

Quality measures for multisource statistics

Ton de Waal^{a,b,*}, Arnout van Delden^b and Sander Scholtus^b

^a*Tilburg University, Tilburg, The Netherlands*

^b*Department of Process Development and Methodology, Statistics Netherlands, 2490 HA The Hague, The Netherlands*

Abstract. The ESSnet on Quality of Multisource Statistics is part of the ESS.VIP Admin Project. The main objectives of that latter project are (i) to improve the use of administrative data sources and (ii) to support the quality assurance of the output produced using administrative sources. The ultimate aim of the ESSnet is to produce quality guidelines for National Statistics Institutes (NSIs) that are specific enough to be used in statistical production at those NSIs. The guidelines aim to cover the diversity of situations in which NSIs work as well as restrictions on data availability. The guidelines will list a variety of potential measures, indicate for each of them their applicability and in what situation it is preferred or not, and provide an ample set of examples of specific cases and decision-making processes. Work Package 3 (WP 3) of the ESSnet focuses on developing and testing quantitative measures for measuring the quality of output based on multiple data sources and on methods to compute such measures. In particular, WP 3 focuses on non-sampling errors. Well-known examples of such quality measures are bias and variance of the estimated output. Methods for computing these and other quality measures often depend on the specific data sources. Therefore, we have identified several basic data configurations for the use of administrative data sources in combination with other sources, for which we propose, revise and test quantitative measures for the accuracy and coherence of the output. In this article we discuss the identified basic data configurations, the approach taken in WP 3, and give some examples of quality measures and methods to compute those measures. We also point out some topics for future work.

Keywords: Administrative data, multi-source statistics, quality measures, survey data

1. Introduction: A short fairy tale

Once upon a time, long ago, in a country far, far away, there lived some good and honest people. These good people were struggling to produce single-source statistics, and did that to the best of their abilities. Still, they were always hoping for more data. They said to each other: “See, if only we had more data, we could have produced more and better statistics.” Their biggest wish in life was to produce multisource statistics. One day a good fairy or wicked witch – the story is not too specific about this – showed up and told them that she saw how hard the good people had to work in order to produce their single-source statistics, and that she had decided to fulfill their biggest wish and give them all the data they ever wanted and

more than that. The good people were very excited and said: “Thank you, dear fairy.” The people immediately started to work on their multisource statistics and were very happy at first. Soon they realized, however, that now they had more problems and had to work harder than ever before. Suddenly they could see measurement errors in their data, and linkage errors, and coverage errors, and new estimation problems, and so on and so on. They were even starting to think that perhaps it was a wicked witch after all who had given them all these data. Nevertheless, they worked hard, even harder and still even harder, and finally managed to produce multisource statistics. Proudly, they went to their king to show him their results. The king, a good and wise man, was impressed by the work they had done. However, after he had expressed his admiration he then asked the question: “And what can you tell me about the quality of your statistics?” That is where this fairy tale ends for now, and the ESSnet on Quality of Multisource Statistics starts.

*Corresponding author: Corresponding author: Ton de Waal, Statistics Netherlands, The Hague, The Netherlands. Tel.: +31 703374930; E-mail: t.dewaal@cbs.nl.

The ESSnet on Quality of Multisource Statistics (also referred to as Komuso) is part of the ESS.VIP Admin Project. The main objectives of the ESS.VIP Admin Project are (i) to improve the use of administrative data sources and (ii) to support the quality assurance of the output produced using administrative sources.

Partners in Komuso are 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 first Specific Grant Agreement (SGA) of the ESSNet lasted from January 2016 until April 2017. At the time of writing this article we are finalizing SGA 2. This second SGA started in May 2017 and lasts until mid-October 2018. A potential SGA 3 might start at the end of 2018. In the first and second SGAs of Komuso, Work Package 3 (WP 3) 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. Statistics Netherlands was project leader of this work package in both SGA 1 and SGA 2.

The problem encountered in WP 3 is not so much in defining the quality measures that one would like to use. For instance, with respect to the quality dimension “accuracy” most National Statistical Institutes (NSIs) would like to use bias, variances and/or mean squared errors of their estimates as quality measures. The main problem is rather how these quality measures should be computed for a given set of input data sets and a certain procedure for combining these input data sets. In other words, the main problem is describing a recipe for calculating quality measures for a given multisource situation.

This problem is much more complicated for multisource statistics than for single source statistics. We give two reasons for this. The first reason is that the procedure for combining the various input data sets has to be accounted for in the quality measure(s). Such procedures may involve many different processing steps and can be very complicated.

The second reason is that, due to the abundance of data, errors become much more visible. For example, when one has two data sets with (supposedly) the same variable, the values of this variable in the different data sets may differ due to measurement errors. A correction procedure for measurement error is then highly de-

sired and should be accounted for in the quality measures for output based on these data sets. Another example is when one has two data sets that are supposed to cover the same target population. In such a case one will often notice that they actually do not. That is, coverage problems can become visible simply by comparing units in different data sets. Also, linkage problems often occur when trying to link units in different data sets. Again, correcting for coverage and linkage errors is highly desired and should be accounted for in the quality measures for output based on the involved data sets. In contrast, in single-source statistics, one often focusses on sampling errors only, since due to the lack of data other kinds of errors are hard or even impossible to detect and correct.

In this article we discuss WP 3 of Komuso and some of the results obtained. Section 2 describes the approach taken in WP 3. Section 3 gives some examples of quality measures and methods to compute them. All examples are based on work (partly) done by Statistics Netherlands. Section 4 concludes the article with a brief discussion.

2. Approach taken in WP 3

In WP 3 we have subdivided the work into three consecutive steps:

1. We carry out a literature review or suitability test. In a literature review we study and describe existing quality measures and recipes to compute them. In a suitability test we go a step further and also test quality measures and recipes to compute them, either already known ones or newly proposed ones. In such a suitability test we examine practical and theoretical aspects of a quality measure and the accompanying calculation recipe.
2. We produce so-called Quality Measures and Computation Methods (QMCMs). Such a QMCM is a 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.
3. We provide hands-on examples to some of the QMCMs.

In SGA 1 of Komuso, the focus in WP 3 was on carrying out literature reviews and suitability tests for the quality dimension “accuracy” (principle 12 in the European Statistics Code of Practice, see [1]). In SGA 2 the focus is on producing QMCMs and a selected

number of hands-on examples for the literature reviews and suitability tests from SGA 1, and on carrying out suitability tests for the quality dimension “coherence” (principle 14 in the European Statistics Code of Practice, see [1]). In a potential SGA 3 we hope to produce more examples and produce some QMCMs for “coherence” and other quality dimensions, such as “reliability”.

WP 3 is strongly related to WP 1 of Komuso in SGA 2 (and probably also in a potential SGA 3). In WP 1 quality guidelines for multisource statistics are produced. The QMCMs and hand-on examples thereof produced by WP 3 will form an Annex to these quality guidelines for multisource statistics.

Many different situations can arise when multiple data sets are used to produce statistical output, depending on the nature of the data sets and on the kind of output produced. In order to structure the work within WP 3 we use a breakdown into a number of Basic Data Configurations (BDCs) that are most commonly encountered in practice at NSIs. The aim of the BDCs is to provide a useful focus and direction for the work to be carried out. In Komuso we have identified 6 BDCs:

- BDC 1: Multiple non-overlapping cross-sectional microdata sets that together provide a complete data set without any under-coverage problems;
- BDC 2: Same as BDC 1, but with overlapping variables and units between different data sets;
- 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 sets that need to be reconciled over time.

