Spanish Journal of Statistics
Vol. 2 No. 1 2020, Pages 79–107
doi:https://doi.org/10.37830/SJS.2020.1.05

Regular Article

Commonly used methods for measuring output quality of multisource statistics

Ton de Waal[1], Arnout van Delden[2], Sander Scholtus[3]
[1]Statistics Netherlands and Tilburg University, t.dewaal@cbs.nl
[2]Statistics Netherlands, a.vandelden@cbs.nl
[3]Statistics Netherlands, s.scholtus@cbs.nl

Received: November 4, 2020. Accepted: March 3, 2021.

Abstract: Estimation of output quality based on sample surveys is well established. It accounts for the effects of sampling and non-response errors on the accuracy of an estimator. When administrative data are used or combinations of administrative data with survey data, more error types need to be taken into account. Moreover, estimators in multisource statistics can be based on different ways of combining data sources. That partly affects the methodology that is needed to estimate output quality. This paper presents results of the ESSnet project Quality of Multisource Statistics that studied methods to estimate output quality. We distinguish three main groups of methods: scoring methods, (re)sampling methods and methods based on parametric modeling. Each of those is split into methods that can be used for both single and multisource statistics and methods that can be applied to multisource statistics only. We end the paper by discussing some of the main challenges for the near future. We argue that estimating output quality for multisource statistics is still more an art than a technique.

Keywords: bias, bootstrap, coherence, data integration, parametric modeling, quality framework, sampling theory, variance

MSC: 62D05, 62D10, 62F40, 62H12, 62P20, 62P25

1 Introduction

The fundamental reason for existence of National Statistical Institutes (NSIs) is that they are responsible for the official figures for policy making. It is therefore crucial that NSIs produce reliable estimates of societal phenomena. That implies that NSIs should be able to monitor the quality of the output that they produce. Estimation of output quality for single source statistics as a function of sampling error is well established. When administrative data sources are used, or a combination of administrative and survey data, more error types, such as measurement and linkage errors, need to be estimated and taken into account. How the effects of those error types can be estimated, the

methodology, will partly depend on how the data are combined. One example is that a target variable with measurement error is available at micro-level in multiple sources. Another example is that estimates of primary statistics, again with measurement error, are reconciled into an integrated set of values that fulfill balancing equations. One needs different methods to measure the output quality in these two situations.

Estimation of output quality of multisource statistics has been studied in the ESSnet project Quality of Multisource Statistics (also referred to as Komuso). The Komuso project lasted from January 2016 until October 2019. It was part of the ESS.VIP Admin Project. The main objectives of that latter project were: (i) to improve the use of administrative data sources, and (ii) to support the quality assurance of output produced using administrative sources. Partners in Komuso were Statistics Denmark (overall project leader of the ESSnet), Statistics Norway, ISTAT (the Italian national statistical institute), Statistics Lithuania, Statistics Austria, the Hungarian Central Statistical Office, the Central Statistical Office of Ireland, and Statistics Netherlands.

The main aim of Komuso was to produce quality guidelines for NSIs that are specific enough to be applied in statistical production by those NSIs. These guidelines take the entire production chain into account (input, process, and output) and cover a variety of situations in which NSIs work: various error types and different basic data configurations (BDCs, see Subsection 2.2). The guidelines list a variety of potential indicators/measures, indicate for each of them their applicability and in what situation they are preferred or not, and provide an ample set of examples of specific cases and decision-making processes.

Work Package 3 (WP 3) of Komuso focused on measuring the quality of statistical output based on multiple data sources. Measuring the quality of statistical output differs fundamentally from measuring the quality of input data since one ideally wants to take into account all processing and estimation steps that were taken to achieve the output. The problem encountered in WP 3 was not so much how to define the quality measures, but rather how these quality measures should be computed for a given set of input datasets and a certain procedure for combining these input datasets. At the moment, there is no all-encompassing theory or framework that can be used as a basis for quality measures for multisource statistics and for methods to calculate such measures. Constructing quality measures for multisource statistics and calculating them is still more of an art than a technical recipe that one can simply follow.

The present paper concentrates on methods to compute output quality measures. Those quality measures and their computational methods were examined and described in WP 3 of Komuso. They form an appendix to the above-mentioned quality guidelines which were developed in WP 1 of Komuso. Section 2 describes the approach taken in WP 3 of Komuso. Section 3 focuses on scoring methods for measuring output quality, Section 4 on methods based on (re)sampling, and Section 5 on methods based on (parametric) modelling. Each of these sections is split into two parts: methods that can be used to measure output quality for single and multisource statistics, and methods that have been developed for multisource statistics only. Section 6 concludes this paper with a brief discussion.

Due to the large variety in situations and methods we consider in this paper, the notation varies slightly over the various (sub)sections. We hope that this will not confuse the reader.

2 Approach taken in Komuso

In WP 3 of Komuso the work was subdivided into three consecutive steps:

1. In the first step a literature review or suitability test was carried out. In a literature review existing quality measures and recipes to compute them were studied and described. In a suitability

test also data were used to test quality measures and the recipes to compute them. Suitability tests were mainly used for newly proposed quality measures, but also for some already known quality measures to learn more about their properties, or for already known quality measures that were applied to a new field. In such a suitability test, practical and theoretical aspects of a quality measure and the accompanying calculation recipe were examined.
2. In order to make the results of Step 1 easily accessible, in Step 2 so-called quality measures and computation methods (QMCMs) were produced. Such a QMCM is a standardized, short description of a quality measure and the accompanying calculation recipe as well as a description of the situation(s) in which the quality measure and accompanying recipe can be applied. In total, 32 QMCMs were developed in the Komuso project.
3. In Step 3 hands-on examples were developed in the Komuso project for 31 of the 32 QMCMs. The one exception for which no example was provided concerned a general description of error types.

In order to cover different situations of different NSIs, and for ease of finding the results, the quality measures were structured along five classifications:

- quality dimensions;
- BDCs;
- error types;
- general approaches;
- computational methods.

The first four classifications are discussed in Subsections 2.1 to 2.4. The fifth classification is discussed extensively in Sections 3 to 5.

2.1 Quality dimensions

WP 3 of Komuso focused on four quality dimensions: accuracy, timeliness, coherence and relevance. The selected quality dimensions can more or less be quantified. *Accuracy* is "the degree of closeness of computations or estimates to the exact or true values that the statistics were intended to measure" (Eurostat, 2014). *Timeliness* was operationalized as "the time lag between the date of the publication of the results and the last day of the reference period of the estimate of the event or phenomenon they describe" [Komuso (ESSnet Quality of Multisource Statistics) (2019), in line with Eurostat (2014)]. *Coherence* "measures the adequacy of statistics to be combined in different ways and for various uses" (Eurostat, 2014). *Relevance* is defined as "the degree to which statistical outputs meet current and potential user needs" (Eurostat, 2014). "It refers to whether all the statistics that are needed are produced and the extent to which concepts used (definitions, classifications etc.) reflect user needs" (Komuso (ESSnet Quality of Multisource Statistics), 2019).

2.2 Basic data configurations

As already mentioned above, WP 3 of Komuso used a breakdown into a number of BDCs that are most commonly encountered in practice. In Komuso, we identified six BDCs [for more information on BDCs and methods to produce multisource statistics we refer to De Waal et al. (2020)]:

- BDC 1: multiple non-overlapping cross-sectional microdata sources that together provide a complete dataset without any under-coverage problems;
- BDC 2: same as BDC 1, but with overlap between different data sources;

- BDC 3: same as BDC 2, but now with under-coverage 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).

2.3 Error types

There exist many different schemes of error categories in survey and administrative sources; see for instance Zhang (2012). The error categories that are distinguished also depend on the level of detail that is used. Table 1 provides an overview of the error categories that were distinguished in Komuso.

Error category	Type of error included	Survey	Admin
Validity error	Specification error	X	
	Relevance error		X
Frame and source error	Under-coverage	X	X
	Over-coverage	X	X
	Duplications	X	X
	Misclassification in the contact variables	X	
	Misclassification in the auxiliary variables	X	X
Selection error	Error in terms of the selected sampling units	X	
	Unit non-response	X	
	Missing units in the accessed dataset		X
Measurement error and Item missingness	Arising from: respondent, questionnaire, interviewer, data collection	X	
	Fallacious or missing information in admin source		X
Processing error (*)	Data entry error	X	
	Coding or mapping error or misclassification	X	X
	Editing and imputation error	X	X
	Identification error		X
	Unit error		X
	Linkage errors	X	X
Model error (examples, non-exhaustive)	Editing and imputation error, record linkage error, ...	X	X
	Model based estimation error (Small Area Estimation, Seasonal Adjustment, Structural Equation Modeling, Bayesian approaches, Capture-Recapture or Dual System Estimation, Statistical Matching, ...)	X	X

Table 1: Main sources of errors in multisource statistics [from Komuso (ESSnet Quality of Multisource Statistics) (2019)]. (*) Processing errors are errors occurring with manual activities. These include also trivial errors, e.g., typographical errors in writing a procedure or errors in specifying a variable in the program (also in a model). When the processing steps mentioned are done via a model they may result in model errors.

2.4 General strategies to measuring output quality

In Komuso, four different general strategies were identified that one can take with respect to measuring quality: (1) using a quality framework without trying to quantify quality measures, (2) us-

ing generic quality measures that do not rely on any underlying model or design, (3) using non-parametric models to quantify quality measures (including sampling theory), and (4) using parametric models to quantify quality measures.

