Uncertainty in production and communication of statistics: challenges in the new data ecosystem

L'incertezza nella produzione e comunicazione delle statistiche: le sfide nel nuovo ecosistema dei dati

Giorgio Alleva[1], Piero Demetrio Falorsi[2]

Abstract. In the paper we focus on how to measure and communicate to users the accuracy in the new data ecosystem, with estimates based on an integrated system of statistical registers, in order to explicitly consider the main sources of uncertainty that can influence the estimates. We propose feasible computational strategies and discusses several approaches useful for communicating accuracy to users.
Abstract. Nel paper ci concentriamo su come misurare e comunicare agli utenti l'accuratezza di stime basate su un sistema integrato di registri statistici, considerando esplicitamente le principali fonti di incertezza che possono influenzare le stime. Sono proposte strategie computazionali fattibili e discussi diversi approcci utili per comunicare l'accuratezza agli utenti.

Key words: Accuracy, Global Mean Squared Error, Data Integration.

1. Introduction

Statistics is the science of data and uncertainty. Uncertainty is the natural environment in which we operate as statisticians. In this paper we emphasize that Official statistics (OS) should make any effort to support users on how to evaluate the accuracy of the estimates on the parameters of interest also making them able to calculate the accuracy of parameters specified by themselves.

[1] Sapienza University of Rome, Piazzale Aldo Moro 5, IT-00185 Rome, Italy. Email for contact: giorgio.alleva@uniroma1.it.
[2] Former Director of Methodology at Istat and international consultant, piero.falorsi@gmail.com

Why is it crucial to communicate uncertainty in OS? The users should take their decisions fully aware of the uncertainty of estimates. The quality of decisions may suffer if decision makers incorrectly take reported statistics at face value or incorrectly conjecture error magnitudes.
We should be aware of the importance to accompany systematically our estimates with measure of accuracy: not only point estimates, but as much as possible intervals, associated with a predetermined level of confidence/credibility. We should overcome any fear of creating confusion among users or losing credibility. Of course, tailoring the communication on uncertainty in different ways and languages for different targets. Several studies have shown how good communication of uncertainty can be understood and appreciated by users (Van der Laan et al., (2015)).
Regardless of the source of the statistics, some principles for communicating uncertainty and change apply. Overall, OS should provide sufficient and appropriate information to allow users to judge the goodness of fit for their purpose; and to maintain and increase users' confidence in the estimates. For change estimates, the point is to provide the direction and size, level of uncertainty and the estimated trend. OS has long been equipped in communicating uncertainty of statistics and the Code of Practice include recommendations on this issue, in particular on the necessary transparency in communicating estimates and processes to calculate them. In the sample context, the accuracy and communication measures are more consolidated; also for statistics based on administrative sources OS developed guidelines and indicators. In the last review of the Code of Practice (Eurostat, (2018)) reference is also made to the accuracy in estimates based on the integration of different sources.

The remainder of the paper is organized as follows. In Section 2 the main taxonomies of uncertainty are presented and discussed. The measures of accuracy in different statistical and informative contexts are presented and discussed in Section 3. Conclusions are highlighted in Section 4.

2. Does the traditional taxonomies of uncertainty still make sense in OS?

In the literature there are interesting taxonomies of uncertainty, from which measurement and communication strategies have been developed.
A three-level categorizations have been provided over time by different authors for focusing on statistical modelling of risks (Diebold et al., (2010); Spiegelhalter, (2017)). *Aleatory uncertainty***:** the natural randomness in a process, fully expressed by classical probabilities. *Epistemic uncertainty***:** the scientific uncertainty about the structure and parameters of our statistical model of a process, expressed, for example, through Bayesian probability distributions, default parameter values, safety factors, and sensitivity analyses to assumptions (Morgan et al., (2009)). *Ontological uncertainty***:** unrecognised ignorance about the entire modeling process as a description of reality, or failure to comprehend unprecedented circumstances.