BDC 1 can be subdivided into two cases: the split-variable case where the data sets contain different variables for the same units and the split-population case where the data sets contain the same variables for different units. For more information on BDCs and methods to produce multi-source statistics we refer to [2].

3. Examples of QMCMs

In total 23 QMCMs are planned to be produced for WP 3 in SGA 2. The vast majority of the QMCMs relate to BDC 2 (“overlapping variables and units between different data sets”), which appears to be the most common and most important situation with re-

spect to multisource statistics at NSIs. For some other BDCs, such as BDCs 3 and 6 we will produce only a few QMCMs; for instance in the case of BDC 3 we will produce only one QMCM. Reasons for producing few QMCMs for BDC 3 and BDC 6 are either that the situation for multisource statistics is similar to the situation for single-source statistics (BDC 3) or, conversely, that NSIs have only very limited experience with the estimation of output quality in a multisource context (BDC6).

Since a complete description of the work done in WP 3 is impossible given the limited length of this article, we limit ourselves to giving some examples of QMCMs.

3.1. Mean squared error of level estimates affected by classification errors – BDC 1 “multiple non-overlapping cross-sectional microdata sets”

In this example we assume that we have several non-overlapping data sets together covering the entire target population and that the only source of errors are classification errors. In this section we will assume that the data are on businesses, which are classified by industry code (main economic activity). The unobserved true industry code of unit i is denoted by s_i ; the observed industry code that is prone to errors is denoted by \hat{s}_i . The set of possible industry codes is denoted by \mathcal{H} .

Let $\theta = f(y_1, \dots, y_N, s_1, \dots, s_N)$ denote a target parameter and y_i stand for the value of a target variable for unit i . Based on the observed data, this parameter is estimated by $\hat{\theta} = f(y_1, \dots, y_N, \hat{s}_1, \dots, \hat{s}_N)$. We are interested to estimate the mean squared error of $\hat{\theta}$ as affected by classification errors. We assume that the values of the target variable, y_1, \dots, y_N , are error-free.

[3] assumes that a business i with a true industry code g is classified as falling into class h with probability p_{ghi} due to classification error. They propose the following method for computing the mean squared error of $\hat{\theta}$. The first step is to estimate the probabilities p_{ghi} . This requires an independent collection of data on the classification variable, where observed and cleaned versions of those data are needed. Options to obtain such data are:

- draw a specific audit sample that is cleaned from errors;
- use information from the editing step in the regular process of producing statistics where units are verified on classification errors;

- the classification variable may also be present in a central business register or in a central population register. A regular quality assessment procedure of the register may then be used.

Together the estimated probabilities \hat{p}_{ghi} form a transition matrix $\hat{\mathbf{P}}_i = (\hat{p}_{ghi})$. The transition matrix, per unit, can be modelled as a function of background variables. See [3] for an example.

The second step is to estimate bias and variance of $\hat{\theta}$ by drawing bootstrap samples from $\hat{\mathbf{P}}_i$. For each unit we draw a new value for the industry code, given the original observed industry code \hat{s}_i , according to $\hat{\mathbf{P}}_i$. Based on the results for this draw, denoted by \hat{s}_{ir} , we compute $\hat{\theta}_r = f(y_1, \dots, y_N, \hat{s}_{1r}, \dots, \hat{s}_{Nr})$. We repeat this procedure R times, thus $r = 1, \dots, R$, and use the set of outcomes $\hat{\theta}_r$ to compute estimates of the bias and variance of $\hat{\theta}$:

$$\hat{B}_R(\hat{\theta}) = m_R(\hat{\theta}) - \hat{\theta},$$

$$\hat{Var}_R(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^R \left\{ \hat{\theta}_r - m_R(\hat{\theta}) \right\}^2.$$

with $m_R(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_r$.

When the target parameter θ concerns a vector of stratum totals (level estimates), one can also estimate the bias and variance-covariance matrix through analytical formulae, with $R \rightarrow \infty$: (cf. [3]):

$$\hat{B}_\infty^*(\hat{\mathbf{y}}) = \sum_{i=1}^N \left(\hat{\mathbf{P}}_i^t - \mathbf{I} \right) \hat{\mathbf{a}}_i y_i,$$

$$\hat{Var}_\infty^*(\hat{\mathbf{y}}) = \sum_{i=1}^N \left\{ \text{diag} \left(\hat{\mathbf{P}}_i^t \hat{\mathbf{a}}_i y_i^2 \right) - \hat{\mathbf{P}}_i^t \text{diag} \left(\hat{\mathbf{a}}_i y_i^2 \right) \hat{\mathbf{P}}_i \right\} \quad (1)$$

where superscript t indicates taking the transpose, $*$ indicates the analytical expression, \mathbf{I} is an identity matrix, $\hat{\mathbf{y}} = \sum_{i=1}^N \hat{\mathbf{a}}_i y_i$ stands for a vector of estimated stratum totals and $\hat{\mathbf{a}}_i = (\hat{a}_{1i}, \dots, \hat{a}_{|\mathcal{H}|i})^t$ is a vector of dummy variables that describes in which stratum unit i is observed ($\hat{a}_{hi} = 1$ if $\hat{s}_i = h$ and $\hat{a}_{hi} = 0$ otherwise). In particular, it follows that the bias of the estimated total in stratum h , $\hat{Y}_h = \sum_{i=1}^N \hat{a}_{hi} y_i$, can be estimated by

$$\hat{B}_\infty^*(\hat{Y}_h) = \sum_{i=1}^N \left\{ (\hat{p}_{hhi} - 1) \hat{a}_{hi} y_i + \sum_{\substack{g \in \mathcal{H}, \\ g \neq h}} \hat{p}_{ghi} \hat{a}_{gi} y_i \right\}$$

and its variance can be estimated by

$$\hat{Var}_\infty^*(\hat{Y}_h) = \sum_{i=1}^N \sum_{g \in \mathcal{H}} \hat{p}_{ghi} (1 - \hat{p}_{ghi}) \hat{a}_{gi} y_i^2$$

[4] developed a method to estimate quarterly turnover by Dutch consumers on European webshops. To that end, they use the tax file returns in the Netherlands, which includes selling of goods and services by European webshops to Dutch customers. From the tax data, quarterly turnover figures can be derived for companies that exceed a certain turnover threshold. The smaller companies are not considered. The challenge of the methodology is to identify which of the records in those tax data concern European webshops. These records are identified using a binary classifier, based on machine-learning, which separates the companies into those that belong to European webshops and those that do not. The predictions by the algorithms are not error-free, they make classification errors. In this example we explain how the uncertainty of the estimated quarterly turnover due to those classification errors is estimated.

[4] constructed a training data set of 180 companies and a test set of 79 companies by manually classifying them. Within the training data set, 76 webshops were identified and within the test set of 79 companies 13 webshops were identified.

Using the training set, the parameters of the machine-learning algorithms were estimated. These trained algorithms were used to predict the scores \hat{s}_i for the units in the test set and compare them with the true scores s_i . For the final algorithm, the result is shown in the transition matrix below, with s_i given as the rows and \hat{s}_i as the columns. The top row concerns having a webshop, the bottom row having no webshop.

$$\hat{\mathbf{P}} = \begin{pmatrix} 8/13 & 5/13 \\ 4/66 & 62/66 \end{pmatrix}$$

[4] assumed that this transition matrix is the same for all units i in the population.