1. A quality framework is often used for measuring the quality of an entire statistical production chain, including data collection and data processing steps, since constructing statistical models for all steps in the statistical production chain is generally not feasible. A quality framework does not rely on statistical distributions nor on distribution-free quality measures. Instead, a quality framework often tries to combine various pieces of information on the quality of the produced output, such as expert opinions, into a set of quality measures.
2. Generic quality measures that do not rely on any underlying model or design are often used when one wants to measure quality for a single step in the statistical production chain, rather than measure the quality of the entire statistical production chain as a quality framework aims to do. Such quality measures are often used for steps in the statistical production that are hard or impossible to capture in a statistical model. An example is the difference between an earlier estimate and the latest revisions, which quantifies the effect of revisions (see Subsection 3.1.2). The QMCMs using generic quality measures that do not rely on any underlying model or design that have been developed in Komuso all concern the dimension "coherence".
3. Sampling theory, including resampling techniques, is the most common form of non-parametric models for measuring output quality, in any case within the Komuso project. Generally, (re)sampling theory is used to estimate bias and variance for estimators based on random samples.
4. Parametric models are often used when one wants to measure quality for a single step in the statistical production chain, and such a step can be captured in a statistical model. Examples of such methods are constrained optimization, latent class modeling, and structural equation modeling.

Within each general strategy one can use different methods to calculate quality measures. An overview of some of the methods encountered in Komuso is given in Sections 3 to 5.

3 Scoring methods: "Art is born of the observation and investigation of nature"

Scoring methods basically just base any quality measure directly on the observations, without relying, for instance, on statistical models. Scoring methods remind us of a quote by Marcus Tullius Cicero (Roman statesman, lawyer and scholar): *"Art is born of the observation and investigation of nature."*

3.1 Scoring methods for single and multisource statistics

3.1.1 Qualitative methods

The usual objective of qualitative methods is to collect non-numerical data such as reasons, (expert) opinions, and motivations. Examples of qualitative methods are individual interviews and group discussions.

Example: Two-phase and three-phase error framework

A multisource production process may consist of several administrative registers that are linked and harmonized by means of micro-integration [see, e.g., Bakker (2011)] for constructing a statistical register and deriving the variable of interest and related variables. The steps in this production process are usually complicated. Zhang (2012) developed a two-phase error framework for the situation where data from multiple sources are integrated to create a statistical micro dataset (see Table 2).

	Measurement dimension	**Representation dimension**
Phase 1	Validity error	Frame error
	Measurement error	Selection error
	Processing error	Missing/redundancy
Phase 2	Relevance error	Coverage error
	Mapping error	Identification error
	Comparability error	Unit error

Table 2: Error sources in the two-phase error framework.

The first phase shows the stages and error sources during the construction of each input source (e.g., an administrative register). For the process along the measurement perspective, the errors are validity error, measurement error, and processing error; along the representation perspective, the errors are frame error, selection error, and missing/redundancy. The model of this phase is adapted from the total survey error model of Groves et al. (2004).

The second phase starts with the target concept and target population, defined according to the purpose of the integrated statistical data. These targets are typically different from the first phase where the targets were defined and generated according to the purpose of the data owner. Because of this, it is sometimes even necessary to "swap" the two measurement and representation dimensions when moving from the first stage to the second stage; e.g., employment can be considered either as representation (of the population of employed people) or as measurement (of employment in the labor force population). During the second phase, the errors related to the measurement dimension are relevance error, mapping error, and comparability error; for the representation dimension the errors are coverage error, identification error, and unit error.

Reid et al. (2017) proposed an extension of this two-phase error framework with a third phase: the estimation phase. According to Reid et al. (2017), phase 2 ends with a unit-level records file containing units and values of the variables. The third phase then describes inaccuracies that can be made during the actual estimation process, in which one may try to correct for errors made during the first two phases. Furthermore, phase three includes estimation of the quality of the output estimates.

To what extent errors are treated can be measured as a simple proportion in this framework, where 1 stands for complete treatment of all the error sources and 0 if none are treated. These values may be based on expert knowledge. Alternatively, any Likert scale measure can be defined subjectively by the expert. Reid et al. (2017) give three examples of how their three-phase error framework was used to qualitatively compare different possible designs for statistical output, treating both single-source and multisource statistics. An example of giving scale measures to errors sources can be found in Biemer et al. (2014). Rocci et al. (2018) also applied an error framework to a multisource statistic, and for different error types they estimated the fraction of units in which that error type occurred.

3.1.2 Descriptive summary statistics

A descriptive summary statistic quantitatively describes features of collected data. A descriptive statistic aims to summarize the observed data and is generally quite simple. Commonly used de-

scriptive summary statistics are the minimum and maximum values of the variables, the means, medians and modes of the variables, the standard deviation and variances of the variables and the correlation between two variables.

In the Komuso project several descriptive summary statistics were examined. Below we give examples of three such statistics. These examples have been developed by ISTAT in the Komuso project.

Example: Cross-domain and sub-annual versus annual statistics coherence

A descriptive summary statistic to measure the coherence of estimates for the same parameter/variable of interest based on cross-domain or sub-annual statistics versus annual statistics is the relative difference between the "main" estimate and the "comparison" estimate. It is computed as

$$I = \frac{y_A - y_B}{y_B} \times 100,$$

where y_A is the main estimate and y_B the comparison estimate. For instance, y_B may be the estimate based on annual statistics and y_A the estimate for the same parameter/variable of interest based on cross-domain or sub-annual statistics. Generally, y_B is based on the most accurate/trustable source, unless there is no reason to consider any of the sources as the most accurate. In the latter case one could consider setting y_B equal to the average of the two estimates.

The above indicator I can be used after the final point estimates have been computed and one or more estimates for the same parameter/variable of interest are available from different sources or from processes with a different frequency.

Example: (Change of) sign, size, bias and variability of revisions and discrepancies

In order to quantify the effect of revisions and discrepancies on statistical estimates one can simply calculate the difference between the latest estimate and earlier estimates in the case of a revision, or the difference between estimates for similar domains in the case of discrepancies. For convenience we will only discuss revisions, but the same holds for discrepancies.

The difference is simply computed as $R^t = L^t - P^t$, where R^t denotes the revision for moment t, L^t the latest estimate for a variable/parameter of interest, and P^t a preliminary estimate for the same variable/parameter of interest. The later calculated estimate L^t is generally considered more reliable than the preliminary estimate P^t. Here, P^t may, for instance, denote the estimated period-on-period growth rate in a certain period, and L^t a later calculated, more accurate, estimated period-on-period growth rate.

Given the calculated revisions R_i^t for several statistics i, the following descriptive summary statistics can, for instance, be estimated: the change of sign due to a revision, the size of the revisions (mean of absolute revisions, median of absolute revisions, mean of relative absolute revisions), bias of the revisions (revision mean and its statistical significance, revision median) and the variability of the revisions (root mean square error, range, min, max, ...).

Seasonally adjusted estimates may be taken into account to calculate descriptive summary statistics. Such seasonally adjusted estimates can, for instance, be obtained by applying available seasonal adjustment software on the unadjusted data.

The descriptive summary statistics are, for instance, applied at ISTAT when monthly seasonally adjusted data of industrial production indices are estimated by means of a direct approach at domain level, and quarterly seasonally adjusted output of the industrial sector is based on disaggregation techniques on annual data with seasonally adjusted industrial production indices. At least two approaches can be applied in such a situation: one can use the quarterly averages of the disseminated seasonally adjusted indices or one can use seasonally adjusted quarterly averages of the unadjusted

monthly indices. The descriptive summary statistics offer some help in choosing between these two (and possibly other) approaches.

Example: Scalar measure of coherence in a reconciled demographic balancing equation
For the situation where estimates related by linear constraints need to be reconciled, descriptive summary statistics are also available. An example of such a situation is when stocks and flows of a population have to be balanced. Another example is when macro-economic figures, for instance figures for the National Accounts, that are connected by accounting equations need to be reconciled.

We will illustrate the descriptive summary statistics that have been examined in the Komuso project by the demographic balancing equation. More in detail, the population sizes in a domain i at times t (P_i^t) and $t+1$ (P_i^{t+1}) (stocks), and flows (migrations, birth and deaths) within the period $[t,\ t+1]$ need to satisfy the demographic balancing equation $P_i^{t+1} = P_i^t + N_i + M_i$, where $N_i = B_i - D_i$ is the natural increase, with B_i and D_i the number of births, respectively the number of deaths in period $[t,\ t+1]$, and $M_i = \sum_j M_{ij}$, where M_{ij} is the number of people who immigrated to domain i from domain j minus the number of people who emigrated from domain i to domain j and the sum is taken over all domains j. Here a domain may be any disjoint partitioning of the population, for instance region by sex by age class.

Let us assume that the estimates for domains i and j are given by $\hat{P}_i^t$, $\hat{P}_i^{t+1}$, $\hat{N}_i$ and $\hat{M}_{ij}$. A simple descriptive summary statistic for the degree of incoherence is then the average over the domains of the differences between the direct estimate for the population $\hat{P}_i^{t+1}$ and the corresponding estimate obtained by the estimates of stock of the population at time t and the flows in period $[t,\ t+1]$, that is by $\hat{P}_i^t + \hat{N}_i + \hat{M}_i$.

An indicator for the degree of incoherence for domain i is

$$d_i = \left|\hat{P}_i^{t+1} - (\hat{P}_i^t + \hat{N}_i + \hat{M}_i)\right|.$$

A descriptive summary statistic for the global measure of coherence is then given by the sum of the differences standardized with respect to the average of the two estimates of the population P_i^{t+1}, i.e., by

$$C = \frac{2}{D} \sum_i \frac{d_i}{\hat{P}_i^{t+1} + \hat{P}_i^t + \hat{N}_i + \hat{M}_i},$$

where D denotes the number of domains and the summation is over all domains $i = 1, \ldots, D$.

The above indicator and descriptive summary statistic for the global measure of coherence do not examine the impact of reconciliation. We will now consider an indicator and descriptive summary statistic for the impact of reconciliation. Let $(\tilde{P}_i^t,\ \tilde{P}_i^{t+1},\ \tilde{N}_i,\ \tilde{M}_{ij})$ be reconciled values that satisfy the demographic balancing equation. Indicators for the impact of reconciliation for the separate variables for each domain i are then given by $\left(\tilde{P}_i^t - \hat{P}_i^t\right)/\hat{P}_i^t$, $\left(\tilde{P}_i^{t+1} - \hat{P}_i^{t+1}\right)/\hat{P}_i^{t+1}$, $\left(\tilde{N}_i^t - \hat{N}_i^t\right)/\hat{N}_i^t$, and $\left(\tilde{M}_i^t - \hat{M}_i^t\right)/\hat{M}_i^t$. The indicators obviously depend on the reconciled values, and hence on the reconciliation method used.

