ESSnet KOMUSO

Quality in Multisource Statistics

http://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics_en

Specific Grant Agreement No 1 (SGA-1)

Framework Partnership Agreement Number 07112.2015.003-2015.226

Specific Grant Agreement Number 07112.2015.015-2015.705

Work Package 3

Framework for the quality evaluation of statistical output based on multiple sources

Version 2017-08-10

Prepared by: Ton de Waal, Arnout van Delden and Sander Scholtus (Statistics Netherlands)

With contributions by ISTAT, Statistics Austria, Statistics Denmark, Statistics Norway, Statistics Lithuania and Statistics Netherlands

ESSnet co-ordinator: Niels Ploug (DST, Denmark), email npl@dst.dk, telephone +45 3917 3951

Contents

1. Introduction

In this report we describe the work done on Work Package 3 ("Framework for the quality evaluation of statistical output based on multiple sources") of *Komuso* (ESSnet on quality of multisource statistics) during the first Specific Grant Agreement. This report is based on longer and more detailed individual literature reviews and suitability tests that are available too.

The aim of Work Package 3 (WP 3) is to produce measures for the quality of the output of multisource statistics. These quality measures primarily focus on the quality dimensions "**accuracy**" and "**coherence**" (principle 12 and 14 in the European Statistics Code of Practice). Within WP 3 we have carried out a critical literature review of existing and currently proposed quality measures. We have also carried out suitability tests. These suitability tests do not directly refer to the suitability of the quality measures (accuracy, coherence) themselves but rather to the methods (or recipes) to estimate them in a given situation. If no methods /recipes exist to estimate a quality measure for a given situation, apparently the quality measure cannot be applied (yet) for that situation.

Many different situations can arise when multiple sources are used to produce statistical output, depending on both the nature of the data sources used and the kind of output produced. In order to structure the work within WP3 we have proposed a breakdown into a number of ***basic data configurations*** that seem most commonly encountered in practice. In practice, a given situation may well involve several basic configurations at the same time. The aim of the basic data configuration is not to classify all possible situations that can occur, but to provide a useful focus and direction for the work to be carried out. Note that the basic data configurations give a simplified view of reality. Nevertheless, many practical situations can be built on these basic configurations, and basic data configurations are a good way to structure the work in our opinion.

Section 2 of this report describes the basic data configurations that we have identified within WP 3. Section 3 focuses on the case where output is based on a single administrative data source. Section 4 describes the work that has been carried out by the partners within WP 3. Section 5 describes an action plan and road map for possible future work in the ESSnet. The action plan and roadmap are based on a gap analysis. The Appendix gives an overview of the suitability tests that have been carried out.

2. Basic data configurations

2.1 Introduction to the basic data configurations

The characterisation of multisource data can be complicated due to the inherent heterogeneous nature of the sources. The following aspects seem relevant in many situations.

- Aggregation level
    - The data sources consist of only microdata
    - The data sources consist of a mix of microdata and aggregated data
    - The data sources consist of only aggregated data
- Units
    - There are no overlapping units in the data sources
    - (Some of the) units in the data sources overlap

- Variables
    - There are no overlapping variables in the data sources
    - (Some of the) variables in the data sources overlap
    - Are the variables categorical or numerical (or both)?
- Coverage
    - There is under-coverage versus there is no under-coverage.
    - There is over-coverage versus there is no over-coverage.
- Time
    - The data sources are cross-sectional
    - The data sources are longitudinal
- Population
    - We know the population (i.e. the set of population units) from a population register
    - We do not know the population from a population register.
- Data type
    - A data source contains a complete enumeration of its target population.
    - A data source is selected by means of probability sampling from its target population.
    - A data source is selected by non-probability sampling from its population.

In Subsection 2.2 we discuss the basic data configurations that we have identified within WP 3, using the above-mentioned aspects.

A prioritisation is equally necessary when it comes to the scope of the outputs. Firstly, one may distinguish between outputs that are

- Population registers (a population register is itself a statistical output when it is compiled from multiple input data sets; see WP2 for quality of population registers)
- Business Registers
- Statistics (Macro data)
- Micro data sets
- Metadata

Within WP3 we focus on the output “Statistics”. Next, one can make a distinction between

- Descriptive statistics such as population and sub-population totals and means
- Analytic statistics such as price index numbers

Within WP3 we focus on the quality of descriptive statistics from the multisource data. Such descriptive statistics may either have computed by using the data sources directly or indirectly, i.e. by means of a model for the true data.

2.2 Identified basic data configurations

We have identified 6 basic data configurations that are quite common in practice. These 6 basic data configurations are given in Table 1 below. In the remainder of this section we will describe these 6 basic data configurations in more detail.

Table 1. The 6 basis data configurations

Basis data configuration	Description
1	Complementary microdata sources
2	Overlapping microdata sources
3	Overlapping microdata sources with under-coverage
4	Microdata and macrodata
5	Only macrodata
6	Longitudinal data

In all identified basic data configurations, measurement errors and errors due to the progressiveness of administrative data sources may play a role, for instance because an administrative data source is slowly being filled.

Basic data configuration 1: complementary microdata sources

The first and most basic configuration is multiple cross-sectional data that together provide a complete data set with full coverage of the target population. Provided they are in an ideal error-free state, the different data sets, or data sources, are ***complementary*** to each other in this case, and can be simply "added" to each other in order to produce output statistics.

Configuration 1 is illustrated in Figure 1. For all illustrations in this document note that:

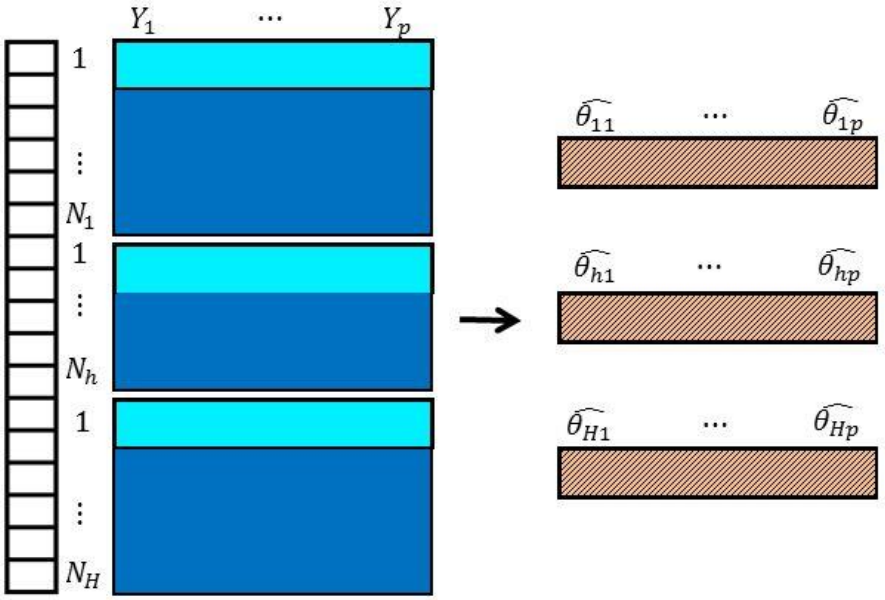
1) The rectangle of white blocks to the left represents the population frame
2) Different blue colours represent different input data sources
3) Orange/brownish colour represent derived output statistics
4) Shaded blocks represent macro data, bright blocks represent micro data

As an example of Configuration 1, consider the situation where we have a Business Register containing all units in the population (the rectangle of white blocks shown at the left). We have administrative data that can only be linked to the smaller and less complex units (the dark blue data source), and we have sample survey data for the remaining units (the light blue data source). In the figure the vertical dimension thus represents the unit and strata (or domains); the horizontal dimension represents the different variables or measurements. Provided unit delineation of the input data is error-free, i.e. the delineation of the population is correct and consistent between data sources, one should be able to match the data to the business units by an identity number. Next, one should be able to tabulate directly statistics on, for instance turnover, export or import, in a coherent manner.

As an example of a potential problem consider NACE classification errors and their effect on the output statistics. For instance, a model for classification error may be used in combination with a resampling method to quantify this effect.

So, in basic data configuration 1 classification errors may occur, besides measurement errors and errors due the progressiveness of administrative data sources.

Figure 1. Combining non-overlapping microdata sources without coverage problems



In Figure 1, as well as in other figures, the numbers N_1, N_h and N_H refer to the number of elements in the strata 1, h and H, respectively. Within stratum h $(h = 1, \dots, H)$, the units are numbered from 1 to N_h. In Figure 1 there are p variables Y_1 to Y_p in the data sources. We try to obtain estimates $\hat{\theta}_{hj}$ for the population total of variable Y_j in stratum h $(h = 1, \dots, H; j = 1, \dots, p)$.

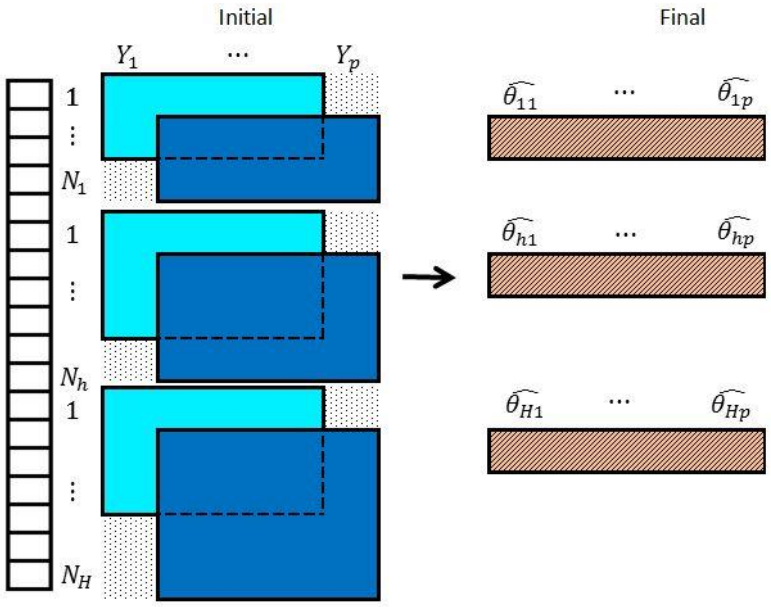
Some people may consider basic data configuration 1 to be too simplistic and optimistic. While we agree that this configuration is simplistic and rather optimistic, we use this as one of our basic data configurations. Basic data configuration 1 is the starting point for all data configurations and even for basic data configuration 1 important errors occur in practice (see Section 3.1 below).

Basis data configuration 1 also encompasses the case where where one only has a single administrative data source covering the entire population.

Basic data configuration 2: overlapping microdata sources

Configuration 2 is characterised by a deviation from configuration 1, by which there exists ***overlap*** between the different data sources. The overlap can concern the units, the measured variables, or both. Configuration 2 is illustrated in Figure 2.

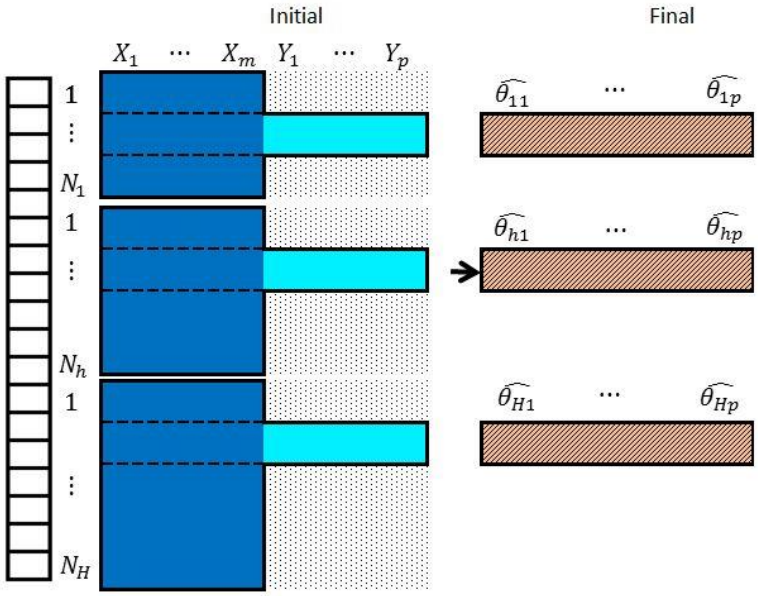
Figure 2. Combining overlapping microdata sources without coverage problems



An example of Configuration 2 arises in education statistics in the Netherlands. In the Netherlands there exist both administrative and survey data on the educational level of a person. The administrative data have a shorter history and do not cover people over a certain age. The survey data, which go further back in time, have full coverage of the current population over the aforementioned age threshold but have missing data due to non-exhaustive sampling. In cases where the same person is found in both sources, the respective education level measurements may not agree with each other. For instance latent class models may be used to quantify this latter effect.

Note that Configuration 2 may be further split into two subcases depending on whether one of the data sources consists of sample data (and where the sampling aspects play an important role in the estimation process) or not. In the former case specific methods should be used in the estimation process, for instance taking the sampling weights into account and considering that sample data may include specific information that is not reported in registers. We refer to this special case as Configuration 2S, displayed in Figure 2S. In Figure 2S the sample is represented by the light blue rectangle and register data by the dark blue rectangle. In this example the sample reports information for both target variables $Y_1, \dots, Y_p$ and auxiliary variables $X_1, \dots, X_m$, while the register contains only information for the auxiliary variables $X_1, \dots, X_m$.

