Output Quality of Multisource Statistics

Danutė Krapavickaitė*
Statistics Lithuania; Vilnius Gediminas Technical University, Vilnius, Lithuania – danute.krapavickaite@stat.gov.lt

Arnout van Delden
Statistics Netherlands, the Hague, the Netherlands – a.vandelden@cbs.nl

Sander Scholtus
Statistics Netherlands, the Hague, the Netherlands – s.scholtus@cbs.nl

Ton de Waal
Tilburg University; Statistics Netherlands, the Hague, the Netherlands – t.dewaal@cbs.nl

Abstract

The problem of data integration from various sources becomes more and more important in official statistics because of the growing needs for statistical information and limited resources to obtain it. The quality of statistical output depends on the statistical methods used to combine the data and on the quality of the data sources used. This paper is devoted to the Work Package 3 of the first Specific Grant Agreement of the ESSnet project Komuso. Its aim is to examine the accuracy aspect of the statistical output depending on the input. The work consists of critical reviews and suitability tests for existing methods and currently proposed methods to estimate the accuracy of statistical results.

Keywords: quality, accuracy, bias, variance, data configuration.

1. Introduction

The accuracy of statistical results in official statistics depends on various factors. The errors of such kind of statistical results depend on:

- the quality of the frames used: unit specification errors, frame undercoverage and overcoverage;
- data content errors;
- missing values in the data sets, nonresponse and methods to overcome it;
- the statistical methods used: model assumptions errors, sampling errors;
- data aggregation errors arising when statistical results are obtained by combining data from multiple sources.

The way to measure the quality of statistical results is to do it through the measurement of the input and statistical process quality. It is considered that high quality input and high quality processing should guarantee high quality statistical output ([1]). The total survey error should be decomposed into components identifying the main sources of error in the statistical process, and their contribution to the total error should be described. The decomposition of the mean squared error (MSE) measure into different sources of bias and variance gives a tool for studying the effects of e.g. measurement errors and frame bias on the precision of estimates.

The aim of this paper is to present a project devoted to collecting the methods for the estimation of the accuracy of statistical results due to frame errors in the case of multisource statistics.

2. Study basis and approach

The ESSnet project on the quality of multisource statistics Komuso, led by Statistics Denmark, is part of the ESS.VIP Admin Project. The main objectives of the ESS.VIP Admin Project are (i) to improve the use of administrative data sources and (ii) to support the quality assurance of the output produced using administrative sources. The first Specific Grant Agreement (SGA1) of Komuso, which started in January 2016 and lasted until April 2017, consisted of four Work Packages (WPs):

- Evaluating the quality of input data (WP 1; leader – Statistics Denmark);
- Methodology for the assessment of the quality of frames for social statistics (WP 2; leader – Statistics Norway);
- Framework for the quality evaluation of statistical output based on multiple sources (WP 3; leader – Statistics Netherlands);
- Communication with respect to the ESSnet (WP 4; leader – Statistics Hungary).

The work done under the WP 3 is presented in this paper. The aim of the WP 3 is to accumulate a collection of methods for the estimation of the accuracy of statistical results obtained using multiple data sources – administrative and sample data. A collection of methods means an overview of methods existing in the literature and in the practical work of the national statistical institutes (NSIs); development of new ideas, proposal of new methods for estimating the accuracy of statistical results based on multiple data sources. Critical reviews of the accuracy estimation methods presented in the literature are done and suitability tests of the application of existing and newly proposed accuracy estimation methods are carried out.

The statistical result is considered as a random variable; its accuracy is estimated by the mean squared error, or another accuracy measure is proposed. The contents of the project are concentrated on the output error terms arising due to the errors in the data sources used. Therefore, the methods proposed are not aimed to take into account other error sources. The backbone classification of the output accuracy estimation methods of multisource statistics is done by data configuration. The following basic data configurations (BDC) are considered:

- BDC 1: Multiple non-overlapping cross-sectional microdata sources that together provide a complete data set without any undercoverage problems;
- BDC 2: Same as BDC 1, but with overlap between different data sources;
- BDC 3: Same as BDC 2, but with undercoverage of the target population;
- BDC 4: Microdata and aggregated data that need to be reconciled with each other;
- BDC 5: Only aggregated data that need to be reconciled;
- BDC 6: Longitudinal data sources that need to be reconciled over time (benchmarking).