A descriptive summary statistic for the impact of reconciliation based on these indicators is the average of the above four indicators over all domains, i.e.,

$$CR = \frac{1}{4D} \sum_i \left(\left| \frac{\tilde{P}_i^t - \hat{P}_i^t}{\hat{P}_i^t} \right| + \left| \frac{\tilde{P}_i^{t+1} - \hat{P}_i^{t+1}}{\hat{P}_i^{t+1}} \right| + \left| \frac{\tilde{N}_i - \hat{N}_i}{\hat{N}_i} \right| + \left| \frac{\tilde{M}_i - \hat{M}_i}{\hat{M}_i} \right| \right).$$

CR can also be seen as a measure of incoherence, since it quantifies the overall change in values required to obtain the reconciled values. Like the four underlying separate indicators, CR depends on the reconciliation method.

When only a subset of the demographic variables $\hat{P}_i^t$, $\hat{P}_i^{t+1}$, $\hat{N}_i$ and $\hat{M}_i$ are reconciled, or when they are reconciled for a subset of domains i only, the descriptive summary statistic should be computed on that subset only.

CR allows one to compare the impact of several reconciliation methods to each other. By zooming in on specific subsets one can study the impact of reconciliation for certain (groups of) domains or on certain variables.

3.2 Scoring methods developed especially for multisource statistics

3.2.1 Dempster-Shafer theory

Dempster-Shafer theory offers a general methodological framework for dealing with uncertainty. It enables one to combine, possibly conflicting, information from different sources. With Dempster-Shafer theory one can take all available information into account and quantify the degree of belief in a certain outcome by means of a belief function. Such a belief function relates the plausibility of a certain answer to a certain question to the plausibilities of answers to a related question. These degrees of belief may be subjective. For instance, they may be based on expert opinions. Dempster-Shafer theory gives rules for combining degrees of belief that are based on independent sources.

Dempster-Shafer theory can be used in many cases. For instance, the theory can be used when one wants to combine information from several experts or when one wants to combine expert opinions with information based on observed data. This makes Dempster-Shafer theory a very broadly applicable methodological tool.

For more on Dempster-Shafer theory, and on an application of Dempster-Shafer theory to the Austrian Population Census, we refer to Berka et al. (2010), Berka et al. (2012), Schnetzer et al. (2015), and Asamer et al. (2016).

4 Methods based on (re)sampling: "Design is the intermediary between information and understanding"

Methods based on (re)sampling generally use the design, for instance the sampling design, by which the data are collected, to base quality measures upon. For such quality measures the "*design is the intermediary between information and understanding*," a quote by the German painter Hans Hoffman.

Sampling theory allows one to compute the sampling variance for a large number of sampling designs (Särndal et al., 1992). Assuming a mechanism for the non-response process, sampling theory may in some cases also be used to estimate non-response variance, besides sampling variance. In sampling theory, one usually derives analytical formulae to compute sampling (and non-response) variance.

For single-source statistics, calculating the sampling (and non-response) variance is often the only realistic way to estimate the quality of the output. Sampling theory is also very useful for multisource statistics. In the case of multisource statistics, different situations may arise for which one may want to estimate sampling (and non-response) variance than for single-source statistics (see, e.g., the first example in Section 4.2).

Resampling can often be used to estimate the variance (and bias) of an estimator. The advantage of resampling methods is that, while analytic variance formulae need to be derived for different kinds of estimators separately and can become quite complex, resampling methods offer a relatively simple computational procedure for obtaining variance estimates that is general enough to be applicable to many estimation problems.

There are several resampling techniques, such as the jack-knife, balanced repeated replication and subsampling (Wolter, 2007). For example, in the jack-knife, one systematically recomputes estimates for the statistic of interest, leaving out one or more observations at a time from the dataset. From the obtained set of replicates of the statistic, estimates for the bias and variance of the statistic can be obtained.

One of the most frequently used resampling methods is the bootstrap (Efron and Tibshirani, 1993). Bootstrapping is a method of repeated sampling from either a sample (non-parametric bootstrapping) or from an estimated parametric distribution (parametric bootstrapping). Under certain conditions, the variance over the set of bootstrap outcomes is an approximately unbiased estimator for the variance of the original estimator. Likewise, the difference between the mean of the bootstrap estimates and the estimate derived from the original sample is often an approximately unbiased estimator of the bias of the original estimator.

For examples of applications of non-parametric bootstrapping in a multisource context, see, e.g., Kuijvenhoven and Scholtus (2011) and Scholtus and Daalmans (2020). In these papers, the bootstrap is used to estimate the variance of an estimated frequency table involving the highest attained level of education based on combined administrative and survey data, where missing values of education in the target population are accounted for either by weighting (the first reference) or mass imputation (the second reference). An example of an application of a parametric bootstrap method will be given in Section 4.2.

4.1 Methods based on (re)sampling for single and multisource statistics

In this section, we give an example of a case where sampling theory can be applied to both single and multisource statistics. This example was developed by Statistics Lithuania in the Komuso project.

Example: Effect of frame under-coverage / over-coverage on the estimator of a total and its accuracy measures

Sampling theory can be applied to a case where a sample is taken from a frame that is kept constant for a longer time. This may occur in business statistics: some NSIs use a business register that is "frozen" for a year, i.e., certain population changes are stored during the year and they are effectuated only once, at the beginning of a year.

In the case of intra-annual estimators (month, quarter), the corresponding population is likely to have changed compared to the sampling frame. With respect to the true population, the sampling frame suffers both from under- and over-coverage. Assume that an up-to-date administrative source is available that does not suffer from coverage errors. Using this administrative source, an adjusted estimator can be calculated. Next, metrics on differences between the original and the adjusted estimator quantify the sensitivity of the original estimator to coverage errors in the frozen sampling frame.

Suppose we have a sampling frame of a population U of size N, divided into non-overlapping strata U_h of size N_h, with $h = 1,\ldots,H$. From each stratum U_h a random sample s_h is taken of size n_h. For the sample units we collect information on a target variable y. For instance, we are interested in quarterly gross earnings of enterprises (y) by economic sector h, estimated by a sample survey drawn from a frozen business register U. Furthermore we have an auxiliary variable x, for instance the number of employees per enterprise. This variable is assumed to be available for all units in the sampling frame U as well as in a population V of a social insurance inspection data base which contains an up-to-date population for the number of employees. Population V consists of non-overlapping strata

V_h of size M_h, with $h = 1,\ldots,H$. As a result of under- and over-coverage, we find that $U \setminus V \neq \emptyset$ and $V \setminus U \neq \emptyset$.

Now assume that we are interested to study the effect of the frozen register on an estimate of the true population total t_y, the quarterly gross earnings. Based on the sampling frame U we obtain the Horvitz-Thompson estimator for the total $\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh}$, with $\hat{t}_{y_h} = \frac{N_h}{n_h'} \sum_{i=1}^{n_h'} y_{hi}$ where n_h' denotes the number of responding (and alive) enterprises ($n_h' \leq n_h$). For the auxiliary variable x, we know the totals for the frozen register: $t_x = \sum_{h=1}^{H} t_{xh}$ and $t_{xh} = \sum_{i=1}^{N_h} x_{hi}$. Similar to variable y, the Horvitz-Thompson estimators of the totals for x are given by $\hat{t}_x = \sum_{h=1}^{H} \hat{t}_{xh}$ with $\hat{t}_{xh} = \frac{N_h}{n_h'} \sum_{i=1}^{n_h'} x_{hi}$. Furthermore, we have more up-to-date totals based on V: $\tilde{t}_x = \sum_{h=1}^{H} \tilde{t}_{xh}$ and $\tilde{t}_{xh} = \sum_{i=1}^{M_h} x_{hi}$.

Economic sector	$\hat{R}_t$		$\hat{R}_{\text{Var}}$	
	Q1	Q4	Q1	Q4
A: Agriculture	−0.64	−1.03	−0.46	−0.18
B: Mining and quarrying	−1.04	−1.21	−2.08	−2.41
C: Manufacturing	−0.52	−0.73	−0.78	−1.26
D: Electricity, gas, steam and air conditioning supply	−0.59	−1.17	+9.66	+16.33
E: Water supply; sewerage; waste management and remediation activities	−0.25	−0.51	+0.70	+1.36
F: Construction	−1.49	−2.42	−2.77	−4.48
G: Wholesale and retail trade; repair of motor vehicles and motorcycles	−1.34	−1.71	−2.68	−3.40
H: Transportation and storage	−0.94	−1.19	−0.58	−0.21
I: Accommodation and food service activities	−1.36	−2.08	−2.67	−4.07

Table 3: Changed totals ($\hat{R}_t$) and changed variances ($\hat{R}_{\text{Var}}$) for separate ratio estimators of quarterly gross earnings for the first and fourth quarter of 2015 in Lithuania for a selection of economic sectors, obtained from Krapavickaitė and Šličkutė-Šeštokienė (2017).

We can now use a separate ratio estimator or a combined ratio estimator for t_y and compare the original estimator (based on U) with an updated version (based on V). The original separate ratio estimator and its updated version are given by:

$$\hat{t}_y^{(sep)} = \sum_{h=1}^{H} t_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}},$$

$$\tilde{t}_y^{(sep)} = \sum_{h=1}^{H} \tilde{t}_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}}.$$

Likewise, the original combined ratio estimator and its updated version are given by

$$\hat{t}_y^{(comb)} = t_x \frac{\hat{t}_y}{\hat{t}_x},$$

$$\tilde{t}_y^{(comb)} = \tilde{t}_x \frac{\hat{t}_y}{\hat{t}_x}.$$

Note that when the population has not changed, i.e., when V is the same as U, then $t_{xh} = \tilde{t}_{xh}$ (for $h = 1,...,H$), and the separate ratio and combined ratio estimators based on V are indeed equal to the corresponding estimators based on U.