Figure 2S. Special case of Figure. 2: Combining overlapping microdata sources without coverage problems and with at least one of the sources being sample data



Configuration 2S may arise for instance when using a projection estimator, or alternative methods, to attach synthetic data for the target variables $Y_1, \ldots, Y_p$ to the register. In configuration 2S the estimation will often apply calibration weighting, since the same variables are known both from the administrative source and the survey data source. In the remainder of the document we will not make a distinction between basic data configuration 2 and 2S.

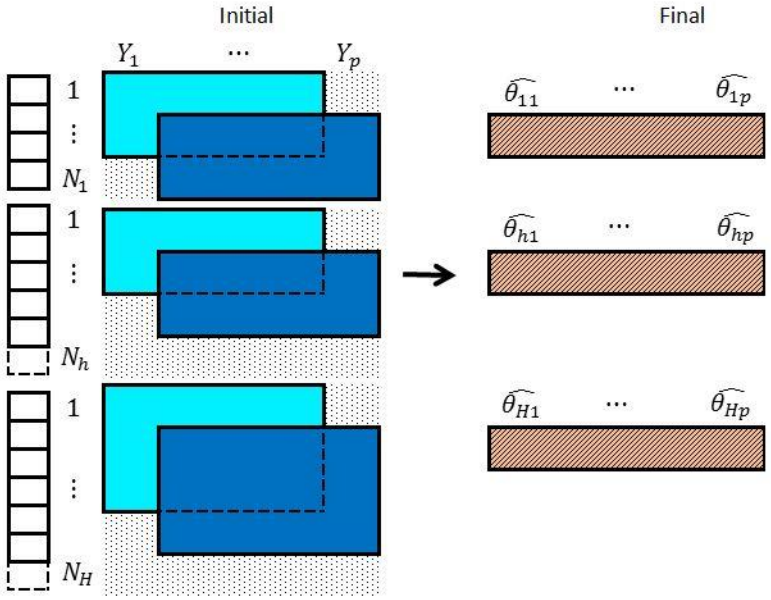
In basic data configuration 2 linkage errors may occur, besides the already mentioned types of errors for basic data configuration 1. If sample survey data are involved, also sampling errors occur.

Basic data configuration 2 is by far the most commonly encountered situation when producing multisource statistics in practice. Most suitability tests therefore focus on this basic data configuration.

Basic data configuration 3: overlapping microdata sources with under-coverage

Configuration 3 is characterised by a further deviation from configuration 2, by which the combined data entail ***under-coverage*** of the target population in addition, even when the data are in an ideal error-free state. Configuration 3 is illustrated in Figure 3.

Figure 3. Combining overlapping microdata sources with under-coverage



The dashed cells in the left row in Figure 3 indicate under-coverage in the population register. Such under-coverage may be noticed from available administrative data sources. For instance, medical data from hospitals may contain information on people not registered in the country.

A possible example of Configuration 3 is a population census and a post-enumeration survey (PES). Both the census and the PES entail under-count of the target population. The binary measurement of being enumerated is the overlap between the two. For instance capture-recapture techniques may be used to quantify the under-coverage and the effect on quality of statistical estimates.

For another example, consider the education data mentioned in Configuration 2, but now allow for under-coverage in the Population Register that provides the frame for both administrative and survey data.

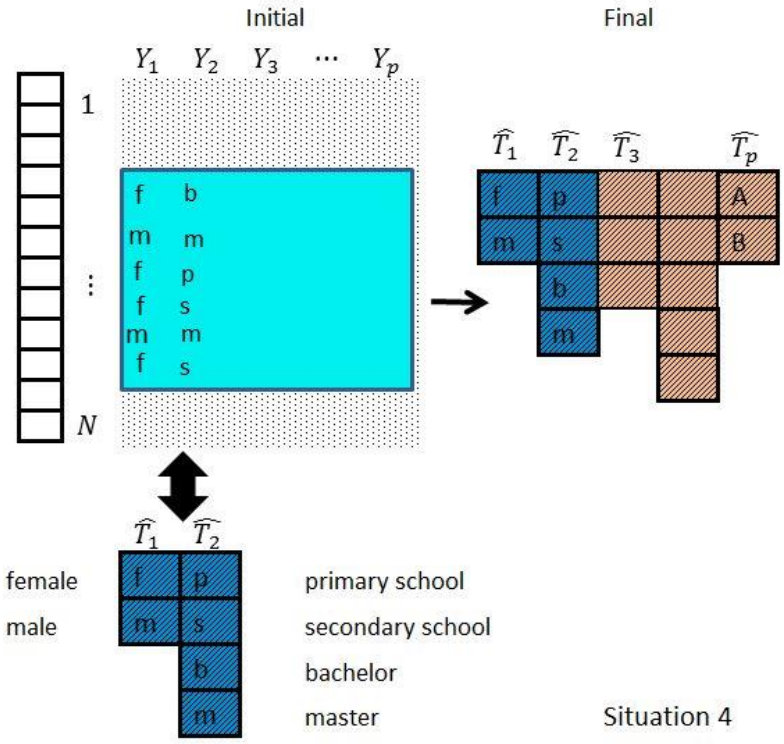
Note that, similar to Configuration 2, Configuration 3 can also be split up into two cases depending on whether one of the data sources consists of sample data or not.

In basic data configuration 3 under-coverage errors occur, besides the already mentioned types of errors for basic data configuration 2.

Basic data configuration 4: microdata and macrodata

Configuration 4 is characterised by a variation of Configuration 2, by which ***aggregated data*** are available besides micro data. There is still overlap between the sources, from which there arises the need to reconcile the statistics at some aggregated level. Of particular interest is when the aggregated data are estimates themselves. Otherwise, the conciliation can be achieved by means of calibration which is a standard approach in survey sampling. Configuration 4 is illustrated in Figure 4.

Figure 4. Combining a microdata source with a macrodata source



An example of Configuration 4 is the Dutch virtual census based on a number of administrative data sources and surveys. Population totals, either known from an administrative data source or previously estimated, are imposed as benchmarks provided they overlap with additional survey data set that is needed to produce new output statistics. In such a case one is interested in the quality of the new output. For instance analytical variance formulas or resampling methods may be used to quantify the quality of the output.

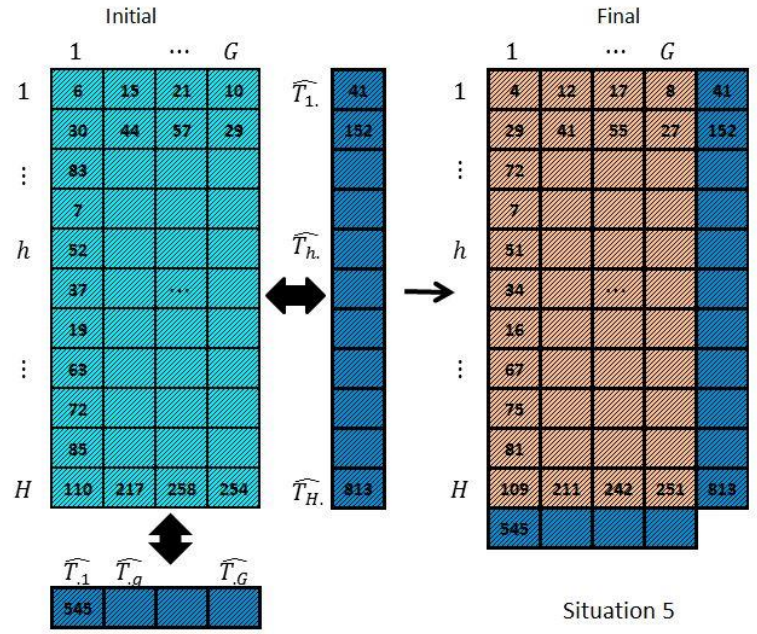
In basic data configuration 4 measurement errors and errors due to the progressiveness of administrative data sources may occur. If sample survey data are involved, also sampling errors may occur.

Basic data configuration 5: only macrodata

Configuration 5 is the complete macro-data counterpart of Configuration 2, where only aggregated data overlap with each other and need to be reconciled. Configuration 5 is illustrated in Figure 5.

Reconciliation of supply and use (SU) tables of National Accounts provides a perfect example for Configuration 5. It is important to be able to estimate the quality of the reconciled SU tables in a comprehensive manner. In some cases analytical formulas are available for this.

Figure 5. Combining macrodata sources

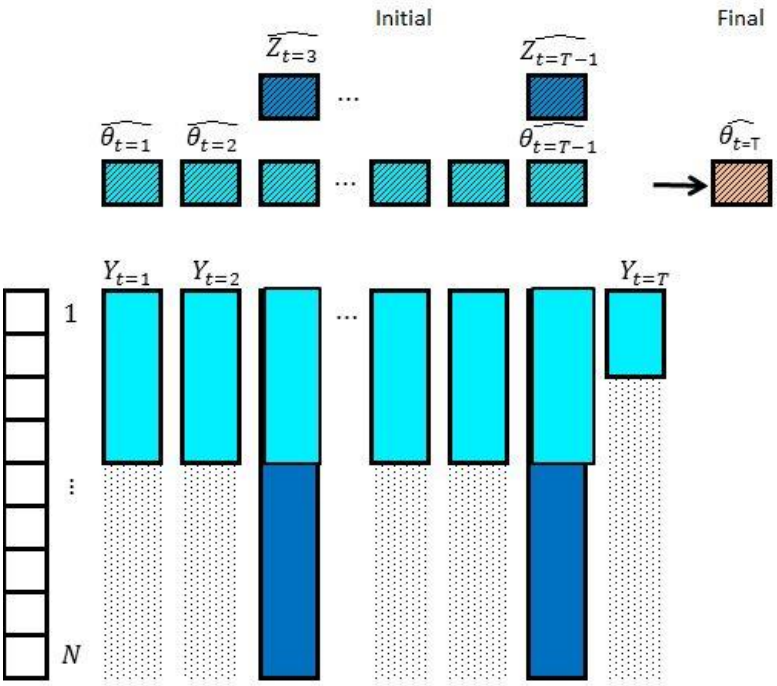


In basic data configuration 5 measurement errors and errors due to the progressiveness of administrative data sources may occur. If sample survey data are involved, also sampling errors may occur.

Basic data configuration 6: longitudinal data

Finally, longitudinal data are introduced in Configuration 6. We limit ourselves to the issue of reconciling time series of different frequencies and qualities, as illustrated in Figure 6.

Figure 6. Combining longitudinal data sources



An example of Configuration 6 is when turnover data of enterprises are available on a monthly basis from a survey and on a quarterly basis from the Tax Office. The monthly survey data and the quarterly administrative data may be combined to produce a new time series, in which case one is interested in the accuracy and coherence of the new series, compared to the two existing ones. Alternatively, one may produce a now-casting estimate of a current month ($\hat{\theta}_{t=T}$), using time series models based on the monthly survey data, and be interested in its quality compared with the benchmarked estimate from the combined sources that is only available some months later. The difference between the now-casting estimate and the (later) benchmarked estimate can be used as a measure for the quality of the now-casting estimate.

Note that a variation on the above example is when the administrative data consist of units that are reporting either monthly or quarterly. The monthly reporters then often form a selective part of the population.

In basic data configuration 6 linkage errors may occur, besides measurement errors and errors due to the progressiveness of administrative data sources. If sample survey data are involved, also sampling errors may occur.

3. A single administrative data source

The situation where we use a single administrative data source to base statistics upon deserves special attention as it is a starting point for other situations involving administrative data. It can be seen as a special case of Configuration 1 “complementary microdata sources”.

However, measuring the quality of a single administrative data source, without any further data is almost impossible as for most kinds of errors that can possibly occur one needs more than one data source to recognise them and/or to quantify their effect on statistical output. For instance, measurement errors are hard to detect, apart from obvious measurement errors, such as logical inconsistencies (e.g. “pregnant males”) and clearly outlying values. For less clear measurement errors, one needs additional data.

In many cases, measuring the output quality of statistics based on a single administrative data source is even more complicated than measuring the output quality of statistics based on a single survey data source. Survey data are generally based on drawing a sample with a well-defined sampling design. For many sampling designs, analytical formulas for the variance due to sampling error are available. In other cases, the variance due to sampling error can often be estimated by means of resampling techniques, such as the bootstrap or the jackknife. Often, surveys are based on relatively small samples, and it is assumed that the effect of non-sampling errors is small compared to the sampling variance. A single administrative data source is often not based on a sampling design, and the units in an administrative often form a selective part of the target population. It is often unclear what the exact ‘selection procedure’ is that determines which units are in the actual administrative data set. Neither analytical formulas nor resampling techniques lead to valid variance estimates for such cases. Moreover, the output quality of statistics based on administrative data often depends mainly on non-sampling errors which are not reflected in a sampling variance formula.

When only a single administrative data source is available, one often has to rely on measures and indicators for the quality of the input (see Work Package 1) or on quality indicators obtained from processing the data, for instance quality indicators derived from the statistical data editing process. Options for measuring the quality of output based on a single administrative data source are, unfortunately, very limited.

When estimates are based on a single administrative data source the following options are available for measuring the quality of output based on this data source:

- If one can safely assume that the administrative data is a random sample from the population, sampling theory can be used to estimate the sampling error. This may, for instance, be the case if the administrative data source is itself based on data obtained from municipalities and a random part of the municipalities did not deliver the data for this administrative data source on time. However, when a selective part of the municipalities is missing, sampling theory will not provide valid results.
- When the administrative data is supposed to be a complete enumeration and a high-quality population register is available, one can compare the administrative data to this register in order to estimate the error due to under- and/or over-coverage. When the units in the administrative data source can be linked to the population register, the error due to over-coverage can easily be corrected by removing the units that do not belong to the population. The effect of under-coverage on a certain output variable may in some cases be estimated by assuming a model for

this output variable based on available background variables in the population register. The quality of the estimated effect depends on the fit of the model. Unfortunately, the fit of the model can only be assessed for the part of the population on which data are available in the administrative source, and not for the part that is missing due to under-coverage.
When the units in the administrative data source cannot be linked to the population register, the effect of under-coverage and/or over-coverage can basically only be measured in terms of the number of target units that are not found in the administrative data source (under-coverage) and/or the number of units in the administrative data source that do not belong to the target population as far as that can be seen from the administrative data set itself (over-coverage). For instance, if the target population consists of the people residing in the Netherlands and according to the administrative data source a person lives in Belgium, then he/she does not belong to the target population. The effect on other output variables is hard to quantify.

