ST2_3 Misclassification in several administrative sources

Statistics Austria

Data configuration

Our quality framework is designed for situations of data configuration 2, i.e. it uses overlapping microdata to compute a quality indicator for the output variable. It was developed and used for the Austrian register-based census 2011. This was a full enumeration from several administrative sources.

For the framework, each source has to deliver on microdata level. For our test, we assume, since we consider the central population register (CPR) as the population basis, that we have no under-coverage. Apart from CPR, the data sources in our test deliver only parts of the target population resp. covers only parts of the target variables, e.g. the tax register contains just the subpopulation of taxable persons and further the variable Place of Birth is not contained in the tax register. So, although the CPR covers the whole population, other registers contains only some but overlapping parts of the population and therefore we have configuration 2. In addition we have configuration 2 also in the sense that some of the attributes (variables) are overlapping, i.e. for one unit there are two (or more) data sources for the same attribute. However, the presented quality framework should be also applicable to situations of configuration 1, 2S and 3.

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, with an example in the appendix. The assessment of imputations for a register-based statistic is discussed in Schnetzer et al. 2015. The procedure allows to track changes in quality during data processing. This is drawn in Asamer et al. 2016a. 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 in general. Furthermore, potentials for analysis a register-based statistic by the framework results are demonstrated in Asamer et al. 2016b.

Type of error and aim of the test

Type: Classification errors

We distinguish between two cases of errors in the output data.

1) The output value comes from a data source. There are misclassifications in all data sources and so (to a lower extent) also in the final output data.
   Via various quantitative indicators for input quality and an additional comparison with external data (e.g. LFS), we derive an output quality indicator.
2) The output value was imputed (i.e. it was not delivered by any data source). When missing values in the final output data are estimated, then a misclassification caused by imputation will occur in general. Such classification errors are computed by the model classification rate and the quality of the predictors (as computed previously). This yields the output quality indicator for the imputation situation.

Aim: Screen and describe all properties of the quality framework, using as an example the recently finished register-based labour market statistics 2014 (RBLMS).

Data used

The actual data processing for the RBLMS is conducted in three stages that have to be considered in the quality assessment: the raw data (i.e. the administrative register i), the combined dataset of these raw data (Central Data Base, CDB) and the final dataset, which includes imputations (Final Data Pool, FDP). Figure 1 illustrates the data processing, beginning with the delivery of raw data from the various administrative data holders. The four hyperdimensions (HD^D, HD^P, HD^E, HD^I) aim to assess the quality for different types of attributes at all stages (Documentation, pre-processing, external source, and Imputation) of the data processing. This yields quality indicators which 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.

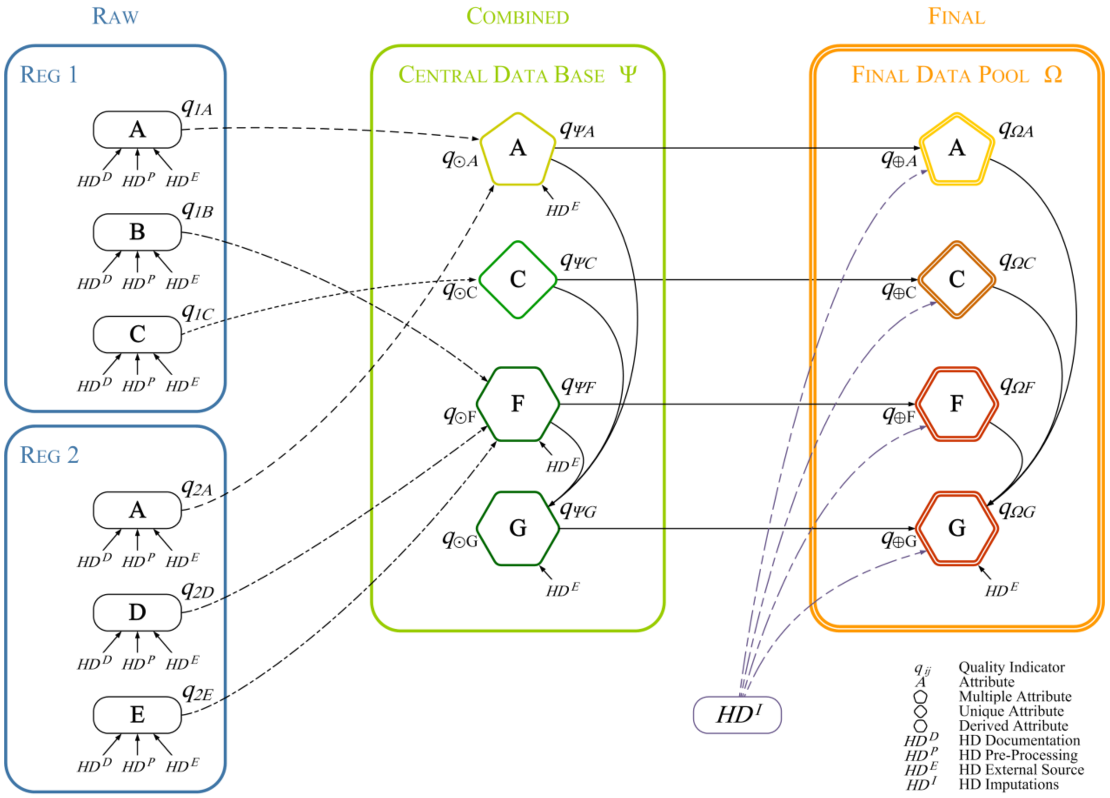


Figure 1: Schematic overview of the quality framework for register based statistics

Raw data level:

Quality information at the raw data level (see left boxes in Figure1) is obtained via three hyperdimensions: Documentation HD_{ij}^D, Preprocessing$HD_{ij}^P$and External SourceHD_{ij}^E for every attribute j in each administrative register i. The score for HD_{ij}^D comes from scored questions in a questionnaire assigned by the data holders (see table1). So HD_{ij}^D describes quality-related processes as well as the data documentation (metadata) at the administrative authorities. The scored questions can be divided into four subgroups, covering data historiography, definitions, administrative purpose and data treatment as it is illustrated by the following table 1.

Table 1: Scored questions -HD Documentation

Data Historiography	Level of measurement
Can we detect data changes over time?	dichotomous
Is the information available for the cut-off date? (reference date)	dichotomous
Definitions	
Are the data definitions for the attribute compatible to those of Statistics Austria?	dichotomous
Administrative Purpose	
Is the attribute relevant for the data source keeper?	dichotomous
Does a legal basis for the attribute exist?	dichotomous
Data Treatment	
How fast are changes edited in the register?	ordinal
Are the data verified on entry?	dichotomous
Are technical input checks applied?	dichotomous
How good is the data management, i.e. ex post consistency checks?	ordinal

For a description of HD_{ij}^{P} and HD_{ij}^{E} see table 2. Given these three quality indicators, an overall quality indicator q_{ij} for each attribute j on the register i level can be derived as the weighted average. The indicator q_{ij} is of course an indicator for input quality, but it is needed for the further calculations of the output quality.

CDB-level:

We distinguish between three types of attributes to compute for each statistical unit n the quality $q_{\Psi j}^{n}$ on the CDB-level:

Unique attributes, i.e. the attribute j exists in exactly one register[1] i; let $x_{ij}(n)$ be the value of the unit n for the attribute j in the register i.

$$q_{\Psi j}^{n} = \begin{cases} q_{ij}, \text{if } x_{ij}(n) \text{ is a vaild value} \\ 0, \text{if } x_{ij}(n) \text{ is not vaild, but it should}^{2} \\ \text{omitted, if } n \text{ is justifiably}^{3} \text{ not delivered by } j \end{cases}$$

Multiple attributes, i.e. j show up in several registers $i_1, \dots, i_s$. To combine different quality indicators $q_{i_1 j}, \dots, q_{i_s j}$, they are interpreted as beliefs in the degree of correctness of each data source. In such a setting, the Dempster-Shafer theory combines the existing evidence and takes all available information from the registers into account to form one quality-indicator on the CDB level, denoted $q_{\odot j}^{n}$ for each statistical unit n (see Berka et al. (2012) for a detailed description).

Derived attributes, i.e. j is derived from different attributes. This may happen if the attribute is not available in any register, but can create from several variables in several registers. For example the

[1] Note that j has not necessary be precisely on the same unit-level in the source and in the statistic. E.g. for the highest level of education, which is an example of an unique attribute, the units are persons, whereas the register of educational attainment is maintained on the level of educational attainment.

[2] For almost every attribute a person needs a valid value (age, sex, …) .

[3] E.g. in Austria, the school education starts for children at the age of 6 years. Hence it is justified that persons of age 5 years and lower are not recorded in the register of enrolled pupils & students.

attributes sex, marital status and age are used to create the attribute household status. Then $q^{n}_{\odot j}$ is the weighted average of the qualities of the input attributes regarding n. The weights are not static in the sense that they are overall weights. More precisely they depend on the ruleset, i.e. on n.

In an optional further step, the values in the CDB are compared to the external source $HD^{E}_{\mathrm{CDB}j}$ to address possible quality issues in the process of combining the raw data to the CDB (cf. Asamer et al. (2016a)). The weighted average yields the final quality indicator $q^{n}_{\Psi j}$ on CDB-level. If this option is omitted, then $q^{n}_{\Psi j} \coloneqq q^{n}_{\odot j} \; \forall n$.

FDP-level:

For the FDP the imputations have to be assessed. Based on a classification rate Φ^{m} (see Schnetzer et al. (2015)) of the imputation model m as well as the final quality $q^{n}_{\Omega j_k}$ of the predictors $j_1, \dots, j_s$, Schnetzer et al. compute a quality indicator $HD_j^{I,m,n}$ by $HD_j^{I,m,n} = \frac{\Phi^{m}}{s} \cdot \sum_{k=1}^{s} q^{n}_{\Omega j_k}$.

The missing values on CDB level, which were assumed to have a quality of zero, are imputed on the FDP level and obtain as quality $HD_j^{I,m,n}$, i.e. $q^{n}_{\oplus \mathrm{j}} = \begin{cases} q^{n}_{\Psi j}, \text{if } j \text{ is not imputed} \\ HD_j^{I,m,n}, \text{if } j \text{ is imputed} \end{cases}$. The number Φ^{m} is computed by applying the imputation model m to already existing complete data and compare the results of the imputation process with the true values of these observations. The classification rate is one fixed number for the model m. The computation depends on the scale of measure (nominal resp. ordinal). We will apply the framework on the variable legal family status LMS of the RBLMS 2014, which was created from ten source registers:

- unemployment register (UR)
- register of public servants of the federal state and the laender (RPS)
- family allowance register (FAR)
- central social security register (CSSR)
- central register of foreigner (CRF)
- chambers register (CHR)
- hospital for public servants register (HPSR)
- register of social welfare recipients (SWR)
- tax register (TR)
- central population register (CPR)

This attribute (legal family status) is an optimal example, since it has not too much different values and the quality indicators differ between the process. The same attribute was used to track changes in quality during data processing in the register-based census 2011 (Asamer et al. 2016a) in much more detail. So in some parts this is an abridged version of a similar investigation. The quality indicators (see table 2[4]) computed throughout the process shows different quality aspect. These are important additional information to "understand" the final quality indicator and how it can be (possible) improved (see Asamer et al. 2016a or Statistics Austria 2014 for a detailed description of these indicators).

[4] Recall, that the quality indicators are standardized between zero and one, where a higher value indicates better data quality.

Table 2: Different quality indicators in the framework from raw data *i* to combined data to final data for the attribute *j* of the unit *n*

	Step	Notation	Brief description	Meaning for the attribute	Formula	Average quality of $j = LMS$
RAW	1	HD_{ij}^{D}	describes quality-relevant processes at the register authority (via questionnaire to the data holders , cf. Statistics Austria 2014)	assess the <u>suitability</u> for the statistical purpose	$\frac{\text{obtained (weighted)score}}{\text{achievable (weighted)score}}$	See table 3
	2	HD_{ij}^{P}	HDP deals with formal errors in the data, i.e. coverage rate of the variable (inside the source i)	indicates the <u>usability</u>	$\frac{\text{usable records}}{\text{total number of records}}$	See table 3
	3	HD_{ij}^{E}	data quality in comparing with an (error-free) external data source (e.g. microcensus LFS)	indicates the <u>accuracy</u> of the raw data	$\frac{\text{number of consistent values}}{\text{total number of linked records}}$	See table 3
	4	q_{ij}	weighted average of HD^{D}, HD^{P}, HD^{E}	overall quality in the register	$v^{D} HD_{ij}^{D} + v^{P} HD_{ij}^{P} + v^{E} HD_{ij}^{E}.$	See table 3
CDB	5	$q_{\odot j}$	unique attribute: j exists in exactly one register i	quality for the final values for each delivered unit	$q_{\Psi j}^{n} = \begin{cases} q_{ij} \\ 0 \\ \text{omitted} \end{cases}$ (see above)	-
			multiple attribute: If the attribute is delivered from multiple sources, the quality q is combined (Dempster-Shafer)		cf. Berka et al. 2012 or Statistics Austria 2014	0.937
			derived attribute: j is derived from different attributes		average of the qualities of the input attributes regarding n	-
	6	$HD_{\mathrm{CDB}j}^{E}$	(optional) comparing the final value with an (error-free) external data source	indicates the <u>accuracy</u> of the predefined ruleset	$\frac{\text{number of consistent values}}{\text{total number of linked records}}$	0.979
	7	$q_{\Psi j}$	weighted average of HD_{CDB}^{E} and $q_{\odot}$	quality for each delivered unit	$u_{1} HD_{\mathrm{CDB}}^{E} + u_{2} q_{\odot}$	0.943
FDP	8	HD_{j}^{I}	model classification rate times average quality of the independent variables	indicates a quality for each imputed value	cf. Schnetzer et al. 2015 or Statistics Austria 2014	0.734
	9	$q_{\oplus j}$	$q_{\odot} \cup HD^{I}$	quality for each unit n	$q_{\oplus \mathrm{j}}^{n} = \begin{cases} q_{\Psi j}^{n}, \text{if } j \text{ is not imputed} \\ HD_{j}^{I,m,n}, \text{if } j \text{ is imputed} \end{cases}$	0.955
	10	$HD_{\mathrm{FDP}j}^{E}$	(optional) comparing the final value with an (error-free) external data source	indicates the <u>accuracy</u> of the predefined ruleset (if Step 6 is not used)	$\frac{\text{number of consistent values}}{\text{total number of linked records}}$	-
	11	$q_{\Omega j}$	weighted average of HD_{FDP}^{E} and $q_{\oplus}$	Final quality for each unit	$w_{1} HD_{\mathrm{FDP}}^{E} + w_{2} q_{\oplus}$	0.955

Table 3: Calculation of the quality indicator for the (LMS) for the registers

Register *i*	$HD_{i\,LMS}^{D}$	$HD_{i\,LMS}^{P}$	$HD_{i\,LMS}^{E}$	$q_{i\,LMS}$
UR	0.683	0.895	0.979	0.852
RPS	0.864	0.960	0.951	0.925
FAR	0.937	0.955	0.962	0.951
CSSR	0.841	0.492	0.938	0.757
CRF	0.444	0.836	0.857	0.713
CHR	0.778	0.512	0.864	0.718
HPSR	0.706	0.697	0.951	0.785
SWR	0.777	0.925	0.934	0.879
TR	0.937	0.861	0.880	0.892
CPR	0.810	0.778	0.969	0.852

Issues (Cost-benefit analysis)

Costs:

- No specific IT-tools are needed (just common used, e.g. SAS, R, SPSS, xls, …)
- No separate audit sampling necessary, but it could be helpful. It is possible to use already existing surveys like the LFS for at least some variable (in the event of the lack of a suitable survey, one can substitute it by a so called expert view).
- Slightly respondence burden for the data holders for metadata information.

