**Quality Measures and Indicators for Multisource Statistics**

Ton de Waal, Statistics Netherlands & Tilburg University, t.dewaal@cbs.nl

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

Sander Scholtus, Statistics Netherlands, s.scholtus@cbs.nl

**Abstract**

*The ESSnet on Quality of Multisource Statistics is part of the ESS.VIP Admin Project. The main objectives of that 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 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. Work Package 3 (WP 3) of the ESSnet focuses on developing and testing quantitative measures/indicators for measuring the quality of output based on multiple data sources and on methods to compute such measures/indicators. WP 3 focuses on non-sampling errors. Examples of such quality measures/indicators are bias and variance of the estimated output. Methods for computing these and other quality measures/indicators 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/indicators for the accuracy and coherence of the output. In this paper we discuss the identified basic data configurations, the approach taken in WP 3, and give some examples of quality measure/indicators and methods to compute those measures/indicators. We also point out some topics for future work.*

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

**1. 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 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.

In the first and second Specific Grant Agreements (SGA) of Komuso, Work Package 3 (WP 3) focuses 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, and is complicated by several factors, such as:

* When measuring the quality of statistical output one ideally wants to take into account all processing and estimation steps that were taken to achieve the output;
* Data can be combined in a large variety of ways. How the output quality can be quantified depends on the input data as well as on how they are combined.

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 would like to use bias and variances 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. In other words, the main problem is describing a recipe for calculating quality measures/indicators for a given multisource situation. This problem is complicated, because we focus on non-sampling errors in WP 3. In this paper, a quality measure gives a quantitative value based on observed data, and a quality indicator either gives a qualitative indication of the quality or a quantitative value that is (partly) based on subjective information.

In this paper we discuss WP 3 and some of the results obtained. Section 2 describes the approach taken in WP 3. Section 3 gives some examples of quality measures/ indicators and methods to compute those measures/indicators. Section 4 gives an example of how one of those quality measures can be applied in a practical situation. Section 5 concludes the paper with a brief discussion.

**2. Approach taken in WP 3**

In WP 3 we have subdivided the work into three steps, given by:

1. We carry out a literature review or suitability test. In a literature review we study and describe existing quality measures/indicators and recipes to compute them. In a suitability test we go a step further and also test quality measures/indicators 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/indicator and the accompanying calculation recipe.
2. We produce Quality Guidelines. Such a Quality Guideline is a short description of a quality measure/indicator and the accompanying calculation recipe as well as a description of the situation(s) in which the quality measure/indicator and accompanying recipe can be applied.
3. We provide hands-on examples to some of the Quality Guidelines.

In SGA 1 of Komuso, the focus was carrying out literature reviews and suitability tests for the quality dimension “accuracy” (principle 12 in the European Statistics Code of Practice). In SGA 2 the focus is on producing Quality Guidelines 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).

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 WP3 we use a breakdown into a number of Basic Data Configurations (BDCs) that are most commonly encountered in practice. 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 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 and the split-population case where the data sets contain different units. For more information on BDCs and methods to produce multi-source statistics we refer to De Waal, Van Delden and Scholtus (2017).

**3. Examples of Quality Guidelines**

In total about 25 Quality Guidelines are planned to be produced for WP 3 in SGA 2. Three Quality Guidelines will relate to all BDCs, one to BDC 1 only, thirteen to BDC 2, one to BDC 3, two to BDC 4, four to BDC 5 and two BDC 6. Since a complete description of the work done is impossible given the limited length of this paper and presentation, we limit ourselves to giving some examples.