- The over-coverage error can be estimated, and corrected for, by checking for duplicates in the administrative data source. After correcting for these duplicates, one usually assumes that there is no remaining over-coverage error. Sometimes there are auxiliary variables in the administrative data set that can be used to approximate the target population, using deterministic rules. For instance, one may derive a population of people with a job from administrative data with information on hours worked and on social benefit payments for employees.
- Given sufficient resources (and time) one can construct an audit file with data of high quality. An approach such as the one described in Section 4.1.2 can then be used to quantify the measurement error. More generally, for statistics that are published regularly, one could conduct an audit or try to link the single-source data to other data sources and estimate the output quality once every x production cycles. An estimate of the output quality for the other periods could then be derived under the assumption that the effects of various error sources remain fairly stable over time.

4. Work carried out

This section is organized as follows. For each configuration, short presentations of the suitability tests that have been carried out by the members of the ESSnet are reported, according to a common structure. The full descriptions of the suitability tests are available as separate documents. Each suitability test is given a code starting with ST (abbreviation for Suitability Test) followed by the number of the data configuration and a number of the specific test. For instance, ST2_3 refers to the third suitability test for data configuration 2. We hope that this code helps retrieving the correct papers quickly.

With respect to the literature review, a list of topics and references is reported at the end of each subsection. Again, each literature review is given a code. This code starts with LR (abbreviation for Literature Review) followed by the number of the data configuration and a number of the specific review.

We have carried out one general literature review that applies to all basic data configurations considered in the current report:

Topic The use of a questionnaire with open questions to assess the quality of (multisource data involving) administrative data.
Title of literature review/file: LR0_1
"Quality Assessment Tool for Administrative Data"
Paper reviewed Marck (2014)
Keywords Quality assessment, All kinds of errors, All kinds of data, Expert knowledge

4.1 Basic data configuration 1: complementary microdata sources

4.1.1 Classification errors in the Business Register

Title of suitability test/file: ST1_1
"Suitability Test of Employment Rate for Employees (Wage Labour Force) (ERWLF)" (Statistics Denmark 2017)

Data configuration In ST1_1 a suitability test with respect to basic data configuration 1.

Keywords Categorical data, Cross-sectional data, Business register, Measurement error, Classification error, Correction models for classification errors

Type of error and aim of the test A certain amount of businesses are classified wrongly with regards to activity codes in the Business Register (BR). Wrong classifications affect all statistics based on the BR, for instance published figures from the Employment rate for the Wage Labour Force (ERWLF).

When businesses are founded, the businesses themselves specify NACE classifications in the BR register. In some countries, such as Switzerland or Croatia, a rather extensive check of registrations is carried out. In Denmark, however, there is only a limited check of new registrations in the BR, but often when businesses participate in surveys errors are observed. Such errors are reported back to the BR, where they are corrected.

Frozen versions of the BR make it possible to calculate effects of corrected NACE codes in the BR. By using old surveys and rates of corrections from newer surveys Statistics Denmark (2017) obtained estimates of the level of misclassifications of the NACE codes by size of businesses.

The ERWLF statistics give the total number of employees and the number of employees by NACE sections. When all data are gathered the total number of employees is fixed. That is, the total will *not* be affected by wrong NACE codes. However, due to reallocations, the estimated numbers of wage earners within each business group are affected by wrong NACE coding.

Data used The ERWLF statistics are indicative for the activity in the Danish labour market. The purpose is to create a fast and efficient indicator rather than creating precise numbers of the Danish labour force.

The results from the ERWLF are presented on aggregated levels of activity groups. The monthly statistics are published on level 0 with 10 groups. The quarterly statistics is published on level 1 with 21 activity groups.

- Level 0: 10 sections identified by alphabetical letters numbers 1 to 10.
- Level 1: 21 sections identified by alphabetical letters A to U. X is unknown activity.

Sources used for the ERWLF are the E-income register, population register and the BR.

- The E-income register is the backbone of the ERWLF statistics. All income taxes from wage earners in Denmark are collected through the E-income register.
- The BR is used to connect business activity with businesses, so it is possible to calculate the number of wage earners by business activity groups.
- Numbers of wage earners are also presented on a geographical level. The population register is used to place wage earners geographically in cases where they cannot be placed through their working place.
- It is planned for the ERWLF to also present wage earners on other demographical break downs, which will require an intensified use of the population register.

Work carried out

Statistics Denmark (2017) carried out a suitability test, in order to examine the quality of the ERWLF as a function of the quality of the BR. The test was intended to reveal the margin of error of number of wage labourers within activity groups due to wrong NACE classifications in the BR. The test was carried out by simulations, where distributions of activity codes were simulated and the effect on published figures from the ERWLF examined.

The input needed for the simulations is data on enterprise level with number of employees and NACE classifications, and the expected number of misclassifications by size of enterprise.

Statistics Denmark (2017) conducted three so-called recoding projects where the aim was to correct NACE codes: two surveys with 3,000 enterprises in 2006 and 2009, and one survey with 50,000 enterprises in 2007. Even though no surveys have been conducted with the same purpose since 2009, the impression from experienced employees at the BR department at Statistics Denmark is that the proportions of misclassifications are roughly the same today as when the above-mentioned surveys were conducted. This impression is primarily based on questions regarding activity, which are a part of any business survey conducted at Statistics Denmark.

In order to simulate accuracy on activity sections it is not enough to know the proportion of wrongly classified enterprises. It is also necessary to know which activity sections wrongly coded enterprises are likely to belong to. Hence a confusion matrix, with the expected distribution of wrongly coded businesses, is required. Each row in a confusion matrix will add to 1 and the values on the diagonal reflect the probability of correct coding within each activity group.

The confusion matrix is constructed by letting the diagonal be the proportion of correctly classified enterprises. The sum of each row equals 1 and corresponds to the probabilities for correct business sections. The remaining probability mass is distributed over the other sections, taking similarities between sections and size of sections into consideration. Another way to construct the confusion matrix would be to observe movements between sections over time and thereby construct an evidence based confusion matrix rather than a confusion matrix based on subjective relationships between business sections.

A simulation study has been performed where the activity section is simulated by using the observed activity and confusion matrices.

Cost-benefit analysis
An audit sample with correct data on the industry codes is needed for this approach. It is quite costly and time-consuming to construct such data.

Gap analysis
A way to collect data on the classification error size is needed that is less costly. Such data might possibly be derived from editing activities in regular statistical production. This would simplify the use of the approach in practice.

4.1.2 Errors in NACE classification

Title of suitability test/file: ST1_2
"Analytical Expressions for the Accuracy of Growth rates as Affected by Classification Errors" (Scholtus, Van Delden and Burger 2017).

Data configuration
SR1_2 considers the situation where there is a Business Register (BR) containing all units in the population. In this situation administrative (VAT) data on turnover, which are available for administrative units, are examined. These data are linked to the statistical units (enterprises) in the BR. For the larger and most complex enterprises, the related administrative units in the VAT data cannot be linked uniquely to enterprises in the BR and census survey data is used instead. For regular estimates of quarterly turnover, the non-response error is small so mainly non-sampling errors occur. That means that quarterly turnover growth rates can easily be computed, which is an important business cycle indicator.

Keywords Categorical data, Cross-sectional data, Business register, Measurement error, Classification error, Correction models for classification errors

Type of error and aim of the test
Growth rates are broken down by main economic activity of the enterprises, using the NACE code classification, further referred to as industries. The industry code is a characteristic that is known for

all units in the population; it is derived from various sources underlying the BR. The correct main activity is the activity that would have been found when the formal rules for deriving an industry code are correctly applied and based on error-free input. The observed industry code in the BR is prone to errors, since it is difficult to derive a correct main activity, and enterprises and their industry code(s) change over time but not all of those changes enter the BR. Van Delden, Scholtus and Burger (2016b) developed a bootstrap method to estimate the accuracy of level estimates under classification errors in the industry code. The aim of the suitability test was two-fold. The first aim was to further develop and test the bootstrap method as given in Van Delden, Scholtus and Burger (2016b) to also estimate the accuracy of growth rates (within data configuration 1) as affected by classification errors in the BR. The second aim was to derive analytical formulas that can approximate these accuracy estimates. Here, the aim is to derive expressions that can be more easily applied in practice by production staff. The ultimate aim is to have a method for estimating the accuracy of statistical outcomes under a variety of non-sampling errors that can be applied in the practice of official statistics.

Data used

Quarterly turnover data from administrative data (VAT) and survey data (only for the largest and most complex enterprises) are used for a case study of car trade. The data refer to a period from the first quarter of 2014 to the last quarter of 2015.

Work carried out

Audit data were obtained for a sample of enterprises in order to estimate errors in *changes* of NACE codes in the BR. An audit sample of 300 enterprises was taken for a case study of car trade for which previously NACE code errors for one time period were estimated (Van Delden, Scholtus and Burger, 2016a). Experts strove to determine the correct NACE code at two time moments 12 months apart. The audit results were used to estimate a model that describes the probability of NACE code changes. Next, a bootstrap simulation approach was used that repeatedly draws from the probability model of classification errors, and each replicate produces a sequence of quarterly turnover growth rate estimates. In particular, 100 bootstrap replicates were drawn and used to compute the bias and variance of the turnover estimates, as affected by classification errors (Van Delden, Scholtus and Burger 2016a). Furthermore, bias and variance were partitioned into contributions from NACE transition groups (transitions from inside or outside the target industries).

Van Delden, Scholtus and Burger (2016a) derived analytical formulas to assess the accuracy of level and growth rate estimates as affected by industry code errors, using a number of simplifying assumptions (see SR1_2: Scholtus, Van Delden and Burger, 2017). The assumptions limit the number of parameters on classification errors that are needed as input for the computations. The analytical formulas can also simplify the computations. Both aims support the implementation of the method in the practice of statistical institutes. We have estimated the input parameters that are needed for the analytical formulas concerning the level estimates. A comparison between the outcomes of the formulas and those of the bootstrap suggests that the variance estimates are similar. The analytical bias estimates were found to be larger and more variable than those of the bootstrap. More research is needed to understand these results. One possible explanation is that the number of bootstrap replications used so far is insufficient to obtain stable estimates of accuracy. We intend to increase the number of replications in future work.

Cost-benefit analysis
In this study an audit sample to estimate classification error size was used. The reason for using such an audit sample is that additional data on the industry codes is needed, because only the BR is available as a source for NACE codes of enterprises. The audit sampling approach turned out to be costly and time consuming (see also "gap analysis" below).

The whole model is rather elaborate: a large set of probabilities need to be estimated. A question remains whether part of the classification error model can be approximated while still obtaining acceptable results.

There may be a lack of resources for carrying out the bootstrap analysis, because (i) most people work on the actual production of these statistics and (ii) high-skilled personnel is required for carrying out such analyses. The analytical formulas hopefully provide a solution for this issue.

Gap analysis
A way to collect data on the classification error size is needed that is less costly and easier combined with daily production. An idea is to derive those data from editing activities in regular statistical production. This is crucial for the implementation of the method, as it would make the approach much easier to apply in practice.

Statistics Netherlands does not only want to quantify variance and bias of current estimates, but would also like to *reduce* them. One gap in this context is whether it is possible to correct current estimates for bias. Another gap is what editing strategy can be used to reduce the effect of classification errors (see Van Delden et al., 2016b for an illustration why this is not straightforward.) This is a field that has received little attention in data editing up till now. This would make the approach more generally applicable.

Statistics Netherlands is not only interested in classification errors, but in the effect of the most important error types on accuracy (for data configuration 1). The question is how to extend the approach to other error types, such as measurement errors. Again, this would make the approach more generally applicable.

4.1.3 Literature review
Topic Accuracy of estimates affected by classification errors
Title of literature review/file: LR1_1
"Effect of classification errors on domain level estimates in business statistics"
Papers reviewed
- Van Delden, Scholtus and Burger (2015)
- Van Delden, Scholtus and Burger (2016b)

Keywords Categorical data, Cross-sectional data, Business register, Measurement error, Classification error, Correction models for classification errors

4.2 Basic data configuration 2: overlapping microdata sources

4.2.1 Overlapping numerical variables without a benchmark
Title of suitability test/file: ST2_1
"Overlapping Numerical Variables without a Benchmark: Integration of Administrative Sources and

Survey Data through Hidden Markov Models for the Production of Labour Statistics" (Filipponi and Guarnera 2017)

Data configuration
ST2_1 have examines basic data configuration 2, in particular when there are overlapping units as well as overlapping variables in the data sources to be combined.

Two possible situations can be distinguished. In the first situation, which is examined in this suitability test (for the second situation, see Section 4.2.2 below), multiple administrative and survey sources provide the value of a same variable of interest for the entire target population or part of it and all the measures are assumed to be imperfect, that is none of them can be considered as a benchmark. In this case, a Latent Class Model approach can be used to estimate the true value. In this framework, it is possible to classify the variables in three groups:

1. Variables Y^* representing the "true" target phenomenon. These are the variables that one would observe if data were error free. In general, Y^* are considered latent variables because they are not directly observed.