The paper of Manski (2015) provided a new significant boost to the measurement and communication of uncertainty. Starting from the traditional classification of errors in sampling and non-sampling ones, Manski identified three main sources of uncertainty for economic statistics, namely: transitory, permanent and definitional.
Transitory statistical uncertainty arises because data collection takes time. Agencies sometimes release a preliminary estimate of an OS in an early stage of data collection and revise the estimate as new data arrive (a typical example is GDP). *Permanent statistical uncertainty* derives from incompleteness or inadequacy of data collection that does not diminish with time. In survey research, considerable permanent uncertainty may stem from non-response and from the possibility that some respondents may provide inaccurate data. *Definitional uncertainty* arises from incomplete understanding of the information that OS provide about well-defined concepts or from lack of clarity in the concepts themselves. Thus, conceptual uncertainty concerns the interpretation of statistics rather than their magnitudes.
When communicating uncertainty, it is interesting to distinguish two fundamental levels of uncertainty (Van der Bles *et al*., (2019)). *Direct uncertainty* about the fact, number or hypothesis. This can be communicated either in absolute quantitative terms, say a probability distribution or confidence interval, or expressed relative to alternatives, such as likelihood ratios, or given an approximate quantitative form, verbal summary and so on. *Indirect uncertainty* in terms of the quality of the underlying knowledge that forms the basis for any claim about the event, number or hypothesis. This will generally be communicated as a list of warnings about the underlying sources of evidence, possibly blended into a qualitative or ordered categorical scale.

In the new ecosystem of OS, estimates are based on an integrated system of statistical registers fed in a systematic and continuous way by surveys, administrative archives and new sources. Does the distinction between transitional and permanent uncertainty still make sense? Doesn't the revision process typically planned for national accounts during the progressive consolidation of sources now represent the way forward for any statistics produced through an integrated system of registers? How to communicate this sort of continuous review to users? Does it still make sense to distinguish between direct and indirect uncertainty? We have some doubts. In anyhow, an integral part of the challenge of producing OS based on an integrated system of statistical registers is the measurement and communication of uncertainty.
The approaches for communicating uncertainty in OS are very different depending on the type of data and information contexts of the dissemination. Leaving out uncertainty is current practice for most of the contexts. De Jonge (2020), underlines that communicating uncertainty in *statistical visualizations*, even if it is not the most widespread practice, adds value to users and clarifies that statistical offices produce statistics. The main approaches for tabular data deriving from survey sampling is that of avoiding dissemination of very unreliable figures, or making the users aware of the uncertainty with specific graphic signals (e.g. an asterisk). Special studies on non-coverage or measurement errors are used for Census data or register-based statistics. Alleva *et al*. (2021) suggest a feasible calculation strategy for register-based statistics allowing a dynamic calculation of the Global Mean Squared Error (GMSE) which

could allow to release the statistics along with the related GMSE, thus improving the relevance, transparency and confidence of official statistics.

3. Measuring the uncertainty

Given the need to compute the statistical errors in disseminated data, it is necessary to determine the measure of accuracy to be calculated and communicated to users. To make it simple, we introduce this topic for the total $Y = \sum_{k \in U} y_k$, of the variable y within the population U, where y_k is the true value of the variable y for unit k. Let $\hat{Y}$ be the estimation of Y. There are multiple sources of error (ranging from sampling errors to coverage errors, etc.). Each specific approach to inference focuses on different sources of variability and bias in the definition of the measure of accuracy; these are related to what is treated as fixed or random in the specific inferential approach. For instance, the *design based* (Cochran, (1977)) or the *model assisted* approaches (Särndal *et al.,* (1992)) treat the population values y_k as unknown constants and the sample selected, with the sample design P, is the only source of randomness; therefore, they develop their inference considering only the variability of the sampling design. The *model-based* approach (Chambers and Clark, (2015)) considers the sample as *fixed* and the y_k values as random variables generated according to the model, M.