*3.1. BDC 2: Correcting for measurement error on the unit level*

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 $Y=\left(Y\_{1},Y\_{2},…,Y\_{s}\right)^{'}$ 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 that all variables $Y\_{j}$ and $X$ have the same set of categories, say $\{1,…,L\}$. The basic LC model is given by

$$P\left(Y=y\right)=\sum\_{x=1}^{L}P\left(X=x\right)\prod\_{j=1}^{s}P\left(X=x\right)$$

The model can 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:

$P\left(Y=y\right)=\frac{P\left(X=x\right)\prod\_{j=1}^{s}P\left(X=x\right)}{\sum\_{x^{,}=1}^{L}P\left(X=x^{,}\right)\prod\_{j=1}^{s}P\left(X=x^{,}\right)}$ (1)

Edit restrictions, for instance the edit restriction that someone who receives rent benefit cannot be a house owner, can be imposed in the following way

$$P(X="owner" │Y="rent benefit")=0$$

The so-called MILC method (see Boeschoten, Oberski and De Waal, 2017) 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 dataset.

2. Create an LC model for every bootstrap sample.

3. Multiply impute latent "true" variable $X$ for each bootstrap sample. $m$ empty variables $\left(W\_{1},…,W\_{m}\right)$ are created and imputed by drawing one of the categories using the estimated posterior membership probabilities (1) from the LC model.

4. Obtain estimates of interest from the imputed variables.

5. Pool the estimates using Rubin's rules for pooling (Rubin, 1987). An essential aspect of these pooling rules is that an estimated variance of the estimates is obtained. This estimated variance is a quality measure for the reconciled data.

In Section 4 we will illustrate the procedure with an example.

*3.2. BDC 2: Timeliness: when to publish the results?*

In this example we focus on the quality dimension “Timeliness” and assume that input data sets for producing statistics are updated at several moments: “reference time + $t\_{1}$”, “reference time + $t\_{2}$”, etc. In all data sets we have the same target variable(s), but the quality of the available data is improving over time. We are interested in the bias of the statistics using available input data at “reference time + $t\_{i}$”.The bias estimate is used to decide how long to wait for the input register until the production process is started.

In order to estimate bias the input registers are run through the production system and the output statistics are produced. This is repeated for all moments: ”reference time + $t\_{1}$”, “reference time + $t\_{2}$” etc. The following approach is proposed by Fosen (2017). Let $\hat{S}\_{r,t}$ denote the estimated output variable at reference time $r$ when measured based on the updated input data at time $t$. The difference estimator between the two time points $t\_{1}$ and $t\_{2}$ is denoted $X\_{r,t\_{1},t\_{2}}=\hat{S}\_{r,t\_{1}}-\hat{S}\_{r,t\_{2}}$. The difference between the outputs (at $t\_{1}$ versus $t\_{2}$) is observed for several reference times $r\_{i}$ $\left(i=1,2,…\right)$. Under the assumption that the observations $X\_{r\_{i},t\_{1},t\_{2}}$ are identically distributed(which holds under some assumptions, such as constancy of the production process), their average value is taken as the bias due to delay.

Under the further assumptions that the $X\_{r\_{i},t\_{1},t\_{2}}$ are independent and normally distributed and sufficient $X\_{r\_{i},t\_{1},t\_{2}}$ can be computed, a t-test can be used to test the hypothesis that the average bias (over the reference times $r\_{i}$) $\overbar{X}\_{t\_{1},t\_{2}}$ differs from zero. The lower the value of the t-test statistic, the more likely it is that $\overbar{X}\_{t\_{1},t\_{2}}$equals zero, and the better the estimate at moment $t\_{1}$ is assumed to be. To some extent, the above assumptions can be tested (Fosen, 2017).

*3.3. BDC 4 and 5: Measuring the quality of accounting equations*

Many macro-economic 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, so that the adjusted estimates can be presented in terms of the reconciled estimated accounting equations. We consider such a system of estimated accounting equations as a single entity and define scalar uncertainty measures that capture the adjustment effect as well as the relative contribution of the various input estimates to the final estimated account. Mushkudiani, Pannekoek and Zhang (2017) discuss two approaches: the covariance approach and the deviation approach. Below we sketch the covariance approach.