BDC 1 can be subdivided into two cases: the split-variable case, where the data sources contain different variables, and the split-population case, where the data sources contain different units. Literature reviews and suitability tests are carried out for all identified BDCs. For more information on the results of SGA1, BDCs and methods of producing multisource statistics, refer to the SGA1 final report [4]. The report on the work under WP3 is prepared by the NSIs of Austria, Denmark, Italy, Lithuania, the Netherlands, and Norway. Examples for each data configuration are presented further.

3. Some results

BDC 1

Errors in the classification of businesses attracted attention of several project partners.

Accuracy of growth rates due to classification errors. The study of Statistics Netherlands is devoted to the estimation of the bias and variance of quarterly and annual growth rates under the business activity classification errors. The ideal unknown NACE code of the enterprise is considered as true, and the enterprise activity code in the business register fixed for one year is considered as possibly erroneous. Let U denote a target population of units, which is an enterprise population in the case study. Suppose that two data sets are observed, where the first data set contains a variable y^r for all

units of a subpopulation $U^r \subset U$, and the second data set contains a variable y^q for all units of a subpopulation $U^q \subset U$. For units in the intersection $U^{r,q} = U^r \cap U^q$, both variables y^r and y^q are available. It is assumed that this intersection is relatively large, i.e. that the two subpopulations have a large overlap. The situation may arise in a repeated survey of the population which is changing over time.

Suppose that both subpopulations are divided into strata, where the set of possible stratum codes is denoted by $\{1, \dots, M\}$. They are NACE codes in this study. Let s_i^r be the true stratum of unit $i \in U^r$ in the first data set, and s_i^q the true stratum of unit $i \in U^q$ in the second data set (which may be different from s_i^r, for instance, because the two data sets refer to different points in time). Let the indicator $a_{hi}^r = 1$ if $s_i^r = h$ and 0 otherwise, and similarly let $a_{hi}^q = 1$ if $s_i^q = h$ and 0 otherwise. Let Y_h^r be the total of variable y^r in the stratum h, and Y_h^q the stratum total for the variable y^q, with $Y_h^r = \sum_{i \in U^r} a_{hi}^r y_i^r$ and $Y_h^q = \sum_{i \in U^q} a_{hi}^q y_i^q$. The statistic of interest is the ratio $R_h^{q,r} = Y_h^q / Y_h^r$. Unfortunately, the classification of units into these strata is prone to errors, and, instead of s_i^r and s_i^q, $\hat{s}_i^r$ and $\hat{s}_i^q$ are observed, which may contain errors. Let $\hat{a}_{hi}^r = 1$ if $\hat{s}_i^r = h$ and 0 otherwise, and similarly let $\hat{a}_{hi}^q$ be the observed version of a_{hi}^q. The stratum totals and their ratio is estimated by $\hat{Y}_h^r = \sum_{i \in U^r} \hat{a}_{hi}^r y_i^r$, $\hat{Y}_h^q = \sum_{i \in U^q} \hat{a}_{hi}^q y_i^q$, and $\hat{R}_h^{q,r} = \hat{Y}_h^q / \hat{Y}_h^r$ respectively.

In order to derive an analytic expression for the approximate bias and approximate variance of the estimated ratio $\hat{R}_h^{q,r}$, the second-order Taylor expansion is used for this estimator. The case of quarter-on-quarter growth rate within the *same* year is relatively simple because the enterprise register is fixed for the same year. The case of quarter-on-quarter growth rate for different years' quarters and the annual growth rate is more complicated.

Some assumptions concerning the origin of the classification errors need to be made. The classification errors in $\hat{s}_i^{q-1}$ (and in $\hat{s}_i^q$ for units that occur only in U^q) are described by the level matrix $\mathbf{P}_i^{OL} = (p_{ghi}^{OL})$, with the elements $p_{ghi}^{OL} = P\left(\hat{s}_i^{q-1} = h \middle| s_i^{q-1} = g\right)$.