Complexity:

- If the output variable is delivered by just one data sources, the computation is simple.
- If the output variable is delivered by more than one data sources, the implementation of the Dempster-Shafer-theory can require much work.
- In practice, simple usable Method (if it is already implemented)

Time:

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 sources and the recyclability of existing code. All numbers are approximations, assuming that the Dempster-Shafer theory is already implemented:

- RAW: each attribute from each source requires 6 hours; further attributes from the same source requires 3 hours (e.g. a multiple attribute from two never assessed sources S1, S2 requires about 6 hours. If S1 and S2 were already been involved in a previous data assessment process this time can be reduced to 3 hours)
- CDB: Unique/ Multiple/ Derived attributes requires 0/ 5/ 5 hours. Using existing code reduces to about 0/ 3/ 4 for further attributes. Derived attributes takes little more time since existing code has to be reworked more often.
- FDP: Unique/ Multiple/ Derived attributes requires 5/ 5/ 5 hours. Using existing code reduces to about 4/ 4/ 4 for further attributes.

Assumptions: The main conditions for an application of the quality framework:

- The data sources should be (administrative) independent by each other; i.e. there is no data exchange between the sources for the relevant attribute (otherwise a stochastic overweighting can occur by using the Dempster-Shafer theory), which means that each register is processed by separate systems (before they enter the NSI. E.g. the register REG1 adopts the attribute *date of birth* directly from REG2. After that, both registers deliver to the NSI. Then one has to consider the attribute *date of birth* as delivered from just one single register or at least for the framework, the sources should be viewed as one single data source.
- Each variable in the final output data which has to be assessed, needs an external data source which contains the variable (i.e. the audit data; e.g. a sample survey like LFS for the employment variables, e.g. a survey of enterprises for the place of work), or a substitute so called expert view.
- The external source can be linked by a unique key variable on unit-level.
- The external source is assumed to be error-free
- The external source has a reference day "very close" to the reference day of the output data.
- There must be "enough" linkable (and hence comparable) units in the audit data
- Different sources can deliver values of different aggregation-level for the same unit. So both sources are true on some level. Respectively, the source value might be higher aggregated than the output-value. This fact should be kept in mind. (e.g. register 1 has the value "former Yugoslavia" and register 2 has the value "Croatia". So both sources contribute to the same value on a higher aggregation level for the variable Place of birth. This respect is important for the correct application of the Dempster-Shafer theory). On the other hand, this is also important for the correct calculation of the indicator HD_{ij}^{E} (e.g. register *i* has the value "former Yugoslavia" and the external source has the value "Croatia".)

Advantages:

- The approach is generic: we expect that it can be applied to other multiple-source statistics of configuration 2 (moreover the framework should be also applicable to situations of configuration 1, 2S and 3).
- Usable for different kinds of administrative data.
- The method has modular design (i.e. it can be used for a variety of purposes.)
- The quality is comparable between different processing stages, data revisions, registers and even single attributes.
- Data processes can be evaluated without influencing them
- Individual observations at every step, so the quality can be tracked for data subsets.

Disadvantages:

- Before one can apply the technique (especially if more than one source delivers the attribute), a certain time and work is required to understand and implement the Dempster-Shafer-theory.

Gaps

- Comparing with an external source (HDE) can be improved. E.g. one can compute a classification rate for each value, instead of an average value for the whole attribute or one may choose a suitable aggregation-level for comparison.

References

Asamer E., Astleithner F., Ćetković P., Humer S., Lenk M., Moser M. and Rechta H. (2016a): Quality Assessment for Register-based Statistics - Results for the Austrian Census 2011. Austrian Journal of Statistics Vol. 45, No. 2, pp. 3-14, retrieved April 10, 2017 from http://www.ajs.or.at/index.php/ajs/article/view/vol45-2-1

Asamer E., Rechta H., Waldner C. (2016b), 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., Humer S., Lenk M., Moser M., Rechta H. and Schwerer E. (2010), *A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011*. Austrian Journal of Statistics, Volume 39, Number 4, 299-308, retrieved April 10, 2017 from http://www.stat.tugraz.at/AJS/ausg104/104Berka.pdf%20

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

Schnetzer M., Astleithner F., Cetkovic P., Humer S., Lenk M. and Moser, M. (2015), *Quality Assessment of Imputations in Administrative Data*, Journal of Official Statistics, Vol. 31, No. 2, pp. 231–247, retrieved April 10, 2017 from 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

ST2_4 Effect of the under-coverage of the classification variable on the domain estimates of the total in social statistics

D. Krapavickaitė and V. Vasilytė, 2017

Data configuration: **2.**

Data used: Combination of
Case A) two administrative data sources
Case B) an administrative data source and survey data set

Statistic of interest: population total and domain totals.

Type of errors: Under-coverage or non-response of the classification variable sitting in the administrative data source

Quality measure: Relative bias and relative increase of variance of the estimator for a domain total due to classification variable under-coverage.

Key words: administrative data source, decomposition of the accuracy measures, bias, variance.

Abstract. One of the problems in nowadays official statistics is estimation of the accuracy of statistical results. This accuracy may be affected by various sources of errors. Administrative data sources are often used together with sample data, and the errors present in administrative data sources influence the accuracy of the estimators obtained. Laitila and Holmberg in [2] presented the decomposition of the mean squared error (MSE) measure into different sources of bias and variance, giving a tool for studying the effects of e.g. measurement errors and frame bias on the precision of estimates. In the current study, 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 or missing values. Population totals of a study variable should be estimated for these domains. The bias and variance of the estimator obtained is not only due to sampling but also due to the under-coverage of the classification variable and methods used to overcome this problem. The solution to this problem is given analytically, and the results of a simulation study are presented.

Work carried out. **Case A)** Data structure:

- administrative data source with a study variable;
- another administrative data source with the classification variable suffering from under-coverage or non-response.

The method uses three basic steps:

1) Simulation algorithm of the missing values for a classification variable.
2) Derivation of the analytical formulas to estimate population total and variance of the estimator.

Step 1) Let us take an administrative data source Q_1 having all values of a study variable and another administrative data source Q_2 having all values of the classification variable A with m different levels a_1, a_2, …,a_m. Let us make $\alpha = 0.1$ (or other) proportion of the values for a classification variable unknown in the following way: choose a random variable ξ with the values in the interval (0,1), find its quantile q_α for $\alpha = 0.1$. For any data source Q_1 unit generate a value of the random variable ξ, and if $\xi < q_\alpha$ then assume A to be unknown, otherwise consider a value of A as known. Distribution of ξ chosen for this study is the uniform distribution on the interval (0,1).

The union $U = Q_1 \cup Q_2$ is considered as a survey population with under-coverage (non-response) for a classification variable A.

Step 2) Let us suppose the values of the classification variable are unknown for the sub-population $U^{(n)}$ of size $N^{(n)}$ and known for the sub-population $U^{(a)}$ of size $N^{(a)}$, $U = U^{(a)} \cup U^{(n)}$, $N = N^{(a)} + N^{(n)}$. Using a variable A the sub-population $U^{(a)}$ is classified into m domains $U^{(a)} = U_1^{(a)} \cup U_2^{(a)} \cup ... \cup U_m^{(a)}$ of size $N_1^{(a)}, N_2^{(a)},..., N_m^{(a)}$, $N^{(a)} = N_1^{(a)} + N_2^{(a)} + ... + N_m^{(a)}$. Let us study a variable y defined on U and denote its domain means for the known sub-population $U^{(a)}$ by

$$\mu_g^{(a)} = \frac{1}{N_g^{(a)}} \sum_{k \in U_g^{(a)}} y_k \,,\; g = 1,2,...,\; m.$$

Classification of $U^{(n)}$. Let us draw an $n^{(n)}$ size simple random reference sample $s^{(n)}$, $s^{(n)} \subset U^{(n)}$, get the values of the classification variable A for the sampled elements and estimate proportions

$$p_g^{(n)} = \frac{N_g^{(n)}}{N^{(n)}} \text{ by } \hat{p}_g^{(n)} = \frac{n_g^{(n)}}{n^{(n)}} \,,\; g = 1,2,...,\; m\;.$$

This sample is used as a kind of audit file in order to estimate the proportions $p_g^{(n)}$, $g = 1,2,...,\; m$.

Here $n_g^{(n)}$ is the number of elements in the sample $s^{(n)}$ with the value a_g of the classification variable A, learned from the sampled elements. Then the domain totals of a study variable y, $t_{yg} = \sum_{k \in U_g} y_k$ are estimated by

$$\hat{t}_{yg}^{(1)} = \sum_{k \in U_g^{(a)}} y_k + N^{(n)} \hat{p}_g^{(n)} \mu_g^{(a)} \,,\; g = 1,2,...,\; m\;; \tag{1}$$

$$Var\,(\hat{t}_{yg}^{(1)}) = N^{(n)2} \mu_g^{(a)2} Var\,(\hat{p}_g^{(n)})\,,\;\; g = 1,2,...,\; m\;; \tag{2}$$

$$V\hat{a}r(\hat{t}_{yg}^{(1)}) = N^{(n)2} \mu_g^{(a)2} V\hat{a}r(\hat{p}_g^{(n)})\;; \tag{3}$$

$$c\hat{v}(\hat{t}_{yg}^{(1)}) = \frac{\sqrt{V\hat{a}r(\hat{t}_{yg}^{(1)})}}{\hat{t}_{yg}^{(1)}}\;.$$

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

Bias of the estimator $\hat{t}_{yg}^{(1)}$ for the total hence mainly depends on the difference $\left(\mu_g^{(a)} - \mu_g^{(n)}\right)$ in domain means for the classified and non-classified population.

Variance (2) and its estimator (3) show the variance of the estimated domain total t_{yg} arising due to under-coverage (or non-response) of the classification variable *A*.

Group sizes N_g are separate cases of the population totals for $y_k \equiv 1$, $k = 1,2,...,\; N$, estimated by

$$\hat{N}_g = N_g^{(a)} + N^{(n)} \hat{p}_g^{(n)},\; Var\,(\hat{N}_g) = N^{(n)2} Var\,(\hat{p}_g^{(n)})\,, \tag{4}$$

$$Var\,(\hat{p}_g^{(n)}) = \left(1 - \frac{n^{(n)}}{N^{(n)}}\right) \frac{N^{(n)}}{N^{(n)} - 1} \frac{p_g^{(n)}\left(1 - p_g^{(n)}\right)}{n^{(n)}}\,, \tag{5}$$

$$V\hat{a}r(\hat{p}_g^{(n)}) = \left(1 - \frac{n^{(n)}}{N^{(n)}}\right) \frac{N^{(n)}}{N^{(n)} - 1} \frac{\hat{p}_g^{(n)}\left(1 - \hat{p}_g^{(n)}\right)}{n^{(n)}}\,. \tag{6}$$

$c\hat{v}(\hat{N}_g) = \dfrac{N^{(n)}\sqrt{V\hat{a}r(\hat{p}_g^{(n)})}}{\hat{N}_g}$ is a measure for the error in the estimate of the domain size N_g due to under-coverage (non-response) of the classification variable.

Case B) Data structure:

- sample data with a study variable

- administrative data source with the classification variable suffering from under-coverage (or non-response)

Simulation of the missing values for a classification variable is carried out in the same way as for Case A.

Derivation of the analytical formulas to estimate population total and variance of the estimator is as follows. Let us suppose a simple random sample $s^{(a)}$ is drawn from the population Q_1 and domain means $\mu_g^{(a)}$ are estimated by $\hat{\mu}_g^{(a)}$ from this sample, $g = 1,2,..., m$. Then instead of (1), t_{yg} are estimated by

$$\hat{t}_{yg}^{(2)} = \hat{t}_{yg}^{(a)} + N^{(n)}\hat{p}_g^{(n)}\hat{\mu}_g^{(a)}, \ g = 1,2,..., m. \qquad (7)$$

Here $\hat{t}_{yg}^{(a)} = N_g^{(a)}\hat{\mu}_g^{(a)} = N_g^{(a)}\dfrac{1}{n_g^{(a)}}\sum_{k\in s^{(a)}\cap U_g^{(a)}} y_k$, $g = 1,2,..., m.$

Expression for bias of $\hat{t}_{yg}^{(2)}$.

$$Bias\left(\hat{t}_{yg}^{(2)}\right) = E\hat{t}_{yg}^{(a)} + N^{(n)}E\hat{p}_{yg}^{(n)}\hat{\mu}_{yg}^{(a)} - t_{yg}^{(a)} - N^{(n)}p_{yg}^{(n)}\mu_{yg}^{(n)} =$$

$$= E\left(\hat{t}_{yg}^{(a)} - t_{yg}^{(a)}\right) + N^{(n)}E(\hat{p}_{yg}^{(n)}\hat{\mu}_{yg}^{(a)} - p_{yg}^{(n)}\hat{\mu}_{yg}^{(a)} + p_{yg}^{(n)}\hat{\mu}_{yg}^{(a)} - p_{yg}^{(n)}\mu_{yg}^{(n)} - p_{yg}^{(n)}\mu_{yg}^{(a)} + p_{yg}^{(n)}\mu_{yg}^{(a)}) =$$

$$= \left(E\hat{t}_{yg}^{(a)} - t_{yg}^{(a)}\right) + N^{(n)}\left(E\left(\hat{p}_{yg}^{(n)} - p_{yg}^{(n)}\right)E\hat{\mu}_{yg}^{(a)} + p_{yg}^{(n)}\left(E\hat{\mu}_{yg}^{(a)} - \mu_{yg}^{(a)}\right) + p_{yg}^{(n)}\left(\mu_{yg}^{(a)} - \mu_{yg}^{(n)}\right)\right).$$

If the estimators $\hat{t}_{yg}^{(a)}$ and $\hat{p}_g^{(n)}$ are unbiased, then the bias of $\hat{t}_{yg}^{(a)}$ depends only on the difference $\mu_{yg}^{(a)} - \mu_{yg}^{(n)}$.