2. Variables Y^g $(g = 1, \dots, G)$ representing imperfect measurements of the target phenomenon. These variables are the ones actually observed from *G* different data sources.

3. Covariates associated respectively with the latent variables Y^* and the measurements Y^g .

Given that the variable of interest is categorical and the data are longitudinal, appropriate models are Hidden Markov Models (HMM).

In this approach, accuracy measures are naturally provided by the conditional distribution of the latent *true* variable given the available information (e.g., the posterior variance).

Keywords Categorical data, Longitudinal data, Survey data, Administrative data, Measurement error, Latent Class Analysis, Hidden Markov Models, Mixture models

Type of error and aim of the test
In this approach measurement error, and in particular the error in classifying individuals with respect to employment status, is accounted for both in the survey and administrative measurement processes, thus considering a symmetric situation for the two data sources. Point estimates for the bias and root mean squared error and distributions of the estimation errors are provided, under different scenarios.

The aim of the test is to assess the robustness of the methodology with respect to departures from the model assumptions.

Data used
The goal of the study was to combine administrative and survey data in order to build a "labour register" to be used for producing estimates on the employment status at detailed level. To this aim, data from the Italian Labour Force Survey (LFS) and administrative data strongly associated with the target phenomenon were used.

The Italian LFS is a continuous survey carried out during every week of the year. Each quarter, the LFS collects information on almost 70,000 households in 1,246 Italian municipalities for a total of 175,000 individuals. The reference population of the LFS consists of all household members officially resident

in Italy, even if temporarily abroad. Households registered as resident in Italy who habitually live abroad and permanent members of collective facilities (hospices, children's homes, religious institutions, barracks, etc.) are excluded.

The LFS provides quarterly estimates of the main aggregates of labour market data (employment status, type of work, work experience, job search, etc.), disaggregated by gender, age and territory (up to regional detail).

Several administrative sources have been used for the application. In particular, social security data have been used for employees, whereas Chambers of Commerce, social security and fiscal data are the main sources of information for self-employed. Since all the administrative data have a linked employer-employees (LEED) structure, a pre-processing activity was necessary in order to combine the different data sources. Specifically, administrative information is arranged in a data set where each record corresponds to the statistical unit, i.e. the worker. Finally, some work was needed in order to take into account the longitudinal nature of the available information. In particular, since the *month* was chosen as time interval for analysis, a complex processing step was necessary in order to transform the information originally available in each source to a *monthly value* of the target variable.

The variable of interest for the application is the employment status. According to the Labour force status individuals are classified in three categories as *employed*, *unemployed* or e*conomically inactive*. The definitions of the employment status used in the LFS are fully compatible with the International Labour Organisation (ILO) guidelines and form the basis concept for labour market statistics. Differently, the employment status derived by the available administrative sources categorizes individuals as employed or not employed based on some administrative signals like social security contributions or earned income in the reference period.

Work carried out

A latent class model has been developed. Results show that the employment status in the LFS data is measured very accurately. The overall amount of error in the administrative data is larger than in the survey data. The *true* values of the target variable can be predicted for all individuals in the sample, using the posterior probability of having a certain employment status at each time point.

In order to take into account the longitudinal structure of the available data, an evaluation of the used Hidden Markov model methodology has been carried out by means of a Monte Carlo study. The analysis is based on completely simulated data. For each Monte Carlo iteration, a sample of error free data has been generated according to a Markovian assumption with different hypotheses on the parameters. Two measurement processes have been simulated according to different sets of misclassification errors. Finally, missing values have been introduced in such a way as to reproduce the missing rate and structure observed in the LFS data.

For each experiment and for each Monte Carlo run, an EM algorithm has been used to fit different Hidden Markov models: the one used to generate the data and others with different levels of complexity. If one assumes that the parameter of interest is the population total of the target variable, this quantity can be computed using on the sum of the model predictions, i.e. the posterior probability of the true data conditional on data observed from the different sources.

Then, for each method, the relative bias is estimated by averaging the estimates over the iterations. Similarly, and the relative root mean square error is estimated by averaging the squares of the estimation errors over the iterations.

The results indicate that the estimation works well both in terms of bias and variance even in presence of incomplete data. However, if the estimating model does not coincide with the generating model, the estimates can deviate significantly from the true value.

The posterior variance of the conditional distribution of the latent variable given the observed measurements is a natural accuracy measure of the predicted values.

Cost-benefit analysis
Considering that estimated probabilities need to be taken into account in the conditional distributions, calculating the posterior variance may not be an easy task. Highly skilled statisticians and specialized software may be needed for carrying out such calculations.

Gap analysis
A gap is a simple and reliable method to calculate standard errors for parameter estimates for the developed approach. This would simplify the use of the approach in practice.

An addition source of variability (and thus of complications) may arise when the reference population is unknown and need to be estimated. This additional source of variability is not yet taken into account in this approach.

4.2.2 Overlapping numerical variables with a benchmark

Title of suitability test/file: ST2_2
"Overlapping Numerical Variables with a Benchmark" (Fasulo and Solari 2017)

Data configuration
ST2_2 examines basic data configuration 2 for the situation where there are overlapping units as well as overlapping variables in the data sources to be combined and one of the data sources (typically a survey measurement) is considered as error free and the administrative measurements are merely used as auxiliary information.

Thus, differently from the situation in Section 4.2.1, the error free variable (Y) is modelled as a *response* variable and all the other measures (X) are considered *covariates*. This supervised approach can be adopted in a model based inference approach as well as in a design based inference approach. In the latter case, the covariates can be used to specify a *working* model (model assisted approach). The choice of the methodological approach depends both on the informative content and the quality of the available data sources.

Keywords Categorical data, Cross-sectional data, Survey data, Administrative data, Measurement error, Model-based approach, Design-based approach, Projection estimator

Type of error and aim of the test
In this approach both survey administrative data are considered error free. The error type considered in this application is the measurement error associated to the estimation of the employment status.

Aim of the test is to analyse the feasibility of the application of a projection estimator to employment status, using as auxiliary variables information from administrative sources, under different models, by evaluating bias and mean square errors associated to the different models.

Data used

Again the goal of the study is to combine administrative and survey data in order to build a "labour register" to be used for producing estimates on the employment status at detailed level. To this aim, again data from the Italian Labour Force Survey (LFS) and administrative data strongly associated with the target phenomenon are used. For more information, see Section 4.2.1. The experimental study was conducted on data coming from the Italian regions Trentino-Alto Adige and Marche.

Work carried out

In this situation, where administrative data play the role of auxiliary variables, evaluation of accuracy is primary based on the estimation of sampling error.

If we assume that administrative and survey data have an asymmetric role, then adequate statistical models can be applied only to the units belonging to survey samples and the corresponding estimates can then be used to predict the phenomena over the unsampled population units, either using a model-based (not considering sample weights) or a design-based approach (considering information derived from the survey sampling scheme).

The choice of a projection unit level estimator is connected with (i) the purpose of producing a microdata archive which can be used to spread information at the required territorial levels, such as the municipal level and (ii) and the need to produce coherent estimates with those provided by different ISTAT surveys dealing with the same thematic areas of interest (employment area in this case).

Also, in this setting, one can use the employment status provided by the LFS, which provides the target variable that need to be estimated, and the administrative data as auxiliary determinants in estimating the phenomena of interest.

Different options for the estimator have been tested. The first is the model assisted linear estimator proposed by Kim and Rao (2012). Other estimators are the composite estimator deriving from the linear mixed model where the random effects are the Labour Market Areas (LMAs).

In order to analyse the feasibility of the application of a projection estimator to employment LFS data considering as auxiliary variables the administrative information already mentioned before, with the aim to produce different territorial levels estimates (provinces, macro LMAs, LMAs and municipalities) a simulation scheme has been implemented. This simulation scheme consists of drawing different samples, estimating the model over the samples and projecting the results over the entire population. The performance is assessed in terms of bias and mean squared errors by means of:

$$\mathrm{MARE} = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{200}\hat{y}_{rd} - Y_d\right|$$

$$\mathrm{ARRMSE} = \frac{1}{D}\frac{1}{R}\sum_{d=1}^{D}\sum_{r=1}^{R}\frac{\sqrt{(\hat{y}_{rd}-Y_d)^2}}{Y_d}.$$

The following model specifications have been used in the experimental study.

Projection on pooled sample

Full model:	Marital status, educational level, citizenship, not in labour force, cross classification gender-age, register variable by region
Reduced model:	Marital status, educational level, citizenship, cross classification gender-age, register variable by region

Projection on register

Minimal model:	Marital status, citizenship, cross classification gender-age, register variable by region

Cost-benefit analysis
Highly skilled statisticians and specialized software may be needed for carrying out the calculations for this approach.

Gap analysis
A very useful aspect to be accounted for in the employment status estimation at a certain time *t*, is the employment status information of the same person from previous moments in time. Taking the longitudinal aspect into account is not always trivial. An interesting research topic we propose to develop in the future is the utilization of longitudinal models in the case of not balanced samples and in a projection model-assisted approach for microdata. This would make the developed approach more generally applicable.

4.2.3 Misclassification in several administrative sources

Title of suitability test/file: ST2_3
"Misclassification in Several Administrative Sources" (Statistics Austria 2017)

Data configuration
We have analysed a quality framework developed by Statistics Austria (ST2_3 (Statistics Austria 2017); see also Asamer et al. 2016, Berka et al. 2010 and Berka et al. 2012) for basic data configuration 2. This quality framework uses overlapping microdata to compute a quality indicator for the output variable. The framework was originally developed and used for the Austrian register-based census 2011. This was a full enumeration from several administrative sources.

In Berka et al. (2010) the content of the framework is briefly introduced, especially the assessment of registers. The main technical part, the Dempster-Shafer Theory of Evidence, is described in detail in Berka et al. (2012). The assessment of imputations for a register-based statistic is discussed in Schnetzer et al. (2015). The procedure allows one to track changes in quality during data processing. This is examined in Asamer et al. (2016). All these papers mentioned, concentrate on the example of the Austrian register-based census 2011. A documentation of methods (Statistics Austria 2014) collects the overall results of the framework in detail. This can be viewed as the main source on the framework in general. Furthermore, some potential for analysis of a register-based statistic via the framework is demonstrated in Asamer, Rechta and Waldner (2016).

In principle, the quality framework should be also applicable to basic data configurations 1, 2S and 3.

Keywords Categorical data, Numerical data, Cross-sectional data, Survey data, Administrative data, Expert knowledge, Measurement errors, Dempster-Shafer Theory of Evidence, Fuzzy logic

Type of error and aim of the test
In the framework, and in the test, two kinds of errors are distinguished: classification errors in the observed data and errors due to imputation.

So, the two cases are:

- An output value comes from a data source. There are misclassifications in all data sources and so (to a lower extend) also in the final output data. Via different quantitative indicators for input quality and an additive indicator based on a comparison with external data (e.g. from the Labour Force Survey) we derive an output quality indicator.
- An output value was imputed. If missing values for the final output data are estimated, then a misclassification by imputation will occur in general. Such classification errors due to imputation are computed by a classification rate of the model and the quality of the predictors (as computed previously). This yields an output quality indicator for the imputation situation.

The aim of the test was to examine and describe all properties of the quality framework.

Data used
For the test, data from the recently finished register-based labour market statistics 2014 (RBLMS) was used.

The actual data processing for the RBLMS is conducted in three stages that are considered in the quality assessment framework: the raw data (i.e. the administrative registers), the combined data set of these raw data (Central Data Base, CDB) and the final data set, which includes imputations (Final Data Pool, FDP).

In particular, we applied the framework on the variable “Legal family status” of the RBLMS 2014, which was created from ten source registers:

- unemployment register
- register of public servants of the federal state and the Länder (lands)
- family allowance register
- central social security register
- central register of foreigner
- chambers register
- hospital for public servants register
- register of social welfare recipients
- tax register
- central population register

Four hyperdimensions aim to assess the quality for different types of attributes at all stages of the data processing. This yields quality measures that are standardized between zero and one, where a higher value indicates better data quality. The individual data lines are matched via a unique personal key and merged to data cubes in the CDB. Finally, missing values in the CDB are imputed in the FDP where every attribute for every statistical unit obtains a certain quality indicator.

In the computations, quality information at the raw data level, the CDB level and the FDP level are computed and combined.

Work carried out

We have computed quality indicators throughout the process in order to show the different quality aspects. These quality indicators provide important additional information to "understand" the final quality indicator and how it can (possibly) be improved (see Asamer et al. 2016 or Statistics Austria 2014 for a detailed description of these indicators).

Cost-benefit analysis

- No specific IT-tools are needed (just common used, e.g. SAS, R, SPSS, xls, …)
- No separate audit sampling is necessary, but it could be helpful. It is possible to use already existing surveys like the Labour Force Survey for at least some variables (in the event a suitable survey is lacking, one can substitute it by a so-called expert view).
- A slight increase in response burden for the data holders for metadata information.
- If the output variable is delivered by just one data source, the computation is simple.
- If the output variable is delivered by more than one data source, the implementation of the Dempster-Shafer-theory can require much work.
- In practice, the approach can be used easily (once it is already implemented)
- If we assume that the Dempster-Shafer-theory is already implemented, the necessary time for applying the actually framework depends on the type of attribute, the data sources and the recyclability of existing code.

Gap analysis

In the framework, the comparison of the data with an external source can be refined. For instance, one can compute a classification rate for each value, instead of an average value for the whole attribute as is currently done or one may choose a suitable aggregation-level for comparison.