Next, the machine-learning model was used to predict \hat{s}_i for the units not in the training set. Denote \mathbf{a}_i for the 2-vector $[\text{Ind}(s_i = 1), \text{Ind}(s_i = 0)]^t$, where Ind stands for an indicator variable. The authors of [4] were interested in the aggregate turnover $\mathbf{y} = \sum_{i \in C} \mathbf{a}_i y_i$ where C stands for the population of companies. The turnover in the first class, i.e. with $\text{Ind}(s_i = 1)$, is the target parameter. For the units in the training set, denoted by C_M , the values of the vector \mathbf{a}_i are deter-

Table 1
Final results in millions of euros

Year	y_{M1}	\hat{y}_{A1}	\hat{y}_1	$\hat{B}_\infty^*(\hat{y}_{A1})$	$\widehat{Sd}(\hat{y}_1)$
2014	405	495	837	63	97
2015	565	586	1,132	21	101
2016	725	667	1,372	19	110

mined manually and considered to be error-free. For the remaining units \mathbf{a}_i is predicted by the machine-learning algorithm. The total aggregated turnover is estimated by

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{y}_M + \hat{\mathbf{y}}_A - \hat{B}_\infty^*(\hat{\mathbf{y}}_A) = \mathbf{y}_M + (2\mathbf{I} - \hat{\mathbf{P}}^t)\hat{\mathbf{y}}_A \\ &= \sum_{i \in C_M} \mathbf{a}_i y_i + (2\mathbf{I} - \hat{\mathbf{P}}^t) \sum_{i \in C \setminus C_M} \hat{\mathbf{a}}_i y_i.\end{aligned}$$

where \mathbf{y}_M stands for turnover of units in the training set which are checked manually, \mathbf{y}_A stands for the true turnover of units not in the training set but which are classified by the algorithm, $\hat{\mathbf{y}}_A$ stands for the estimate of \mathbf{y}_A , and $\hat{B}_\infty^*(\hat{\mathbf{y}}_A)$ is the estimated bias of $\hat{\mathbf{y}}_A$ [cf. expression (1)]. The variance of the final turnover estimate $\hat{\mathbf{y}}$ is estimated in [4] by $(2\mathbf{I} - \hat{\mathbf{P}}^t)\widehat{Var}_\infty^*(\hat{\mathbf{y}}_A)(2\mathbf{I} - \hat{\mathbf{P}}^t)^t$, with $\widehat{Var}_\infty^*(\hat{\mathbf{y}}_A)$ given by Eq. (1). This variance estimate ignores the additional uncertainty due to estimating \mathbf{P} . Note that \mathbf{y}_M does not contribute to the variance.

The final results on companies estimated to be European retailers with a webshop can be found in Table 1 (taken from [4]). The manually checked turnover of European webshops, y_{M1} , consists of about half the value of the total estimated turnover, \hat{y}_1 . The estimated bias of \hat{y}_{A1} varied considerably over the years with the largest values in 2014. The estimated standard deviation of \hat{y}_1 suggests that the margin around the final estimate is still rather large. If we assume a normal distribution, the 95% confidence interval around \hat{y}_1 in 2014 would be 837 ± 190 . This corresponds to a relative margin of 22%.

[5,6] propose an extension of the approach in this section in order to quantify the effect of classification errors on the accuracy of growth rates in business statistics, rather than the effect on the accuracy of level estimates.

3.2. Variance of estimates based on reconciled microdata containing measurement errors – BDC 2 “overlapping variables and units”

The situation in this example is that a categorical target value is measured for each individual unit (with measurement error) in several data sources. We assume

that a Latent Class (LC) model is used to estimate the true values of this target variable. The quality of estimates based on the reconciled microdata is then measured by the estimated variance of these estimates.

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_s)^t$ denote a vector of observed categorical variables that measure the same conceptual variable of interest (for instance, in s different data sources). The true value with respect to the variable of interest is represented by a latent class variable X . We assume for convenience that all variables Y_j and X have the same set of categories, say $\{1, \dots, L\}$. The marginal probability $\Pr(\mathbf{Y} = \mathbf{y})$ of observing a particular vector of values $\mathbf{y} = (y_1, y_2, \dots, y_s)^t$ can be expressed as the sum of the joint probabilities

$$\begin{aligned}\Pr(X = x, \mathbf{Y} = \mathbf{y}) &= \Pr(X = x) \\ &\quad \Pr(\mathbf{Y} = \mathbf{y} | X = x)\end{aligned}$$

over all possible latent classes:

$$\begin{aligned}\Pr(\mathbf{Y} = \mathbf{y}) &= \sum_{x=1}^L \Pr(X = x) \\ &\quad \Pr(\mathbf{Y} = \mathbf{y} | X = x)\end{aligned}\quad (2)$$

A common assumption in LC analysis is that the classification errors in different observed variables are *conditionally independent*, given the true value (local independence), i.e.

$$\begin{aligned}\Pr(\mathbf{Y} = \mathbf{y} | X = x) &= \Pr(Y_1 = y_1 | X = x) \\ &\quad \Pr(Y_2 = y_2 | X = x) \dots \\ &\quad \Pr(Y_s = y_s | X = x).\end{aligned}$$

In combination with Eq. (2), this leads to the basic LC model

$$\begin{aligned}\Pr(\mathbf{Y} = \mathbf{y}) &= \sum_{x=1}^L \Pr(X = x) \prod_{j=1}^s \\ &\quad \Pr(Y_j = y_j | X = x)\end{aligned}$$

Estimating the LC model amounts to estimating the probabilities in this expression. Probabilities of the form $\Pr(Y_j = y_j | X = x)$ provide information about classification errors in the observed data. For example, units that in reality belong to the first category of the target variable ($X = 1$) are misclassified on observed variable Y_j with probability $\Pr(Y_j \neq 1 | X = 1) = 1 - \Pr(Y_j = 1 | X = 1)$. This can be seen as a quality measure for the quality of the j -th observed variable.

The model can also be used to estimate, for each unit in the data, the probability of belonging to a particular

latent class, given its vector of observed values. Using Bayes' rule, it follows that:

$$\Pr(X = x | \mathbf{Y} = \mathbf{y}) = \frac{\Pr(X = x) \prod_{j=1}^s \Pr(Y_j = y_j | X = x)}{\sum_{x'=1}^L \Pr(X = x') \prod_{j=1}^s \Pr(Y_j = y_j | X = x')} \quad (3)$$

Edit restrictions, for instance the edit restriction that someone who receives rent benefit cannot be a home owner, can be imposed by setting certain conditional probabilities equal to zero; for instance:

$$\Pr(X = \text{owner} | Y = \text{rent benefit}) = 0$$

The so-called MILC method (see [7]) takes measurement errors into account by combining Multiple Imputation (MI) and LC analysis. The method starts with linking all data sets on the unit level, and then proceeds with 5 steps.

1. Select m bootstrap samples from the original combined data set.
2. Create an LC model for every bootstrap sample.
3. Multiply impute latent "true" variable X for each bootstrap sample. That is, create m empty variables (W_1, \dots, W_m) and impute them by drawing one of the categories using the estimated posterior membership probabilities Eq. (3) from the LC model.
4. Obtain estimates of interest from each data set with imputed variables.
5. Pool the estimates using Rubin's pooling rules for multiple imputation (see [8]). An essential aspect of these pooling rules is that an estimated variance of the pooled estimate is obtained. This estimated variance is a quality measure for the reconciled data.