Expression for variance of $\hat{t}_{yg}^{(2)}$.

Let us denote by R the sampling design distribution for reference sample $s^{(n)}$ in the non-classified sub-population Q_2 , and by C the sampling distribution for the classified subpopulation Q_1 . We find expression for $Var\left(\hat{t}_{yg}^{(2)}\right)$ assuming that $\hat{t}_{yg}^{(a)}$ and $\hat{p}_g^{(n)}$ are unbiased.

Using expression for a variance of a sum of two random variables we obtain:

$$Var\left(\hat{t}_{yg}^{(2)}\right) = Var\left(\hat{t}_{yg}^{(a)} + N^{(n)}\hat{\mu}_{yg}^{(a)}\hat{p}_g^{(n)}\right) =$$

$$= Var\ (\hat{t}_{yg}^{(a)}) + N^{(n)2}Var\left(\hat{\mu}_{yg}^{(a)}\hat{p}_g^{(n)}\right) + 2N_g^{(a)}N^{(n)}Cov\left(\hat{\mu}_{yg}^{(a)}, \hat{p}_g^{(n)}\hat{\mu}_{yg}^{(a)}\right) =$$

$$= Var\left(\hat{t}_{yg}^{(a)}\right) + N^{(n)2}I_1 + 2N_g^{(a)}N^{(n)}I_2 . \qquad (7`)$$

The expressions for I_1 and I_2 will be found. The formula for conditional and unconditional variances and expectations is applied:

$$I_1 = Var\left(\hat{\mu}_{yg}^{(a)}\hat{p}_g^{(n)}\right) = Var_R E_C\left(\hat{\mu}_{yg}^{(a)}\hat{p}_g^{(n)} \mid R\right) + E_R Var_C\left(\hat{\mu}_{yg}^{(a)}\hat{p}_g^{(n)} \mid R\right) =$$

$$= Var_R\left(\mu_{yg}^{(a)}\hat{p}_g^{(n)}\right) + E_R\hat{p}_g^{(n)2}Var_C\hat{\mu}_{yg}^{(a)} .$$

Now, using a property of a variance, in the same way as in the proof of Theorem 1 on pages 23 and 24 of Bethlehem, 2008, we get:

$$Var_R\left(\hat{p}_g^{(n)}\right) = E_R\left(\hat{p}_g^{(n)2}\right) - \left(E_R\hat{p}_g^{(n)}\right)^2 = E_R\left(\hat{p}_g^{(n)2}\right) - p_g^{(n)^2} .$$

Hence:

$$I_1 = \mu_{yg}^{(a)2}Var_R\left(\hat{p}_g^{(n)}\right) + Var_C\hat{\mu}_{yg}^{(a)}\left(Var_R\left(\hat{p}_g^{(n)}\right) + p_g^{(n)2}\right).$$

The formula for conditional and unconditional variances and expectations is used once more:

$$I_2 = Cov\left(\hat\mu_{yg}^{(a)}, \hat p_g^{(n)}\hat\mu_{yg}^{(a)}\right)=$$

$$= E_R\left(Cov_C\left(\hat\mu_{yg}^{(a)}, \hat p_g^{(n)}\hat\mu_{yg}^{(a)}\middle|R\right)\right)+Cov_R\left(E_C\left(\hat\mu_{yg}^{(a)}\middle|R\right), E_C\left(\hat\mu_{yg}^{(a)}\hat p_g^{(n)}\middle|R\right)\right)=$$

$$= E_R\left(\hat p_g^{(n)}\right)Var_C\left(\hat\mu_{yg}^{(a)}\right)+Cov_R\left(\mu_{yg}^{(a)}, \hat p_g^{(n)}\mu_{yg}^{(a)}\right)= p_g^{(n)}Var_C\left(\hat\mu_{yg}^{(a)}\right)+0\ .$$

Finally, having in mind, that $Var\left(\hat t_{yg}^{(a)}\right)=Var_C\left(\hat t_{yg}^{(a)}\right)$, and inserting expressions obtained for I_1 and I_2 into (7`), it is obtained:

$$Var\left(\hat t_{yg}^{(2)}\right)=Var_C\left(\hat t_{yg}^{(a)}\right)+$$
$$+N^{(n)2}\mu_{yg}^{(a)2}Var_R\left(\hat p_g^{(n)}\right)+N^{(n)2}Var_C\left(\hat\mu_{yg}^{(a)}\right)\left(Var_R\,\hat p_g^{(n)}+p_g^{(n)2}\right)+2N_g^{(a)}N^{(n)}p_g^{(n)}Var_C\left(\hat\mu_{yg}^{(a)}\right)=$$

$$=Var_C\left(\hat t_{yg}^{(a)}\right)+\left(2N_g^{(a)}N^{(n)}p_g^{(n)}+N^{(n)2}p_g^{(n)2}\right)Var_C\left(\hat\mu_{yg}^{(a)}\right)+N^{(n)2}\left(Var_C\left(\hat\mu_{yg}^{(a)}\right)+\hat\mu_{yg}^{(a)}\right)Var_R\left(\hat p_g^{(n)}\right).$$

Using notation $V_{1g}=\left(2N_g^{(a)}N^{(n)}p_g^{(n)}+N^{(n)2}p_g^{(n)2}\right)Var_C\left(\hat\mu_{yg}^{(a)}\right)$ and
$V_{2g}=N^{(n)2}\left(Var_C\left(\hat\mu_{yg}^{(a)}\right)+\hat\mu_{yg}^{(a)}\right)Var_R\left(\hat p_g^{(n)}\right)$
it follows from the previous expression that

$$Var\left(\hat t_{yg}^{(2)}\right)=Var_C\left(\hat t_{yg}^{(a)}\right)+V_{1g}+V_{2g}\ . \qquad (8)$$

The increase of variance $Var\left(\hat t_{yg}^{(2)}\right)$ due to under-coverage (or non-response) of the classification variable is presented in the term V_{2g} on the right hand of (8). It may be high if the proportion estimates $\hat p_g^{(n)}$ from the reference sample have high variances.

For estimation of $Var\left(\hat t_{yg}^{(2)}\right)$ variances and proportions in (8) should be estimated:

$$V\hat ar\left(\hat t_{yg}^{(2)}\right)=V\hat ar_C\left(\hat t_{yg}^{(a)}\right)+$$
$$+\left(2N_g^{(a)}N^{(n)}\hat p_g^{(n)}+N^{(n)2}\hat p_g^{(n)2}\right)V\hat ar_C\left(\hat\mu_{yg}^{(a)}\right)+N^{(n)2}\left(V\hat ar_C\left(\hat\mu_{yg}^{(a)}\right)+\hat\mu_{yg}^{(a)}\right)V\hat ar_R\left(\hat p_g^{(n)}\right) \qquad (9)$$

or

$$V\hat ar\left(\hat t_{yg}^{(2)}\right)=V\hat ar_C\left(\hat t_{yg}^{(a)}\right)+\hat V_{1g}+\hat V_{2g}\ .$$

Indicators

$$R\hat V_{2g}=\frac{V_{2g}}{Var\left(\hat t_{yg}^{(2)}\right)},\ g=1,2,\ldots,\ m\ ,$$

show relative input of the estimation of the domain size proportion in the non-classified subpopulation to the compound variance $Var\left(\hat t_{yg}^{(2)}\right)$.

4. Simulation study. Let us use for the simulation study the data of the Lithuanian survey on income and living conditions, and take a classification variable “profession” from an administrative data source – a social insurance institution. Profession levels:

1 – leaders,
2 – specialists,
3 – technicians and junior specialists,
4 – officers,
5 – employees of the service sector and salespeople,
6 – qualified employees of agriculture, forest and fishery,
7 – qualified workers and tradespeople,
8 – operators and assemblers for devices and machines,

9 – non-qualified workers.

After merging two administrative data sources, we obtain the population U of size N = 5051. To obtain a non-classified subpopulation, a random variable with the uniform distribution on the interval (0,1) is generated for each population element, and for values of this random variable less than $\alpha = 0.1$, the level of the classification variable is considered to be unknown. Subpopulations of sizes $N^{(a)}$ = 4559, $N^{(n)}$ = 492 are obtained, the number of levels for the classification variable A is m = 9. The study variable y is the self-estimated minimal amount of money (euro) needed for living during a month per household member.

A simple random sample of size n=800 has been selected, the totals of the study variable for sex and profession groups have been estimated by (7), and the estimates of the variances for the estimators of totals are calculated by (9). The procedure is repeated k=100 times, k estimates $\hat{t}_{ygi}^{(2)}$, $i = 1,2,\ldots, k$, of the estimator $\hat{t}_{yg}^{(2)}$ are obtained, and the empirical averages of the estimates

$$\bar{\hat{t}}_{yg}^{(2)} = \frac{1}{k}\sum_{i=1}^{k} \hat{t}_{ygi}^{(2)} \ , \ \overline{V\hat{a}r}\left(\hat{t}_{yg}^{(2)}\right) = \frac{1}{k}\sum_{i=1}^{k} \hat{t}_{ygi}^{(2)} \ , \ g = 1,2,\ldots, m \ ,$$

are calculated. The relative empirical biases of total/mean estimates and relative increase in variance due to the missing values of the classification variable are given by

$$RBias\ \left(\hat{t}_{yg}^{(2)}\right) = \frac{\bar{\hat{t}}_{yg}^{(2)} - t_{yg}}{t_{yg}} = \frac{\bar{\hat{\mu}}_{yg}^{(2)} - \mu_{yg}}{\mu_{yg}} \ , \ \bar{\hat{\mu}}_{yg}^{(2)} = \bar{\hat{t}}_{yg}^{(2)} / N_g \ , \ \hat{R}\hat{V}_{2g} = \frac{\bar{\hat{V}}_{2g}}{V\hat{a}r\left(\hat{t}_{yg}^{(2)}\right)}$$

and are presented in Table 1. We see that the under-coverage of the classification variable defining domains increases the variance of the estimator of the total for domains. It means that one should be careful with respect to the quality of the sampling frame.

Table 1. Simulation results, n=800. Relative increase in variance for estimator of mean due to under-coverage of the classification variable

Sex	Profession	N_g	$\bar{\hat{\mu}}_{yg}^{(2)}$	$RBias\ \left(\bar{\hat{\mu}}_{yg}^{(2)}\right)$	$R\hat{V}_{2g}$	$\bar{\hat{\mu}}_{yg}^{(2)}$	$RBias\ \left(\bar{\hat{\mu}}_{yg}^{(2)}\right)$	$\hat{R}\hat{V}_{2g}$
			$\alpha = 0.1$			$\alpha = 0.18$		
M	1	252	343	0.012	0.297	340	0.004	0.441
F	1	212	350	0.015	0.274	345	-0.001	0.381
M	2	339	314	-0.010	0.298	315	-0.007	0.430
F	2	931	330	-0.010	0.117	333	-0.001	0.183
M	3	195	326	0.022	0.276	326	0.023	0.366
F	3	273	324	0.023	0.315	318	0.005	0.399
M	4	53	286	-0.021	0.262	287	-0.017	0.301
F	4	137	284	-0.025	0.185	284	-0.025	0.344
M	5	155	281	0.014	0.184	282	0.016	0.349
F	5	461	277	0.006	0.304	276	0.002	0.404
M	6	158	253	-0.001	0.136	251	-0.006	0.225
F	6	129	225	-0.008	0.226	228	0.007	0.374
M	7	545	267	-0.022	0.278	265	-0.032	0.373
F	7	203	258	0.007	0.348	254	-0.006	0.471
M	8	457	308	-0.004	0.164	304	-0.014	0.218
F	8	61	283	0.081	0.529	286	0.090	0.568
M	9	231	265	-0.014	0.249	263	-0.024	0.447
F	9	259	272	-0.014	0.278	267	-0.034	0.368
Average	Average	281	291	0.003	0.262	290	-0.001	0.369

Conclusions

The estimators used are unbiased, which can be seen from the simulation results: the average of the relative bias for domain mean equals 0.03 and -0.01. Therefore the size of the mean squared error of the estimator of the mean (total) depends only on the variance of this estimator. The term $\hat{R}\hat{V}_{2g}$ demonstrates relative increase of the mean squared error for the estimator of the mean (total) due to under-coverage of the classification variable. For the level of under-coverage of the classification variable $\alpha = 0.1$ more than one quarter of the relative MSE (0.262) is due to under-coverage of the classification variable. Relative increase in MSE is from 0.262 to 0.369, when the level of under-coverage for classification variable is increases from $\alpha = 0.1$ to $\alpha = 0.18$.

Use of the reference sample gives a possibility to estimate domain size proportion in the non-classified population, but it has a shortage that it means additional expenses. Saving funds, size of the reference sample may be small, and estimates for the domain proportions of the non-classified part of the population may have high variances and make high input into the total error of the domain estimates.

Log-linear models can be used to estimate proportions in the unclassified population part instead of the reference sample. It would be cheaper in practice.

Much work should be done in future in order to find methods to estimate the quality of the sampling frames and incorporate accuracy measures for frames into the accuracy estimators of statistical results.

References

Bethlehem, J. (2008). *How accurate are self-selection web surveys*? ISSN: 1572-0314. Statistics Netherlands, The Hague/Heerlen.