4.2.4 Two merged administrative data sources with missing values in a classification variable (household data)

Title of suitability test/file: ST2_4

"Effect of the Under-coverage of the Classification Variable on the Domain Estimates of the Total in Social Statistics" (Krapavickaitė and Vasilytė 2017a)

Data configuration

ST2_4 describes a suitability test in which it is supposed that the classification variable used to define the population domains is taken from an administrative data source and has population under-coverage and/or missing values. The aim is to estimate population totals of a target variable for these domains. The bias and variance of the estimator obtained is not only due to sampling but also due to the under-coverage and/or missingness of the classification variable and methods used to overcome this problem.

Keywords Numerical data, Classification error, Sampling theory, Analytical formulas, Cross-sectional data

Type of error and aim of the test

The population total for the target variable per profession category suffers from missing values in

profession category. The bias and variance for estimators of the target variable per profession category are derived by analytical formulas. The results of these formulas are compared with simulated values, where the complete set of data is known from all elements in the population. The accuracy of the estimators (that use different methods to correct for missingness) is compared.

Data used

A data set on the budget used for subsistence per household is used. These data are based on an existing survey on living conditions. Also, an administrative tax inspection data base is used that contains the variable profession.

Work carried out

In the suitability test an analytical solution to the problem is given, and results of a simulation study are presented.

In the suitability test two situations are distinguished. In case A the target variable is contained in an administrative data source covering the complete population. In case B the target variable is contained in a survey data set. Case B is more complicated than case A as the sampling error of the survey has to be taken into account.

For case A, let us suppose that the values of the classification variable are unknown for a sub-population $U^{(b)}$ of size $N^{(b)}$ and known for a sub-population $U^{(a)}$ of size $N^{(a)}$. Using a variable A the sub-population $U^{(a)}$ is classified into m domains $U^{(a)} = U_1^{(a)} \cup U_2^{(a)} \cup \ldots \cup U_m^{(a)}$ of size $N_1^{(a)}, N_2^{(a)}, \ldots, N_m^{(a)}$. Let us study a target variable y defined on $U = U^{(a)} \cup U^{(b)}$ and denote its domain means for the known sub-population $U^{(a)}$ by

$$\mu_g^{(a)} = \frac{1}{N_g^{(a)}} \sum_{k \in U_g^{(a)}} y_k$$

for $g = 1, \ldots, m$.

In order to classify $U^{(b)}$, let us draw a simple random reference sample $s^{(b)}$ of size $n^{(b)}$ from $U^{(b)}$, obtain the values of the classification variable A for the sampled elements, and estimate the proportions

$$p_g^{(b)} = \frac{N_g^{(b)}}{N^{(b)}}$$

by

$$\hat{p}_g^{(b)} = \frac{n_g^{(b)}}{n^{(b)}},$$

where $g = 1, \ldots, m$, and $n_g^{(b)}$ is the number of elements in the sample $s^{(b)}$ with value A_g for classification variable A. The sample $s^{(b)}$ is used as a kind of audit file in order to estimate the proportions $p_g^{(b)}$.

The domain totals of the target variable y, $t_{yg} = \sum_{k \in U_g} y_k$ are estimated by

$$\hat{t}_{yg}^{(1)} = \sum_{k \in U_g^{(a)}} y_k + N^{(b)} \hat{p}_g^{(b)} \mu_g^{(a)}$$

$$\widehat{\text{Var}}\left(\hat{t}_{yg}^{(1)}\right) = \left(N^{(b)}\right)^2 \left(\mu_g^{(a)}\right)^2 \widehat{\text{Var}}(\hat{p}_g^{(b)})$$

$$\text{Bias}\left(\hat{t}_{yg}^{(1)}\right) = N^{(b)}(\mu_g^{(a)}\left(E\left(\hat{p}_g^{(b)}\right) - p_g^{(b)}\right) + p_g^{(b)}(\mu_g^{(a)} - \mu_g^{(n)}))$$

for $g = 1, \ldots, m$.

In order to derive analytical formulas for case B we assume that also a simple random sample $s^{(a)}$ is drawn from the population and domain means $\mu_g^{(a)}$ are estimated by $\hat{\mu}_g^{(a)}$ from this sample, $g = 1, \ldots, m$. We can then estimate t_{yg} with

$$\hat{t}_{yg}^{(2)} = \hat{t}_{yg}^{(a)} + N^{(b)} \hat{p}_g^{(b)} \mu_g^{(a)},$$

where $\hat{t}_{yg}^{(a)} = N_g^{(a)} \hat{\mu}_g^{(a)} = N_g^{(a)} \frac{1}{n_g^{(a)}} \sum_{k \in s^{(a)} \cap U_g^{(a)}} y_k$ (for $g = 1, \ldots, m$).

Analytical expressions for the bias and variance of $\hat{t}_{yg}^{(2)}$ can be derived, and are given in ST2_4 (Krapavickaitė and Vasilytė 2017a).

The quality of the estimators for bias and variance have been tested by means of a simulation study.

Cost-benefit analysis

The developed method requires the collection of additional data, which implies additional expenses.

Gap analysis

The method requires an additional source, namely data on estimates for the missing values of the classification errors. The question remains how this kind of information can be obtained in practical situations. Loglinear models to estimate the proportions $p_g^{(b)}$ can be applied. Alternatively, information from the editing process (and other internal processes) can be used for this.

4.2.5 Accuracy of an estimator of a total as affected by frame-errors (business data)

Title of suitability test/file: ST 2_5

"Effect of the Frame Under-Coverage / Over-Coverage on the Estimator of Total and Its Accuracy Measures in the Business Statistics" (Krapavickaitė and Šličkutė-Šeštokienė 2017)

Data configuration

ST2_5 considers a situation with a Business Register (BR) and two data sources. Besides the BR with all units in the population, there is an up-to-date data source with auxiliary data (say x) for all (actual) units in the population and a survey sample data set with target variable y. This sample and the current estimate to be published are based on a frozen version of the BR. That means that throughout the year the same set of units is used. Based on this frozen version of the data set a target parameter (aggregated from variable y) is published. This target parameter is estimated using a ratio estimator (making use of the ratio between values for target variable y and auxiliary variable x). The problem is to estimate the effect of using a frozen BR (rather than the actual BR) on the accuracy of the target parameter.

First of all the original target parameter might be biased because a frozen BR is used rather than the actual population of businesses, and second, the variance of the original survey variable of the frozen BR might differ from the variance of the survey variable based on the actual population of businesses.

Keywords Numerical data, Cross-sectional data, Administrative data, Survey data, Business register, Measurement error, Classification error, Sampling theory

Data used Business data are used, in particular a quarterly survey on earnings from 2015. The target variable is the average monthly gross earnings, the auxiliary variable is the number of employees in full time units.

Type of error and aim of a test

The aim of the test is to measure impact of the changes in the enterprise population on the accuracy of the estimator of the total. We assume that a stratified sample design is used for a survey. A separate ratio estimator and a combined ratio estimator are used to estimate the total of a target variable from a population suffering from under-coverage, over-coverage and/or missingness.

Work carried out. Suppose we have a population U of size N, divided into the strata U_h of sizes N_h $(h = 1,2,\dots,H)$ $(\sum_h N_h = N)$. A simple random stratified sample $s = s_1 \cup \dots \cup s_H$ of sizes $n_1, \dots, n_H$ $(\sum_h n_h = n)$ is selected from U. Let the administrative data source V be divided into the same strata with sizes N_h' $(h = 1,2,\dots,H)$. Let the administrative data source V have both under-coverage and over-coverage of the population U. We consider V as an evolution of the population U over time. As a result of under-coverage and non-response in V the sample s reduces to $s' = s_1' \cup \dots \cup s_H'$ with $s_h' \subset s_h$ of sizes n_h', $\sum_h n_h' = n'$, $n_h' \leq n_h$.

Let a target variable y be obtained from U, and an auxiliary variable x defined from V. Denote $t_{yh} = \sum_{k\in U_h} y_k$ and $t_{xh} = \sum_{k\in U_h\cap V_h} x_k$ for $h = 1,2,\dots,H$, and $t_{yh} = \sum_{k\in U} y_k = \sum_{h=1}^{H} t_{yh}$.

Let variable z be defined for V. Restricted to U variable z coincides with y, i.e. $z_k = y_k$ for $k \in U$. We define

$$t_z = \sum_{k\in V} z_k = \sum_{h=1}^{H}\sum_{k\in V_h} z_k = \sum_{k\in U\cap V} y_k + \sum_{k\in V\setminus U} z_k.$$

We are interested in estimators of t_z. Denote

$t_x = \sum_{k\in U\cap V} x_k = \sum_{h=1}^{H} t_{xh} = \sum_{h=1}^{H}\sum_{k\in U_h\cap V_h} x_k$, $\tilde{t}_x = \sum_{k\in V} x_k = \sum_{h=1}^{H} \tilde{t}_{xh} = \sum_{h=1}^{H}\sum_{k\in V_h} x_k$,
$\hat{t}_{yh} = \frac{N_h^*}{n_h'}\sum_{k\in s_h'} y_k$, $\hat{t}_{xh} = \frac{N_h^*}{n_h'}\sum_{k\in s_h'} x_k$ for $h = 1,2,\dots,H$.

Here N_h^* $(N_h^* \leq N_h)$ is the estimated stratum size after subtraction of $N_h - N_h^*$ units from U that are no longer in V. It is estimated taking non-response into account.

Finally, we define

$$\hat{t}_{y,\text{str}} = \sum_{h=1}^{H} \hat{t}_{yh},\ \hat{t}_{x,\text{str}} = \sum_{h=1}^{H} \hat{t}_{xh}.$$

Three estimators are considered.

(i) <u>*The separate ratio estimator for a total t_x with a fixed sampling frame*</u>. When changes in the population are not taken into account this estimator is given by:

$$\hat{t}_{y,\text{fixed}}^{(1)} = \sum_{h=1}^{H} t_{xh}\frac{\hat{t}_{yh}}{\hat{t}_{xh}} \qquad \text{for known totals } t_{x1}, \dots, t_{xH}.$$

When changes in the population frame are taken into account this estimator is given by

$$\hat{t}_{z,\text{adj}}^{(1)} = \sum_{h=1}^{H} \tilde{t}_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}} \qquad \text{where } \tilde{t}_{xh} = \sum_{k \in V_h} x_k \text{ is known.}$$

(ii) *The combined ratio estimator for a total t_x with a fixed sampling frame*. When changes in the population are not taken into account this estimator is given by:

$$\hat{t}_{y,\text{fixed}}^{(2)} = t_x \frac{\hat{t}_{y,\text{str}}}{\hat{t}_{x,\text{str}}}.$$

When changes in the population frame are taken into account this estimator is given by

$$\hat{t}_{z,\text{adj}}^{(2)} = \tilde{t}_x \frac{\hat{t}'_{y,\text{str}}}{\hat{t}'_{y,\text{str}}},$$

where $\tilde{t}_x$ is known, $\hat{t}'_{y,\text{str}} = \sum_{h=1}^{H} \frac{N_h^*}{n'_h} \sum_{k \in s'_h} y_k$, $\hat{t}'_{x,\text{str}} = \sum_{h=1}^{H} \frac{N_h^*}{n'_h} \sum_{k \in s'_h} x_k$.

(iii) The estimator $\hat{t}_{z,\text{adj}}^{(2)}$ can be replaced by the post-stratified estimator of the total or by the calibrated estimator of the total, which may take into account the population changes after a time period.

For case (i), the relative difference between $\hat{t}_{y,\text{fixed}}^{(1)}$ and $\hat{t}_{z,\text{adj}}^{(1)}$ and the relative difference between the variances of these estimators are measures for the effect of under-coverage, over-coverage and missingness problems in the (evolved) population frame. For cases (ii) and (iii) the same quality measures are used for the corresponding estimators. The quality measures have been tested in a case study.

Cost-benefit analysis

The method assumes that we have an administrative data source with perfect population coverage containing an auxiliary variable level, which is correlated with the target variable, of which values are available for a sample. If this assumption is met, the method is easy to apply.

Gap analysis

The method can be extended by replacing the ratio estimator with other estimators using auxiliary information: the post-stratified, regression and calibrated estimators.

The method can be adapted to the estimators of ratios. After the method proposed is used to estimate totals of two variables, it can be applied to estimate the ratio of these totals in order to obtain an estimator of the ratio adjusted for the frame changes.

4.2.6 Accuracy of an estimator as affected by classification errors (business data)

Title of suitability test/file: ST2_6

"Effect of Stratum Changes, Joining and Splitting of the Enterprises on the Estimator of a Total" (Krapavickaitė and Vasilytė 2017b)

Data configuration

ST2_6 describes a situation with an administrative data source with a target variable y and a classification variable z. The classification variable z suffers from errors. There is also a second administrative data source with classification variable z, but that in this second data source the classification variable suffers much less from errors (say it is perfect). Note that this situation might occur in practice when the second data source is received too late to be used at the time of first compilation of the estimator. This second data source might be used to benchmark when it comes

available later in time. There is a target variable y with (original) inclusion probabilities that are stratified by the classification variable z in data source one. Now a population parameter is published that is based on a Horvitz-Thompson (HT) estimator using variable y, stratified by variable z.

The problem consists of estimating the changes of the enterprise population after the sample selection due to the changes in their kind of activity, enterprise joining and enterprise splitting. The parameter estimation method proposed is based on the re-calculation of the inclusion probabilities. The variability of the difference between the sampling design-based estimator and re-calculation plus the design-based estimator shows the impact of the frame shortcomings on the accuracy of the final estimates.