Approximations for the bias and variance of the industry growth rates due to errors in the industry classification code are obtained analytically. The level matrix should be estimated. A simulation study is carried out.

Accuracy of the domain totals due to enterprise activity classification errors. As presented by Statistics Denmark, the employment rate for the wage labour force statistics (ERWLF) gives the total number of employees and the number of employees by NACE section. It is a case of register-based statistics. ERWLF is based on the so-called e-income register, population register and business register. The purpose of such statistics is to create fast and efficient indicators rather than precise figures of the Danish labour force. A wrong classification of businesses affects all statistics based on the BR. Statistics Denmark carried out a suitability test in order to examine the quality of the ERWLF as a function of the quality of the BR. The test was intended to reveal the error margin of the number of wage labourers within activity groups due to wrong NACE classifications in the BR. The test was carried out by simulations, where the distributions of activity codes were simulated and the effect on published figures from the ERWLF was examined.

Accuracy of the estimator of the total due to changes in enterprise composition and classification. This study is carried out by Statistics Lithuania. Let us suppose that all statistical units (enterprises) are registered in a business register (BR). This BR is used as a sampling frame for a stratified simple random sample without replacement, where the strata are formed by the NACE code for economic activity and size (number of employees). The sampled units are used to estimate the overall population total for a variable y. After the survey data of enterprises from the selected sample had been obtained, information was received from some of the respondents that the sampling units (SU) had changed their characteristics:

a) the SU was joined with other enterprises, possibly from different strata;
b) another value for the classification variables (for example, NACE code or size group) of the SU was reported.

These changes may occur because of errors in the classification variable taken from an administrative data source or because of changes in the population which occurred between the sample selection and data collection. It is assumed that all information about the enterprise changes is known.

The initial population U is evolved to the population U' at the time of the observation, with the variable y replaced by the variable y', and the population total $t_y = \sum_{k \in U} y_k$ replaced by the population total $t'_y = \sum_{k \in U'} y'_k$. The selected sample ω, $\omega \subset U$ is replaced by the observed sample ω', $\omega' \subset U'$. The problems mentioned above are solved in the following ways:

a) first- and second-order inclusion probabilities are calculated for the sample ω';
b) a model for the size group or NACE code changes is assumed.

The Horvitz–Thompson estimator for the totals and variances of the estimates is applied. Relative biases and relative variances are used as accuracy measures, as opposed to the case when changes in the population are not taken into account. A simulation study is carried out. Simulation results show that the relative bias and the relative variance for the estimator of the total are increasing with the increasing size of changes in the observed population.

BDC 2

Several microdata sources with undercoverage have been associated with social studies. The examples for BDC 3 and BDC 4 are also directed to social statistics.

Quality framework for register-based statistics to derive a quality indicator. Statistics Austria presents a quality framework, which has been developed for a register-based census 2011. It was a full enumeration from several administrative data sources.

Each source has to deliver data at the micro level. The central population register is considered as the population basis, which has no undercoverage. The data sources are also overlapping with respect to the units and to the variables. A procedure for the quality evaluation of statistical output starts from the assignment of quality indicators for every variable in every register used for input. These quality indicators are quantitative functions with values in the interval (0,1). A higher value of the quality indicator means higher quality of the variable. Three quality dimensions are evaluated at the starting step: documentation, preprocessing, and external source. Some of the variables are unique for a data source, but some of the variables occur in multiple data sources. When the initial data sets are merged into the Central Database (CDB), the quality indicators for the multiple variables are combined using the Dempster–Shafer theory. Some of the variables are derived from other variables, imputations are done, and quality indicators for each imputed value are calculated. In a further step, the CDB is compared to an external source – to check for the quality changes during data combining, and a final quality indicator is derived.