For both estimators, i.e., the separate ratio estimator and the combined ratio estimator, we can quantify the effect of the change of the frame by looking into the extent that the totals have changed, and the extent that their variances have changed:

$$\hat{R}_t = \frac{\tilde{t}_y - \hat{t}_y}{\hat{t}_y},$$
$$\hat{R}_{\text{Var}} = \frac{\widehat{\text{Var}}\left(\tilde{t}_y\right) - \widehat{\text{Var}}\left(\hat{t}_y\right)}{\widehat{\text{Var}}\left(\hat{t}_y\right)},$$

where we have omitted the superscripts "(sep)" and "(comb)". The variances can be derived analytically with standard sampling theory; see, e.g., Särndal et al. (1992, p. 253 and pp. 270-271). Note that $\hat{\Delta}_y = \hat{t}_y - \tilde{t}_y$ is an estimate of the bias of the total $\hat{t}_y$ due to the use of a frozen sampling frame. Note further that $\hat{\Delta}_y / \sqrt{\widehat{\text{Var}}(\hat{\Delta}_y)}$ could be used to test whether the difference between the two estimators is significant.

The indicators $\hat{R}_t$ and $\hat{R}_{\text{Var}}$ for the separate ratio estimator have been applied to the Lithuanian survey on earnings in 2015. Outcomes for the first (Q1) and last quarter (Q4) are shown in Table 3 for the first nine economic sectors; more results can be found in Krapavickaitė and Šličkutė-Šeštokienė (2017). The results on $\hat{R}_t$ show that the estimator becomes more sensitive to frame updates from Q1 to Q4 when coverage errors have increased.

4.2 Methods based on (re)sampling developed especially for multisource statistics

In this section, we give two examples of methods based on (re)sampling theory that have been developed especially for multisource statistics.

Example: Variance of cell values in estimated frequency tables
The Repeated Weighting (RW) estimator was developed to ensure numerical consistency among tables estimated from different combinations of administrative data and sample surveys (Houbiers, 2004). The basic idea of RW is to use the regression estimator to calibrate table estimates to any marginal tables that have been estimated previously. Calculation of the variances of the resulting estimates can be rather complicated.

In general, the RW estimation procedure consists of the following three steps:

1. Specify the set of target tables to be estimated, and order them in descending order of available information.
2. Estimate each table separately from an appropriate subset of the available data, called a block. In general, a block will consist of the largest survey or combination of surveys in which all variables of the table are observed. It is assumed that a regression estimator is used in this step, based on auxiliary variables that are observed throughout the population.
3. Perform reweighting: adjust each table consecutively using the regression estimator, in the order specified in step 1, so that numerical consistency is achieved for any part of the table (including its margins) that overlaps with a previously estimated table.

It should be noted that tables that are estimated from the same block using the same regression estimator (as is done in Step 2) are automatically numerically consistent. Reweighting is applied when not all tables can be estimated from the same block.

As an illustration, consider the following example taken from Knottnerus and van Duin (2006). There are three categorical variables, X, Y and Z. The available data consist of one register and two without-replacement samples S_1 and S_2 of sizes n_1 and n_2, respectively. The auxiliary variable X is observed in the register for all N units in the population. The target variable Y is observed only in S_2; the target variable Z is observed in S_1 and S_2. We want to find consistent estimates for two tables: t_Z and $t_{Z\times Y}$. Note that the first table is actually a margin of the second table. Since more information is available for estimating t_Z than for estimating $t_{Z\times Y}$, the two tables will be estimated in this order from different blocks. [Technically, we are using the so-called "splitting up" variant of RW here (Knottnerus and van Duin, 2006).]

Let S_{12} denote the union of S_1 and S_2. We begin by deriving initial estimates for the two tables using the regression estimator with X as auxiliary information. This yields: $\hat{t}_Z^{REG(S_{12})}$, estimated from S_{12}, and $\hat{t}_{Z\times Y}^{REG(S_2)}$, estimated from S_2. In general, the estimated margin for Z from $\hat{t}_{Z\times Y}^{REG(S_2)}$ will be numerically inconsistent with $\hat{t}_Z^{REG(S_{12})}$. In the third step of the RW procedure, $\hat{t}_{Z\times Y}^{REG(S_2)}$ is therefore reweighted with respect to its Z-margin, using the regression estimator. This yields:

$$\hat{t}_{Z\times Y}^{RW} = \hat{t}_{Z\times Y}^{REG(S_2)} + \hat{\boldsymbol{B}}'_{w;Z}\left(\hat{t}_Z^{REG(S_{12})} - \hat{t}_Z^{REG(S_2)}\right).$$

Here, $\hat{\boldsymbol{B}}_{w;Z}$ denotes a matrix of estimated regression coefficients, and $'$ denotes the transpose. More generally, if there was also additional information available outside S_2 about the Y-margin, then the table would simultaneously be reweighted with respect to this margin, leading to a third term in the above expression for $\hat{t}_{Z\times Y}^{RW}$.

To estimate the variance of this RW estimator, Knottnerus and van Duin (2006) noted that, under certain regularity assumptions, $\hat{t}_{Z\times Y}^{RW}$ can be approximated by

$$\hat{t}_{Z\times Y}^{RW} = t_{Z\times Y} + \hat{t}_{e(Z\times Y)}^{HT(S_2)} + \boldsymbol{B}'_Z\left(\hat{t}_{e(Z)}^{HT(S_{12})} - \hat{t}_{e(Z)}^{HT(S_2)}\right) + O_p\left(N/n_2\right),$$

where $e(.)$ denotes a vector of residuals from a regression of $(.)$ on the register variable X, the superscript HT denotes a Horvitz-Thompson estimator, and $\boldsymbol{B}_Z$ is the matrix of population regression coefficients estimated by $\hat{\boldsymbol{B}}_{w;Z}$. Now, assuming for simplicity that $1 \ll n_1, n_2 \ll N$ and that the two samples S_1 and S_2 are independent, the variance-covariance matrix of $\hat{t}_{Z\times Y}^{RW}$ can be estimated by

$$\widehat{\mathrm{cov}}\left(\hat{t}_{Z\times Y}^{RW}\right) = \sum_{i\in S_1}\left(d_i^{(S_1)}\right)^2 \epsilon_{1i}\epsilon'_{1i} + \sum_{i\in S_2}\left(d_i^{(S_2)}\right)^2 \epsilon_{2i}\epsilon'_{2i},$$

where $d_i^{(S_k)}$ is the design weight of unit i in sample S_k $(k = 1, 2)$, $\epsilon_{1i} = \lambda_1 \boldsymbol{B}'_Z e_i(Z)$, $\epsilon_{2i} = e_i(Z\times Y) - \lambda_1 \boldsymbol{B}'_Z e_i(Z)$, and λ_1 is a weighting factor that reflects the relative reliability of S_1 in S_{12}. A simple choice that is often made is to set $\lambda_1 = n_1/(n_1+n_2)$. In particular, the square roots of the diagonal elements of $\widehat{\mathrm{cov}}\left(\hat{t}_{Z\times Y}^{RW}\right)$ provide standard errors for the cells of the estimated table $\hat{t}_{Z\times Y}^{RW}$.

The variables ϵ_{1i} and ϵ_{2i} in the above expression are examples of "super-residuals", which are linear combinations of ordinary regression residuals. More generally, Knottnerus and van Duin (2006) showed that the variance-covariance matrix of an RW estimator for a frequency table can always be approximated by means of super-residuals. If the above assumptions that $n_1, n_2 \ll N$ and/or that the two samples are independent do not hold, other variance estimators from sample survey theory can be used (Knottnerus and van Duin, 2006). Unlike the above variance estimator, these other variance estimators require that all second-order inclusion probabilities are known,

which may be difficult to achieve in practice.

Example: Bias and variance of parameter estimates affected by classification errors
NSIs often publish statistics that are obtained by aggregating numerical variables separately for each domain defined by a classification variable. For instance: the total turnover of businesses by type of economic activity, or the average hourly income of employed persons by highest attained education level. If the numerical variables are observed accurately for all units in the target population (e.g., in an administrative dataset), then the main issue affecting the quality of these statistics may be errors in the assignment of units to the right domain (classification errors). Van Delden et al. (2016) proposed a parametric bootstrap method to evaluate the bias and variance of statistics due to classification errors, under the simplifying assumption that these are the only errors that occur.

Let $i = 1,\ldots,N$ denote the units in the target population. Given the classification of interest, each unit has an unknown true code $s_i \in \{1,\ldots,H\}$, where H is the total number of classes. For each unit in the population, we observe a code $\hat{s}_i \in \{1,\ldots,H\}$ which may or may not be equal to the true code. We suppose that random classification errors occur, independently across units, according to a (possibly unit-specific) transition matrix $\boldsymbol{P}_i = (p_{ghi})$, with $p_{ghi} = \mathrm{P}(\hat{s}_i = h \mid s_i = g)$. The true codes are considered fixed.

In general, we write the target parameter as a function $\theta = f(y_1,\ldots,y_N;s_1,\ldots,s_N)$, where y_i denotes the value of a numerical target variable for unit i (or, more generally, a vector of numerical variables). For instance, a domain total can be written as $\theta_g = \sum_{i=1}^{N} y_i I\{s_i = g\}$, where $I\{.\}$ equals 1 if its argument is true and 0 otherwise. We assume that all $y_1,\ldots,y_N$ are known. An important special case where this is trivially true occurs when the target parameter is the number or proportion of units per domain (i.e., a domain total with $y_i \equiv 1$ and $y_i \equiv 1/N$, respectively). Given the assumption that no errors occur besides classification errors, θ can be estimated from the observed data by $\widehat{\theta} = f(y_1,\ldots,y_N;\hat{s}_1,\ldots,\hat{s}_N)$. We are interested in the bias and variance of $\widehat{\theta}$ as an estimator for θ.