Laitila, T. & Holmberg, A. (2010). Comparison of Sample and Register Survey Estimators via MSE Decomposition. *Proceedings of Q2010 European Conference on Quality in Official Statistics*, Statistics Finland and Eurostat, Helsinki, Finland. https://q2010.stat.fi/sessions/special-session-34

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 D., and M. Šličkutė-Šeštokienė, 2017

Data configuration: **1.**

Data used: Sample selected from a sampling frame suffering from under-coverage / over-coverage and administrative data source with the perfect coverage of the survey population

Type of errors: Frame under-coverage and over-coverage.

Quality measures. *Relative difference* between the straightforward and ratio estimators of total, the *relative difference* between the estimators of variances for the estimators of total.

Introduction. The sampling frame (business register) has been fixed for a subsequent year. The administrative data source is updated regularly during a subsequent year, and includes the changes of the enterprise population. The changes are in the size of a data source (resulting in an under-coverage of the sampling frame because of new-established enterprises and over-coverage due the closed enterprises) and in the contents of some variables. If the population is considered as fixed then the quarterly estimates will be biased due to the real population changes.

The aim of a test is to measure impact of the changes in the enterprise population on the accuracy of the estimator of the total. The stratified sample design is used for a survey. The separate ratio estimator and combined ratio estimator are used to estimate the total of a study variable from the population suffering from under-coverage / over-coverage with the auxiliary variable from the administrative data source and perfect coverage.

Work carried out. Suppose we have population *U* of size *N*, divided into the strata U_h, *h*=1,2,...,*H*, $U_1 \cup ... \cup U_H = U$, of sizes $N_1,..., N_H$, $N_1 + ... + N_H = N$. A simple random stratified sample $\mathbf{s} = \mathbf{s}_1 \cup ... \cup \mathbf{s}_H$ of sizes $n_1,..., n_H$, $n_1 + ... + n_H = n$, is selected from *U*. Let administrative data source V, $V_1 \cup ... \cup V_H = V$, be divided into the strata according to the same criteria with the strata sizes $N'_1,..., N'_H$, $N'_1 + ... + N'_H = N$. The population V may be considered as evolution of the population U over time. As a result of this evolution, the original sampling frame *U* has under-coverage and over-coverage with respect to the population in V: $U \setminus V \neq \varnothing$ and $V \setminus U \neq \varnothing$. Due to these coverage problems and non-response cases the sample $\mathbf{s}$ reduces to $\mathbf{s}' = \mathbf{s}'_1 \cup ... \cup \mathbf{s}'_H$, $\mathbf{s}'_h \subset \mathbf{s}_h$, $\mathbf{s}'_h \subset V_h$, $h = 1,..., H$ of size $n'_1,..., n'_H$, $n'_1 + ... + n'_H = n'$, $n'_h \leq n_h$.

Let a study variable *y* be defined for *U*, and auxiliary variable *x* defined for *V*. Denote

$$t_{yh} = \sum_{k \in U_h} y_k\ ,\ \ t_{xh} = \sum_{k \in U_h \cap V_h} x_k\ ,\ \ h=1,2,\ldots,H,\ \ t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} t_{yh}\ .$$

Let variable *z* be defined for *V* and restricted for *U* to be coinciding with *y*: $z_k = y_k$ for $k \in U$,

$$t_z = \sum_{k \in V} z_k = \sum_{h=1}^{H} \sum_{k \in V_h} z_k = \sum_{k \in V \cap U} y_k + \sum_{k \in V \setminus U} z_k\ .$$ We are interested in the estimator of t_z. Denote

$$t_x = \sum_{k \in U \cap V} x_k = \sum_{h=1}^{H} t_{xh} = \sum_{h=1}^{H} \sum_{k \in U_h \cap V_h} x_k \,, \quad \tilde{t}_x = \sum_{k \in V} x_k = \sum_{h=1}^{H} \tilde{t}_{xh} = \sum_{h=1}^{H} \sum_{k \in V_h} x_k \,,$$

$$\hat{t}_{yh} = \frac{N_h^*}{n_h'} \sum_{k \in s_h'} y_k \,, \; \hat{t}_{xh} = \frac{N_h^*}{n_h'} \sum_{k \in s_h'} x_k \,, \; h=1,2,\ldots,H.$$

Here N_h^*, $N_h^* \le N_h$ is estimated stratum size after subtraction of $N_h^* - N_h$ closed enterprises. It is estimated taking into account reasons of non-response.

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

Case 1. Separate ratio estimator [1] for a total t_y with a fixed sampling frame and changes in the population size not taken into account:

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

Fixed sampling frame and changes in the population size taken into account (adjusted population) using separate estimator of the ratio gives us

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

Let us denote the differences of the totals $\Delta t_{xh} = \tilde{t}_{xh} - t_{xh}$.

The measure of a change of the estimator of total due to the frame under-coverage / over-coverage proposed is a relative difference between the estimators of total

$$R_{tot.diff}^{(1)} = \frac{\hat{t}_{z.adj}^{(1)} - \hat{t}_{y.fixed}^{(1)}}{\hat{t}_{y.fixed}^{(1)}} = \frac{1}{\hat{t}_{y.fixed}^{(1)}} \sum_{h=1}^{H} \left(\tilde{t}_{xh} - t_{xh}\right) \frac{\sum_{k \in s_h'} y_k}{\sum_{k \in s_h'} x_k} \,. \qquad (1)$$

Expression for approximate variance for the separate estimator of the ratio is presented in [1]. p. 271.

The measure of a change of the estimator of variance for estimator of total due to the frame under-coverage / over-coverage used is the relative difference between the estimators of variances for the estimators of totals:

$$R_{Var.diff}^{(1)} = \frac{Var\,(\hat{t}_{z.adj}^{(1)}) - Var\,(\hat{t}_{y.fixed}^{(1)})}{Var\,(\hat{t}_{y.fixed}^{(1)})} \,. \qquad (2)$$

Case 2. Combined ratio estimator [1] for a total t_y with a fixed sampling frame and changes in the population size not taken into account:

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

Combined ratio estimator adjusted to the changes of frame:

$$\hat{t}^{(2)}_{z.adj} = \tilde{t}_x \frac{\hat{t}'_{y.str}}{\hat{t}'_{x.str}}, \ \tilde{t}_x \text{ known. } \hat{t}'_{y.str} = \sum_{h=1}^{H} \frac{N'_h}{n'_h} \sum_{k \in s'_h} y_k, \ \hat{t}'_{x.str} = \sum_{h=1}^{H} \frac{N'_h}{n'_h} \sum_{k \in s'_h} x_k .$$

Let us denote the difference $\Delta t_x = \tilde{t}_x - t_x = \sum_{k \in V \setminus U} x_k$.

The measure of a change of the estimator of total due to the frame under-coverage / over-coverage is a relative difference between the estimators of total

$$R^{(2)}_{tot.diff} = \frac{\hat{t}^{(2)}_{z.adj} - \hat{t}^{(2)}_{y.fixed}}{\hat{t}^{(2)}_{y.fixed}} . \qquad (3)$$

The measure of a change of the estimator of variance for estimator of total due to the frame under-coverage / over-coverage is the relative difference between the estimators of variances for the estimators of totals:

$$R^{(2)}_{Var.diff} = \frac{Var(\hat{t}^{(2)}_{z.adj}) - Var(\hat{t}^{(2)}_{y.fixed})}{Var(\hat{t}^{(2)}_{y.fixed})} . \qquad (4)$$

We consider the indicators $R^{(1)}_{tot.diff}$, $R^{(1)}_{Var.diff}$ and $R^{(2)}_{tot.diff}$, $R^{(2)}_{Var.diff}$ (1), (2), (3), (4) as measures of influence of frame under-coverage / over-coverage to the estimator of total for t_y .

If the variable *x* is identically equal to unit: $x_k \equiv 1$, $k \in U$, then the estimators proposed are adjusted by updated population stratum sizes (case 1) or updated population size (case 2).

Case 3. The estimator $\hat{t}^{(2)}_{z.adj}$ can be replaced by the post-stratified estimators of the total or by the calibrated estimators of total, which may take into account the population changes after a time period.

Assumption. The method is based on the following assumption:

> Administrative data source with the perfect population coverage contains an auxiliary variable on the micro level or aggregated level, which is correlated with the study variable, and values of which are available for a sample selected from the sampling frame.

Advantages. It is a classical method of survey statistics

Case study. The analysis of influence of under-coverage / over-coverage of the quarter survey population to the estimates of total is accomplished using data of the Quarterly Survey on Earnings 2015 at Statistics Lithuania by the kind of activity. The estimates for the fixed population (4th quarter of 2014) and estimates adjusted for the frame under-coverage / over-coverage (separate ratio estimators for total) are calculated and compared. The study variables are a number of employees in full time units and average gross earnings. Auxiliary variable equals to the quarter average of the enterprise number of employees according to the social insurance inspection data base. Estimation results are presented in Table 1. Relative differences between the separate ratio estimates are presented for the 1st quarter and 4th quarter of 2015. The changes in the population are much higher in the 4th quarter than in the 1st quarter when compared to the end of 2014.

Enterprise kinds of activity used for the case study:

A Agriculture, forestry and fishing
B Mining and quarrying
C Manufacturing
D Electricity, gas, steam and air conditioning supply
E Water supply; sewerage; waste management and remediation activities

F Construction
G Wholesale and retail trade; repair of motor vehicles and motorcycles
H Transportation and storage
I Accommodation and food service activities
J Information and communication
K Financial and insurance activities
L Real-estate operations
M Professional, scientific and technical activities
N Administrative and support service activities
O Public administration and defence; compulsory social security
P Education
Q Human health and social work activities
R Arts, entertainment and recreation
S Other service activities

Table 1. The relative changes between estimates and estimates of their variances for fixed population and adjusted population for key variables of the Lithuanian Quarterly Survey on Earnings 2015, in per cent

	$R^{(1)}_{tot\ .diff}$				$R^{(1)}_{Var\ .diff}$			
	Number of employees in full-time units		Average monthly gross earnings		Number of employees in full-time units		Average monthly gross earnings	
Quarter	2015 Q1	2015 Q4	2015 Q1	2015 Q4	2015 Q1	2015 Q4	2015 Q1	2015 Q4
Total	1,67	2,64	-0,75	-1,05	5,26	7,82	-0,07	0,54
A	1,07	1,85	-0,64	-1,03	3,07	5,87	-0,46	-0,18
B	1,16	1,33	-1,04	-1,21	2,34	2,69	-2,08	-2,41
C	1,17	2,00	-0,52	-0,73	2,65	4,27	-0,78	-1,26
D	0,92	2,32	-0,59	-1,17	12,84	33,46	9,66	16,33
E	0,37	0,76	-0,25	-0,51	2,15	4,25	0,70	1,36
F	3,10	5,38	-1,49	-2,42	6,52	11,42	-2,77	-4,48
G	2,61	3,75	-1,34	-1,71	5,29	7,63	-2,68	-3,40
H	1,92	2,96	-0,94	-1,19	5,46	8,32	-0,58	-0,21
I	4,94	8,59	-1,36	-2,08	10,16	17,98	-2,67	-4,07
J	2,62	4,21	-0,96	-1,04	5,72	9,25	-1,52	-1,54
K	0,77	1,29	-0,54	-0,80	6,73	10,97	3,27	6,03
L	3,45	5,36	-1,70	-2,11	8,00	12,61	-2,53	-2,80
M	4,15	5,89	-1,25	-1,60	11,21	15,83	0,04	-0,25
N	2,82	4,26	-0,81	-1,15	6,88	10,49	-0,62	-0,88
O	0,08	0,10	0,01	0,01	0,07	0,03	-0,08	-0,13
P	0,20	0,36	-0,09	-0,13	0,43	0,59	-0,17	-0,37
Q	0,43	0,65	-0,10	-0,20	2,62	3,67	1,84	3,12
R	1,62	2,54	-0,70	-1,13	3,27	3,56	-1,33	-3,40
S	3,78	6,48	-2,08	-2,92	8,14	14,23	-3,71	-5,05

Results. The measures of change proposed show well how accuracy of the estimates of total decreases (compare 2015 Q1 and 2015 Q4) with increasing under-coverage of the population (time lag increasing from 1 quarter to 4 quarters). Relative changes are positive for *number of employees in full-time units* and mostly negative for *average monthly gross earnings*.

Gap analysis.

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

2) The method can be adapted to the estimators of the 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 to the frame changes.

3) The difference between two separate ratio estimators estimators, adjusted and non-adjusted, $\hat{\Delta}^{(1)} = \hat{t}^{(1)}_{z.adj} - \hat{t}^{(1)}_{y.fixed}$ can be studied. Denote its variance $Var\left(\hat{\Delta}^{(1)}\right)$ estimator by $V\hat{a}r\left(\hat{\Delta}^{(1)}\right)$. The relative difference $\hat{\Delta}^{(1)} / \sqrt{V\hat{a}r\left(\hat{\Delta}^{(1)}\right)}$ might be a useful way to measure how serious is the estimation problem due to under- and over-coverage. In particular, it could be tested then whether the observed difference between the adjusted and non-adjusted ratio estimates is significant. Similar relative difference may be studied for the combined ratio estimator.

References

Särndal C.-E., Swenson B., Wretman J. *Model assisted survey sampling*. New York, Springer Verlag, 1992.

ST2_6 Effect of stratum changes, joining and splitting of the enterprises on the estimator of a total

Krapavickaitė D. and V. Vasilytė (2017)

Data configuration: **1.**
Aim of the test: To measure impact of the frame errors to the bias and variance of the estimator of total in the case enterprises are joining, splitting and changing their kind of activity.
Type of errors: Enterprise identification errors / sampling frame errors.
Data used: Survey data set and updated sampling frame (or *selected sample* and *observed sample*).
Statistic of interest: Estimators of the population total in business statistics.
Quality measure: Relative bias and relative variance of the estimator for a population total.

Description of the work carried out

Introduction. Let us suppose that all statistical units (enterprises) are registered in a Business register (BR). This BR is used as a sampling frame for a stratified simple random sample without replacement, 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 effect on the stratification of the units: 1) errors in the value of the classification variable - NACE code, identifying stratum 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, information has been received from some of the respondents 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 of 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 us assume that all information about the enterprise changes is known. The sample after changes is named observed sample.