The quality framework demonstrates changes in quality during data processing in the register-based census 2011. Although developed for a population census, the method can also be applied for any register-based statistical procedure. The assumptions for the application of the quality framework include independence among administrative data sources, the possibility to link them by a unique key-variable at a unit level, the reference day should be close for all data sets and so on. An analysis of the complexity of the method and time usage is presented.

Accuracy of the observed data due to measurement errors. Statistics Italy has examined multiple administrative and survey sources that provide the value of the same variable of interest for the entire target population or part of it, and all measurements are assumed to be imperfect. In this case, an approach based on a latent class model can be used to estimate the true values. In this approach, the estimates of the probabilities $P(Y^g = i \mid X = i)$, where Y^g is the observed value in the data source g and X is the true (latent) value, can be used to evaluate the accuracy of the data source g. In this approach, accuracy measures are naturally provided by the conditional distribution of the latent true variable given the available information (e.g. posterior variance). The goal of the study is to combine administrative and survey data in order to build a "labour register" to be used for producing estimates on the employment status at a detailed level.

BDC 3

Sensitivity analysis of the population size estimates using capture-recapture models ([3]). Let us suppose two registers I and II of the same population are linked. Some elements are included in both of them; denote their number by m_{11}. Some elements are included in the first register but not in the second one (m_{10}); some elements are included in the second register but not in the first one (m_{01}). Based on that, the number of elements of the population which are not included in any of the registers, m_{00}, can be estimated by $\hat{m}_{00} = m_{10}m_{01}/m_{11}$. Assumptions for this estimator should be made:

1) The inclusion of the element in register I is independent on its inclusion in register II;
2) Inclusion probabilities of the element in at least one of the registers are homogeneous;
3) The population is closed;
4) It is possible to link the elements of registers I and II perfectly.

The first and the second assumptions are usually violated in human populations. This violation should influence the accuracy of the population size estimates obtained. The approach taken by the authors is to use covariates, whose levels have heterogeneous inclusion probabilities in both registers. The loglinear models can be fitted to the contingency table of the inclusion indicators in registers I and II and the auxiliary variable – covariate. The independence assumption of an element to be included in registers I and II is then replaced by the weaker assumption of the conditional independence of the element to be included in register I and register II conditionally on the covariates available. The paper Gerritse et al. (2015) presents a study of the impact of the violation of the independence assumption on the accuracy of the population size estimates. The known level of dependence between the inclusion probabilities in the two registers is created, and the estimates of the population size under an assumption of independence are obtained. The results are compared to the results without any additional inclusion of dependence. The sensitivity of the population size estimates to the violation of the independence assumption is studied by simulation in such a way.

BDC 4

Variance estimation for the estimator obtained by repeated weighting ([5]). Statistics Netherlands has implemented a repeated weighting (RW) estimator in its regular estimation process for the Population Census. This estimator ensures numerical consistency among tables estimated from different surveys. Especially when the tables have some variables in common, this approach appears to be very useful. After a concise summary of the repeated weighting procedure, Knottnerus and Van Duin (2006) give the variance formulas for the repeated weighting estimator. They also give an example from the Dutch Labour Force Survey. First, the set of target tables to be estimated is specified. Next, all margins of a target table are added to the set of tables to be estimated. A marginal table is obtained by (i) aggregating over one or more categorical variables of a multi-way table or (ii) using a less detailed classification of a categorical variable.

In the second step, each table is estimated by means of the regression estimator from the most appropriate data set. The accuracy measure presented – estimation of the variance of reweighted totals.

BDC 5

The methods for BDC 5 and BDC 6 consider operations under aggregated data and can be used in national accounts and in combining other aggregated indicators.

Scalar uncertainty measure for an accounting equation. There are many indicators in official statistics which should satisfy some constraints summing them up (or multiplying them) to some balancing variable, whether fixed or random. Theoretically, these restrictions should be satisfied (for example, turnover for four quarters should sum up to annual turnover). Practically, these indicators can be estimated in various ways, and they have to be benchmarked in order to satisfy a constraint equation, which is also called an accounting equation. The uncertainty measure for the accounting equation is based on a variance-covariance matrix of the estimators for all adjusted estimators and a balancing variable. This approach is taken in the example for BDC 6. The authors of the joint work

done by Statistics Netherlands and Statistics Norway propose an alternative scalar measure of uncertainty for the accounting equation.