The problem is to estimate the accuracy of the population parameter per stratum as affected by the errors in the classification variable z.

Keywords Numerical data, Cross-sectional data, Administrative data, Survey data, Business register, Measurement error, Classification error, Sampling theory

Data used
The approach can be applied to business data.

Work carried out
Consider the situation where all statistical units (enterprises) are registered in a Business Register (BR). This BR is used as a sampling frame for a stratified random sample, where the strata are formed by, for instance, the NACE code for economic activity and size (number of employees). The sampled units are used to estimate an overall population total of a variable y. Now we assume that two types of errors occur that have an effect on the stratification of the units: 1) errors in the value of the classification variable (e.g. errors in the NACE code)and 2) errors in the composition of the enterprise (see below).

Let us introduce two terms: the *selected sample* and the *observed sample*.

After the survey data of sampled enterprises have been obtained, it may be detected from the observed sample that the sampling units (SU) have changed their characteristics. Three types of such changes occur:

(a) the SU have been joined with other enterprises, possibly from different strata;
(b) the SU have been split into multiple new ones with possibly different values for the stratification variables;
(c) another value for the classification variables (for example, NACE code or size group) of the SU has been reported.

These changes may occur because of errors in the classification variable taken from the administrative data source or because of the changes in the population which occurred between the sample selection and data collection.

Let u_i denote unit (= enterprise) i in the selected sample and u_i' the "same" unit i in the observed sample. Note that u_i may differ in composition from u_i', but it concerns the "same" unit in the sense that this is the unit for which the observed data are reported. Between u_i and u_i' the changes (a), (b) and/or (c) might have occurred.

The population U consists of the elements u_k, with $k = 1, \dots, N$, and is stratified into H non-overlapping strata U_h, with U_h the size of stratum h, h= $1, \dots, H$. A simple random sample ω_h of size n_h is drawn from the stratum U_h. We denote the complete sample of size $n = n_1 + \cdots + n_H$ by ω.

After the survey data have been received, it is detected that, in some cases, the observed units are not necessarily the same as the selected ones. Denote the observed sample by ω' with elements u_i', $i = 1, \dots, n'$, where n' may differ from n (and similarly n_h' may differ from n_h).

We are interested in estimating a population total for a variable y. When there would not have been any errors in the population, this total would have been given by

$$t_y = \sum_{k=1}^{N} y_k,$$

where for brevity in the notation we use y_k for observations of variable y for enterprise u_k in the BR at the time of sampling. Since there are errors in the population in our case, t_y has to be estimated, however. Krapavickaitė and Vasilytė (2017b) show that by recalculating the inclusion probabilities, we can estimate t_y, and that the variance of this estimate $\hat{t}_y$ can also be estimated.

A simulation study has been carried out to evaluate the estimator $\hat{t}_y$ and its estimated variance.

Cost-benefit analysis

The methods presented are based on the following assumptions:

- no non-response
- the probability of joining units in the sampling frame or a unit changing from one stratum to another, is completely at random: it is not related to the stratification variable and it is not related to the study variable.

If these assumptions are met, the method is quite straightforward and can be applied easily. If these assumption are not met, the method cannot be applied.

Gap analysis

The method requires an additional source, namely a data source that is free of classification errors. The question remains how this kind of information can be obtained in practical situations. Possibly information from the editing process (and other internal processes) can be used for this.

4.2.7 Suitability test using "a-meldingen"

Title of suitability test/file: ST2_7

"Output Quality for Statistics Based on Several Administrative Sources" (Fosen 2017)

Data configuration and data used

ST2_7 uses a proxy for the new register-based employment status based on the a-ordningen system. The data set is a proxy since, in the production process this data set slightly precedes data set used by the Division for Labour market statistics for producing the statistics. The difference is some minor additional integration

Keywords Categorical data, Numerical data, Cross-sectional data, Survey data, Administrative data, Measurement errors, Progressiveness error

Data used

ST2_7 uses a proxy for the new register-based employment status. It is a proxy in the sense that we

use a statistical data set that comes slightly earlier in the production process than the final data set used for producing the statistics: we use the so-called "L2 data set" which is the statistical register from the perspective of the a-ordningen production system, whereas the Division for Labour market statistics do some further integration based on this L2 data set before they do the final classification of employment status. We will below by "estimated employment rate"' refer to this proxy.

Type of error and aim of the test

The aim is to study the effect of progressiveness, more specifically "delayed arrival of register information" from the employers into the input sources/registers that are used in the production process of the employment statistics.

For a given reference time, messages from the employers arrive at different times, the *arrival times*. This delay has an effect on the classification of employees and thus on the employee-statistics being disseminated. The *measurement time* is the time when we take the last updated input register and put this into the production process, i.e. we measure all the messages arrived so far and get employee rates including all delayed messages arriving before the measurement time. The choice of the measurement time is a trade-off between accuracy and timeliness. A goal of the test is to find out whether and to what extent there is anything to gain by using measurement time t_2 months after the reference time instead of t_1 months after the reference time.

Work carried out

For a register-based statistic, there are at least two time dimensions. The first time dimension is the reference time, i.e. the calendar time that our statistics is intended to describe. As an example, we can consider the employment rate for the week containing 15 February 2016.

For a given reference time, the input registers (input into the production system) are updated several times in the months following the reference time due to the *progressiveness* of the input registers: for a given reference time some of the register information can be received delayed from the employer and arrive after the information should have arrived. For the given reference time: when we choose to start the production, we choose the last updated version. We denote this chosen update time as the *measurement time*. This is the second time dimension. We denote the interval between the reference time and the measurement time as the *relative measurement time* (RMT). The RMT is an indicator of timeliness.. The accuracy increases with longer RMT, but this naturally comes at the cost of loss in timeliness. The purpose of the suitability study was to find an estimate of the effect of the output cause by delay, in order to decide on how to long to wait before starting the production process.

Let $\hat{S}_{r,t}$ denote the estimated employment rate at reference time r based on using the updated register information at measurement time t as input sources to the production process. Consider the difference in estimated employment rates between two different measurement times t_1 and t_2:

$$X_{r,t_1,t_2} = \hat{S}_{r,t_1} - \hat{S}_{r,t_2}$$

In order to model the effect-of-delay estimator, we will take as a starting point the assumption that the delay mechanism, measured by our simple estimator**,** has a constant expectation over the reference time, more specifically we assume that X_{r,t_1,t_2} $(r = r_1, \dots, r_n)$ are independently identically distributed, and thus

$$\mu_{r,t_1,t_2} = \mathrm{E}\left(X_{r,t_1,t_2}\right) = \mu_{t_1,t_2} \text{ and } \sigma^2_{r,t_1,t_2} = \mathrm{Var}\left(X_{r,t_1,t_2}\right) = \sigma^2_{t_1,t_2}$$

This is a strong assumption that is tested by a set of tests suggested in ST2_7 (Fosen 2017). .

The first question is whether there is any substantial gain in waiting until t_2 when producing the register-based employment statistics instead of starting with the measurements at t_1. The second question is how much gain there is in waiting.

If we have a sufficient number of observations, i.e. X_{r,t_1,t_2}, for a large number n of reference times r, $\bar{X}_{,t_1,t_2} \sim N(\mu, {\sigma^2}/{n})$. Since the variance σ^2 is unknown, we have

$$U = \frac{\left(\bar{X}_{t_1,t_2} - \mu\right)}{\mathrm{SE}\left(\bar{X}_{t_1,t_2}\right)} \sim t_{n-1}$$

where t_{n-1} is Student's T-distributed with $n-1$ degrees of freedom. We can then use the standard t-test for testing the null hypothesis H_0: $\mathrm{E}\left(X_{r,t_1,t_2}\right) = 0$ against the alternative H_1: $\mathrm{E}\left(X_{r,t_1,t_2}\right) \neq 0$ for answering the first question of whether there is a gain in waiting.

Under the distributional assumption, the effect of delay is $\bar{X}_{,t_1,t_2}$, thus by waiting until t_2 before starting the production process, the estimated output variable is improved by $\bar{X}_{,t_1,t_2}$.

Cost-benefit analysis

For many production processes at different NSIs, there will probably occur some challenges in obtaining output data based on the different measurement times, since the production process typically only uses data from *one* measurement time and the production is run through only for this measurement time. For the other measurement times, the data acquisition is less straightforward and it may be relatively costly or time-consuming to obtain required output data. Firstly, the input data from different measurement times must be available, and then the NSI has to run through the complete production process for each measurement time, without disturbing all the files produced by the official production based on the selected measurement time.

An advantage of the approach is that the statistical methodology is rather simple and thus transparent. A disadvantage is, however, the strong distributional assumptions that currently is required. It should be possible to develop methods relying on weakened assumptions.

Gap analysis

The separation of delay from other measurement errors is an interesting topic for future work.

Delays from all the sources are treated together and the combined effect of all delays is studied. The effects could be split by kind of message, which would be very interesting not for assessing the output quality but for identifying in which input sources the delays occur. This would make the approach more generally applicable.

4.2.8 Overlapping categorical variables

Title of document/file "Boeschoten Oberski De_Waal"
"Estimating Classification Error under Edit Restrictions in Combined Survey-Register Data"
(Boeschoten, Oberski and De Waal 2016).

This work has not been carried out as part of the ESSnet on Quality in Multisource Statistics (Komuso). We have included a description of the work in this report as we feel it may be of interest to producers of multi-source statistics. The file is made available as a reference document.

Data configuration
SR2_7 examines basic data configuration 2, in particular when there are overlapping units as well as overlapping variables in the data sources to be combined.

Keywords Categorical data, Cross-sectional data, Survey data, Administrative data, Measurement error, Latent Class Analysis

Type of error and aim of the test
In the suitability test measurement errors are examined. In all data sources used measurement errors can occur. The different values for a variable measured for a certain unit in several data sources are seen as indicators for the true value. The true value is estimated. The aim of the test is to estimate the quality of the estimated true value.

Data used
For the suitability test a combined data set is used to measure home ownership. This combined data set consists of data from the LISS (Longitudinal Internet Studies for the Social sciences) panel from 2013, which is administered by CentERdata (Tilburg University, The Netherlands) and an administrative data set from Statistics Netherlands from 2013. From this combined data set, two variables are used, indicating whether a person is a home-owner or rents a home as indicators for the “true" latent variable home-owner/renter or other. The combined data set 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. In a contingency table of the imputed latent variable home-owner/renter and rent benefit, there should be zero persons in the cell “home-owner × receiving rent benefit”.

The three data sets used to combine the data are:

- Registration of addresses and buildings (BAG): A register containing data on addresses and buildings originating from municipalities from 2013. From the BAG a variable was used that indicates whether a person owns or rents the house he or she lives in.
- LISS background study: A survey on general background variables from January 2013. From this survey the variable marital status was used as auxiliary variable.
- LISS housing study: A survey on housing from June 2013. From this survey the variable rent benefit is used, indicating whether someone receives rent benefit or not.

These data sets are linked on unit level, and matching is done on person identification numbers. Not every individual is observed in every data set.

Work carried out
The true value for “home-ownership” has been modelled as a latent variable by means of a latent class model. The method takes visibly and invisibly present errors into account by combining multiple Imputation and latent class (LC) analysis. “Visibly” present errors are those errors that violate specified edit rules, such as that house owners cannot receive rent benefit. “Invisibly” present errors are all other kinds of errors.

The method starts by taking m bootstrap samples from the original combined data set. These bootstrap samples are drawn because we want the imputations we create in a later step to take parameter uncertainty into account. We have multiple data sets linked on a unit level, containing the same variable, which can be used as indicators measuring one latent variable. In this way, we estimate the invisibly present classification errors in the combined data set. The latent variable we aim to estimate can be seen as the "true variable". We denote this latent variable by X. We assume that X is categorical and has C categories.

The LC model is based on three assumptions. The first assumption is that the probability of obtaining marginal response pattern y, $P(\boldsymbol{Y} = y)$, is a weighted average of the C class-specific probabilities $P(\boldsymbol{Y} = y|\boldsymbol{X} = x)$: $\mathrm{P}(\mathbf{Y} = \mathrm{y}) = \sum_{x=1}^{C} P(X = x)P(\boldsymbol{Y} = y|X = x)$. The second assumption is that the observed indicators are independent of each other given an individual's score on the latent "true variable". The third assumption is that the measurement errors are independent of the covariates. For example, a covariate which can help in identifying whether someone owns or rents a house is marital status, this covariate is denoted by Q. These assumptions lead to the model:

$$P(\boldsymbol{Y} = y|Q = q) = \sum_{x=1}^{C} P(X = x|Q = q) \prod_{l=1}^{L} P(Y_l = y_l|X = x)$$

Covariate information can also be used to impose a restriction on the model, to make sure that the model does not create a combination of a category of the "true" variable and a score on a covariate that is in practice impossible. For example, when an LC model is estimated to measure the variable *home ownership* using three indicator variables, and a covariate (denoted by Z) measures *rent benefit*, the impossible combination of owning a house and receiving rent benefit should not be created. Such a restriction can be taken into account by imposing

$$P(X = \text{own a home}|Z = \text{rent benefit}) = 0$$

The next step is to impute latent "true" variable X. The uncertainty caused by the classification errors should be correctly taken into account. Therefore, multiple imputation is used to impute X. m values are created and imputed by drawing one of the categories using the m estimated posterior membership probabilities from the LC model. The m estimates can be pooled by making use of the rules defined by Rubin for pooling (Rubin 1987, p.76).

Besides the uncertainty caused by missing or conflicting data, the developed model also takes parameter uncertainty into account.

In Boeschoten, Oberski and De Waal (2016) a simulation study has been carried out to study the quality of the estimated true values.

