Measuring the Quality of Multisource Statistics

**Keywords:** administrative data, multi-source statistics, quality measures, survey data

# Introduction

The ESSnet on Quality of Multisource Statistics – also referred to as Komuso – 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 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 are expected to take the entire production chain into account (input, process, and output). They also 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 indicators/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.

The first Specific Grant Agreement (SGA) of the ESSnet lasted from January 2016 until April 2017. The second SGA started in May 2017 and lasts until mid-October 2018. A third and final SGA is planned to start mid-October 2018 and end mid-October 2019.

Work Package (WP) 3 of the ESSnet focuses on developing and testing quantitative measures and indicators for measuring the quality of output based on multiple data sources and on methods to compute such measures and indicators. Examples of such quality measures and indicators are bias and variance of the estimated output. Methods for computing these and other quality measures and indicators often depend on the specific situation at hand. Many different situations can arise when multiple sources are used to produce statistical output, depending on the nature of the data sources and on the kind of output produced. 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 and indicators for the accuracy, timeliness and coherence of the output.

In this paper 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. Section 4 concludes this paper with a brief discussion.

# Approach

As already mentioned above, in WP 3 we use a breakdown into a number of Basic Data Configurations (BDCs) that are most commonly encountered in practice. In Komuso we have identified six BDCs (see also [1]):

* BDC 1: Multiple non-overlapping cross-sectional microdata sources that together provide a complete data set 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).

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 use data to 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.

WP 3 is strongly related to WP 1 of Komuso in SGA 2 (and probably also in a potential SGA 3). In WP 1 of SGA 2 (and possibly SGA 3) 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.

# Results

## Introduction

In this section we first give a global overview of the products (QMCMs and hands-on examples) of WP 3. Next, we give two examples of QMCMs (Subsections 3.2 and 3.3).

In total 23 QMCMs and 13 hands-on examples have been produced in SGA 2: 19 QMCMs and 10 examples for the quality dimension “Accuracy”, one QMCM for the quality dimension “Timeliness”, and three QMCMs and two examples for the quality dimension “Coherence”. The vast majority of the QMCMs and examples relate to BDC 2, which appears to be the most common and most important situation with respect to multisource statistics at NSIs. Three of these QMCMS focus on general quality frameworks, six on sampling errors, four on measurement error in target variables, three on linkage and coverage errors, and the remaining two on measurement error in classifying variables (classification errors). In SGA 2 also some suitability tests have been carried out for which no QMCM has been produced yet.

## Mean squared error of level estimates affected by classification errors

In this example (see also [2]), we assume that we have one data source with data on a classifying background variable and one or more (non-overlapping) data sources with data on target variables. The only source of errors are classification errors. 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 is denoted by ; the observed industry code that is prone to errors is denoted by . The set of possible industry codes is denoted by .

Let denote a target parameter and stand for the value of a target variable for unit ; for instance, the stratum total . Based on the observed data, this parameter is estimated by . We are interested to estimate the mean squared error of as affected by classification errors. We assume that the values of the target variable, , are all observed and error-free.

[2] assumes that a business with a true industry code is classified as falling into class with probability due to classification error. [2] proposes the following method for computing the mean squared error of . The first step is to estimate the probabilities . One way to do this is by collecting independent sample data on the classification variable, where observed and cleaned versions of those data are needed. Together the estimated probabilities form a transition matrix . The transition matrix, per unit, can be modelled as a function of background variables.

The second step is to estimate bias and variance of by drawing bootstrap samples from **.** For each unit we draw a new value for the industry code, given the original observed industry code , according to . Based on the results for this draw, denoted by , we compute . We repeat this procedure times, thus , and use the set of outcomes to compute estimates of the bias and variance of :

|  |  |
| --- | --- |
|  |  |
|  |  |

with The mean squared error of is then estimated as . The smaller and , the higher the quality of .

## Validity and measurement bias of observed numerical variables

In this example (see also [3]), 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. 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. Measurement bias can be summarized in terms of *intercept bias* and *slope bias*. Intercept bias indicates a constant shift that occurs for all values. Slope bias indicates a shift that is proportional to the true value. Ideally, the intercept and slope bias are both zero.

Suppose that one has a linked dataset with observed variables from different data sources. The underlying “true” target variables are not observed directly and denoted by latent variables . It is assumed that each observed variable is an indicator of exactly one target variable and that each target variable is measured by multiple (at least two) observed variables. A linear structural equation model (SEM) for these data consists of two sets of regression equations. First, there are *measurement equations* that relate the observed variables to the latent variables:

(1)

Here, denotes a measurement intercept, denotes a slope parameter, and denotes a zero-mean random measurement error that affects . Second, the SEM may also contain *structural equations* that relate different latent variables to each other.

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* of is defined as the absolute value of the correlation between and . It captures the effect of random measurement errors in the observed data. The parameters and provide information about measurement bias in with respect to . If no bias occurs, then it holds that and ; cf. (1). *Intercept* *bias* is indicated by a deviation of from ; *slope bias* is indicated by a deviation of from .

The validity coefficients and the parameters and provide information about the quality of the input data . They can also be used to measure output quality. Suppose that we are interested in the population mean of the true variable , and we have two available estimators: (i) the direct estimator based on a simple random sample without replacement of units, , where the target variable is measured by , , and (ii) an estimator based on a register that covers the entire population, where the target variable is measured by , . Under model (1) for and , the following expressions can be derived for the mean squared error of the two estimators:

Here, and denote the expected population variances of the observed variables under the model. In this example the validity coefficients and intercept and slope bias can be used directly to quantify and compare the accuracy of the two estimators.

As an alternative to using either or one could estimate the population mean by , where are estimates for the . The quality of this estimator could be estimated by a resampling or multiple imputation approach.

# Conclusions

In the first two SGAs of Komuso we have produced a large number of QMCMs and related hands-on examples for a broad range of situations and error types. In the third and final SGA of the ESSnet we aim to take feedback on the produced QMCMs and hands-on examples into account. Depending on that feedback we will, where necessary or useful, produce new QMCMs and examples or update already developed ones, with the goal to enhance the usability and usefulness of the QMCMs and hand-on examples in practice at NSIs.

# References

1. De Waal, T., A. van Delden and S. Scholtus (2017), Multisource Statistics: Basic Situations and Methods. Discussion paper, Statistics Netherlands.
2. QMCM\_A\_11 (2018), Mean Squared Error of Level Estimates Affected by Classification Errors. Deliverable of WP 3 of the ESSnet on Quality of Multisource Statistics (SGA 2).
3. QMCM\_A\_11 (2018), Validity and Measurement Bias of Observed Numerical Variables as Indicators for a Target Variable. Deliverable of WP 3 of the ESSnet on Quality of Multisource Statistics (SGA 2).