Problem formulation [3]. 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 unit keeps succession of the unit u_i and it 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 $k = 1,2,..., N$, it is stratified into H strata: $U = U_1 \cup U_2 \cup ... \cup U_H$, N_h is the size of the stratum h, $h = 1,..., H$. A simple random sample ω_h of size n_h is drawn from the stratum U_h. Denote the whole sample of size $n = n_1 + ... + n_H$ by ω, $\omega = \omega_1 \cup ... \cup \omega_H$, where ω_h consists of the units u_{hi} with $i = 1,2,..., n_h$.

As it has been mentioned, 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' , $i = 1,2,..., n'$, for which nor equality $n = n'$, nor $n_h = n_h'$ may hold.

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

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

where for brevity in the notation we use y_k for observations of variable *y* for enterprises u_k in the BR at time of sampling. Let y'_k denote observations of variable *y* for enterprises u'_k at the time of observation, and for a changed population U' we define a total

$$t'_y = \sum_{k \in U'}^{N} y'_k \ .$$

Our aim is to estimate this total.

Error type (a)

The unit $u'_k \in \omega'$ from the observed sample is a union of some j_k, $j_k \geq 1$, units from the BR: $u'_k = u_k \cup u_{k+1} \cup ... \cup u_{k+j_k-1}$, and at least unit u_k belongs to the sample *ω*. In this case, all the BR units $u'_{k_1},..., u'_{k_j}$ are identified, but the values of the study variable y are not known for these units. Instead, we observe one value y'_k of the study variable *y* for the unit u'_k.

Since the values of the study variable are not known for all the elements of the sample *ω*, and because of the presence of the frame errors, we cannot use the Horwitz-Thompson estimator (Särndal et all, 1992):

$$\hat{t}_y = \sum_{k:u_k \in \omega} \frac{y_k}{\pi_k} \ . \tag{B}$$

So, we are interested to estimate t'_y which is the total t_y corrected for error type (a). Here $\pi_k = P(\omega : u_k \in \omega)$ is an inclusion probability of the element u_k into the sample *ω*. In our stratified sample case $\pi_k = n_h / N_h$ if unit u_k belongs to the stratum U_h. An alternative may be to use the estimator

$$\hat{t}'_y = \sum_{k:u'_k \in \omega'} \frac{y'_k}{\pi'_k} \tag{1}$$

with the inclusion probabilities $\pi'_k = P(\omega' : u'_k \in \omega')$ for the elements of the observed sample ω' given error type (a). This provides us with an unbiased estimate of the population total t'_y since it accounts for the adjusted inclusion probabilities of the units in the sample.

Estimation of the inclusion probabilities π'_k and estimation of the total t'_y.

The unit u'_k is included into the sample ω', if at least one of its compound parts $u_k,..., u_{k+j_k-1}$ is selected into the sample *ω*. Denote by j_{hk}, $j_{hk} \geq 1$, the number of units from the set $\{u_k,..., u_{k+j_k-1}\}$ that belong to the stratum U_h, $h = 1,..., H$. We have $j_k = j_{1k} + ... + j_{Hk}$. The probability for any part of the unit u'_k in the stratum U_h to be not selected into the sample ω_h, is

$$C^{n_h}_{N_h - j_{hk}} \Big/ C^{n_h}_{N_h} = \prod_{s=0}^{n_h - 1} \left(1 - \frac{j_{hk}}{N_h - s} \right),$$

so

$$\pi'_k = 1 - \prod_{\substack{h=1 \\ h: j_{hk} > 0}}^{H} \prod_{s=0}^{n_h - 1} \left(1 - \frac{j_{hk}}{N_h - s} \right), \tag{2}$$

for $N_h > n_h$, $N_h - j_{hk} \geq n_h$. Otherwise we have $\pi'_k = 1$.

Note, that for $j_k = 1$ ($u'_k = u_k$), the formula (2) gives $\pi'_k = n_h / N_h = \pi_k$, if $u_k \in \omega_h$.

Now let us calculate the second order inclusion probabilities $\pi'_{kl} = P(\omega' : u'_k \;\&\; u'_l \in \omega')$. Let the element $u'_k \in \omega'$, $u'_k = u_k \cup ... \cup u_{k+j_k-1}$, be as above a composite unit, and the numbers j_k are defined as above. Assume that the second element $u'_l \in \omega'$, $u'_l = u_{l_1} \cup ... \cup u_{l+m_l-1}$, $m_l \geq 1$, may also be composite. Denote by m_{hl}, $m_{hl} \geq 0$, the number of elements from the set $\{u_l, ..., u_{l+m_l-1}\}$ that belong to the stratum h. Thus we have $m_l = m_{1l} + ... + m_{Hl}$. From the probability theory we know that

$$P(\omega' : u'_k \in \omega' \cap u'_l \in \omega') = P(\omega' : u'_k \in \omega') + P(\omega' : u'_l \in \omega') - 1 + P(\omega' : u'_k \notin \omega' \cap u'_l \notin \omega').$$

Therefore

$$\pi'_{kl} = \pi'_k + \pi'_l - 1 + \prod_{\substack{h=1 \\ h: j_{hk} > 0, \\ m_{hl} > 0}}^{H} \prod_{s=0}^{n_h - 1} \left(1 - \frac{j_{hk} + m_{hl}}{N_h - s}\right) \tag{3}$$

for $N_h > n_h, N_h - j_{hk} - m_{hl} \geq n_h$; and $\pi'_{kl} = 1$ otherwise.
Now we can use the estimator

$$V\hat{a}r\,(\hat{t}'_y) = \sum_{k,l \in \omega'} \left(1 - \frac{\pi'_k \pi'_l}{\pi'_{kl}}\right) \frac{y'_k}{\pi'_k} \frac{y'_l}{\pi'_l} \tag{4}$$

to estimate the variance of the estimator $\hat{t}'_y$, if needed.
The total (A) is estimated by the estimator (1) with the inclusion probabilities (2), its variance is estimated by (4) using (3). The estimates obtained are compared with the estimates obtained using Horvitz-Thompson estimator (B), constructing relative measures of difference between estimators of total and their variance estimators:

$$Rdiff\;\;(\hat{t}'_y) = \frac{\hat{t}'_y - \hat{t}_y}{\hat{t}_y}, \quad RVar\;\;(\hat{t}'_y) = \frac{V\hat{a}r(\hat{t}'_y) - V\hat{a}r(\hat{t}_y)}{V\hat{a}r(\hat{t}_y)}. \tag{5}$$

The expectation of the relative difference $E(Rdiff\;\;(\hat{t}'_y)) = 0$ with respect to the sampling design, because $\hat{t}'_y$ and $\hat{t}_y$ are unbiased, but the relative differences themselves have fluctuations around 0. In the case a statistician does not account for the error of type (a) one would use the unadjusted inclusion probabilities π_k to compute the estimate $\hat{t}^*_y = \sum_{k: u_k \in \omega} \frac{y'_k}{\pi_k}$ of total t_y. Note that the observed values y'_k are included in the estimator, because no other values are available, and they are synthetic for the units belonging to $\omega \setminus \omega'$. The estimator $\hat{t}^*_y$ is biased, because estimator (1) is unbiased for t'_y. The relative difference between the estimators $\hat{t}^*_y$ and $\hat{t}'_y$ is defined as

$$Rdiff\;\;(\hat{t}^*_y) = \frac{\hat{t}^*_y - \hat{t}'_y}{\hat{t}'_y}. \tag{6}$$

At the same time this indicator may be considered as a relative approximate bias $RBias\;\;(\hat{t}^*_y) = Rdiff\;\;(\hat{t}^*_y)$.

Simulation study. Data of the enterprise expenses for environment protection survey is used for a simulation study. Study population consists of 384 enterprises belonging to 6 kinds of activity. This population is stratified by the kind of activity and number of employees (5-9, 10-49, 50-99, >99), stratified simple random sample of size n=100 is selected. At the time of observation it is observed $n', n' < n$, of enterprises because some of them are merged (joining of enterprises is simulated inside the stratum with the probability $p = 0.1$ and $p = 0.05$). Enterprise income is used as a study variable. The relative measures of accuracy (5) and (6) due to joining of the enterprises are presented in the Table 1.

Table 1. Relative measures of accuracy for estimators of totals in the case of enterprise joining

	$Rdiff\ (\hat{t}'_y)$	$RVar\ (\hat{t}'_y)$	$Rdiff\ (\hat{t}^*_y)$	$RVar\ (\hat{t}^*_y)$
$p = 0.05$	0.0105	-0.0027	0.0388	0.0108
$p = 0.1$	0.0771	0.0227	0.3577	1.7430

Conclusion. The relative measures of accuracy for estimators are increasing with the growing up probability of error (probability of enterprise joining). $Rdiff\ (\hat{t}^*_y)$ shows that large bias of the estimator of total $\hat{t}^*_y$ may occur if enterprise joining will not be taken into account at the estimation stage.

The similar topic of the companies joining is studied in Knottnerus, 2011, p. 30.

Error type (b)

Besides the sample units that merge, it is possible to observe other elements that have split since the selected sample was drawn. Information about the study variable y is received not from all of those split elements, and value of the variable y received is not as it would be if the units had not been split. If the observed value is only for the some of the split-part of the unit then splitting can be viewed as a second stage sampling with the second stage inclusion probability $\pi_k^{(2)}$ and the final inclusion probability $\pi'_k = \pi_k \pi_k^{(2)}$. The probability $\pi'_k \leq \pi_k$ is not greater than π_k parallel to error type (a) where $\pi'_k \geq \pi_k$. Increase of variance for estimator due to the second stage sampling should be taken into account. It shows increase of variance due to enterprise splitting.

Another solution to the problem would be to consider the values of a study variable of split-parts, which have not provided data, as missed. Any imputation method can be used to fill in the missed values of a study variable, and then the population total can be estimated. Increase of variance for estimator due to imputation should be taken into account. It also shows increase of variance due to enterprise splitting.

Error type (c)

It is possible that according to the information received from the observed units values for the classification variables (for example, economic activity – NACE code or size group) have been changed since the sample selection for some of them. It will be studied further what influence make these changes to the accuracy of the estimator of total.

Let U_h still means the enterprise population in stratum h, N_h - its size, and a simple random sample ω_h of size n_h is selected from this stratum. Let $\omega_h^{(1)}, \omega_h^{(1)} \subset \omega_h$ be a subsample of enterprises reporting other stratum codes than selected, their number is $n_h^{(1)}$. Further, let $\omega_h \setminus \omega_h^{(1)}$ be a set of enterprises remaining in the stratum h, their number is $n_h - n_h^{(1)}$. Let $\omega_{h-1}^{(2)}$, $\omega_{h-1}^{(2)} \subset \omega_{h-1}$ be the subset of enterprises from the stratum h-1, which reported belonging to the stratum h, their number is $n_{h-1}^{(2)}$. Also, let $\omega_{h+1}^{(2)}$, $\omega_{h+1}^{(2)} \subset \omega_{h+1}$, be a subset of enterprises from the stratum h+1, which reported belonging to the stratum h, their number is $n_{h+1}^{(2)}$. This notation is reasonable in the case of size class errors: sample elements might go one stratum up or down (due to the increase or decrease in number of employees). In the case of NACE code errors, "h-1" and "h+1" can be viewed as any two other NACE codes containing elements which in fact belong to NACE code "h".

Denote by $\omega_h^{(3)} = \left(\omega_h \setminus \omega_h^{(1)}\right) \cup \omega_{h-1}^{(2)} \cup \omega_{h+1}^{(2)}$ - a sample subset that belongs to the stratum *h* at the observation time, its size is $n_h^{(3)} = n_h - n_h^{(1)} + n_{h-1}^{(2)} + n_{h+1}^{(2)}$. Let *y* be a study variable, $t_{yh} = \sum_{k\in U_h} y_k$ - stratum population total at the observation time. Let us estimate it taking into account changes in the population, or, actually, another, domain total, in such a way:

$$\hat{\hat{t}}_{yh} = \sum_{k\in \mathbf{i}_h \setminus \mathbf{i}_h^{(1)}} \frac{N_h}{n_h} y_k + \sum_{k\in \mathbf{i}_{h-1}^{(2)}} \frac{N_{h-1}}{n_{h-1}} y_k + \sum_{k\in \mathbf{i}_{h+1}^{(2)}} \frac{N_{h+1}}{n_{h+1}} y_k = \hat{\hat{t}}_{1h} + \hat{\hat{t}}_{2h-1} + \hat{\hat{t}}_{2h+1}. \qquad (7)$$

Because of independent samples in the strata equality for variances is valid:

$$Var\left(\hat{\hat{t}}_{yh}\right) = Var\,(\hat{\hat{t}}_{1h}) + Var\,(\hat{\hat{t}}_{2h-1}) + Var\,(\hat{\hat{t}}_{2h+1}). \qquad (8)$$

Variance estimator used:

$$V\hat{a}r\left(\hat{\hat{t}}_{yh}\right) = V\hat{a}r(\hat{\hat{t}}_{1h}) + V\hat{a}r(\hat{\hat{t}}_{2h-1}) + V\hat{a}r(\hat{\hat{t}}_{2h+1}). \qquad (9)$$

The estimator for a population total and estimator for the variance of this estimator used is as follows:

$$\hat{\hat{t}}_y = \sum\nolimits_{h=1}^{H} \hat{\hat{t}}_{yh}, \quad V\hat{a}r\left(\hat{\hat{t}}_y\right) = \sum\nolimits_{h=1}^{H} V\hat{a}r\left(\hat{\hat{t}}_{yh}\right) + 2\sum_{\substack{h,l=1\\ l>h}}^{H} C\hat{o}v\left(\hat{\hat{t}}_{yh}, \hat{\hat{t}}_{yl}\right) \approx \sum\nolimits_{h=1}^{H} V\hat{a}r\left(\hat{\hat{t}}_{yh}\right).$$

Estimators $\hat{t}_{yh}$, $h = 1,\ldots, H$, are dependent because each of them may depend on the units belonging to the different strata. For estimation $Var\left(\hat{t}_{yh}\right)$ this dependency is not taken into account.

In order to study changes in variance of the estimator for stratum total due to the change of the values for classification variable, some assumptions on the enterprise stratum change mechanism are made.

Assumption. Let us assume that any enterprise changes the stratum with the same probability p, $p \in (0,1)$. With the probability p enterprise changes stratum *h* to the stratum *h*-1, stratum *h* to the stratum *h*+1 and vice versa. We consider that a stratum *h* has two neighbouring strata, and probability for the enterprise to leave it equals 2*p*, probability to remain is 1-2*p*. We consider that the situation can be described by a two phase sampling design with the Bernoulli second phase sampling, when enterprises from the strata *h*-1 and *h*+1 with the probability *p* are coming to the stratum *h*, and enterprises with the same probability 1-2*p* independently are remaining in the stratum *h*.