Cost-benefit analysis

The model developed can only be applied if the auxiliary variables used are error-free themselves.

The model assumes that all variables that are related to the target variable are used in the latent class model. Additional variables, for instance from data sources that become available later, cannot be included anymore.

Gap analysis

The points mentioned above in the Cost-benefit analysis also describe the gap identified so far: to

extend the approach to the case the auxiliary variables can also contain errors, and to the case where a variable related to the target variable becomes available later. This would make the approach more generally applicable.

4.2.9 Literature review

We have carried out an extensive literature review for basic data configuration 2. We have reviewed the following papers/topics:

Topic Effects of classification errors in categorical data
Title of literature review/file: LR2_1
"Estimating classification errors in administrative and survey variables by latent class analysis"
Papers reviewed

- Boeschoten, Oberski and De Waal (2016)
- Pavlopoulos and Vermunt (2015)

Keywords Classification errors, Measurement errors, Social statistics, Categorical data, Cross-sectional data

Topic Quality assessment for Register-based Statistics
Title of literature review/file: LR2_2
"Quality Assessment for Register-based Statistics - Results for the Austrian Census 2011"
Paper reviewed Asamer et al. (2016)
Keywords Measurement errors, Cross-sectional data, Dempster-Shafer theory

Topic Quality assessment of imputations in administrative data
Title of literature review/file: LR2_3
"Quality Assessment of Imputations in Administrative Data"
Paper reviewed Schnetzer et al. (2015)
Keywords Missing data, Imputation errors

Topic Quality for Census enumeration
Title of literature review/file LR2_4
"A Comparison of Methodologies for Classification of Administrative Records - Quality for Census Enumeration"
Paper reviewed Steeg Morris (2014)
Keywords Classification errors

Topic Classification errors on domain level estimates in business statistics
Title of literature review/file: LR2_5
"Effect of classification errors on domain level estimates in business statistics"
Paper reviewed Guarnera and Varriale (2016)
Keywords Classification errors, Measurement errors, Business statistics, Numerical data, Cross-sectional data

Topic Effects of linkage errors
Title of literature review/file: LR2_6
"Effect of linkage errors using 1-1 linkage on inferences from the linked data"

Papers reviewed

- Chambers et al. (2009)
- Chipperfield and Chambers (2015)
- Fellegi and Sunter (1969)
- Lahiri and Larsen (2005)

Keywords Linkage errors

Topic Validity of observed variable as an indicator for the target variable
Title of literature review/file: LR2_7
"Estimating measurement errors in administrative and survey variables by structural equation models"
Paper reviewed Scholtus, Bakker and Van Delden (2015)
Keywords Measurement error

Topic Quality assessment of register-based census employment status
Title of literature review/file: LR2_8
"Quality assessment of register-based census employment status"
Paper reviewed Fosen and Zhang (2011)
Keywords Measurement error, Micro-integration

Topic Statistical theory for register-based statistics and data integration
Title of literature review/file: LR2_9
"Topics of statistical theory for register-based statistics and data integration"
Paper reviewed Zhang (2012)
Keywords All kinds of errors, All kinds of data, Framework

4.3 Basic data configuration 3: overlapping microdata sources with under-coverage

4.3.1 Literature review

Topic Uncertainty of population size estimates
Title of literature review/file: LR3_1
"Capture-recapture method and log-linear models to estimate register undercoverage"
Papers reviewed

- Gerritse, Van der Heijden and Bakker (2015)
- Van der Heijden et al. (2012)

Keywords Under-coverage errors, Missing data, Linkage errors, Population size estimation, Capture-recapture

Topic Population size estimates
Title of literature review/file: LR3_2
"Domain estimates of the population size"
Papers reviewed Di Consiglio and Tuoto (2015)
Keywords Under-coverage errors, Population size estimation, Capture-recapture

4.4 Basic data configuration 4: microdata and macrodata

4.4.1 Scalar uncertainty measures

Title of suitability test/file: ST45_1

"Uncertainty measures for economic accounts" (Mushkudiani, Pannekoek and Zhang 2017)

Data configuration

Mushkudiani, Pannekoek and Zhang (2017) have examined basic data configuration 4. Many macro-economic figures, for instance arising in the context of national economic and social accounting systems, are connected by known constraints. The actual input estimates, often based on a variety of sources, usually do not automatically satisfy the constraints due to measurement and sampling errors. The estimation of an accounting equation then involves an adjustment or reconciliation step by which the input estimates are modified to conform to the known identity. In Mushkudiani, Pannekoek and Zhang (2017) an accounting equation is considered as a single entity and scalar uncertainty measures are defined. These measures capture more directly the adjustment effect as well as the relative contribution of the various input estimates to the final estimated account.

Keywords Numerical data, Reconciliation

Data used

Simulated data and an empirical data set obtained from the quarterly and annual Dutch Supply and Use (SU) tables were used for the suitability test. These are quarterly time series of 12 quarters of over 2 thousand time series. The yearly values are the benchmarks for these series.

Work carried out

We consider two kinds of account constraints: additive accounting constraints and multiplicative accounting constraints. An additive accounting constraint is given by

$$Y_1 + \cdots + Y_p = Z$$

where the Y_i $(i = 1, \ldots, p)$ are component variables summing up to a total Z. A multiplicative accounting constraint is given by

$$Y_1 Y_2 = Z$$

When an accounting constraint does not for the input data $Y_1, \ldots, Y_p, Z$, the input data are adjusted so the reconciled data do satisfy the accounting constraint.

Two approaches for defining scalar measures for these accounting constraints have been developed: the covariance approach and the deviation approach. For each approach two variants of scalar measures are defined.

Covariance approach

We consider the case of an additive account. In practice we observe the initial input estimates $(\hat{Y}_1, \ldots, \hat{Y}_p, \hat{Z})$ of the true values $(Y_1, \ldots, Y_p, Z)$. The true values and the estimated reconciled values $(\tilde{Y}_1, \ldots, \tilde{Y}_p, \tilde{Z})$ satisfy the account constraint. For notational convenience we combine the $\tilde{Y}_i$ and $\tilde{Z}$ in one vector $\tilde{X}$:

$$\tilde{X} = (\tilde{Y}_1, \ldots, \tilde{Y}_p, \tilde{Z})$$

Due to the accounting constraint the variance-covariance matrix $\Sigma_{\tilde{X}}$ has the following formal structure

$$\Sigma_{\tilde{X}} = \begin{pmatrix} \Sigma_{\tilde{Y}} & \Sigma_{\widetilde{YZ}} \\ \Sigma_{\tilde{Z}\tilde{Y}} & \Sigma_{\tilde{Z}} \end{pmatrix} = \begin{pmatrix} \Sigma_{\tilde{Y}} & \Sigma_{\tilde{Y}}\mathbf{1} \\ \mathbf{1}^t\Sigma_{\tilde{Z}\tilde{Y}} & \mathbf{1}^t\Sigma_{\tilde{Z}}\mathbf{1} \end{pmatrix}$$

where the superscript t denotes the transpose.

Scalar measures for the covariance approach are defined as follows:

$$\tau_1 = \mathbf{1}^t\Sigma_{\tilde{X}}\mathbf{1} = 4(\mathbf{1}^t\Sigma_{\tilde{Y}}\mathbf{1})$$

and

$$\tau_2 = \mathrm{Trace}(\Sigma_{\tilde{X}}) = \mathrm{Trace}(\Sigma_{\tilde{Y}}) + \mathbf{1}^t\Sigma_{\tilde{Y}}\mathbf{1}$$

Deviation approach

Again, we consider the case of an additive account. Let $M_X = (M_1, \ldots, M_p, M_Z)$ be the expectation of $\tilde{X}$ under a posited model. $\tilde{X} - M_X$ then contains the deviation of all the components of the final estimated reconciled account from those of the expected account. Two suitable scalar measures to summarize the component-wise deviation are

$$\delta_1 = \sum_{k=1}^{p} w_k|\tilde{X}_k - M_k|$$

and

$$\delta_2 = \sum_{k=1}^{p} w_k(\tilde{X}_k - M_k)^2$$

where $w_k \geq 0$ are certain user-specified weights.

Mushkudiani, Pannekoek and Zhang (2017) also specified scalar quality measures for multiplicative accounts. The proposed scalar measures have been tested on simulated data and on empirical time series data.

Cost-benefit analysis

The developed method is easy to apply in practice.

The method summarizes the quality of an accounting constraint, or a set of such constraints, in one scalar. In some cases, one scalar is not enough to describe the quality in sufficient detail.

Gap analysis

Further research is needed to understand the properties of the proposed scalar quality measures.

4.4.2 Literature review

Topic Variance estimation after application of repeated weighting

Title of literature review/file: LR4_1

"Effect of reconciliation on estimated tables"

Papers reviewed

- Houbiers et al. (2003)
- Knottnerus and Van Duin (2006)

Keywords Reconciliation, Sampling error, Repeated weighting, Cross-sectional data

Topic Quality measurement using the "confidence image"
Title of literature review/file: LR4_2
"Constructing confidence images based on multiple sources"
Papers reviewed

- Laitila (2013)
- Laitila (2014)

Keywords Reconciliation, Measurement error, Sampling errors, Confidence image

4.5 Basic data configuration 5: only macrodata

4.5.1 Scalar uncertainty measures

The approach developed and tested for basic data configuration 4 can also be applied to basic data configuration 5 (see Section 4.4.1).

4.5.2 Literature review

Topic Estimation of variance of reconciled totals
Title of literature review/file: LR5_1
"Effect of reconciliation on estimated totals"
Paper reviewed Boonstra, De Blois and Linders (2011)
Keywords Reconciliation, Macrodata, Sampling errors

Topic Estimation of variance of reconciled totals
Title of literature review/file: LR 5_2
"Effect of reconciliation on estimated totals"
Paper reviewed Knottnerus (2016)
Keywords Reconciliation, Macrodata, Sampling errors

Topic Estimation of mean squared error for small area estimates
Title of literature review/file: LR5_3
"Area-level small area estimation methods for domain statistics"
Paper reviewed Boonstra et al. (2008)
Keywords Sampling errors, Measurement errors, Small area estimation

Topic Balancing the national accounts
Title of literature review: LR5_4
"Automatic balancing using the "SCM method" with application to e.g. national accounts"
Paper reviewed Chen (2012)
Keywords Reconciliation, Measurement error, Sampling errors

4.6 Basic data configuration 6: longitudinal data

4.6.1 Literature review

Topic Quality of benchmarked data

Title of literature review/file: LR6_1

"Macro Integration: Data Reconciliation"

Papers reviewed

- Denton (1971)
- Bikker, Daalmans and Mushudiani (2011)

Keywords Reconciliation, Benchmarking, Sampling errors, Measurement errors

5. Action plan and roadmap

The ultimate aim of the ESSnet is to produce usable quality guidelines for NSIs that are specific enough to be used in statistical production at those NSIs. The guidelines are expected to cover the entire production chain (input, process, output). The guidelines aim to take the diversity of situations in which NSIs work and the restrictions on data availability into account. The quality of the final output will depend both on the existing data sources and on the use and processing of the data.

It is clear that general decision rules and single thresholds will not suffice. Instead the guidelines will list a variety of potential indicators/measures, indicate for each of them its applicability and in what situation it is preferred or not and provide an ample set of examples of specific cases and decision making processes. For this reason we have identified several basic data configurations for the use of administrative data sources in combination with other sources, for which we proposed, revised and tested some measures for the accuracy of the output in the first Specific Grant Agreement (SGA 1).

As possible future work in the ESSnet we propose to continue this work on indicators/measures, by developing further quality indicators/measures related to process and output needed for the use in practice of the guidelines. The areas of work follow from the gaps that are identified in SGA 1.

We have identified two major gaps with respect to process and output quality measures that need to be bridged before we are able to complete the set of guidelines. These major gaps are discussed below.

5.1 Making indicators more suitable for application in practice

The first major gap is making the quality measures and methods to compute them that are proposed in SGA 1 more suitable for application in practice. This entails two different aspects. First, it entails adaptation of some of the proposed quality measures and, in particular, the methods to compute them. The currently proposed measures and methods are sometimes quite complicated and hard to apply in practice. We therefore propose to seek ways to obtain measures that are easier to compute in practice, using as much as possible data that are readily available during regular statistical production. Second, it entails extending the range of situations in which the quality measures and methods to compute them can be applied. Some of these measures and methods can be applied only if certain specific conditions are met. As possible future work in the ESSnet we propose to relax these conditions and make the quality measures and methods to compute them more broadly applicable to account for the diversity of situations that may occur at NSIs. We propose to carry out this work by means of a critical literature review and suitability tests.

5.2 Coherence

The second major gap is the “coherence” (and comparability) dimension of quality. The quality measures examined in SGA 1 all focus on the “accuracy” dimension of quality. The “coherence” dimension is highly relevant for multisource statistics as well. As possible future work in the ESSnet we propose to examine this quality dimension.

According to principle 14 “coherence and comparability” of the European Statistics Code of Practice statistics should be consistent internally, over time and comparable between regions and countries; it should be possible to combine and make joint use of related data from different sources. Several indicators have been proposed, such as:

- Statistics should be internally coherent and consistent (i.e. arithmetic and accounting identities should be fulfilled).
- Statistics should be comparable over a reasonable period of time. This can be measured by the frequency of changes to standards, classifications and coding of variables. A trade-off between relevance and comparability over time is also considered.
- Statistics should be compiled on the basis of common standards with respect to scope, definitions, units and classifications in the different surveys and sources.
- Statistics from different sources and of different periodicity should be compared and reconciled.