[7] applied the MILC method on a combined data set to measure home ownership. This combined data set consisted of data from the LISS (Longitudinal Internet Studies for the Social sciences) panel from 2013 and a register from Statistics Netherlands from 2013. From this combined data set, they used two variables indicating whether a person is a home-owner or rents a home as indicators for the imputed "true" latent variable home-owner/renter (or other). The combined data set also contained a variable measuring whether someone receives rent benefit from the government. A person can only receive rent benefit if this person rents a house. Moreover, a variable indicating whether a person is married or not was included in the latent class model as a covariate. The three data sets used to combine the data are:

- BAG: A register containing data on addresses and buildings originating from municipalities from 2013. From the BAG, [7] used a variable indicating whether a person "owns"/ "rents (or other)" the house he or she lives in.
- LISS background study: A web survey on general background variables from January 2013. [7] used the variable marital status. They also used a variable indicating whether someone is a "(co-)owner" and "(sub-)tenant or other".
- LISS housing study: A web survey on housing from June 2013. From this survey [7] used the variable rent benefit, indicating whether someone "receives rent benefit", "does not receive rent benefit", or "prefers not to say".

These data sets were linked on a unit level, and matching was done on person identification numbers. Not every individual is observed in every data set. This causes that some missing values are introduced when the different data sets are linked on a unit level. Full Information Maximum Likelihood (see, e.g. [9]) was used to handle the missing values when estimating the LC model. The MILC method was applied to impute the latent variable home owner/renter (or other) by using two indicator variables and two covariates.

As already explained, the MILC method can be used to assess the quality of the input sources. In Table 2 classification probabilities of the models, estimated by means of the MILC method, are given. The higher these probabilities, the higher the quality of the input data.

To give an example of how to measure the quality of – even quite complicated – aspects of the combined data set, [7] used a logit model to predict home ownership by means of marriage. By using Rubin's pooling rules on the imputations produced by the MILC method they obtained the following estimates for the intercept and regression coefficient: 2.7712 and -1.3817 . This means that the estimated odds of owning a home when not married are $e^{-1.3817} = 0.25$ times the odds when married. The 95% confidence interval of the estimated intercept is given by [2.5036; 3.0389], and the 95% confidence interval of the estimated regression coefficient by $[-1.6493; -1.1140]$. These 95% confidence intervals provide quality measures for this aspect of the combined data set. The smaller these confidence intervals, the more accurate the estimates based on the combined data set.

In a similar way, by using Rubin's pooling rules on the imputations produced by the MILC method, variances and confidence intervals for other estimands can be estimated.

Table 2
Classification probabilities for LISS and BAG

	Pr (observed = rent true = rent)	Pr (observed = own true = own)
LISS	0.9344	0.9992
BAG	0.9496	0.9525

3.3. *Validity and measurement bias of observed numerical variables – BDC 2 “overlapping variables and units”*

In this example we again have several indicators (with measurement error) for target variables that we use to estimate the true values of these target variables. In particular, the true distribution of one or more numerical target variables, which are measured (with measurement error) for individual units in several linked data sets, is estimated, as well as the relation between each target variable and its associated observed variables. From this, it can be assessed to what extent each observed variable is a valid indicator of its target variable, and to what extent measurement bias occurs. The quality measure and related calculation method can be seen as equivalents for numerical data of the quality measure and calculation method for categorical data based on latent class analysis given in the previous example.

The *validity coefficient* of an observed variable is defined as the absolute value of its correlation with the associated target variable. In the context of the model used here, this coefficient captures the effect of random measurement errors in the observed data. The validity coefficient lies between 0 and 1, with values closer to 1 indicating better measurement quality (absence of random measurement error).

Measurement bias indicates to what extent values of an observed variable are systematically larger or smaller than the true values of the associated target variable. Under the assumption that the relation between the target and observed variable is linear, the measurement bias is summarized in terms of *intercept bias* and *slope bias*. Intercept bias indicates a constant shift that occurs for all values; for instance, observed values are on average €1,000 too large. Slope bias indicates a shift that is proportional to the true value; for instance, observed values are inflated by on average 5%. Ideally, the intercept and slope bias are both zero. In the case of administrative data, measurement bias can occur in particular due to conceptual differences between the variable that is used for administrative purposes and the target variable that is needed for statistical production (see, e.g. [10]).

Suppose that one has a linked micro data set with observed variables y_1, \dots, y_p from different sources. The underlying “true” target variables are not observed directly and denoted by latent variables η_1, \dots, η_m . For simplicity, it is assumed that each observed variable y_k is an indicator of exactly one target variable $\eta_{j(k)}$. By contrast, it is assumed that the same target variable is measured by multiple (at least two) observed variables (so $p > m$).

A linear structural equation model (SEM) for these data consists of two sets of regression equations. Firstly, there are *measurement equations* that relate the observed variables to the latent variables:

$$y_k = \tau_k + \lambda_k \eta_{j(k)} + \epsilon_k, (k = 1, \dots, p) \quad (4)$$

Here, τ_k denotes a measurement intercept, λ_k denotes a slope parameter (also known as a *factor loading* in this context), and ϵ_k denotes a zero-mean random measurement error that affects y_k . It is often assumed that the random errors ϵ_k and ϵ_l are uncorrelated for $k \neq l$.

Second, the SEM may contain *structural equations* that relate different latent variables to each other:

$$\eta_j = \alpha_j + \sum_{\substack{g=1 \\ (g \neq j)}}^m \beta_{jg} \eta_g + \zeta_j, (j = 1, \dots, m)$$

Here, α_j denotes a structural intercept and ζ_j denotes a zero-mean disturbance term. The coefficient β_{jg} represents a direct effect of variable η_g on variable η_j (with $g \neq j$). In practice, some of these coefficients are usually set equal to zero when the model is specified, based on substantive considerations.

Once the SEM has been estimated, the validity and measurement bias of the observed variables can be assessed from the model parameters. The *validity coefficient VC* of y_k as an indicator for the target variable $\eta_{j(k)}$ is defined as the absolute value of the correlation between y_k and $\eta_{j(k)}$. It can be shown that this correlation is equal to the standardized version of λ_k , say $\lambda_k^{(s)}$, so that:

$$\begin{aligned} VC(y_k) &= |\lambda_k^{(s)}| = |\lambda_k| \frac{sd(\eta_{j(k)})}{sd(y_k)} \\ &= \sqrt{1 - \frac{\text{var}(\epsilon_k)}{\text{var}(y_k)}} \end{aligned}$$

In addition, the parameters τ_k and λ_k provide information about measurement bias in y_k with respect to $\eta_{j(k)}$. If no bias occurs, then it holds that $\tau_k = 0$ and $\lambda_k = 1$; cf. (4). *Intercept bias* is indicated by a deviation of τ_k from 0; *slope bias* is indicated by a deviation of λ_k from 1.

For any SEM, it has to be checked whether all model parameters can be identified from the available data. Here, a distinction occurs between applications where only the validity is estimated and applications where also the intercept and slope bias are estimated. In the first case, the model can be identified with $m \geq 2$ correlated latent variables and at least two indicators for each latent variable, or with $m = 1$ latent variable that has at least three indicators (see [11]). Identification of the model may also be improved by including covariates that are considered to be measured (essentially) without error.