Let I_k denotes indicator for element to belong to the sample $\omega_h \setminus \omega_h^{(1)}$:

$$I_k = \begin{cases} 1, k \in \omega_h \setminus \omega_h^{(1)}, \\ 1, k \notin \omega_h \setminus \omega_h^{(1)} \end{cases}, \qquad k \in U_h.$$

Expression for variance of the estimator $\hat{\hat{t}}_{1h}$ is obtained:

$$Var\,(\hat{\hat{t}}_{1h}) = Var\left(\frac{N_h}{n_h} \sum_{k\in\omega_h\setminus\omega_h^{(1)}} y_k\right) = Var\left(\frac{N_h}{n_h} \sum_{k\in\omega_h} y_k I_k\right) = \frac{N_h^2}{n_h^2}\left(EVar\left(\sum_{k\in\omega_h} y_k I_k \middle| \omega_h\right) + VarE\left(\sum_{k\in\omega_h} y_k I_k \middle| \omega_h\right)\right) =$$

$$= \frac{N_h^2}{n_h^2}\left(E\left(\sum_{k\in\omega_h} y_k^2\, 2p(1-2p)\right) + Var\,(1-2p)\sum_{k\in\omega_h} y_k\right) = \frac{N_h}{n_h} 2p(1-2p)\sum_{k\in U_h} y_k^2 + (1-2p)^2 Var\left(\frac{N_h}{n_h}\sum_{k\in\omega_h} y_k\right) =$$

$$= \frac{N_h}{n_h} 2p(1-2p)\sum_{k\in U_h} y_k^2 + (1-2p)^2 Var\,\left(\hat{t}_{yh}\right). \qquad (10)$$

Let $\hat{t}_{yh}, \hat{t}_{yh-1}, \hat{t}_{yh+1}$ denote estimators for sums $t_{yh}, t_{yh-1}, t_{yh+1}$ in the strata *h*, *h*-1, *h*+1.

$$V\hat{a}r(\hat{\hat{t}}_{1h}) = \frac{N_h^2}{n_h\left(n_h - n_h^{(1)}\right)} 2p(1-2p) \sum_{k\in\omega_h\setminus\omega_h^{(1)}} y_k^2 + (1-2p)^2 V\hat{a}r\left(\hat{t}_{yh}^{\left(n_h - n_h^{(1)}\right)}\right). \qquad (11)$$

Let I_k denotes indicator for unit k to belong to the sample $\omega_{h-1}^{(2)}$. The expression for variance of the estimator $\hat{\hat{t}}_{2h-1}$ is obtained as follows:

$$Var\,(\hat{\hat{t}}_{2h-1}) = Var\left(\frac{N_{h-1}}{n_{h-1}}\sum_{k\in\omega_{h-1}^{(2)}} y_k\right) = Var\left(\frac{N_{h-1}}{n_{h-1}}\sum_{k\in\omega_{h-1}} y_k I_k\right) = \frac{N_{h-1}}{n_{h-1}}\, p(1-p)\sum_{k\in U_{h-1}} y_k^2 + p^2 Var\left(\hat{t}_{yh-1}\right).$$

$$V\hat{a}r(\hat{\hat{t}}_{2h-1}) = \frac{N_{h-1}^2}{n_{h-1}n_{h-1}^{(2)}}\, p(1-p)\sum_{k\in\omega_{h-1}^{(2)}} y_k^2 + p^2 V\hat{a}r\left(\hat{t}_{yh-1}^{\,n_{h-1}^{(2)}}\right). \tag{12}$$

Let I_k denotes indicator for the unit to belong to the sample $\omega_{h+1}^{(2)}$.

$$Var\,(\hat{\hat{t}}_{2h+1}) = Var\left(\frac{N_{h+1}}{n_{h+1}}\sum_{k\in\omega_{h+1}^{(2)}} y_k\right) = Var\left(\frac{N_{h+1}}{n_{h+1}}\sum_{k\in\omega_{h+1}} y_k I_k\right) = \frac{N_{h+1}}{n_{h+1}}\, p(1-p)\sum_{k\in U_{h+1}} y_k^2 + p^2 Var\left(\hat{t}_{yh+1}\right).$$

$$V\hat{a}r(\hat{\hat{t}}_{2h+1}) = \frac{N_{h+1}^2}{n_{h+1}n_{h+1}^{(2)}}\, p(1-p)\sum_{k\in\omega_{h+1}^{(2)}} y_k^2 + p^2 V\hat{a}r\left(\hat{t}_{yh+1}^{\,(n_{h+1}^{(2)})}\right). \tag{13}$$

The following variance estimators are for fixed size sample, and they underestimate the variances to some extent. The better accuracy may be achieved taking random sample size into account.

$$V\hat{a}r\left(\hat{t}_{yh}^{\left(n_h - n_h^{(1)}\right)}\right) = N_h^2\left(1 - \frac{n_h - n_h^{(1)}}{N_h}\right)\frac{\hat{s}_{1h}^2}{n_h - n_h^{(1)}},\ \hat{s}_{1h}^2 = \frac{1}{n_h - n_h^{(1)} - 1}\sum_{k\in\omega_h\backslash\omega_h^{(1)}}\left(y_k - \tilde{\bar{y}}_h\right)^2,\ \tilde{\bar{y}}_h = \frac{1}{n_h - n_h^{(1)}}\sum_{k\in\omega_h\backslash\omega_h^{(1)}} y_k \tag{14}$$

$$V\hat{a}r\left(\hat{t}_{yh-1}^{\,(n_{h-1}^{(2)})}\right) = N_{h-1}^2\left(1 - \frac{n_{h-1}^{(2)}}{N_{h-1}}\right)\frac{\hat{s}_{2h-1}^2}{n_{h-1}^{(2)}},\quad \hat{s}_{2h-1}^2 = \frac{1}{n_{h-1}^{(2)} - 1}\sum_{k\in\omega_{h-1}^{(2)}}\left(y_k - \tilde{\bar{y}}_{h-1}^{(2)}\right)^2,\quad \tilde{\bar{y}}_{h-1} = \frac{1}{n_{h-1}^{(2)}}\sum_{k\in\omega_{h-1}^{(2)}} y_k\,. \tag{15}$$

$$V\hat{a}r\left(\hat{t}_{yh+1}^{\,(n_{h+1}^{(2)})}\right) = N_{h+1}^2\left(1 - \frac{n_{h+1}^{(2)}}{N_{h+1}}\right)\frac{\hat{s}_{2h+1}^2}{n_{h+1}^{(2)}},\quad \hat{s}_{2h+1}^2 = \frac{1}{n_{h+1}^{(2)} - 1}\sum_{k\in\omega_{h+1}^{(2)}}\left(y_k - \tilde{\bar{y}}_{h+1}^{(2)}\right)^2,\quad \tilde{\bar{y}}_{h+1} = \frac{1}{n_{h+1}^{(2)}}\sum_{k\in\omega_{h+1}^{(2)}} y_k\,. \tag{16}$$

Expression for $V\hat{a}r\left(\hat{\hat{t}}_{yh}\right)$ is obtained by inserting (11), (13),...,(16) into (9).

Simulation study. Simulated population of 3674 enterprises belonging to 6 kinds of activity divided into 4 groups by the number of employees is used for a study. It is obtained by duplicating elements of the population used for error type (a) case. *n*=1000 size sample with proportional allocation of a sample size is used for a study. Enterprise income is simulated. Enterprise size changes are simulated with probabilities *p*=0.05, 0.2, 0.4. Stratum totals are estimated, and relative stratum biases and relative stratum variances for estimator, are presented in Table 2. The following expressions are used for relative biases and relative variances for estimator of total:

$$RBias\,\left(\hat{\hat{t}}_{yh}\right) = \frac{\hat{\hat{t}}_{yh} - t_{yh}}{t_{yh}},\quad RV\hat{a}r\left(\hat{\hat{t}}_{yh}\right) = \frac{V\hat{a}r\left(\hat{\hat{t}}_{yh}\right) - V\hat{a}r\left(\hat{t}_{yh}\right)}{V\hat{a}r\left(\hat{t}_{yh}\right)}.$$

Here $V\hat{a}r\left(\hat{t}_{yh}\right)$ is estimator of total in the case of simple random sampling in the strata, if there where no changes.

Table 2. Relative measures of accuracy for estimators of totals in the case of enterprise size changes

Kind of activity	Number of empl	N_h	n_h	p=0.05		p=0.2		p=0.4	
				$RBias\left(\hat{t}_{yh}\right)$	$RV\hat{a}r\left(\hat{t}_{yh}\right)$	$RBias\left(\hat{t}_{yh}\right)$	$RV\hat{a}r\left(\hat{t}_{yh}\right)$	$RBias\left(\hat{t}_{yh}\right)$	$RV\hat{a}r\left(\hat{t}_{yh}\right)$
1	5-9	48	13	0.005	0.262	12.57	1557.41	25.67	4628.61
1	10-49	218	59	0.052	0.108	0.20	2.60	-0.33	3.39
1	50-99	170	46	1.802	442.451	4.61	219.93	19.28	1503.11
1	>100	873	238	-0.036	0.014	-0.10	0.07	-0.38	-0.58
2	5-9	41	11	0.000	0.101	2.94	139.74	4.45	204.20
2	10-49	132	36	-0.021	0.113	-0.05	1.69	-0.38	0.76
2	50-99	49	13	1.078	4.204	4.74	49.10	7.64	60.30
2	>100	190	52	-0.049	0.503	-0.23	0.91	-0.35	0.17
3	5-9	89	24	-0.012	0.119	1.40	15.18	2.02	20.61
3	10-49	159	43	1.134	21.601	0.20	17.00	2.45	138.43
3	50-99	112	31	0.130	3.357	1.27	39.15	1.89	42.45
3	>100	189	51	-0.041	0.129	-0.21	-0.02	-0.35	-0.46
4	5-9	54	15	0.256	3.585	3.42	491.66	3.01	735.82
4	10-49	222	60	-0.046	0.200	-0.39	-0.22	-0.09	2.53
4	50-99	61	17	-0.037	0.231	2.23	45.29	2.24	70.61
4	>100	111	30	0.006	0.016	-0.22	0.24	-0.24	-0.37
5	5-9	83	23	0.457	2.592	0.76	6.25	2.36	32.51
5	10-49	168	46	-0.080	0.278	-0.38	0.29	-0.30	1.38
5	50-99	50	14	0.013	0.530	3.50	30.02	11.76	84.82
5	>100	150	41	0.000	0.061	-0.21	0.19	-0.72	-0.87
6	5-9	92	25	0.193	4.764	0.37	3.95	0.87	36.82
6	10-49	166	45	0.080	0.610	0.22	6.76	0.48	78.74
6	50-99	89	24	0.256	1.142	0.80	7.07	0.92	16.46
6	>100	158	43	-0.069	0.212	-0.21	0.19	-0.26	-0.53
Total $\hat{t}_y$		3674	1000	0.211	20.299	1.55	109.77	3.40	319.12
$RV\hat{a}r\left(\hat{t}_y\right)$					0.259		0.274		0.352

Conclusion. The relative bias and relative variance of the estimator of total is increasing with increasing probability of the enterprise to change its size.

Issues. The methods presented are based on the following assumptions:

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

Mathematically it is a straightforward approach.
The method resulted in plausible estimates of accuracy (bias and variance).

Gap analysis. The problems studied in this paper suggest deeper analysis in the following directions:

- random sample sizes for estimators of variances included in (11), (12), (13) should be taken into account;
- the assumptions in on the study of classification errors could be loosened and replaced, for example, by variable proportions of the enterprises, that move between strata, to make it wider applicable;
- the joined effect of errors in mergers, splits and classification errors could be estimated.

References

1. Särndal C.-E., Swenson B., Wretman J. *Model assisted survey sampling*. New York, Springer Verlag, 1992.

2. Knottnerus, P. 2011. *Panels, business panels. Statistical methods* 201119. Publication by Statistics Netherlands. Retrieved at 20170406 from https://www.cbs.nl/NR/rdonlyres/58E4FA6F-6802-4125-BE72-D3566ECD37B6/0/2011x3719.pdf.

3. Krapavickaitė D., Plikusas A. Some choices of a specific Sampling Design. In: *Official statistics. Methodology and applications in honour of Daniel Thorburn*. Eds.: M. Carlson, H. Nyquist, M. Villani. Stockholm University: Stockholm, 2010.

4. Krapavickaitė D., Vasilytė V. *Effect of stratum changes, joining and splitting of the enterprises on the estimator of a total*. Manuscript. Statistics Lithuania, 2017.

ST2_7 Output Quality for statistics based on several administrative sources

Case: The Norwegian register-based employment statistics and the effect of delays

Johan Fosen, Statistics Norway

1. Method for estimating register-based employment proxy

In Statistics Norway, a register-based employment status is derived for each person in Norway between 15 and 74 years old. This status is derived at the *reference week* which used to be essentially the third week of November, and a brief description of the method that was used is described in Fosen (2012). With the introduction of *a-ordningen,* the new integrated register system for jobs and payments in 2015, the system of employment status classification has been changed in accordance with the improved register situation for employees (whereas self-employed are classified essentially as before). The reference week is now defined for each month as the week containing the 15th day of the month.

In this document we use a proxy for the new register-based employment status. It is a proxy in the sense that we use a statistical dataset that comes slightly earlier in the production process than the final dataset used for producing the statistics: we use the so-called "L2 dataset" 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 dataset before they do the final classification of employment status.[1] We will below by 'estimated employment rate' refer to this proxy.

1.1 Data configuration

Looking at "A-ordningen", it apparently seems like all labour market information is contained within one register, pointing to data configuration 1[2]. However, when investigating more closely, we notice that the register information consists of different input sources being "messages" of different kinds, and the most important kinds of messages are "working relations", "salary relations" and "benefit relations".

A *message* received by Statistics Norway, is the counterpart to what we in surveys denote "the received questionnaire". Instead of questionnaire we will in our situation use the term *information form*. Each type of messages mentioned above corresponds to a different information form, and therefore each type of message should to some extent be regarded as a different source into a-ordningen, and hence a-ordningen is based on the multiple sources "working relations", "salary relations" and "benefit relations". These sources should in a sense be regarded as three administrative registers that are exceptionally fit for integration with each other since they were defined within the same overall a-ordningen framework under a joint project between Statistics Norway, The National Insurance Administration and the Tax Authorities.