If the methodology embedded in the estimator is transparent and does not introduce bias in the estimates, the main advice is to compute at least the leading components of the errors: sampling variance, model variance, or both. *Sampling variance*, which measures the uncertainty deriving from the randomness of the observed set of data, is an adequate measure of accuracy when the construction of statistical indicators is based on the inferential properties of repeated sampling. It may be defined as $V_P(\hat{Y}) = E_P\left[\hat{Y} - E_P(Y)\right]^2$, where E_P and V_P denote the operators of expectation and variance under repeated sampling. *Model variance* is a suitable measure of accuracy when the construction of statistical indicators is based on models using x-auxiliary variables, generating the value of the target variable y for the units in the population. Model variance may be defined as $V_M(\hat{Y}) = E_M\left[\hat{Y} - E_M(Y)\right]^2$, where E_M and V_M denote the operators of expectation and variance under the model M generating the data. However, some statistical indicators can be obtained via statistical procedures that utilize model-based approaches jointly with inference based on sampling design. For these cases, it is suggested to consider *global variance*, $GV_M(\hat{Y}) = E_P E_M\left[\hat{Y} - E_P E_M(Y)\right]^2$ (Wolter, (1985)) as the measure of accuracy.

When statistical data are produced from the census or administrative records, it is also necessary to consider the bias in measuring accuracy. Bias generally derives from the measurement error (based on statistical models) and the coverage error, the latter deriving from erroneous inclusion in the observation of elements extraneous to the population of interest (over-coverage) or from incorrectly excluding certain units from the target population (under-coverage). These types of error can be detected with

special observational techniques (based on double and independent measurements), which may however be costly. In the case of OS, the techniques are implemented only in certain specific cases. Alleva et al. (2021) propose the GMSE as a more general measure of accuracy: $GMSE(\hat{Y}) = E_P E_M(\hat{Y} - Y)^2$.
The GMSE includes, as particular cases, the bias and the measure discussed above (the GV, sampling and model variance). Moreover, the GMSE is an extension of the well-known Mean Squared Error (Biemer, (2010)) taking into consideration all the random components involved in the inferential process for computing the statistics. For instance, we may consider the non-response by defining GMSE as: $GMSE(\hat{Y}) = E_P E_M E_{NR}(\hat{Y} - \hat{Y})^2$, in which E_{NR} indicates the expectation under the models adopted for imputing the non-response in survey data. The GMSE could be accepted as a measure of precision by the main professional families of methodologists within the National Statistical Offices (NSOs): at least those who base their inference only on statistical models and those who use the statistical models as a support for inference which continues to be based essentially on sampling design. The global measure has a number of advantageous qualities, including the following: generality, stability over time and robustness in the case of model failures. GMSE is simple to use and to communicate to users. It is based on the first and second moments of the random distributions of the specific source of uncertainty. Its calculus does not imply the full knowledge of the underlying distributions.

As far regards the measurability of the accuracy, we note that Statistical data may be the result of different statistical surveys where, according to the statistical quality framework followed by Statistics Canada (2009), the term survey includes the following components: (i) a census, which attempts to collect data from all members of a population; (ii) a sample survey, in which data is collected from a (usually random) sample of population members; (iii) a collection of data from administrative records, in which data is derived from records originally kept for non-statistical purposes; (iv) a derived statistical activity, in which data is estimated, modelled, or otherwise derived even integrating a multiplicity of existing statistical data sources. Each of the previous components introduces different sources of uncertainty that should be considered when informing the users on the accuracy.

For instance, the component (i) introduces a possibility of coverage errors, which we can deal with specific statistical models. In this case, the adequate measure of accuracy is the GMSE, which incorporates the coverage-bias. An interesting example of how measuring the components of errors in the GMSE is given in Daddi, *et al.* (2021) for the Italian Census Population Coverage Survey (PCS) carried out each year as a component of the Permanent Census Survey System.