To assess the “true” measurement bias, an additional assumption is needed to fix the “true” scale of each latent variable. (Note that this scale is not relevant for the validity coefficient, since it is defined as a correlation.) Following [12,13], [14] suggests to identify the model in this case by collecting additional “gold standard” data for a small random subsample of the original data set (an *audit sample*).

Figure 1 shows an example of an SEM that is identified by means of an audit sample. The model contains three latent variables, each of which is measured by two ordinary, error-prone observed variables. For the units that are selected in the audit sample, additionally observed variables y_7 , y_8 and y_9 are obtained that are supposed to measure the latent variables without error. Thus, the measurement equations for these observed variables are simply: $y_7 = \eta_1$, $y_8 = \eta_2$ and $y_9 = \eta_3$. The model is divided into two groups: group 1 represents the audit sample and group 2 the remaining units without additional “gold standard” variables. In group 1, the model is identified by means of the error-free audit data. In group 2, the model is identified by restricting all model parameters in this group to be equal to the corresponding parameters in group 1. This restriction is meaningful, because the audit sample has been selected by random subsampling from the original data set.

In practice, the “gold standard” audit data may be obtained by letting subject-matter experts re-edit a random subset of the original observations. Results on simulated data in [15] suggest that only a small audit sample (say, 50 units) is needed.

An alternative way to identify the model, without the need to collect additional audit data, is to assume that

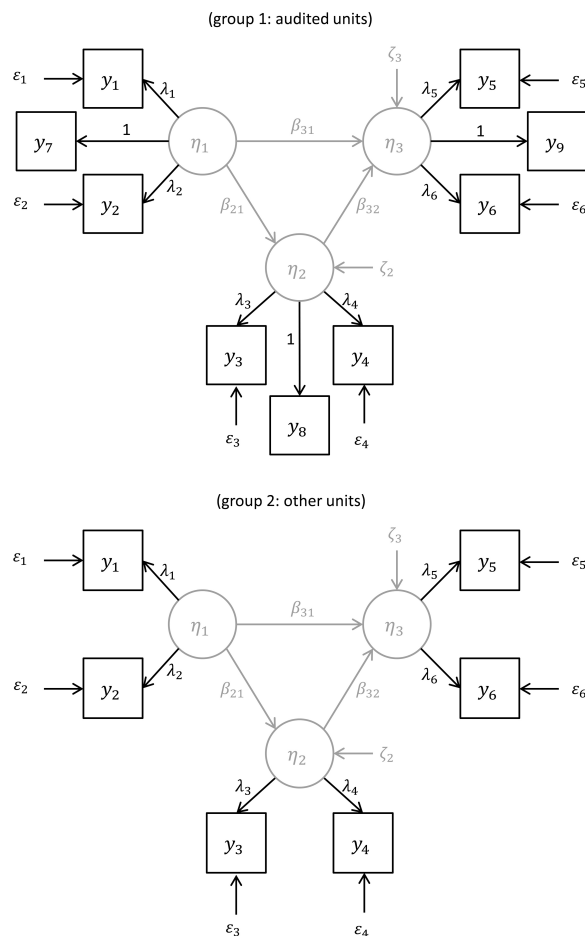


Fig. 1. Example of a two-group SEM identified by an audit sample.

one of the observed variables for each latent variable does not contain systematic measurement errors (while still allowing for random errors). Then the model may be identified by restricting $\tau_k = 0$ and $\lambda_k = 1$ for these variables. In some cases, this assumption may be reasonable, e.g., for survey variables. However, the assumption cannot be tested with the data at hand (the same holds for the assumption that audit data are error-free).

To estimate the parameters of an SEM, a standard approach is to apply maximum likelihood estimation under the assumption that the data consist of independent, identically distributed observations from a multivariate normal distribution (see [11]). A so-called Robust Maximum Likelihood or Pseudo Maximum Likelihood estimation procedure is available which can handle non-normality of the data as well as complex sampling designs for finite populations (see [16]). Details on how an audit sample can be formally incor-

porated into this estimation procedure can be found in [14]. Standard fit measures are available to evaluate whether a fitted SEM gives an adequate description of the observed data, and to compare the fit of different SEMs (see [11]).

The validity coefficients VC and the parameters τ_k and λ_k provide information about the quality of the input data y_k . They can also be used as input to a further procedure to assess the accuracy of output based on these data. As a very simple example, suppose that we are interested in the population mean of the true variable η : $\theta = (1/N) \sum_{i=1}^N \eta_i$. Suppose that we have two available estimators:

- The direct estimator based on a simple random sample without replacement of n units, S , where the target variable is measured by y_1 : $\hat{\theta}_1 = (1/n) \sum_{i \in S} y_{1i}$.
- An estimator based on a register that covers the entire population, where the target variable is measured by y_2 : $\hat{\theta}_2 = (1/N) \sum_{i=1}^N y_{2i}$.

The measurement model for the observed variables y_1 and y_2 is given by Eq. (4). Under this model, with the true values η_i treated as fixed, the following expressions can be derived for the mean squared error of the two estimators:

$$\begin{aligned} MSE(\hat{\theta}_1) &= \{\tau_1 + (\lambda_1 - 1)\theta\}^2 + \\ &\quad \frac{1}{n} \left\{1 - \frac{n}{N} VC^2(y_1)\right\} \sigma_{y1}^2; \\ MSE(\hat{\theta}_2) &= \{\tau_2 + (\lambda_2 - 1)\theta\}^2 + \\ &\quad \frac{1}{N} \{1 - VC^2(y_2)\} \sigma_{y2}^2 \end{aligned}$$

Here, σ_{y1}^2 and σ_{y2}^2 denote the expected population variances of the observed variables under the model. The first terms in the above expression for the mean squared errors are the squares of the corresponding biases, the second terms are the variances.

Thus, in this example the validity coefficients and intercept and slope bias can be used directly to quantify and compare the accuracy of different estimators for the same target parameter. In other situations, it may be too complicated to derive an analytical expression for the mean squared error of a proposed estimator. In that case, the results of the structural equation model could still be used as input for a resampling method (such as the bootstrap) to simulate the effect of measurement errors on the output accuracy.

3.4. Variance-covariance matrix for a vector reconciled by means of macro-integration – BDCs 4 and 5 “microdata and aggregated data”

Many statistical figures, for instance in the context of national economic and social accounting systems, are connected by known constraints. We refer to the equations which satisfy such constraints as accounting equations. Insofar as the initial input estimates need to be based on a variety of sources, they usually do not automatically satisfy the set of accounting equations due to the errors of estimates. An adjustment or reconciliation step is required, by which the input estimates are modified to conform to the constraints. Statistical figures sometimes also have to satisfy inequality constraints besides accounting constraints, and input data may need to be adjusted to satisfy these inequality constraints as well. Macro-integration is a technique used for reconciling statistical figures so they satisfy the known constraints.

Consider, for example, an economic model with five macro-economic quantities: national income y , consumption c , investments i , export x , and import m , which are stacked in a vector β . Suppose that there is one known national accounts identity $y = c + i + x - m$ and one known inequality $x \geq m$. Given is an initial vector of estimates $\beta_0 = (\hat{y}_0, \hat{c}_0, \hat{i}_0, \hat{x}_0, \hat{m}_0)$, where $\hat{y}_0 = 600$, $\hat{c}_0 = 445$, $\hat{i}_0 = 100$, $\hat{x}_0 = 515$ and $\hat{m}_0 = 475$. These input data do not satisfy the accounting equations, since $600 \neq 445 + 100 + 515 - 475$. By means of macro-integration the initial estimates can be reconciled so the constraints are satisfied.