[1] Still, the L2 data set is *near* the end of the production process.

[2] The data configurations as defined in Section 2 of «Intermediate report WP 3 "Framework for the quality evaluation of statistical output based on multiple sources".

There is an overlap between the sources when it comes to employee-information, since salary is crucial in the classification into employee/not employee and salary comes from the salary relation source. Thus we have data configuration 2. We also have an overlap between units covered by these different kinds of messages: most of the work relations from the working relation source corresponds to a salary relation from the salary relation source.

During the "a-ordning production process" within Statistics Norway, the three message types are integrated. Linking the sources by micro-unit is made almost completely due to the close relation between these sources. By micro-integrating these sources of information, a classification into employee/not employee is performed. Notice that since the sources are so much adapted to each other, there is little contradictory information between the kinds of messages. An example of contradictory information is that for some of the persons, salaries exist but there is no working relation.

2. Time dimensions

For a register-based statistic, there is at least two time dimensions and since most of us have an intuition being unfamiliar with handling two (or more) time-dimension simultaneously, we will below look more closely into the relevant time dimensions. We will also try to motivate the study of delay.

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 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. This chosen update time we here denote *measurement time* albeit the intuitive ambiguity of this term. The interval between the reference time and the measurement time we denote *relative measurement time* (RMT). RMT is the second time dimension. The accuracy increases with longer RMT, but this naturally comes at the cost of loss in timeliness. In order to find a sound balance between the accuracy and timeliness, we should have an estimate of the effect of the output cause by delay, and this is the purpose of this paper. We then have to bear in mind which time dimension we are considering whenever we compare something as a function of time: we sometimes fix reference time and compare along the RMT-dimension, we sometimes fix RMT and compare along the reference time dimension, and finally we could consider both time dimensions simultaneously.

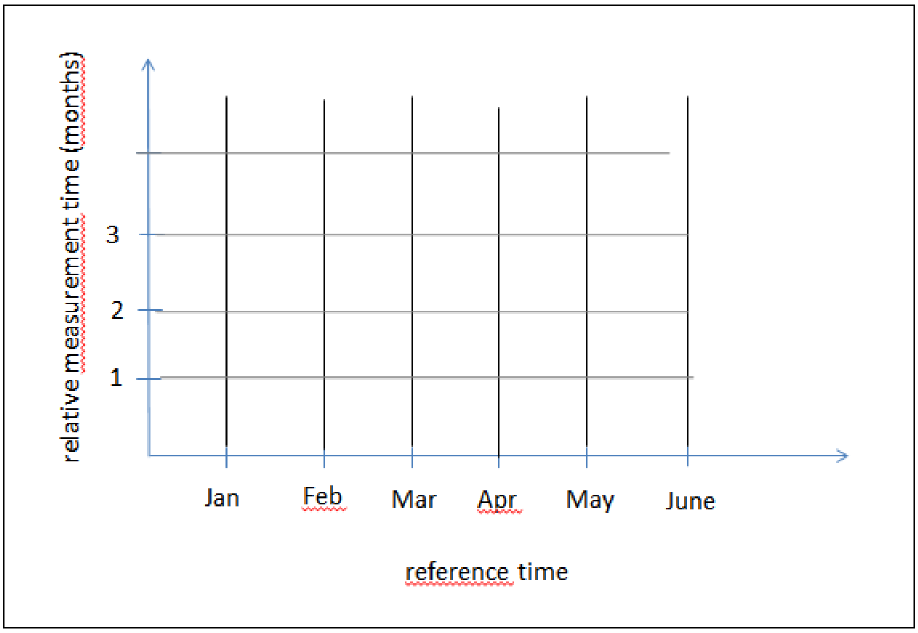


Figure 1. The two time dimensions 'reference time' and 'measurement time'. Measurement time measured as relative measurement time: difference between reference time and measurement time.

In order to conceptually handle the two time-dimensions simultaneously, we can visually represent them in an ordinary Cartesian space (Figure 1). We have put reference time along the x-axis and RMT along the y-axis, so that e.g. the point (1,2) is reference time January and measured in March.

For a given (fixed) reference time point we can consider the corresponding vertical line of Figure 1 and we there have a set of RMT. The output estimates for different RMT along this vertical axis, is the basis for studying the delay process (progressiveness) for a given reference time as a function of RMT. We can study this process in more detail, or we can simplify the problem by finding one suitable estimate for the effects of delay. As an example of a simple effect-of-delay-estimate we could measure the delay mechanism as the difference between the output statistics between RMT=2 months and RMT= 1 month (a measure defined along the vertical line)

Once we have a picture of the delay mechanism as a function of RMT (along a vertical line), either in detail or just a simple effect-of-delay-estimate as in our example above, we would want to see whether the other reference times (other vertical lines) have the same delay mechanism or whether we can find a sound estimate for the development of the delay mechanism as a function of the reference time (horizontal line).

In case of investigating the simple delay effect estimate as a function of reference time, we are only working along the reference time dimension (horizontal) since we have fixed the RMT-dimension since we collapsed the delay process into a single point estimate. On the other hand, assuming a more sophisticated model of the delay process as a function of RMT, when comparing the delay process as a function of reference time, we would jointly consider both time dimensions: reference time (horizontal) and RMT (vertical).

When studying the change of mechanism by reference time (along the horizontal line), we will use a simple measure of delay mechanism as the starting point for our analysis. Below we use the simple measure of our example above: the difference of the output statistic between RMT2 and RMT1. The approach below can however be replaced with any single or set of point estimates. Notice that we instead use a detailed process model, this might imply that we are considering the two time dimensions jointly (cf. above).

The study of the delay mechanism as a function of reference time is an important question: if we now assume that we have found a sound estimate for the change pattern of our simple delay estimate as a function of reference time (horizontal line), we can disseminate the more timely statistics based on RMT=1 and predict how much the estimate would change if we instead waited for RMT=2. We then would have increased the timeliness and we would also have an estimate of the loss of accuracy due to the increased timeliness. We notice here that only a month afterwards we will be able to confirm the quality of our prediction.

Consider the situation that we want to confirm our belief that, along the vertical line, there is a gain in waiting two instead of one month before doing the measurement. Our simple difference between the output statistics between RMT 2 months and RMT 1 month is just what we need in this situation, here we don't need any deeper understanding of the delay mechanism as a function of RMT (i.e. along the vertical axis).

In the situation above or in another situation, given our simple or sophisticated delay mechanism estimate, we study the sequence of estimates for each reference time (i.e. along the horizontal line). This sequence of estimates is the observations X that we are studying in the following sections.

In the analyses outlined above, we analyse along one of the time dimensions conditionally on the other time dimension and vice versa. It is our experience that it is rather common to confuse measurement time with reference time and vice versa. A mental image like that of Figure 1 may help the intuition to separate these dimensions.

3. Some considerations about delay

In the previous section we have assumed that any change in the output statistic due to using input registers at RMT=2 instead of RMT=1 as caused by delay, but this is an assumption that we will try to justify below. Further, we have used *delay* to mean *net delay from all sources together*, and below we will also elaborate on why we choose this approach.

3.1 Delay or other causes

Along the way from the employer's administrative system for employees and salaries, until the register-based employment statistics is disseminated, many errors may occur. Zhang (2012) has given an overview over these two errors along the two phases of the production system: the first phase is the construction of the administrative registers, typically outside the NSI. The second phase is the data integration phase where several sources are integrated into the output statistics.

Disentangling these errors empirically is challenging. However, what we need is somewhat simpler: to disentangling the changes of these errors between RMT 1 and RMT2, thus how the error changes just because the measurement time is taken as one month later. Consider the scenario that all errors are the same at these two measurement time except for the error due to delay. Then we have solved the problem and the change in the output variable between RMT 1 and RMT 2 is completely due to delay. So are these error constant over RMT? We assume that the production system is constant over RMT, the only change in the production between these two times is the input where delayed messages are included.

First we consider editing errors, i.e. errors added by NSI during the editing process. It is reasonable to assume that the editing errors are fairly unchanged between RMT 1 and RMT 2. One way that there could be a change is that the editing system might erroneously change the new observations. Another way is the following: an editing system will often be *parameter-driven*, meaning that parameters are estimated during the process and these estimates will decide the editing performed. The new observations might change these parameter estimates but we assume these changes not to lead to large changes in the editing process, at least not for mange reference dates. This assumption could be a topic for further studies but there is a problem that any change on the output variable "employee/not employee" being caused by change of editing, is difficult to assess: the derivation of the output variable is done late in the production process. Therefore, any effect of editing-change is only possible to measure on some very rough output variable proxy.

So, assuming that the output difference between RMT1 and RMT2 is due to delay, we need to make more precise what delay is. In our case where the output variable is register-based employment statistics, a delayed message is of three types defined by its consequences on the output variable: delays that implies that the corresponding person

1. moves from initial stage 0 "not employee" to 1 "employee"
2. moves from initial stage 1 to stage 0
3. moves back to the initial stage

Notice that type 3 requires at least three measurement times: the initial time (giving the initial stage), another time leading to the other stage, and a third time leading to a return to the initial stage. So, for our example with only two measurement times, type 3 is not possible.

The three types are all delays from a registration point of view. However, if we consider the process measuring an individual's registered employee status as function of measurement time, type 3 is not delay but error (Zhang & Fosen 2012, pp…**…..**).

A study of type 1 and type 2 separately is an interesting topic when studying the nature of the delay processes involved, and was studied in Zhang & Fosen (2012) for the main register source being available before a-ordningen was introduced: The *National employer/employee register* (NEER). Our focus is on the overall effect of delay onto the output variable and we therefore only consider *net delay*. For any other situation than RMT1 versus RMT2 this also means that type 3 is a part of the net delay. For NEER, Zhang & Fosen (2012) showed that the amount of type 3 is ignorable compared to type 1 and type 2 but this might be different for a-ordningen: whereas a-ordningen is a *cross-sectional register system* that requires all employers to report all employees for every reference time (ideally reported before RMT1 but often delayed), NEER was an event-based register system where the employers only reported changes in the employee relations between the employer and the employee.

We have now given arguments for why is it not unreasonable to consider the changes of the output variable from RMT1 to RMT2 as being caused by delay. In the next section we are then ready to model the effect of delay.

4. Method for studying the output quality caused by delay

We are now ready to model the effect of delay onto the output variable and we will continue to use the register-based employment as our example. In section 2(?) we outlined different modelling strategies when it comes to involvement of the two time dimensions 'reference time' and 'measurement time'. We will use a simple effect-of-delay estimator, the value at RMT2 versus RMT1, and thus we have fixed the measurement time dimension (or equivalently the RMT-dimension). We will to study this estimator along the reference time dimension, and could be aiming at three challenges:

1. what is the distribution of our simple effect-of-delay estimator?
2. how is the effect-of-delay mechanism (using our simple estimator) changing as a function of reference time, specifically we want to see whether the mechanism is constant over RMT.
3. based on these results we want to see whether we can use this pattern as a function of RMT to predict the effect of delay at measurement time 1 month after reference time (RMT1), when we could be tempted to start the production of the output statistics that we are going to disseminate. But if we were to disseminate based on RMT1, we would like to predict the effect of delay, i.e. how much we would have gained by waiting another month (RMT2).

Let $\hat{S}_{r,t}$ denote the estimated employment rate at time r when measured based on the updated register information at time t. Consider the difference in estimated employment rates between different measurement times t_1, t_2 ,

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

In order to model the effect-of-delay estimator, i.e. answer challenge 1 above, we will take as a starting point the initial assumption that the mechanism, measured by our simple estimator **(1),** has a constant expectation over the reference time, more specifically we assume that $X_{r,t_1,t_2}, r = r_1, \cdots r_n$ are iid, and thus $\mu_{r,t_1,t_2} = E\left(X_{r,t_1,t_2}\right) = \mu_{t_1,t_2}$ and $\sigma^2_{r,t_1,t_2} = Var\left(X_{r,t_1,t_2}\right) = \sigma^2_{t_1,t_2}$. This is a strong assumption that we will have to test below, but if it turns out to be a fair assumption, we are able to predict according to challenge 3 above.

Below we first want to test 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 .

4.1 Any gain in waiting until t_2 instead of t_1 ?

Should we produce the register-based employment statistics based on the registers updated until t_1 or should we rather wait for the registers updated at t_2 ? Is there a gain? Notice that this question moves along the vertical lines of Figure 1, but for answering this question we need to consider the distribution of X_{r,t_1,t_2} , i.e. move along the horizontal lines of Figure 1.

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 , $X_{r,t_1,t_2} \sim N(\mu, \sigma^2)$ and since the variance is unknown,

$$Z = \frac{(\bar{X}_{t_1,t_2} - \mu)}{\sigma / \sqrt{n}} \overset{Approx}{\sim} N(0,1) \ ,$$

but since the variance σ^2 is unknown, we have

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

where t_{n-1} is the student's T-distribution 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 $H_1 : \mathrm{E}\left(X_{r,t_1,t_2}\right) \neq 0$. Since t_{n-1} is a symmetric distribution, we should reject H_0 when $|U| > t_{n-1}^{1-\alpha/2}$ where $t_{n-1}^{1-\alpha/2}$ is the $1-\alpha/2$ -quantile in the t_{n-1} -distribution, and the significance level α is the probability for rejecting the test given H_0. The test then becomes to reject H_0 if

$$\overline{X}_{t_1,t_2} < t_{n-1}^{\alpha/2} SE\left(\overline{X}_{t_1,t_2}\right) \text{ or } \overline{X}_{t_1,t_2} > t_{n-1}^{1-\alpha/2} SE\left(\overline{X}_{t_1,t_2}\right) \quad (1.1)$$

The p-value is the smallest α that leads to a rejection of H_0 with our data and is the most interesting value in the test.