As a preliminary step towards evaluating the bias and variance due to classification errors, we need to estimate the probabilities in the transition matrix $\boldsymbol{P}_i$. Typically, this requires the collection of additional data on the classification variable. Possible approaches include:

- Draw a random audit sample of units for which, in addition to $\hat{s}_i$, the true code s_i is obtained, e.g., by manual verification.
- Use process information from an editing step during regular production, where $\hat{s}_i$ may have been checked and corrected for certain units.
- Use multiple measurements of s_i from different, independent sources (e.g., a population register, an external administrative source and a sample survey). Under certain conditions, the error probabilities can be estimated from these multiple measurements using latent class analysis (see also Section 5.2.2).

In general, some model assumptions have to be introduced to reduce the number of unknown parameters. In this way, the unit-specific transition matrix $\boldsymbol{P}_i$ can be estimated as a function of a limited number of background variables. An example can be found in Van Delden et al. (2016). In applications where the codes $\hat{s}_i$ are predicted from a machine-learning algorithm, an estimate for $\boldsymbol{P}_i$ may be obtained naturally when the quality of the algorithm is evaluated. See, e.g., Meertens et al. (2020) for an example of such an application.

Having obtained an estimated transition matrix $\widehat{\boldsymbol{P}}_i = (\hat{p}_{ghi})$, we can apply the bootstrap method. For each unit, we draw a new code $\hat{s}^*_{ir}$ given the original observed code $\hat{s}_i$, using probabilities that mimic (our best estimate of) the original process by which $\hat{s}_i$ was generated from s_i:

$$\mathrm{P}\left(\hat{s}^*_{ir} = h \mid \hat{s}_i = g\right) \equiv \widehat{\mathrm{P}}(\hat{s}_i = h \mid s_i = g) = \hat{p}_{ghi}.$$

Based on the obtained values $\hat{s}^*_{1r},\ldots,\hat{s}^*_{Nr}$, we compute the bootstrap replicate $\widehat{\theta}^*_r = f(y_1,\ldots,y_N;\hat{s}^*_{1r},\ldots,\hat{s}^*_{Nr})$. This procedure is repeated R times, yielding replicates $\widehat{\theta}^*_1,\ldots,\widehat{\theta}^*_R$. From these replicates, the bias and variance of $\widehat{\theta}$ are estimated as follows [see also Efron and Tibshirani (1993)]:

$$\hat{B}^*_R\left(\widehat{\theta}\right) = m_R\left(\widehat{\theta}^*\right) - \widehat{\theta},$$
$$\hat{V}^*_R\left(\widehat{\theta}\right) = \frac{1}{R-1}\sum_{r=1}^{R}\left\{\widehat{\theta}^*_r - m_R\left(\widehat{\theta}^*\right)\right\}^2,$$

with $m_R\left(\widehat{\theta}^*\right) = R^{-1}\sum_{r=1}^{R}\widehat{\theta}^*_r$. In the bootstrap literature, it is often recommended to take $R \geq 200$ for variance estimates and $R \geq 1000$ for bias estimates.

An advantage of the bootstrap is that the above algorithm can be applied to many different types of estimators in the same way. For instance, θ could be a regression coefficient or a median. For certain simple target parameters, it is possible to derive an explicit formula for the analytical bias and variance estimators to which $\hat{B}^*_R\left(\widehat{\theta}\right)$ and $\hat{V}^*_R\left(\widehat{\theta}\right)$ converge for $R \to \infty$. As an example, Van Delden et al. (2016) obtained the following formulas for the estimated bias and variance of an estimated domain total $\widehat{\theta}_h = \sum_{i=1}^{N} y_i I\{\hat{s}_i = h\}$:

$$\hat{B}^*_\infty\left(\widehat{\theta}_h\right) = \sum_{i=1}^{N} y_i\left[(\hat{p}_{hhi}-1)\,I\{\hat{s}_i = h\} + \sum_{g=1,g\neq h}^{H}\hat{p}_{ghi}I\{\hat{s}_i = g\}\right],$$
$$\hat{V}^*_\infty\left(\widehat{\theta}_h\right) = \sum_{i=1}^{N} y_i^2\sum_{g=1}^{H}\hat{p}_{ghi}(1-\hat{p}_{ghi})I\{\hat{s}_i = g\}.$$

It can be shown that, in general, these bias and variance estimators are biased for the true bias and variance of $\widehat{\theta}_h$; hence, the same holds for the above bootstrap estimators based on a finite number of replicates. For simple target parameters such as a domain total, it is possible to correct for this bias in the estimated bias and variance, although this typically leads to bias and variance estimates that are less stable. See Van Delden et al. (2016) and Kloos et al. (2020) for more details.

5 Methods based on (parametric) modelling: "Artists can spend a lifetime searching for a perfect model"

Finally, we consider methods for calculating quality measures based on (parametric) modelling. Such methods usually construct a model for the target variable(s) to be estimated, and then use properties of the estimated model, such as the bias and variance of an estimator based on the model, as (basis for) quality measures. In some cases, methods based on (parametric) models give excellent results. In other cases, a suitable model may be difficult or even impossible to construct: "*Artists can spend a lifetime searching for a perfect model*," a quote by the American painter Robert Liberace.

5.1 Methods based on (parametric) modelling for single and multisource statistics

Besides methods that are used for measuring output quality directly, there are also some supporting methods, which do not measure output quality directly but are often used in combination with other methods that do measure quality directly. Examples of such supporting methods are estimating equations, log-linear modelling and mixture models.

Estimating equations specify how the parameters of a statistical model should be estimated. Examples of estimating equations are the method of moments, minimum distance methods like least squares estimation, Bayesian methods and (in some cases) maximum likelihood estimation [see, for example, Van der Vaart (1998)]. The idea of the estimation equations method is to find a set of simultaneous equations, involving observed data and model parameters of a statistical model, that need to be solved in order to find estimates of the model parameters.

In a log-linear model [see, for example, Bishop et al. (1975)], a logarithm of a certain variable equals a linear combination of the parameters of a statistical model. The technique is often used to study the relationship between several categorical variables. It can be used to build a statistical model as well as for statistical hypothesis testing.

Mixture models are often used in statistics when there are several subpopulations with different characteristics within the population [see, for example, McLachlan and Peel (2000)]. When using such a mixture model, it is not necessary to identify to which subpopulation each individual observation belongs. Mixture models can, for instance, be used when there are different subpopulations within the population with different kinds or different rates of measurement errors.

Below we discuss constrained optimization, a (parametric) modelling method that can be used for both single and multisource statistics.

5.1.1 Constrained optimization

Constrained optimization aims to optimize an objective function with respect to some variables, given constraints on those variables. These constraints can be either hard constraints, i.e., constraints that need to be satisfied, or soft ones, i.e., constraints that preferably – but not necessarily – should be satisfied. Soft constraints are often taken into account by incorporating them into the objective function, and penalizing the violation of such soft constraints.

Constrained optimization can, for instance, be used to adjust the values of some variables so they satisfy (or nearly satisfy) certain hard or soft constraints. Constrained optimization can also be used to benchmark data over time, i.e., to ensure that high-frequency time series data are reconciled with low-frequency time series data. These kinds of problems are quite common in, for instance, National Accounts.

Constrained optimization can also be used for single-source statistics, for instance when one wants to impute missing data such that constraints within individual records are satisfied (De Waal et al., 2011, Chapter 10).

Example: Macro integration / Data reconciliation

Macro-integration is often used for National Accounts and other kinds of statistical accounts. It can be carried out by means of several methods. We start by describing Stone's method [see, e.g., Stone et al. (1942) and Bikker et al. (2011)]. After describing Stone's method, we will focus on the univariate Denton method and then on the multivariate Denton method.

Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ is a vector of high frequency, say quarterly, numerical data. We denote the corresponding estimated covariance matrix by $\boldsymbol{V}$. Let us assume that our aim is to ensure by means of macro-integration that the reconciled components of $\boldsymbol{x}$ sum up to the values of low frequency, say annual, data $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)'$, where $n = 4m$. We then have to fulfill the following constraints

$$\sum_{j=4(k-1)+1}^{4k} x_j = b_k, \qquad \text{for } k = 1, \ldots, m. \tag{1}$$

These constraints are generally violated by the initial high frequency data $\boldsymbol{x}$. The vector $\boldsymbol{x}$ is therefore adjusted so the adjusted vector $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)'$ does satisfy (1).

In matrix notation (1) can be written as

$$\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{A}$ is an $m \times n$ matrix given by

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{j} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{j} & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{j} \end{pmatrix}$$

with $\boldsymbol{j} = (1,1,1,1)$ and $\boldsymbol{0} = (0,0,0,0)$.

In Stone's method the adjustment is done by minimizing the quadratic distance function

$$\min_{\tilde{\boldsymbol{x}}} (\tilde{\boldsymbol{x}} - \boldsymbol{x})' \boldsymbol{V}^{-1} (\tilde{\boldsymbol{x}} - \boldsymbol{x}) \tag{3}$$

subject to the constraints (1). The quadratic optimization problem (3) subject to (1) can be solved by the Lagrange multiplier method. The thus reconciled vector $\tilde{\boldsymbol{x}}$ is given by

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{V}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1} (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \tag{4}$$

and the variance $\tilde{\boldsymbol{V}}$ of $\tilde{\boldsymbol{x}}$ is

$$\tilde{\boldsymbol{V}} = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{V}. \tag{5}$$

(5) is a measure for the quality of the reconciled data.

A drawback of distance function (3) in the case of quarterly time series and annual time series is that discontinuity may arise between the last quarter of one year and the first quarter of the next year. The univariate Denton method (Denton, 1971) aims to avoid this discontinuity by minimizing a quadratic function based on differences between the first order differences, i.e., on $\Delta^{(1)} x_j = \Delta\tilde{x}_j - \Delta x_j$ where $\Delta\tilde{x}_j = \tilde{x}_j - \tilde{x}_{j-1}$ and $\Delta x_j = x_j - x_{j-1}$, rather than on differences between the levels of original and reconciled time series. The underlying idea is to preserve as much as possible the original quarter to quarter changes (the movement preservation principle). Note that Δx_1 is undefined and a value needs to be specified for Δx_1. Denton proposed to use $\Delta x_1 = x_1$. So, the univariate Denton method consists of solving

$$\min_{\tilde{\boldsymbol{x}}} \sum_{j=1}^{n} \left(\Delta\tilde{x}_j - \Delta x_j\right)^2 \tag{6}$$

subject to the boundary condition $\Delta x_1 = x_1$.