Macro-integration can be used to reconcile aggregated data as in the above example. It can also be used to reconcile microdata with aggregated data by first transforming the microdata into estimates on an appropriate aggregated level, for instance by means of weighting these microdata. This approach is, for instance used for the Dutch Population Census (see, e.g. [17,18]).

[19] considers the following general macro-integration problem

$$\begin{aligned} \min_{\beta} \frac{1}{2} (\beta - \hat{\beta}_0)^t V^{-1} (\beta - \hat{\beta}_0) \\ \text{subject to } T\beta = g \text{ and } R\beta \geq h \end{aligned} \quad (5)$$

where $\hat{\beta}_0$ is an unbiased initial vector of estimates for the (numerical) quantities of interest obtained from several data sources with a known (or estimated) variance-covariance matrix V . T and R are matrices and g and h are vectors defining the constraints

that have to be satisfied by the quantities of interest. Problem Eq. (5) is a so-called Quadratic Programming problem.

Our aim is to calculate the variance of the vector of estimates β after reconciliation. The calculation formula for this variance depends on the type of restrictions that have to be obeyed. Four different cases are distinguished. These cases are discussed below.

3.4.1. Only equality restrictions

If there are only equality restrictions in Eq. (5), i.e. $T\beta = g$, the solution to Eq. (5) is given by

$$\hat{\beta}_{QP} = \hat{\beta}_0 + VT^t(TVT^t)^{-1}(g - T\hat{\beta}_0)$$

and

$$\text{Var}(\hat{\beta}_{QP}) = (I_k - VT^t(TVT^t)^{-1}T)V$$

where I_k is the $k \times k$ identity matrix.

3.4.2. Only one inequality restriction and no equality restrictions

When there is only one inequality restriction, the vector h reduces to a scalar h and the solution to Eq. (5) is given by

$$\hat{\beta}_{QP} = \begin{cases} \hat{\beta}_0 & \text{if } R\hat{\beta}_0 \geq h \\ \hat{\beta}_0 + K(h - R\hat{\beta}_0) & \text{if } R\hat{\beta}_0 < h \end{cases}$$

$\hat{\beta}_{QP}$ can be re-written as

$$\hat{\beta}_{QP} = \beta + K(s_{h+} - R\beta) + u \tag{6}$$

where $s_{h+} = \max(R\hat{\beta}_0, h)$, $u = \hat{\beta}_0 - E(\hat{\beta}_0|s) = (I_k - KR)(\hat{\beta}_0 - \beta)$, and $K = VR^t(RVR^t)^{-1}$. s_{h+} follows a censored normal distribution $CN(R\beta, RVR^t, h, \infty)$ and is independent of u . Hence,

$$\begin{aligned} \text{Var}(\hat{\beta}_{QP}) &= K\text{Var}(s_{h+})K^t + \text{Var}(u) \\ &= (I_k - (1 - d_2^2)KR)V \end{aligned} \tag{7}$$

The parameter d_2^2 follows from the variance formula for a censored normal distribution (see [19,20]).

3.4.3. Multiple inequality restrictions and no equality restrictions

When there are $r \geq 2$ inequality restrictions and no equality restrictions, [19] proposes to use Eq. (6) as an approximation for $\hat{\beta}_{QP}$ in order to evaluate the variance-covariance matrix. Furthermore, [19] assumes that the correlation coefficients between the elements of s_{h+} differ not too much, say less than 0.05,

from those between the elements of $R\hat{\beta}_0$. $\text{Var}(\hat{\beta}_{QP})$ can then be approximated by

$$\begin{aligned} \text{Var}(\hat{\beta}_{QP}) &\approx \text{Var}(Ks_{h+}) + (I_k - KR)V \\ &\approx KD_2RVR^tD_2K^t + (I_k - KR)V \end{aligned}$$

where $D_2 = \text{diag}(d_{21}, \dots, d_{2r})$ and the parameters d_{21}, \dots, d_{2r} again follow from the variance formula of a censored normal distribution.

3.4.4. Multiple equality and inequality restrictions

When we have a set of t equality restrictions $T\beta = g$ and a set of r inequality restrictions $R\beta \geq h$ two steps have to be carried out in order to estimate the variance-covariance matrix of the final reconciled vector $\hat{\beta}_{QP2}$. In the first step, we calculate

$$K_1 = VT^t(TVT^t)^{-1}$$

$$\hat{\beta}_{QP1} = \hat{\beta}_0 + K_1(g - T\hat{\beta}_0) \tag{8}$$

$$V_1 \equiv \text{Var}(\hat{\beta}_{QP1}) = (I_k - K_1T)V$$

In the second step we find the final solution $\hat{\beta}_{QP2}$ by solving

$$\begin{aligned} \min_{\beta} &\frac{1}{2}(\beta - \hat{\beta}_{QP1})^t V_1^{-1}(\beta - \hat{\beta}_{QP1}) \\ &\text{subject to } R\beta \geq h \end{aligned}$$

The variance-covariance matrix $\text{Var}(\hat{\beta}_{QP2})$ can now be estimated in the same way as for the case where we have only inequality restrictions, with $\hat{\beta}_0$ and V in Eq. (5) replaced by $\hat{\beta}_{QP1}$ and V_1 .

We return to our example with the five macro-economic quantities, y, c, i, x and m . We assume that the standard errors of the initial estimates are given by $se(\hat{y}_0) = 30.0$, $se(\hat{c}_0) = 22.0$, $se(\hat{i}_0) = 7.1$, $se(\hat{x}_0) = 28.3$ and $se(\hat{m}_0) = 28.3$. We also assume that the covariances are zero. Applying Eq. (8) to this example, we first obtain the following QP1 solution: $\hat{y}_{QP1} = 595.55$, $\hat{c}_{QP1} = 447.39$, $\hat{i}_{QP1} = 100.25$, $\hat{x}_{QP1} = 518.96$ and $\hat{m}_{QP1} = 471.04$.

This QP1 solution obeys the equality $y = c + i + x - m$. According to Eq. (8), the standard errors of $\hat{\beta}_{QP1}$ are given by: $se(\hat{y}_{QP1}) = 24.74$, $se(\hat{c}_{QP1}) = 20.02$, $se(\hat{i}_{QP1}) = 7.01$, $se(\hat{x}_{QP1}) = 24.00$ and $se(\hat{m}_{QP1}) = 24.00$.

This solution satisfies the inequality $x \geq m$. So, the solution $\hat{\beta}_{QP2}$ to the second step mentioned below Eq. (8) equals $\hat{\beta}_{QP1}$. We still need to take the effect of the inequality on the variances and standard errors into account, though. In order to do so, the variances

and the standard errors have to be adjusted according to Eq. (7) with $R = (0, 0, 0, 1, -1)^t$ and $V = V_1$ where V_1 is defined by Eq. (8). This yields the standard errors of the estimates: $se(\hat{y}_{QP2}) = 24.15$, $se(\hat{c}_{QP2}) = 19.81$, $se(\hat{i}_{QP2}) = 7.00$, $se(\hat{x}_{QP2}) = 23.61$ and $se(\hat{m}_{QP2}) = 23.61$.