The component (ii) includes the sampling errors. According to the inferential process adopted for the predictions, either the sampling variance or the model variance may be adequate. These can be computed with the standard statistical techniques.

The components (iii) and (iv) comprise the uncertainty derived by models adopted for building the predictions at the unit level. Alleva *et al.* (2021) propose computing the GMSE by adopting an approach based on linearization techniques. Scholtus (2019) proposes an approach for computing the GMSE based on replication methods.

4. Conclusions

Official statistics should make any effort to support users on how to evaluate the accuracy of the estimates on the parameters of interest. We should overcome any fear of creating confusion among users or losing credibility. In the new ecosystem of OS, with estimates based on an integrated system of statistical registers (ISSR) fed in a systematic and continuous way by surveys, administrative archives and new sources, we propose the GMSE to take into account a plurality of sources of uncertainty. A strategic choice is whether to make the use of ISSR limited and allow the dissemination of only planned outputs having a certified accuracy or make the system more flexible for the users allowing different users to produce their own statistics from the ISSR. We suggest to opt for the second option which makes OS more relevant for its users but which obliges the NSOs to impose a policy for reducing the risk of inappropriate use of the data.

References

1. Alleva, G., Falorsi, P. D., Petrarca, F., Righi, P. Measuring The Accuracy Of Aggregates Computed From A Statistical Register, Journal of Official Statistics. Accepted for publication (2021)
2. Biemer, P.P.: Total Survey Error Design, implementation, and evaluation. Public Opinion Quarterly, Vol. 4, No. 5, pp. 817-848. (2010)
3. Chambers, R.L., Clark, R.G.: An Introduction to Model-Based Sampling with Applications. Oxford Statistical Science. 37. (2015)
4. Cochran, W. G.: Sampling techniques. Third Edition. New York Wiley. (1977)
5. Daddi S., Falorsi P.D., Fiorella E, Massoli P., Righi P., Terribili. M.D. Optimal sampling for the Population Coverage Survey of the new Italian Register Based Census. *Journal of Official Statistics*, Accepted for publication (2021)
6. De Jonge, E.: Communicating uncertainties in official statistics. A review of communication methods, European Commission. (2020)
7. Diebold, F.X., Doherty, N.A.,Herring, R.J.: The Known, the Unknown, and the Unknowable in Financial Risk Management: Measurement and Theory Advancing Practice. Princeton, NJ: Princeton Univ. Press. (2010)
8. Eurostat: European statistics Code of Practice, Publ. Office of the EU (2018)
9. Manski, C.: Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. J Econ Lit, 53:631–653. (2015)
10. Morgan, G., Dowlatabadi, H., Henrion. M., Keith, D., Lempert, R.: Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Climate Decision Making. Silver Spring, MD: Natl. Ocean. Atmos. Organ. (2009)
11. Morgan, G., Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R.: Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Climate Decision Making. Silver Spring, MD: Natl. Ocean. Atmos. Organ (2009)
12. Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer-Verlag (1992).
13. Scholtus, S.: A bootstrap method for estimators based on combined administrative and survey data. NTTS Conference 2019 (2019)
14. Spiegelhalter, D.: Risk and Uncertainty Communication, Annu. Rev. Stat. Appl. 2017.4:31-60 (2017)
15. Statistics Canada Statistics Canada Quality Guidelines, 5th edn. (2009)
16. Van der Bles, A.M., Van der Linden, S., Freeman, A.L.J., Mitchell, J., Galvao, A.B., Zaval, L., Spiegelhalter, D.J.: Communicating uncertainty about facts, numbers and science, R. Soc. open sci. 6: 181870. http://dx.doi.org/10.1098/rsos.181870 (2019)
17. Wolter, K.M.: Some Coverage Error Models for Census Data. Journal of the American Statistical Association, 81, 338 - 346 (1986)