That Δx_1 needs to be fixed to a value can be seen as a disadvantage of the univariate Denton method. This disadvantage can be overcome by using the Cholette adaptation. For the Cholette adaptation we first rewrite (6) as

$$\min_{\tilde{\boldsymbol{x}}} (\tilde{\boldsymbol{x}} - \boldsymbol{x})' (\boldsymbol{D}'\boldsymbol{D}) (\tilde{\boldsymbol{x}} - \boldsymbol{x}), \tag{7}$$

where

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

That is, we use $\boldsymbol{V}^{-1} = \boldsymbol{D}'\boldsymbol{D}$ in formula (3). In the Cholette adaptation, the first line from the matrix $\boldsymbol{D}$ is deleted and the generalized inverse of $\boldsymbol{D}'\boldsymbol{D}$ needs to be taken to find $\boldsymbol{V}$ in (4) and (5).

Denton's idea can be taken further: the quadratic distance function can also be based on the second order differences $\Delta^{(2)}x_j = \Delta^{(1)}\tilde{x}_j - \Delta^{(1)}x_j$ (Sax and Steiner, 2013), i.e., instead of minimizing (6) one can minimize

$$\min_{\tilde{\boldsymbol{x}}} \sum_{j=1}^{n} \left(\Delta^{(1)}\tilde{x}_j - \Delta^{(1)}x_j\right)^2, \tag{8}$$

or even on third or higher order differences. Again, the Cholette adaptation can be used for the alternative distance function (8) as well.

Thus far we discussed *additive* Denton methods. Besides additive Denton methods also proportional Denton methods exist. We first define

$$\Delta_{prop}\left(x_j\right) = \frac{\tilde{x}_j - x_j}{x_j} - \frac{\tilde{x}_{j-1} - x_{j-1}}{x_{j-1}} = \frac{\tilde{x}_j}{x_j} - \frac{\tilde{x}_{j-1}}{x_{j-1}}.$$

Instead of solving (6), we then solve

$$\min_{\tilde{\boldsymbol{x}}} \sum_{j=1}^{n} \left(\Delta_{prop}\left(x_j\right)\right)^2.$$

In a similar way a proportional version of (8) can be formulated.

The univariate Denton method can be extended to multiple variables. The multivariate Denton method allows linear restrictions for separate variables as well as linear restrictions involving several variables. So, besides reconciliation of quarterly data to annual data, relationships between different variables can be taken into account.

By $\boldsymbol{x}$ we now denote a vector consisting of M variables that are observed k times per year for T years, i.e., $\boldsymbol{x}$ is an MkT-dimensional vector. The multivariate Denton method consists of solving a quadratic optimization problem (3) subject to (2). The only difference with Stone's method is that $\boldsymbol{x}$ now consists of several variables, and that $\boldsymbol{A}$ and $\boldsymbol{b}$ are now a matrix, respectively vector, that together describe the reconciliation constraints and constraints between different variables. $\boldsymbol{V}$ is again the covariance matrix of vector $\boldsymbol{x}$. The solution is given by (4) with variance given by (5). For more details on the multivariate Denton method we refer to Di Fonzo and Marini (2003) and Bikker et al. (2011).

5.2 Methods based on (parametric) modelling developed for multisource statistics only

In this section we discuss five commonly used methods for measuring output quality that can be used for multisource statistics only, and were examined in Komuso.

5.2.1 Capture-recapture methodology

Capture-recapture methodology is commonly used to estimate the size of a population. The methodology originated in ecology where it is used to estimate an animal population's size. In that context, a number of animals are captured, marked and then released. Later, again a number of animals are captured. The number of marked animals in the second sample can then be used to obtain an estimate for the total number of animals.

The methodology is also used by NSIs to estimate the size of a population. To estimate the total number of individuals that possess a certain characteristic, one records the individuals with that

characteristic occurring in a certain dataset, for example, a census. Next, one counts how many of the recorded individuals occur in another dataset, for example, a post-enumeration survey. This allows one to obtain an estimate for the total number of individuals with this characteristic in the population. Instead of a census or survey, administrative data may also be used.

Example: Capture-recapture methodology in its basic form
We will describe the basic form of capture-recapture methodology. We assume that two datasets A and B of the same fixed population size N are linked. We also assume that the following five technical assumptions are satisfied:

1. inclusion of an element into dataset A is independent of its inclusion in dataset B;
2. inclusion probabilities of units in at least one of the datasets are homogeneous, i.e., all units have an equal probability to be included in this dataset;
3. the population is closed;
4. it is possible to link the elements of datasets A and B perfectly;
5. The datasets do not contain units that do not belong to the target population ("erroneous captures"), nor do they contain duplicates.

Table 4 below describes how many units in datasets A and B occur in both datasets (n_{AB}), in dataset A only (n_A), in dataset B only (n_B), and how many units in the population occur in neither of the two datasets (n_{00}). The value of n_{00} is unknown and has to be estimated. Once the value of n_{00} is estimated, the population size can easily be estimated.

		Dataset B	
		Yes	No
Dataset A	Yes	n_{AB}	n_A
	No	n_B	n_{00}

Table 4: Numbers of units in datasets A and B.

When all five above-mentioned assumptions are valid, n_{00} can be estimated by means of the Petersen estimator [see, e.g., Sekar and Deming (1949)]. The Petersen estimate for n_{00} is

$$\hat{n}_{00} = \frac{n_A n_B}{n_{AB}}.$$

The Petersen estimate for the population size is then given by

$$\hat{N} = n_A + n_B + n_{AB} + \hat{n}_{00}.$$

An estimator for the variance of $\hat{N}$ is given by [see, e.g., Sekar and Deming (1949) and Bishop et al. (1975)]:

$$\widehat{\mathrm{Var}}\left(\hat{N}\right) = \frac{(n_A + n_{AB})(n_B + n_{AB})\, n_A n_B}{(n_{AB})^3}.$$

Van der Heijden et al. (2012) and Gerritse et al. (2015) consider more complicated approaches involving covariates. Those approaches are based on log-linear modelling instead of the Petersen estimator to estimate the unknown population size.

In general, the saturated log-linear model for a contingency table as in Table 4 would be given by

$$\log(n_{00}) = \mu, \tag{9}$$
$$\log(n_A) = \mu + \mu_A, \tag{10}$$
$$\log(n_B) = \mu + \mu_B, \tag{11}$$
$$\log(n_{AB}) = \mu + \mu_A + \mu_B + \mu_{AB}, \tag{12}$$

where μ_A, μ_B, and μ_{AB} indicate that the number of units in the corresponding cell depends on dataset A, on dataset B, or on both. However, in our case, equation (12) of the saturated log-linear model has to be replaced by

$$\log(n_{AB}) = \mu + \mu_A + \mu_B, \tag{13}$$

since the interaction term μ_{AB} cannot be identified. Assuming that datasets A and B are independent (i.e., the first technical assumption made above) this term can be set to zero.

We can estimate the model parameters μ, μ_A, and μ_B using those relations for which we know the cell totals, i.e., (9), (10) and (11) in our case. In general, we can estimate the model parameters by means of maximum likelihood estimation. In our simple example we can compute μ, μ_A and μ_B directly. From (9) to (11) and (13), we directly obtain $n_{00} = \exp(\mu)$, $n_A = \exp(\mu + \mu_A)$, $n_B = \exp(\mu + \mu_B)$ and $n_{AB} = \exp(\mu + \mu_A + \mu_B)$. This means that

$$\hat{n}_{00} = \exp(\hat{\mu}) = \frac{\exp(\hat{\mu} + \hat{\mu}_A)\exp(\hat{\mu} + \hat{\mu}_B)}{\exp(\hat{\mu} + \hat{\mu}_A + \hat{\mu}_B)} = \frac{n_A n_B}{n_{AB}}.$$

That is, the saturated log-linear model under the assumption that datasets A and B are independent gives the same estimate as the Petersen estimator.

An advantage of using log-linear models instead of the Petersen estimator is that they are easy to extend to more general cases, such as:

- three or more datasets instead of two;
- using auxiliary data besides the cell counts, for instance using an available auxiliary variable "gender" to differentiate between counts for women and counts for men, which may improve the quality of the final estimate for the population size;
- adding interaction terms between counts and available auxiliary variables.

For a given situation one can base one or more log-linear models on substantive knowledge, and then select the "best" model. What is considered "best" in a given situation may depend on the model fit (e.g., one can use a chi-square distribution where observed values are compared to expected values), the number of model parameters, and substantive considerations.

As already mentioned, the parameters of log-linear models can be estimated by means of maximum likelihood estimation. In this estimation procedure it may be necessary to set some model parameters to zero beforehand, since otherwise some parameters cannot be identified.

The variance of population size estimates based on a log-linear model can be estimated by means of a bootstrap procedure. For an example where log-linear models are used to estimate the population size, and bootstrapping is used to estimate the variance of the estimated population size, we refer to Van der Heijden et al. (2012) and Gerritse et al. (2015). For more on log-linear modelling in general, see, e.g., Bishop et al. (1975) and Agresti (2013).

5.2.2 Latent variable/class modeling and structural equation modeling

A latent variable model is a statistical model that relates a set of observable variables to a set of non-observable variables. The non-observable variables are called latent variables; the observable variables manifest variables. In a latent class model (Hagenaars and McCutcheon, 2002; Biemer, 2011), the latent variables are categorical.

Latent variable modeling is strongly related to structural equation modeling. In structural equation modeling, one also relates one or more unobserved latent variables to a set of observed variables. Structural equation modeling can be applied to both categorical and numerical variables. In fact, latent class analysis can be considered as a type of structural equation modeling for categorical data.

In both latent class modeling and structural equation modeling, true values can be fitted as a function of background variables or they can be modeled over time. In latent class modeling one estimates the probability that a certain value is observed given the true value. In structural equation modeling, each observed value is considered to be a function of the latent true value plus an error.