As could be expected, these standard errors are smaller than the standard errors of $\hat{\beta}_{QP1}$ mentioned above, because the inequality restricts the region of the allowed values. Since there is only one inequality restriction in this example, these standard errors are exact under the assumptions of [19].

3.5. *Scalar measures of uncertainty in (economic) accounts – BDC 5 “aggregated data”*

We again consider accounting systems connected by known constraints. The uncertainty in such an accounting system is, in principle, given by a variance-covariance matrix of all variables involved in the accounting system. However, [21] considers such a system of accounting equations as a single entity and aims to define uncertainty measures that capture the adjustment effect as well as the relative contribution of the various input estimates to the final estimated account in a single scalar. [21] discussed two approaches: the covariance approach and the deviation approach. Below we sketch the covariance approach.

Consider the additive account $A = [Y_1 + \dots + Y_i + \dots + Y_p = Z]$. Let $\Sigma_{\tilde{X}}$ be the variance-covariance matrix of the adjusted estimates $\tilde{X} = (\tilde{Y}_1, \dots, \tilde{Y}_p, \tilde{Z})$. This matrix can be partitioned as follows: $\Sigma_{\tilde{X}} = \begin{pmatrix} \Sigma_{\tilde{Y}} & \Sigma_{\tilde{Y}\tilde{Z}} \\ \Sigma_{\tilde{Z}\tilde{Y}} & \Sigma_{\tilde{Z}} \end{pmatrix}$.

One scalar measure, denoted here by $\gamma(A)$, proposed in [21] is defined as the sum of the variances of the components of \tilde{X} , that is

$$\gamma(A) = \text{Trace}(\Sigma_{\tilde{X}}) = \text{Trace}(\Sigma_{\tilde{Y}}) + 1^t \Sigma_{\tilde{Y}} \quad (9)$$

$$1 = \sum_{i=1}^p \text{Var}(\tilde{Y}_i) + \text{Var}(\tilde{Z})$$

To illustrate the covariance approach let us suppose that we have normally distributed input estimates $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_p) \sim N_p(\mu, \Sigma_Y)$, where $\mu = (\mu_1, \dots, \mu_p)$ and Σ_Y is a diagonal matrix with diagonal values $\sigma_k^2 = \text{Var}(Y_k) = \sigma^2 \mu_k$ for $k = 1, \dots, p$.

Suppose we have an additive accounting equation of the form

$$Y_1 + \dots + Y_p = Z$$

where $\tilde{Z} = z$ is treated as fixed. The original values $\hat{Y}_1, \dots, \hat{Y}_p$ do not necessarily satisfy this accounting equation. Consider a common benchmarking method, which yields the adjusted estimates

$$\tilde{Y}_k = \hat{Y}_k + \left(z - \sum_{j=1}^p \hat{Y}_j \right) v_k$$

where the v_k are adjustment weights that sum up to 1 for $k = 1, \dots, p$. By this benchmarking method the total difference is simply apportioned to each component estimator. Suppose one would like to compare two choices: $v_k = 1/p$ and $v_k = \mu_k / \sum_{j=1}^p \mu_j$, which yield, respectively, the adjusted estimates

$$\tilde{Y}_{1k} = \hat{Y}_k + \frac{1}{p} \left(z - \sum_{j=1}^p \hat{Y}_j \right) \quad (10)$$

and

$$\tilde{Y}_{2k} = \hat{Y}_k + \frac{\mu_k}{\sum_{j=1}^p \mu_j} \left(z - \sum_{j=1}^p \hat{Y}_j \right) \quad (11)$$

We are interested in which of the two choices, Eq. (10) or Eq. (11), yields statistical data of higher quality. Denote the reconciled estimated account based on Eq. (10) by A_1 and the one based on Eq. (11) by A_2 .

It can be shown (see [21]) that the random variables $\tilde{Y}_k - M_k$ ($k = 1, \dots, p$) where $M_k = E(\tilde{Y}_k) = \mu_k + (z - \sum_{j=1}^p \mu_j) v_k$, are negatively correlated with each other and have normal distributions $N(0, \tilde{\sigma}_k^2)$ with

$$\tilde{\sigma}_k^2 = v_k^2 \sum_{j=1}^p \text{Var}(Y_j) + \text{Var}(Y_k) (1 - 2v_k)$$

$$= v_k^2 \sigma^2 \sum_{j=1}^p \mu_j + \sigma^2 \mu_k - 2v_k \sigma^2 \mu_k$$

Since $\text{Var}(\tilde{Z}) = 0$ because $\tilde{Z} = z$ is fixed, measure $\gamma(A)$ (see Eq. (9)) is by given by

$$\gamma(A) = \sum_{k=1}^p \text{Var}(\tilde{Y}_k) = \sum_{k=1}^p \tilde{\sigma}_k^2$$

We examine the difference between $\gamma(A_1)$ and $\gamma(A_2)$. This difference is given by

$$\begin{aligned}
\gamma(A_1) - \gamma(A_2) &= \sum_{k=1}^p \sigma^2 \\
&\left(\frac{1}{p^2} \sum_{j=1}^p \mu_j - \frac{2}{p} \mu_k - \left(\frac{\mu_k}{\sum_{j=1}^p \mu_j} \right)^2 \right. \\
&\quad \left. \sum_{j=1}^p \mu_j + \frac{2\mu_k}{\sum_{j=1}^p \mu_j} \mu_k \right) \\
&= \sum_{k=1}^p \sigma^2 \left(\frac{1}{p^2} \sum_{j=1}^p \mu_j - \frac{2}{p} \mu_k + \frac{\mu_k^2}{\sum_{j=1}^p \mu_j} \right) \\
&= \sum_{k=1}^p \sigma^2 \left(\frac{\sqrt{\sum_{j=1}^p \mu_j}}{p} - \frac{\mu_k}{\sqrt{\sum_{j=1}^p \mu_j}} \right)^2 \geq 0
\end{aligned}$$

This implies that $\gamma(A_1) \geq \gamma(A_2)$ with strict inequality unless all μ_k are equal. So, we conclude that adjustment method Eq. (10) always leads to more uncertainty than method Eq. (11) according to quality measure γ – unless all μ_k are equal in which case both adjustment methods are equivalent – and that reconciliation method Eq. (11) should therefore be preferred.

3.6. Other BDCs

With respect to BDC 3 (“under-coverage”) we have studied extensions of the well-known Petersen capture-recapture estimator (also referred to as the Petersen-Lincoln estimator). This estimator can be used to estimate the size of a population. In the simplest case of this estimator two random samples A and B from the same target population are linked. By n_{11} we denote the number of units that are observed in both samples A and B, by n_{10} the number of units that are observed in sample A only, by n_{01} the number of units that are observed in sample B only, and by n_{00} the number of population units that are not observed in either of the samples. The unknown population size equals $n_{11} + n_{10} + n_{01} + n_{00}$. This quantity is unknown since the value of n_{00} is not observed and unknown. The Petersen estimator for n_{00} is given by

$$\hat{n}_{00} = \frac{n_{10}n_{01}}{n_{11}}$$

Its variance (see, e.g., [22]), and hence the variance of the estimator for the population size $n_{11} + n_{10} + n_{01} + \hat{n}_{00}$, is given by

$$\text{Var}(\hat{n}_{00}) = \frac{(n_{10} + 1)(n_{01} + 1)(n_{10} - n_{11})(n_{10} - n_{11})}{(n_{11} + 1)^2 (n_{11} + 2)}$$

This variance estimate provides a quality measure for the population size estimation. In the literature review we carried out for WP 3, we studied how the Petersen estimator can be applied when the samples A and B are not obtained by survey sampling, but are instead based on administrative data (see, e.g. [23,24]).