Consider the additive account $A=[Y\_{1}+\cdots +Y\_{i}+\cdots +Y\_{p}=Z]$. Let $Σ\_{\tilde{X}}$ be the variance-covariance matrix of the adjusted estimates $\tilde{X}=(\tilde{Y\_{1}},…,\tilde{Y\_{i}},…,\tilde{Y\_{p}},\tilde{Z})$. This matrix can be partitioned as follows: $Σ\_{\tilde{X}}=\left(\begin{matrix}Σ\_{\tilde{Y}}&Σ\_{\tilde{Y}\tilde{Z}}\\Σ\_{\tilde{Z}\tilde{Y}}&Σ\_{\tilde{Z}}\end{matrix}\right)$.

Two scalar measures of this variance-covariance matrix are defined as the following quantities: the sum of all elements (variances and covariances) and the sum of the diagonal elements (variances) only, that is

$$τ\_{1}\left(A\right)=1^{T}Σ\_{\tilde{X}} 1=4\left(1^{T}Σ\_{\tilde{Y}} 1\right)=Var\left(2\tilde{Z}\right)$$

and

$$τ\_{2}\left(A\right)=Trace\left(Σ\_{\tilde{X}}\right)=Trace\left(Σ\_{\tilde{Y}}\right)+1^{T}Σ\_{\tilde{Y}} 1=\sum\_{i=1}^{p}Var\left(\tilde{Y\_{i}}\right)+Var\left(\tilde{Z}\right)$$

*3.4. All BDCs: a framework for quality assessment*

We have analyzed and tested a quality framework developed by Statistics Austria (see Asamer et al., 2016) that can be used when several data sets with possibly conflicting values for common variables are available. The quality framework models errors in variables in these data sets as well as systematically uses expert knowledge. The framework distinguishes three levels:

* the raw individual data sets. The raw data are assessed by simple key figures.
* the combined data set of these raw data sets. The crucial step for the combined dataset is the application of the Dempster-Shafer theory to combine the different indicators and the expert knowledge to an overall quality indicator.
* the final data set, which includes imputations. The imputed values are assessed by a classification rate of the imputation model.

For the framework, each data source has to deliver on unit level. In addition, the same unique key must be available in each source, so that they can be linked. Let $j$ denote the attribute of interest and $i$ a data source containing $j$. An (error-free) external data source, which may be too small or be available too late to use as a direct source, should be usable (i.e. linkable) to the raw data to assess the accuracy of $j$ (in $i$). Otherwise an expert view can be used to substitute the external source. The whole data editing process must be accessible, since the indicator is computed stepwise.

The quality information at the raw data level is obtained via three hyperdimensions: Documentation $HD\_{ij}^{D}$, Preprocessing $HD\_{ij}^{P}$ and External Source $HD\_{ij}^{E}$, by the formulas

$$HD\_{ij}^{D}=\frac{obtained score for j in i}{achievable score for j in i}$$

$$HD\_{ij}^{P}=\frac{usable records for j in i}{total number of records of i}$$

$$HD\_{ij}^{E}=\frac{number of consistent values for j in i}{total number of linked records of i}$$

The score in the first equation comes from scored questions in a questionnaire assigned by the data holders. Given these three quality indicators, an overall quality indicator $q\_{ij}$ for each attribute $j$ in the source $i$ is derived as a weighted average.

Afterwards, each $j$ is classified as one of three possible types (unique, multiple, derived) of attributes. Depending of these types, a further indicator $q\_{⊙j}^{n}$ is calculated (here $n$ denotes the statistical unit). In this step the “multiple” type is the most challenging case since it requires the Dempster-Shafer theory. This because one has to combine different indicators and the most meaningful way to do this is to interpret these in terms of beliefs of correctness of each data source. In the remaining cases “unique” respectively “derived” $q\_{⊙j}^{n}$ is just $q\_{ij}$ resp. an average of $q\_{ik\_{s}}$ over certain attributes $k\_{1},k\_{2},…$ .