In the context of measuring the quality of multisource statistics, latent class modeling and structural equation modeling can be used in a situation where one has several datasets measuring the same target variable with measurement error. One can then see these error-prone measurements as observed indicators for an unobserved (latent) variable that represents the true values. Quality assessments based on these latent class models or structural equation models can then be used to assess the quality of output based on the observed indicators. If the model is trusted sufficiently, one could also correct output for measurement error using the predicted latent variable.

Example: Variance of estimates based on microdata reconciled by means of latent class analysis
Suppose that we have observed data on $S \geq 3$ categorical variables for the same units, where all variables are intended to measure the same categorical target variable. These multiple measurements could be obtained by linking data from different sources (e.g., administrative datasets, or a combination of administrative datasets and a survey) or by asking multiple questions about the same construct in a single survey. We use $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_S)'$ to denote the observed variables in general and $\boldsymbol{y} = (y_1, y_2, \ldots, y_S)'$ for a particular realization of values. The underlying target variable that these variables attempt to measure is denoted by X (with a particular value x) and is considered to be unobserved (latent) for all units. For simplicity we assume here that the set of categories is the same for all variables Y_j and X, denoted by $\{1, \ldots, L\}$.

Using some standard rules of probability theory, the marginal probability of observing $\boldsymbol{Y} = \boldsymbol{y}$ can be written as

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{L} \mathrm{P}(X = x, \boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{L} \mathrm{P}(X = x)\,\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} \mid X = x),$$

where $\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} \mid X = x)$ denotes the conditional probability of observing $\boldsymbol{Y} = \boldsymbol{y}$ when the true value of the target variable is x.

In latent class analysis, it is often assumed that each Y_j is measured independently of the other observed variables. Thus, the errors in different observed variables for the same unit are assumed to be independent. This assumption is known as "local independence" or "conditional independence". Under this assumption, we can write:

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} \mid X = x) = \mathrm{P}(Y_1 = y_1 \mid X = x)\ \mathrm{P}(Y_2 = y_2 \mid X = x) \cdots \mathrm{P}(Y_S = y_S \mid X = x).$$

Applying the local independence assumption to the above expression for $\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$, we obtain the formula that describes the basic latent class model:

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{L} \mathrm{P}(X = x) \prod_{j=1}^{S} \mathrm{P}\left(Y_j = y_j \mid X = x\right).$$

To estimate the latent class model, we need to estimate the unknown probabilities on the right-hand side of this expression from the observed values for the probabilities on the left-hand side. This can be done, for instance, by maximum likelihood estimation for incomplete data (Hagenaars and McCutcheon, 2002; Biemer, 2011). Without additional assumptions, the latent class model is not identified with $S < 3$ observed variables.

Note that each factor $\mathrm{P}\left(Y_j = y_j \mid X = x\right)$ can be interpreted as a model for classification errors in one of the observed variables. Hence, estimates of these probabilities can provide information about the quality of each observed variable as an indicator for the true target variable X. For instance, the probability that a unit with true value $X = 1$ is misclassified on observed variable Y_j is given by $\mathrm{P}\left(Y_j \neq 1 \mid X = 1\right) = 1 - \mathrm{P}\left(Y_j = 1 \mid X = 1\right)$. In this way, a latent class model could provide input for the parametric bootstrap method discussed in Section 4.2.

The estimated model also provides predictions for the probability that a unit with a certain combination of observed values belongs to a particular category of the true target variable. These predictions can be obtained by the formula

$$\mathrm{P}(X = x \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\mathrm{P}(X = x) \prod_{j=1}^{S} \mathrm{P}\left(Y_j = y_j \mid X = x\right)}{\sum_{x'=1}^{L} \mathrm{P}(X = x') \prod_{j=1}^{S} \mathrm{P}\left(Y_j = y_j \mid X = x'\right)},$$

which follows from the previous expressions by Bayes' rule.

Boeschoten et al. (2017) proposed the MILC method to construct an error-corrected estimator based on the probabilities $\mathrm{P}(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$, along with a variance estimate for this estimator. The acronym MILC stands for the combination of multiple imputation (MI) and latent class (LC) analysis. An application of the MILC method consists of the following steps:

1. From the original dataset containing all observations of $\boldsymbol{Y}$, select M bootstrap samples.
2. For each bootstrap sample, estimate the latent class model. Denote the predicted probabilities $\mathrm{P}(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$ from the parameter estimates for the m^{th} bootstrap sample by $\widehat{\mathrm{P}}_m(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$.
3. In the original dataset, construct M multiply imputed versions of X, based on the predicted probabilities from the bootstrap samples. That is, create M empty variables $(W_1, \ldots, W_M)$ and impute variable W_m by drawing one category from $\{1, \ldots, L\}$ for each unit based on $\widehat{\mathrm{P}}_m(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$.
4. Obtain M estimates for the parameter of interest based on the imputed variables $W_1, \ldots, W_M$.
5. Apply Rubin's rules for multiple imputation to pool the estimates from the previous step [see Rubin (1987)]. These pooling rules yield both a final estimate and an associated variance estimate. This variance estimate reflects the uncertainty about the true value of the parameter of interest due to classification errors in the observed variables and, if relevant, also due to sampling error.

The basic latent class model as described here can be extended in many ways. For instance, auxiliary variables can be added to the model if these are available. Depending on the available data, the local independence assumption can sometimes be relaxed to a certain extent. Boeschoten

et al. (2017) also discuss how to incorporate edit restrictions – e.g., that certain values for X are impossible given a particular value for an auxiliary variable – into the MILC method so that these are automatically satisfied by the imputed values.

So far, we have assumed that all variables are categorical and refer to the same point in time. A particular type of latent class model can be applied when multiple measurements are available over a period of time (e.g., each month or each quarter) and the latent variable can also change over time. This is known as a Hidden Markov Model (HMM). See, e.g., Pavlopoulos and Vermunt (2015) for an application of an HMM to model classification errors in linked data from two sources. Finally, when the observed and latent variables are numerical, structural equation models can provide a similar approach; see, e.g., Scholtus et al. (2015) and Oberski et al. (2017).

5.2.3 Statistical hypothesis testing

In statistical hypothesis testing, one uses observed data to determine the likelihood that a posited hypothesis holds true. In order to do so, one must formulate a null hypothesis and the alternative hypothesis, which says that the null hypothesis is not valid (in a particular way). One then computes how likely the observed data are, assuming the null hypothesis. The likelihood that the observed data were obtained as a realization of the null hypothesis is used as a measure for the validity of the null hypothesis and its alternative.

Hypothesis testing can, for instance, be used to test whether and to which extent the quality of revised estimates improves. In that case, the null hypothesis would be that there is no change in the quality of the estimates. One would then attempt to reject this hypothesis in favour of the alternative hypothesis that there is an improvement in the quality of the estimates. For instance, Fosen (2017) describes a test applied to revised employment statistics that are derived from gradually completing register data. As another example, suppose that a new estimation method has been proposed to replace an existing method, but it is not clear a priori whether the new method is an improvement. Here, the null hypothesis would again be that there is no difference in the quality of the estimates between the two methods, but now the alternative hypothesis may be that the quality with the new method either increases or decreases (i.e., a two-sided alternative).

5.2.4 Using estimated model parameters

Using estimated model parameters is a mix of descriptive summary statistics and statistical hypothesis testing. When using estimated model parameters for assessing quality, one uses a statistical model as in statistical hypothesis testing, but one does not use a posited statistical model to test a hypothesis. Instead, one directly uses estimated model parameters and calculates some descriptive summary statistics for them. For example, one can use a model to estimate the probability that a target variable in a certain unit has an incorrect value. One can then, for instance, take the average of these probabilities over all units in the dataset as an overall quality measure.

5.2.5 Small area estimation

Small area estimation is an umbrella term for several statistical techniques aiming to estimate parameters for small areas [see, e.g., Rao (2003) for an excellent introduction to small area estimation and descriptions of many small area estimators]. The main problem in small area estimation is usually the lack of observations for such small areas, which prevents one from using more standard estimation techniques, such as standard survey weighting [see, e.g., Särndal et al. (1992)].

In general, the term "small area" refers to a small geographical area such as a municipality, but it may also refer to other kinds of "small domains", such as small groups of individuals in the population.

6 Discussion

In this paper we have focused on the work with respect to the measurement of output quality that has been carried out in the Komuso project. In the Komuso project, we examined a large number of different basic data configurations, error types, and methods to assess output quality for some important quality dimensions (accuracy, timeliness, coherence, and relevance). We hope that the QMCMs that were produced in the Komuso project – of which we gave several examples in the current paper – are directly useful for many practical cases the readers of this paper are confronted with, and in other cases may form a source of inspiration to develop similar methods.

In the introduction to this paper, we mentioned that constructing quality measures for multisource statistics and calculating them is still more of an art than of a technical recipe. An illustration of this point is that, even with all the QMCMs and corresponding hands-on examples that have been developed in the Komuso project, one still needs to use one's instinct, or expert knowledge, to decide on what the most important error sources could be in a given situation. For instance, in some cases, measurement error may affect data quality the most, whereas in other cases sampling error or linkage error may affect data quality the most. Besides using one's instinct or expert knowledge, exploratory analyses and the use of scoring methods (see Section 3) may also provide some insight in the most important error sources. Something similar holds for the interaction between different kinds of errors. In many cases it may be reasonable to assume that the different kinds of errors are more or less independent, whereas in other cases this is definitely not the case. Again, using one's instinct or expert knowledge is important in order to distinguish between these situations. Although expert knowledge is not directly quantifiable, it can be valuable because it may capture years of experience. This may be illustrated by an anecdote about Pablo Picasso. When Picasso was asked by an admirer to scribble something on a napkin, Picasso complied and asked a large amount of money. The admirer was astonished by the large sum: "But you did that in thirty seconds." He replied: "No, it has taken me forty years to do that."

