof. Richard Valliant, whose work I have always read with great interest, for his very relevant reflections that shed additional light on the future of official statistics. If I may summarize his thoughts: official statistics must face increasing

nonresponse, data integration problems, to produce statistics faster by taking advantage of new data sources while being attentive to confidentiality issues. These challenges can only be met with highly skilled statisticians. I fully share this view. Nonresponse is also increasing in Europe. This is a very worrying issue. There are only two options: reduce the statistical burden by using administrative data as much as possible, build trust between official statistics and the citizen.

I would like to add, however, that the situation in most European countries is somewhat different from that in the United States. Indeed, many European countries have been building population registers, sometimes for decades, with a compulsory declaration of domicile. In these countries, censuses have been abolished and a large part of the population statistics are based on registers. Of course, these registers are supplemented by surveys, but the registers form a reliable basis for comparison. Moreover, when social security systems cover the whole population, it is also possible to obtain reliable data on pensions and health without resorting to surveys. The valorization of administrative files is therefore easier in Europe than in the United States.

References

Alleva, G., E. Baldacci, and P.D. Falorsi. 2021. "Methodological, Technological and Design Challenges in the New Multisource Statistics Ecosystem." *Statistical Journal of the IAOS* 37: 245–254. DOI: https://doi.org/10.3233/SJI-210801.

Tillé, Y., and M. Wilhelm. 2017. "Probability sampling designs: balancing and principles for choice of design." *Statistical Science* 32(2): 176–189. DOI: https://doi.org/10.1214/16-STS606.

Received September 2020
Revised January 2021
Accepted February 2021