Let $Y_1,...,Y_p,Z$ be statistical variables which should satisfy the equation $f(Y_1,...,Y_p,Z)=0$ for some aggregation function f, for example, $Y_1+...+Y_p=Z$. Unfortunately, the values of these variables are unknown, they are estimated by $\hat{Y}_1,...,\hat{Y}_p,\hat{Z}$ and, subsequently, $f(\hat{Y}_1,...,\hat{Y}_p,\hat{Z})\neq 0$. Then the values of the estimates $\hat{Y}_1,...,\hat{Y}_p,\hat{Z}$ should be adjusted, e.g. replaced: $\hat{Y}_1\to\tilde{Y}_1,...,\hat{Y}_p\to\tilde{Y}_p,\hat{Z}\to\tilde{Z}$, in order to satisfy the accounting equation $f(\tilde{Y}_1,...,\tilde{Y}_p,\tilde{Z})=0$. A scalar uncertainty measure for the estimated account is defined by the authors as

$$\Delta A=E\delta,\quad \delta=\sum_{k=1}^{p}w_k\left|\tilde{Y}_k-E\tilde{Y}_k\right|^{\alpha}+w_{p+1}\left|\tilde{Z}-E\tilde{Z}\right|^{\alpha},$$

where w_k are some positive weights and $\alpha=1;2$. The higher the value of the measure ΔA, the more uncertain is the estimated account. The properties of this uncertainty measure are studied. The uncertainty measure can be used for the comparison of several adjustment methods applied to the same accounting equation.

BDC 6

Covariance matrix for a reconciled low frequency and high frequency aggregated data set ([2]). Let us suppose that low frequency aggregated data of high accuracy are available and are considered further as fixed. For example, annual estimates of some indicator. High frequency aggregated data are available from another source, and sums over their subsets should coincide with the elements of the low frequency data. For example, the sums of quarterly indicators should be equal to the values of annual indicators. Here are the data sets obtained over time. The problem is to replace the high frequency data with new values which satisfy the restrictions of summation to the low frequency data and differ from the initial high frequency data as little as possible in the sense of a quadratic distance function. The procedure used is called a reconciliation procedure and is formulated as a quadratic optimisation problem with linear restrictions.

The dependence of the accuracy of the reconciliation result on the accuracy of the input data is studied. The variance-covariance matrix (or vector of variances) is taken as a measure of accuracy (Bikker et al. (2011)). Several versions of the Denton method presented in the paper show the same: dependence of the covariance matrix of the reconciled data set on the covariance matrix of the initial data set.

4. Conclusions

The results of the work done demonstrate a sequence of ways to estimate the output accuracy of statistical results depending on the input quality. More methods can be found to satisfy specific output quality estimation needs. On the other hand, the methods studied under the SGA1 should be adopted to practical situations to be applicable to the routine work of the NSIs. The last task is the aim of the SGA2.

References

1. Agafiţei, M.; Gras, F.; Kloek, W.; Reis, F.; Vâju, S. (2015). Measuring output quality for multisource statistics in official statistics: Some directions. *Statistical Journal of the IAOS*, 31, pp. 203–211. DOI 10.3233/SJI-150902. http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji902

2. Bikker R., Daalmans J., Mushudiani N. (2011) Macro Integration. Data reconciliation. *Statistical methods* (201104). The Hague/Heerlen, Statistics Netherlands.

3. Gerritse, S., P.G.M. van der Heijden, B.F.M. Bakker (2015), Sensitivity of Population Size Estimation for Violating Parameter Assumptions in Log-linear Models. *Journal of Official Statistics*, 31, pp. 357-379.

4. European Comission. *ESSnet on quality of multisource statistics – Komuso*.

https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en

5. Knottnerus, P. and C. van Duin (2006), Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics 22*, pp. 565–584.