The QMCMs developed in the Komuso project provide quality measures and methods to calculate them for separate steps, or building blocks, in the statistical production process. We hope that in the, hopefully near, future, an all-encompassing theory or framework to base quality measures for multisource statistics upon will be developed. Such an all-encompassing theory or framework should be able to handle several different types of error sources at the same time and, preferably, use the same statistical theory to treat these error sources. Possible examples of approaches that may able to deal with several error sources simultaneously are, for instance, based on Bayesian techniques [see, e.g., Bryant and Graham (2015)] and over-imputation (Blackwell et al., 2015a,b).

We see two potential paths towards the development of such an all-encompassing theory or framework. The first potential path is the further development of Total Survey Error frameworks [see, e.g., Amaya et al. (2020), Biemer (2010), Biemer et al. (2014), Reid et al. (2017), Rocci et al. (2018), and Zhang (2012)], and the development of quality measures and methods to calculate them for the separate steps in such a framework.

The second – fundamentally different – potential path we see is not to specify and examine all the separate error sources, but develop a quality measure that covers several error sources at once. A landmark paper for such an approach is Meng (2018) [see also Rao (2020)]. Meng (2018) focuses

on the inclusion of units in the datasets under consideration, and hence on sampling error, inclusion error, non-response error, et cetera; measurement error and related error types are not considered. In the approach by Meng (2018) basically only three factors determine the quality of a certain estimated target parameter. In Meng's terminology these factors are referred to as "data quantity" (i.e., the amount of data), "problem difficulty" (i.e., the variation in the target variable), and "data quality" (i.e., the correlation between the target variable and possible inclusion in the data source).

Biemer and Amaya (2018) have extended the approach by Meng (2018) to include measurement error. For multisource statistics, it may be useful to further extend the approach to include linkage error.

A major challenge appears to be the application of Meng's approach in practical situations. In particular, the "data quality", and to a lesser extent the "problem difficulty", can be (very) hard to estimate in a practical situation.

Personally, we feel that this latter approach proposed by Meng (2018) may be the most promising path of the two paths towards an all-encompassing theory to base quality measures for multisource statistics upon as it avoids having to consider all possible error sources and their interactions. As noted above, despite the promising nature of Meng's approach, quite some research needs to be done before it can be applied in the day-to-day practice at, for instance, an NSI.

In any case for the next few years, we expect that measuring the output quality of multisource statistics will remain a field in motion, and a field that still is more an art than a technique.

Acknowledgments

We sincerely thank our colleagues in the ESSnet on Quality of Multisource Statistics. It has been a pleasure and a privilege for us to work with them in the ESSnet. Without the work done in that project we could not have written this paper.

We also thank a reviewer of our paper for carefully reading it.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Part of the work underlying this paper has been carried out as part of the ESSnet on Quality of Multisource Statistics, funded by the European Commission (FPA 07112.2015.003-2015.226: SGA 07112.2015.015-2015.705, SGA 07112.2016.019-2017.144 and SGA 07112.2018.007-2018.0444).

References

Agresti, A. (2013). *Categorical data analysis* (3 ed.). Hoboken, New Jersey: John Wiley & Sons.

Amaya, A., P. P. Biemer, and D. Kinyon (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology 8*, 89–119.

Asamer, E., F. Astleithner, P. Ćetković, S. Humer, M. Lenk, M. Moser, and H. Rechta (2016). Quality assessment for register-based statistics - results for the Austrian census. *Austrian Journal of Statistics 45*(2), 3–14.

Bakker, B. F. M. (2011). Micro-integration: State of the art. In *ESSnet on Data Integration, Report on WP1*, pp. 77–107. Available at http://ec.europa.eu/eurostat/cros/content/essnet-di-final-report-wp1_en.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer (2010). A quality framework for statistics based on administrative data sources using the example of the Austrian census 2011. *Austrian Journal of Statistics 39*(4), 299–308.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer (2012). Combination of evidence from multiple administrative data sources: Quality assessment of the Austrian register-based census 2011. *Statistica Neerlandica 66*(1), 18–33.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly 74*, 817–848.

Biemer, P. P. (2011). *Latent class analysis of survey error*. Hoboken, New Jersey: John Wiley & Sons.

Biemer, P. P. and A. Amaya (2018). A total error framework for hybrid estimation. Paper presented at the BigSurv18 conference, Barcelona, Spain.

Biemer, P. P., D. Trewin, H. Bergdahl, and L. Japec (2014). A system for managing the quality of official statistics. *Journal of Official Statistics 30*(3), 381–415.

Bikker, R., J. Daalmans, and N. Mushkudiani (2011). Macro integration. Data reconciliation. Technical report, Statistical Methods (201104), Statistics Netherlands, The Hague. Available at https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/output/output/macro-integration-data-reconciliation.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Blackwell, M., J. Honaker, and G. King (2015a). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods & Research 46*(3), 342–369.

Blackwell, M., J. Honaker, and G. King (2015b). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research 46*(3), 303–341.

Boeschoten, L., D. Oberski, and T. de Waal (2017). Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *Journal of Official Statistics 33*, 921–962.

Bryant, J. R. and P. Graham (2015). A Bayesian approach to population estimation with administrative data. *Journal of Official Statistics 31*(3), 475–487.

De Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Hoboken, New Jersey: John Wiley & Sons.

De Waal, T., A. van Delden, and S. Scholtus (2020). Multisource statistics: Basic situations and methods. *International Statistical Review 88*, 203–228.

Denton, F. T. (1971). Adjustment of monthly to quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association 66*(333), 99–102.

Di Fonzo, T. and M. Marini (2003). Benchmarking systems of seasonally adjusted time series according to Denton's movement preservation principle. Technical report, University of Padova. Available at http://www.oecd.org/dataoecd/59/19/21778574.pdf.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

Eurostat (2014). ESS handbook for quality reports, 2014 edition. Technical report, Eurostat. Available at http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf.

Fosen, J. (2017). Output quality for statistics based on several administrative sources. Technical report. Deliverable of WP 3 of the ESSnet on Quality of Multisource Statistics (SGA 1), available at https://ec.europa.eu/eurostat/cros/system/files/st2_7.pdf.

Gerritse, S., P. G. M. van der Heijden, and B. F. M. Bakker (2015). Sensitivity of population size estimation for violating parameter assumptions in log-linear models. *Journal of Official Statistics 31*(3), 357–379.

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey methodology*. New York: John Wiley & Sons.

Hagenaars, J. A. and A. L. McCutcheon (Eds.) (2002). *Applied latent class analysis*. New York: Cambridge University Press.

Houbiers, M. (2004). Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics 20*(1), 55–75.

Kloos, K., Q. Meertens, S. Scholtus, and J. Karch (2020). Comparing correction methods to reduce misclassication bias. Paper presented at the BNAIC/Benelearn conference, Leiden, The Netherlands.

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*(3), 565–584.

Komuso (ESSnet Quality of Multisource Statistics) (2019). Quality guidelines for multisource statistics (QGMSS). Technical report. Available at https://ec.europa.eu/eurostat/cros/content/quality-guidelines-multisource-statistics-qgmss_en.

Krapavickaitė, D. and M. Šličkutė-Šeštokienė (2017). Effect of the frame under-coverage / over-coverage on the estimator of total and its accuracy measures in the business statistics. Technical report. Deliverable of WP 3 of the ESSnet on Quality of Multisource Statistics (SGA 1), available at https://ec.europa.eu/eurostat/cros/system/files/st2_5.pdf.

Kuijvenhoven, L. and S. Scholtus (2011). Bootstrapping combined estimators based on register and sample survey data. Technical report, discussion paper, Statistics Netherlands, The Hague. Available at http://www.cbs.nl/nl-nl/achtergrond/2011/39/bootstrapping-combined-estimator-based-on-register-and-sample-survey-data.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Meertens, Q. A., C. G. H. Diks, H. J. van den Herik, and F. W. Takes (2020). A data-driven supply-side approach for estimating cross-border internet purchases within the European Union. *Journal of the Royal Statistical Society Series A 183*(1), 61–90.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics 12*(2), 685–726.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association 112*, 1477–1489.

Pavlopoulos, D. and J. K Vermunt (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology 41*, 197–214.

Rao, J. N. K. (2003). *Small area estimation*. Hoboken, New Jersey: John Wiley & Sons.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B: The Indian Journal of Statistics*, In press.

Reid, G., F. Zabala, 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.

Rocci, F., R. Varriale, and O. Luzi (2018). A proposal of an evaluation framework for processes based on the use of administrative data. Paper presented at the UNECE Workshop on Statistical Data Editing, NeuchÃćtel, Switzerland. Available at https://www.unece.org/index.php?id=47802.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Sax, C. and P. Steiner (2013). Temporal disaggregation of time series. *The R Journal 5*(2), 80–87.

Schnetzer, M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk, and M. Moser (2015). Quality assessment of imputations in administrative data. *Journal of Official Statistics 31*(2), 231–247.

Scholtus, S., B. F. M. Bakker, and A. van Delden (2015). Modelling measurement error to estimate bias in administrative and survey variables. Technical report, discussion paper, Statistics Netherlands, The Hague. Available at https://www.cbs.nl/nl-nl/achtergrond/2015/46/modelling-measurement-error-to-estimate-bias-in-administrative-and-survey-variables.

Scholtus, S. and J. Daalmans (2020). Variance estimation after mass imputation based on combined administrative and survey data. *Journal of Official Statistics*. Accepted for publication.

Sekar, C. C. and W. E. Deming (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association 44*(245), 101–115.

Stone, J. R. N., D. A. Champernowne, and J. E. Maede (1942). The precision of the national income accounting estimates. *Review of Economic Studies 9*, 111–125.

Van Delden, A., S. Scholtus, and J. Burger (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics 32*(3), 619–642.

Van der Heijden, P. G. M., J. Whittaker, M. J. L. F. Cruyff, B. F. M. Bakker, and H.N. van der Vliet (2012). People born in the middle east but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics 6*, 831–852.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wolter, K. M. (2007). *Introduction to variance estimation* (2 ed.). New York: Springer.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica 66*(1), 41–63.