These indicators were originally defined having single source statistics in mind. As possible future work in the ESSnet we propose to carry out a similar literature review and suitability tests as we have carried out for the accuracy dimension in order to examine to what extent such indicators can also be used for multisource statistics and to what extent new indicators for multisource statistics are required.

5.3 Other suggestions

At the Workshop organized by *Komuso* in Budapest in March 2016 it was also suggested to extend the number of basic data configurations examined in the ESSnet, for instance to consider the situation where longitudinal microdata are produced. This would, however, water down the focus of the Komuso project. We therefore propose not to examine more basic data configurations, unless that is deemed necessary for the two major gaps mentioned above or if there is strong evidence that a basic data configuration that is highly important for NSIs is missing and should be examined.

For the somewhat further future research we mention the effect of timeliness (or the lack thereof) of administrative data on quality measures. In particular, the progressiveness of administrative data, i.e. the fact that administrative data sources generally contain more and/or higher quality data as time passes, deserves more attention than it has received so far.

5.4 Summary of identified gaps

In Table 2 below we summarize the gaps that are identified thus far. These gaps make the possible future work suggested in Sections 5.1 and 5.2 more concrete. The numbers in the column “Examples” refer to the gap analysis in the corresponding sections.

5.5 Roadmap

The roadmap – when will what aspect of the action plan be carried out? – is dependent on decisions with respect to a possible SGA 2 (and a possible later SGA 3) of the ESSnet, and therefore cannot be described at this moment. If a possible SGA 2 is granted, a roadmap for the work envisaged for that SGA will be drawn up as soon as possible.

Table 2. Summary of identified gaps

Conf.	Gap	Aim	Examples
Dimension: Accuracy			
1	Using of information from the editing process (and other internal processes)	Simplify and reduce costs	4.1.2, 4.2.4 and 4.2.6
1	Considering more types of errors; Separating different kinds of errors	Extend range of application	4.1.2 and 4.2.8
2	Developing simple and reliable methods to calculate bias and variance; Developing and simplifying analytical formulas to calculate bias and variance	Simplify and extend range of application	4.2.1 and 4.2.6
2	Using longitudinal information	Extend range of application	4.2.2
2	Improving comparison with other data sources	Extend range of application	4.2.3
2	Taking errors in auxiliary data into account	Extend range of application	4.2.7
2	Measuring quality of relations with variables becoming available later	Extend range of application	4.2.7
4/5	Developing uncertainty measures for complex systems	Extend range of application	4.4.1
4/5	Improving the understanding of the practical and theoretical properties of developed methods	Simplify and extend range of application	4.4.1
Dimension: Coherence			
1/2/3 4/5/6	Developing measures for internal coherence and consistency	Develop	-
1/2/3 4/5/6	Developing measures for comparability over time	Develop	-
1/2/3 4/5/6	Developing measures with respect to the use of common standards	Develop	-
1/2/3 4/5/6	Developing measures with respect to the comparison and reconciliation of data from different sources/periodicity	Develop	-

References

Asamer, E., F. Astleithner, P. Ćetković, S. Humer, M. Lenk, M. Moser and H. Rechta (2016): Quality Assessment for Register-based Statistics - Results for the Austrian Census 2011. *Austrian Journal of Statistics 45*, pp. 3-14. http://www.ajs.or.at/index.php/ajs/article/view/vol45-2-1

Asamer E., H. Rechta H. and C. Waldner (2016), *Quality Indicators for the Individual level – Potential for the Assessment of Subgroups*. Conference contribution to the European Conference on Quality in Official Statistics, Madrid, 31 May-3 June 2016.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta and E. Schwerer (2010), A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011. *Austrian Journal of Statistics 39*, pp. 299-308. Retrieved April 29, 2016 from http://www.stat.tugraz.at/AJS/ausg104/104Berka.pdf

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

Bikker, R., J. Daalmans and N. Mushkudiani (2013). Benchmarking Large Accounting Frameworks: A Generalized Multivariate Model. *Economic Systems Research 25*, pp. 290-408.

Boeschoten, L., D. Oberski and T. de Waal (2016), *Latent Class Multiple Imputation for Multiply Observed Variables in a Combined Dataset*, Discussion paper, Statistics Netherlands..

Boonstra, H.J.C., C. De Blois and G. J. Linders (2011), Macro-integration with Inequality Constraints: An Application to the Integration of Transport and Trade Statistics. *Statistica Neerlandica 65*, pp. 1-25.

Boonstra, H.J., van den Brakel, J., Buelens, B., Krieg, S. and M. Smeets (2008), *Towards Small Area Estimation at Statistics Netherlands*. *Metron* 66, pp. 21-49.

Chambers, R., J. Chipperfield, W. Davis and M. Kovacevic (2009), *Inference Based on Estimating Equations and Probability-Linked Data*. Centre for Statistical & Survey Methodology Working Paper Series.

Chen, B. (2012), A Balanced System of U. S. Industry Accounts and Distribution of Aggregate Statistical Discrepancy by Industry. Journal of Business and Economic Statistics 30, pp. 202-211.

Chipperfield, J.O. and R.L. Chambers (2015), Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data. *Journal of Official Statistics 31*, pp. 397–414.

Denton, F.T. (1971), Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization. *Journal of the American Statistical Association 66*, pp. 99-102.

Di Consiglio L., and T. Tuoto (2015), Coverage Evaluation on Probabilistically Linked data, *Journal of Official Statistics 31*. pp. 415–429.

Fasulo, A. and F. Solari (2017), *ST2_2 Overlapping Numerical Variables with a Benchmark*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Filipponi, D. and U. Guarnera (2017), *ST2_1 Overlapping Numerical Variables without a Benchmark: Integration of Administrative Sources and SurveyDdata through Hidden Markov Models for the Production of Labour Statistics*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Fosen (2012), *Register-Based Employment Statistics. Micro-Integration and Quality-Perspective Life-Cycle. A Case Study*. In: Essnet Data Integration report on WP4 Case Studies.

Fosen, J. (2017), *ST2_7 Output Quality for Statistics Based on Several Administrative Sources: The Norwegian Register-Based Employment Statistics and the Effect of Delays*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Fosen, J. and L.-C. Zhang (2011), *Quality Assessment of Register-Based Census Employment Status.* Proceedings of the International Statistical Institute, World Congress, Dublin.

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

Guarnera U. and R. Varriale (2016), Estimation from Contaminated Multi-Source Data Based on Latent Class Models. *Statistical Journal of the IAOS 32*, pp. 537–544.

Houbiers, M., P. Knottnerus, A.H. Kroese, R. Renssen, R.H and V. Snijders (2003), Estimating Consistent Table Sets: Position Paper on Repeated Weighting. Discussion Paper, Statistics Netherlands.

Kim J. and J.N.K. Rao (2012), Combining Data from two Independent Surveys: A Model-Assisted Approach. *Biometrika 99*, pp. 85–100.

Knottnerus, P. (2016), On New Variance Approximations for Linear Models with Inequality Constraints. *Statistica Neerlandica 70*, pp. 26-46.

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

Krapavickaitė D., and M. Šličkutė-Šeštokienė (2017), *ST2_5 Effect of the Frame Under-Coverage / Over-Coverage on the Estimator of Total and its Accuracy Measures in the Business Statistics*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Krapavickaitė D. and V. Vasilytė (2017a), *ST2_4 Effect of the Under-Coverage of the Classification Variable on the Domain Estimates of the Total in Social Statistics*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Krapavickaitė D. and V. Vasilytė (2017b), *ST2_6 Effect of the Change of the Kind of Activity and Joining of the Enterprises on the Estimator of Total*. Statistics Lithuania, Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Laitila, T. (2013), *Communicating Accuracy of Register Statistics*. Paper prepared for Nordiskt Statistikermöte, Bergen 2013. URL: https://tinyurl.com/n6wls6r (retrieved on April 2, 2017).

Laitila, T. (2014), *Constructing Confidence Intervals based on Register Statistics*. Paper prepared for European Conference on Quality in Official Statistics, Vienna 2014. URL: https://tinyurl.com/me4l7xb (retrieved on April 2, 2017).

Marck, P.S. (2014), *Quality Assessment Tool for Administrative Data*. U.S. Census Bureau, Research and Methodology Directorate, Quality Program Staff, Conference contribution to the European Conference on Quality in Official Statistics, Vienna, 2-5 June 2014

Mushkudiani N., J. Pannekoek and L-C. Zhang (2017), *Uncertainty measurement for economic accounts*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Pavlopoulos, D. and J. K. Vermunt (2015). Measuring Temporary Employment. Do Survey or Register Data Tell the Truth? *Survey Methodology 41*, pp. 197-214.

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

Scholtus, S., B.F.M. Bakker and A. van Delden (2015), *Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables*. Discussion Paper 2015-17, Statistics Netherlands, The Hague. Available at: www.cbs.nl.

Scholtus, S., A. van Delden and J. Burger (2017). *ST1_2 Analytical Expressions for the Accuracy of Growth Rates as Affected by Classification Errors*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Schnetzer, M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk, and M. Moser (2015), Quality Assessment of Imputations in Administrative Data. *Journal of Official Statistics 31*, pp. 231–247, http://dx.doi.org/10.1515/JOS-2015-0015

Statistics Austria (2014), *Quality Assessment of Administrative data - Documentation of Methods*. Provided via http://www.statistik.at/web_de/static/documentation_of_methods_077211.pdf

Statistics Austria (2017), *ST2_3 Suitability Test from WP3 of Komuso*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Statistics Denmark (2017), *ST1_1 Suitability Test of Employment Rate for Employees (Wage Labour Force) (ERWLF)*. Report on a suitability test for WP3 of SGA 1 of the ESSnet Quality of Multi-Source Statistics.

Steeg Morris , D., *A Comparison of Methodologies for Classification of Administrative Records*. JSM 2014 - Survey Research Methods Section, pp. 1729-1743. http://ww2.amstat.org/sections/SRMS/Proceedings/y2014/files/311864_88281.pdf

Van Delden, A., S. Scholtus, and J. Burger (2015), Quantifying the Effect of Classification Errors on the Accuracy of Mixed-Source Statistics. Discussion Paper 2015-10. Available at : https://www.researchgate.net/publication /281450992_Quantifying_the_effect_of_classification_errors_on_ the_accuracy_of_mixed-source_statistics.

Van Delden, A., S. Scholtus and J. Burger (2016a). *Exploring the Effect of Time-Related Classification Errors on the Accuracy of Growth Rates in Business Statistics*. Paper presented at the ICES V Conference, 21–24 June 2016, Geneva.

Van Delden, A., S. Scholtus and J. Burger (2016b). Accuracy of Mixed-Source statistics as Affected by Classification Errors. *Journal of Official Statistics 32*, pp. 619–642.

Van der Heijden, P.G.M., J. Whittaker, M.J.L.F. Cruyff, B.F.M. Bakker and H.N. Van der Vliet (2012), People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Estimates and the Role of Active and Passive Covariates. *The Annals of Applied Statistics 6*, pp. 831–852.

Zhang, L.-C. (2012), Topics of Statistical Theory for Register-Based Statistics and Data Integration. *Statistica Neerlandica 66*, pp. 41–63.

Appendix: Suitability tests

This appendix gives an overview of the suitability tests that have been carried out.

Conf./ (Section) / (Title)	Partner	Error	Method	Example	Quality measure
1 / (4.1.1) / (ST1_1)	Stat. Denmark	Classification errors in business register	Bootstrap	Classification errors affect total number of wage earners (employees) per NACE code using administrative data	Bias, variance
1 / (4.1.2) / (ST1_2)	Stat. Nether-lands	Classification error in business register	Bootstrap and analytical formulae	Quarterly turnover growth rates based on register and survey data	Bias, variance, indirect measure
2 / (4.2.1) / (ST2_1)	ISTAT	Classification error in target variable	Latent Class Analysis	Employment status derived from Labour Force survey data and administrative data	Bias, variance
2 / (4.2.2) / (ST2_2)	ISTAT	Classification error in target variable	Model-assisted estimator / synthetic estimator (analytical)	Employment status from LFS with administrative data as auxiliary variables	Bias, variance
2 / (4.2.3) / (ST2_3)	Stat. Austria	Classification error and imputation error in target variables	Dempster-Shafer method	Legal family status based on different administrative data sets	Qualitative indicator of quality
2 / (4.2.4) / (ST2_4)	Stat. Lithuania	Missingness in a classification (target) variable	Analytical formulas	Missingness in profession class based on census data and administrative data	Bias, variance
2 / (4.2.5) / (ST2_5)	Stat. Lithuania	Delayed frame population (unit errors)	Analytical formulas	Monthly gross earnings (survey) affected by using a frozen population	Bias, variance
2 / (4.2.6) / (ST2_6)	Stat. Lithuania	Classification errors in target variable (delayed correct information)	Analytical formulas	Sample survey stratified by e.g. NACE code, suffers from NACE code errors.	Bias, variance of sample estimator adjusted for NACE errors
2 / (4.2.7) / (ST2_7)	Stat. Norway	Delayed reporting of changes in register data	Methods for specific processing steps that include uncertainty estimation	Effect of measurement time on accuracy of employment status based on register data	Bias, variance
2 / (4.2.8) / (Boe-schoten Oberski De_Waal)	Stat. Nether-lands	Measurement error in target variable combined survey-register data	Latent Class Analysis	Home-ownership status	Accuracy (confidence intervals)
4+5 / (4.4.1) / (ST45_1)	Stat. Netherlan ds	Inconsistent macro data (aggregates)	Method proposed by Mushkudiani, Pannekoek and Zhang	Quarterly and annual supply and use tables	scalar uncertainty measure