With respect to BDC 6 (“longitudinal data”), we have, for instance, carried out a literature review of [25]. [25] presents an approach to estimate (and possibly correct for) the amount of classification error in longitudinal microdata of categorical variables by estimating a so-called Hidden Markov Model for the underlying “true” distribution. A Hidden Markov Model is a special type of latent class model suitable for longitudinal data. Besides assumptions on the latent class model for each time point also assumptions on the longitudinal structure of the data are needed. In the approach proposed in [25], the true distribution of a categorical target variable, which is measured (with measurement error) for individual units in several linked data sets at several time points, is estimated at each time point. The observed variables are seen as error-prone indicators of a latent true variable, with some assumptions about the distribution of measurement errors. From this, the misclassification rate in each observed variable at each time point can be estimated. The accuracy of observed higher-order properties, such as observed transition rates between categories over time, can be estimated. The approach proposed in [25] can be seen as a longitudinal variation on the approach proposed in [7], which was described in Section 3.2. We refer to [25] for more information about their approach.

4. Discussion

We hope that we have succeeded in giving a flavor of the work that has been and is being done in WP 3 of the ESSnet on Quality of Multisource Statistics and the results that have been achieved with respect to quality measures for output based on multiple data sources.

Future work could focus on improving the developed quality measures, and make them easier to apply

in practical situations as well as extend the range of situations in which they can be applied. Future work could also focus on developing quality measures and related computation methods that have not yet been considered in WP 3 of the ESSnet on Quality of Multisource Statistics.

Another important topic for future work is the further development of a systematic framework for situations, methods and quality measures that can arise in a multisource context. For instance, for single-source statistics where the focus is generally on sampling error, survey sampling theory offers the basics (and usually a lot more than just the basics) for computing survey estimates (by means of weighting) and variance estimates thereof. We feel that NSIs should aim for similar generally applicable theories that can be used for measurement errors, coverage errors, linkage errors etc. in a multisource context. In our opinion, NSIs should also aim for a general framework enabling one to combine all these separate aspects in an overall quality measure for output based on multiple data sets.

5. Back to the fairy tale

We do not know how the fairy tale from the beginning of our article will end, but we hope that it will end along the following lines.

The good and wise king looked at the work that had been done in the ESSnet, and told his people that he was once again impressed by the work they had done. He said: "I know this is not the final answer. Only the future may provide us the final answer. However, what you have done in this project is really useful and an important step forward. It will help not only us but also others in their efforts to determine the quality of multisource statistics." The king and his people lived happily ever after, constantly improving the quality measurement of their statistics.

Acknowledgments

This work has partly been carried out as part of the ESSnet on Quality of Multisource Statistics (Framework Partnership Agreement Number 07112.2015.003-2015.226), and has been funded by the European Commission.

An earlier version of this paper has been presented at the Q2018 conference, Krakow, Poland, June 2018.

References

- [1] Eurostat, European Statistics Code of Practice; 2018, Available at <https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142>.
- [2] De Waal T, Van Delden A, Scholtus S, Multisource statistics: basic situations and methods. Discussion paper, Statistics Netherlands; 2017, Available at www.cbs.nl.
- [3] Van Delden A, Scholtus S, Burger J, Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics* 2016; 32(3): 1-25.
- [4] Meertens QA, Diks CGH, Van den Herik HJ, Takes FW, Data-driven supply-side approach for measuring cross-border internet purchases; 2018, Available at: [arXiv1805.06930v1 \[statAP\]](https://arxiv.org/abs/1805.06930v1), accessed at 17 May 2018.
- [5] Van Delden A, Scholtus S, Burger J, Exploring the effect of time-related classification errors on the accuracy of growth rates in business statistics; 2016, Paper presented at the ICES V conference, 21–24 June 2016, Geneva.
- [6] Scholtus S, Van Delden A, Burger J, Analytical expressions for the accuracy of growth rates as affected by classification errors; 2017, Deliverable of SGA1 of the ESSnet on Quality of Multisource Statistics.
- [7] Boeschoten L, Oberski D, De Waal T, Estimating classification error under edit restrictions in combined survey-register data using multiple imputation latent class modelling (MILC), *Journal of Official Statistics* 2017; 33(4): 921-962.
- [8] Rubin DB, Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York; 1987.
- [9] Arbuckle JL, Full information estimation in the presence of incomplete data. In: Advanced structural equation modeling, (eds). Marcoulides G.A., Schumacker, R.E., Mahwah NJ. Lawrence Erlbaum Associates, Inc.; 1996, p. 243-277.
- [10] Van Delden A, Banning R, De Boer A, Pannekoek J, Analysing correspondence between administrative and survey data. *Statistical Journal of the IAOS* 2016; 32(4): 569-584.
- [11] Bollen KA, Structural equations with latent variables. John Wiley & Sons, New York; 1989.
- [12] Bielby WT, Arbitrary metrics in multiple-indicator models of latent variables. *Sociological Methods and Research* 1986; 15(1-2): 3-23.
- [13] Sobel ME, Arminger G, Platonic and operational true scores in covariance structure analysis. *Sociological Methods and Research* 1986; 15: 44-58.
- [14] Scholtus S, Bakker BFM, Van Delden A, Modelling measurement error to estimate bias in administrative and survey variables. Discussion Paper 2015-17; 2015, Available at: www.cbs.nl.
- [15] Scholtus S, Explicit and implicit calibration of covariance and mean structures. Discussion Paper 2014-09; 2014, Available at: www.cbs.nl.
- [16] Muthén BO, Satorra A, Complex sample data in structural equation modeling. *Sociological Methodology* 1995; 25: 267-316.
- [17] Houbiers M, Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics* 2016; 20(1): 55-75.
- [18] Knottnerus P, Van Duin C, Variances in repeated weighting with an application to the Dutch labour force survey. *Journal of Official Statistics* 2006; 22(3): 565-584.
- [19] Knottnerus P, On new variance approximations for linear models with inequality constraints. *Statistica Neerlandica* 2016; 70(1): 26-46.

- [20] Johnson NL, Kotz SN, Balakrishnan N, Continuous univariate distributions Vol. 1. Wiley, New York; 1994.
- [21] Mushkudiani N, Pannekoek J, Zhang LC, Uncertainty measures for economic accounts. Deliverable of the ESSnet on Quality of Multisource Statistics; 2017.
- [22] Chao A, Pau H-Y, Chiang, S-C, The Petersen-Lincoln estimator and its extension to estimate the size of a shared population. *Biometrical Journal* 2008; 50(6): 957-970.
- [23] Gerritse S, Van der Heijden PGM, Bakker BFM, Sensitivity of population size estimation for violating parameter assumptions in log-linear models. *Journal of Official Statistics* 2015; 31(3): 357-379.
- [24] Van der Heijden PGM, Smith PA, Cruyff M, Bakker B, An overview of population size estimation where linking registers results in incomplete covariates, with an application to node of transport of serious road casualties. *Journal of Official Statistics* 2018; 34(1): 239-263.
- [25] Pavlopoulos D, Vermunt JK, Measuring temporary employment: do survey or register data tell the truth? *Survey Methodology* 2015; 41(1): 197-214.