In general, the final indicator $q\_{Ωj\_{ }}^{n}$ is just$ q\_{⊙j}^{n}$. But, if imputations occur, they have to be assessed. First the accuracy of the imputations is assessed in a classification rate measure $Φ$. Based on a classification rate $Φ$ as well as the final quality $q\_{Ωj\_{k}}^{n}$ of the predictors$ j\_{1},…,j\_{s}$, Asamer et al. (2016) compute a quality indicator $HD\_{j}^{I,n}$ by$ HD\_{j}^{I,n} =\frac{Φ^{ }}{s}⋅\sum\_{k=1}^{s}q\_{Ωj\_{k}}^{n}$.

**4. A hands-on example**

Boeschoten, Oberski and De Waal (2017) apply the MILC method on a combined dataset to measure home ownership. This combined dataset consists 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 dataset, they use 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 dataset also contains a variable measuring whether someone receives rent benefit from the government. A person can only receive rent benefit if this person rents a house. A research question for such combined data could be whether married individuals more often live in a house they own compared to non-married individuals. Therefore, a variable indicating whether a person is married or not is included in the latent class model as a covariate. The three datasets used to combine the data are:

* BAG: A register containing data on addresses and buildings originating from municipalities from 2013. From the BAG, Boeschoten, Oberski and De Waal (2017) used a variable indicating whether a person "owns"/ "rents (or other)" the house he or she lives in.
* LISS background study: A survey on general background variables from January 2013. Boeschoten, Oberski and De Waal (2017) 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 survey on housing from June 2013. From this survey Boeschoten, Oberski and De Waal (2017) used the variable rent benefit, indicating whether someone "receives rent benefit", "does not receive rent benefit", or "prefers not to say".

These datasets were linked on a unit level, and matching was done on person identification numbers. Not every individual is observed in every dataset. This causes that some missing values are introduced when the different datasets are linked on a unit level. Full Information Maximum Likelihood was used to handle the missing values. The MILC method is applied to impute the latent variable home owner/renter by using two indicator variables and two covariates.

The MILC method can be used to assess the quality of the input sources. In Table 1 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.

*Table 1. Classification probabilities for LISS and BAG*

|  |  |  |
| --- | --- | --- |
|  | $$P(observed=rent|true=rent)$$ | $$P(observed=own|true=own)$$ |
| LISS | $$0.9344$$ | $$0.9992$$ |
| BAG | $$0.9496$$ | $$0.9525$$ |

To give an example of how to measure the quality of – even quite complicated –aspects of the combined data set, Boeschoten, Oberski and De Waal (2017) 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 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.

**5. Discussion**

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

In potential future work, we aim to produce final Quality Guidelines and hands-on examples for (some of) the suitability tests carried out in SGA 2, in particular those related to the quality dimension “coherence”.

An important topic for future work is the further development of a systematic framework for situations, methods and quality measures/indicators that can arise in a multisource context.

**6. References**

Asamer E.-M., Astleithner F., Cetkovic P., Humer S., Lenk M., Moser M. and Rechta H. (2016), Quality Assessment for Register-based Statistics – Results for the Austrian Census 2011. Austrian Journal of Statistics 45, pp. 3-14.

Boeschoten, L., Oberski D. and De Waal T. (2017), Estimating Classification Error under Edit Restrictions in Combined Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC). Journal of Official Statistics 33, pp. 921–962.

Fosen J. (2017), Output Quality for Statistics Based on Several Administrative Sources. Case: The Norwegian Register-Based Employment Statistics and the Effect of Delays. Deliverable of the ESSnet on Quality of Multisource Statistics.

De Waal T., Van Delden A. and Scholtus S. (2017), Multisource Statistics: Basic Situations and Methods. Discussion paper, Statistics Netherlands.

Mushkudiani N., Pannekoek J. and Zhang L.-C. (2017), Uncertainty Measures for Economic Accounts. Deliverable of the ESSnet on Quality of Multisource Statistics.

Rubin, D. B. (1987), Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.