For our register-based employment status proxy described above (based on statistics data, L2), we have currently only seven observations of X_{r,t_1,t_2} : for reference period March 2016 until September 2016, and we have $t_1 = 1, t_2 = 2$ (relative measurement time at one and two months after the reference time. This is way too few observations to make any conlusions but we will outline the method which will be more interesting as we for each month get a new data point. We have $\overline{X}_{t_1,t_2} = -0.011$ and $SE\left(\overline{X}_{t_1,t_2}\right) = 0.0020$, giving $\overline{X}_{t_1,t_2} / SE\left(\overline{X}_{t_1,t_2}\right) = -5.422$ which is the 0.08 percentile of the t_{n-1} distribution and then $\alpha = 0.0008 * 2 = 0.0016$ is the smallest α leading to rejection and hence this is our p-value.

The 95% confidence interval $\left[\overline{X}_{t_1,t_2} \pm t_{0.975,n-1} \cdot SE\left(\overline{X}_{t_1,t_2}\right)\right]$ turns out to be $[-0.015, -0.006]$, and tells us that we are 95% confident that the expected gain in waiting does not exceed 1.5 percentage points.

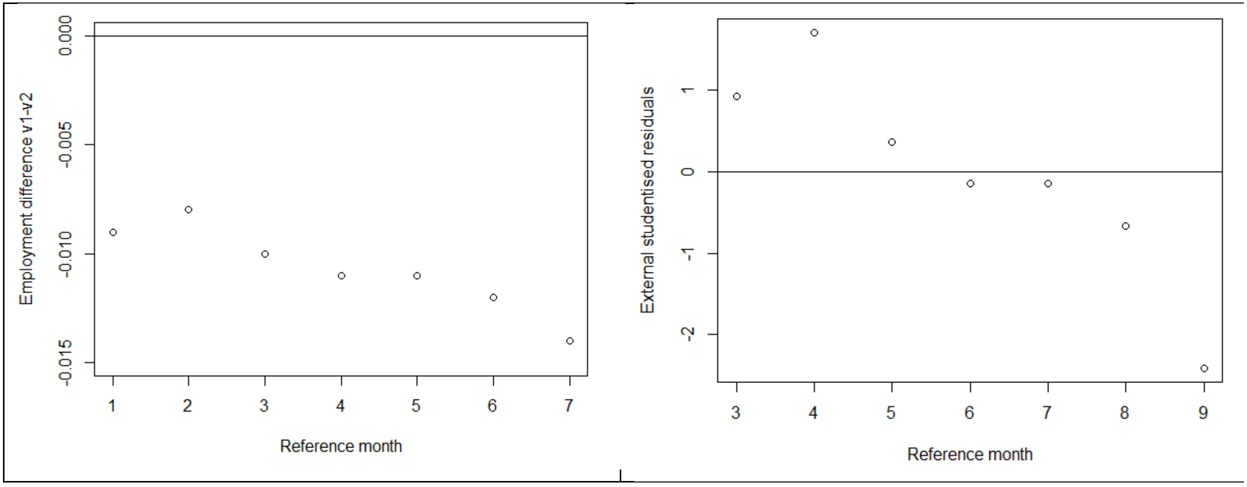


Figure 2. The difference between register-based employment between the relative measurement times 2 months and 1 month (left panel), and the corresponding exernal studentized residuals (right panel). Reference months 3-9 in 2016.

4.2 The iid assumption

The test and confidence interval in the previous section relies on the iid-assumption. Because of the low number of observations we in fact also needs each observation to be approximately normal distributed, but this assumption can be weakend as the number of observation increases. Below we will consider the iid-assumption.

4.2.1 The indentical distribution assumption

The left panel of Figure 2 shows the estimated differences $X_{r,1,2}$ between the employment rate measured two months after the reference date and one month after the reference date. Is it compatible with the identical distribution assumption?

The externally studentized residuals are $\varepsilon^{st}_{r,t_1,t_2} = \dfrac{X_{r,t_1,t_2} - \overline{X}_{t_1,t_2}}{\hat{\sigma}_{-r}\sqrt{1-h_{rr}}}$ where $\hat{\sigma}_{-r}$ is the estimated standard deviation of $\{X_{i,1,2}\}$ when datapoint r has been removed from the data, and h_{rr} in element rr of the *hat matrix* or the *projection matrix* $C(C'C)^{-1}C'$ where C is the design matrix which in our case is only the identity matrix with dimensions $n,1$ representing the intercept term since our model have no covariates. Since of response **X** is already assumed to be iid, our regression model has no covariates, only the intercept term. For our model we also have $h_{rr} = 1/n$. Studentized residuals are t-distributed with $n-p-2$ degrees of freedom, where p is the number of parameters in the regression model (Beckman and Trussel 1974). Thus we have $n-3$ degrees of freedom. So, if our model is correct the qq-plot of the ordered residuals against the corresponding quantiles of the t_{n-3} -distribution should resemble a straight line. We notice in Figure 3 that the this approximately seems to be the case.

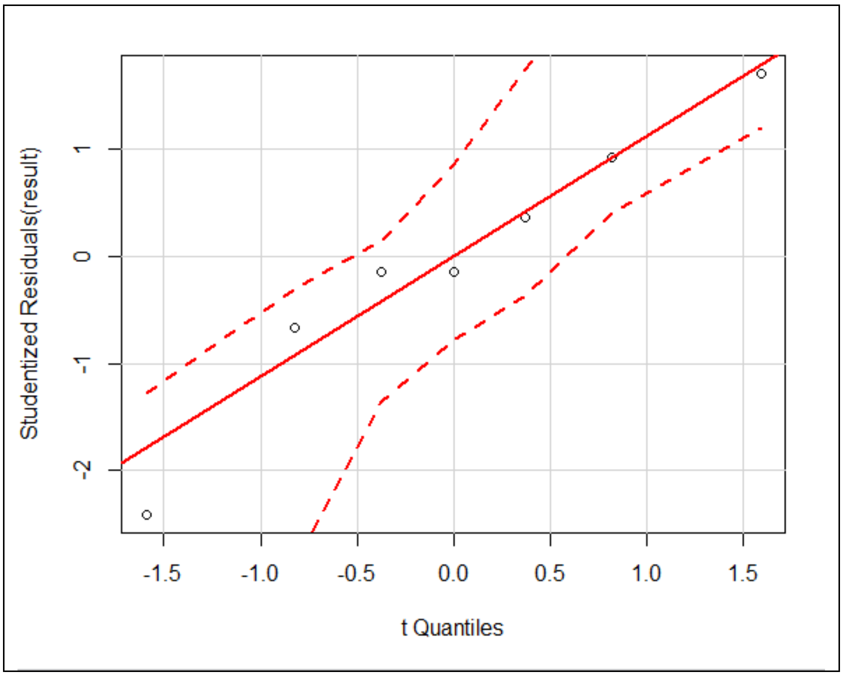


Figure 3. Quantile-quantile plots of the external studentized residuals against the t-distribution.. Using r-function qq.plot (…check whether this R-function uses the external studentized residuals and the t-distribution….)

A formal test for confirming that the residuals are t-distributed would be based on the null hypothesis that the data are not from the same distribution. Such test does not exist, and we have to use a test with H_0 is "identical distribution". The Kolmogorov-Smirnoff test (ks.test in R) compares the cumulative distribution (cdf) function of the residuals with cdf of the t-distribution by considering the maximum vertical distance between the two cdf's. The test is based on the null hypothesis that the data *are* from the identical distribution. This test gives a p-value of 0.99, thus giving rejection of the null hypothesis for any choice of significance level up to 0.99, indicating that the null hypothesis is true. However, to give a more precise answer to that question, we would have to consider the power function, i.e. the probability of rejecting the null hypothesis for a certain value of "not identical distributed" given our test at a reasonable significance level such as 0.05.

4.2.2. Assumption of uncorrelated residuals

We can not test for independence between the data, but there are several tests for rejecting an assumption of no autocorrelation between the residuals. One simple test is Durbin-Watson-test (dwt in R, under the car package). H_0 is that there is no autocorrelation and the test statistic is simply $\sum(\varepsilon_{t+1}-\varepsilon_t)^2 / \sum \varepsilon_t^2$ and is discussed in e.g. Chatfield (2003; p. 69). For our data we get a p-value of 0.01 which is a strong indication of autocorrelation. Another test is the Ljung-Box test (Ljung & Box 1978) which for our data gives a p-value of 0.21 (Box.test in R, specify type="Ljung") and thus indicating that there might very well be no autocorrelation.

Acknowledgements

Many thanks especially to Li-Chun Zhang for very valuable discussions on methodology. Many thanks also to Ole Villund for his role in explaining the data and discussing the mechanisms behind the a-ordningen data. Further Ingvild Johansen and Sølve Mikal Nerland are thanked for helping in finding the data and explaining the variables. Thanks also to Håkon Grini for providing background insight into the a-ordningen system, and to Christoph Waldner at Statistics Austria for reviewing this document.

References

Chatfield, M. (2003), The Analysis of Time Series: An Introduction, Sixth Edition, Chapman & Hall/CRC Texts in Statistical Science)

Beckman, R.J. & Trussel, H.J. (1974), The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple Regression, JASA, Vol 69, No. 345, pp. 199-201.

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.

LJung, G. M. & Box, G. E. P. (1978), On a measure of lack of fit in time series models College of Business Administration, Biometrika, vol. 65, no. 2, pp. 297-303.

Zhang, L.-C. & Fosen, J. (2012), "A modeling approach for uncertainty assessment of register-based small area statistics", Journal of the Indian Society of Agricultural Statistics, vol. 66, pp. 91-104.

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

ST1_1 Suitability Test of Employment Rate for Employees (Wage Labour Force) (ERWLF)

A certain amount of businesses are classified wrongly with regards to activity codes in the business register (BR). Wrong classifications affect the published figures from the ERWLF.

We will perform a suitability test, in order to examine the quality of the ERWLF as a function of the quality of the BR. The test is intended to reveal the margin of error of number of wage laborers within activity groups due to wrong NACE classifications in the Business Register (BR). We will make simulations, where we will simulate distributions of activity codes and examine the effect on published figures from the ERWLF.

Background

The Employment rate for Wage Labor Force (ERWLF) statistics is an indicative statistics reflecting the activity in the Danish labor market. The purpose is creating a fast and efficient indicator rather than creating precise numbers of the Danish employment (labor force). All income taxes from wage earners in Denmark are collected through the E-income register.

The results from the ERWLF are presented on aggregated levels of activity groups. The monthly statistics is published on level 0 with 10 groups. The quarterly statistics is published on level 1 with 20 activity sections.

- Level 0: 10 sections identified by numbers 1 to 10.
- Level 1: 19 sections identified by alphabetical letters A to S. There are 20 sections if X, unknown activity, is included.

Activity is linked to ERWLF through the BR. In order to secure possible reproduction of the statistic, it is based on a frozen version of the BR. The early publication of ERWLF forces the statistic to use the preliminary version of the BR.

Output: Number of Danish wage earners, in total, by NACE activity groups and by regional breakdowns.

Sources: E-income register, population register, business register (BR).

- The E-income register is the backbone of the ERWLF statistics.
- The BR is used to connect business activity with businesses, so it is possible to calculate number of wage earners on business activity groups.
- The first preliminary frozen version of BR is used.
- 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, that will require an intensified use of the population register.

Undercoverage:

- Self-employed persons who never get their wage through e-income.
- Other wage earners who do not get their wage through e-income. Danish wage earners working in e.g. Sweden or Norway do not get their wage through e-income and are not covered in the ERWLF statistics.
- Unreported work. Due to high taxing level unreported work is somewhat common in Denmark, but only a very small fraction of the working force is believed to work fully unreported and therefore not covered by the e-income register at all.

Overcoverage:

Non-Danish employees payed by Danish employers through E-income register. These wage earners are by definition counted with in the ERWLF figures, even though they strictly taken not are Danish wage earners.

Misclassifications:

When businesses are founded, the businesses themselves complete NACE classifications in the BR register. There is 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.

Small businesses with less than one full time employee (FTE) can be wrongly classified for a long period or even permanently. The quality of 'medium size businesses' (2-10 employees), is better than that of small businesses, but does still contain a proportion of misclassifications. For 'larger businesses' (More than 10 employees), it is believed that misclassifications always will be corrected at some point in the BR. All larger businesses participate in a series of surveys conducted by Statistics Denmark and NACE misclassifications will eventually be discovered when business are selected for a survey, it not is meant to be a part of. Misclassifications for larger businesses, is primarily due to delays in updatings, when businesses merge, split or change activity. Updating of NACE codes in the BR is sometimes, but not always aligned with events of this kind.

Table 1. Number of enterprises and employees by size of enterprise (Enterprises with employees in march 2016).

	Number of enterprises	Number of employees	Share of enterprises	Share of employees
1 or fewer employees	51,961	44,274	27.4%	1.7%
2 - 10 employees	91,189	406,424	48.2%	15.6%
More than 10 employees	46,156	2,160,508	24.4%	82.7%
Total	189,306	2,611,206	100%	100%

Table 1 shows that by far the most wage earners are employed at enterprises with more than 10 employees. ‘Small businesses’ (0-1 employees) only represent a small fraction of the total and only 1.7% of the employees work at enterprises with 1 or fewer fulltime employees. Hence the relatively large misclassification of small business does only have a limited influence on totals.

Statistics Denmark has previously, conducted surveys for measuring the quality of NACE codes in the Danish BR. These surveys are not up to date, but might still give a good impression of the correctness of the NACE codes. Feedback from both continuous and ad-hoc surveys is used to correct information in the BR. Larger enterprises participate in more surveys and the information on larger enterprises is in general of better quality than the one on smaller enterprises.

Delays in updates in the BR

In some cases businesses are split into smaller businesses or merged with other businesses. In other cases businesses simply change main branch. Updates in BR are not always performed at exactly the same moment as they occur. Misclassifications caused by timeliness issues are included in misclassifications mentioned earlier and will not be treated separately.

By using the old surveys and rates of corrections from newer surveys Statistics Denmark has estimates of the level of misclassifications over time (t) of the NACE codes split on size of businesses.

Frozen versions of the BR are saved, so it is possible to see the effect of corrections over time. This is sometimes referred to as the progressiveness of the register. It is assumed that the Business Register never reaches perfect classifications, but larger enterprises participating in national surveys are expected at some time the achieve correct classification in the BR. We will examine impact on results from the ERWLF using older frozen versions and compare with results when using newer and more updated versions for the same reference period, always assuming that the most updated versions are the most valid ones.

Handling of quality issues in ERWLF

Clearly there are quality issues regarding ERWLF that can be addressed. Since ERWLF primarily is meant to be a fast indicator of the activity in Denmark, many of these issues are dealt with in the definition of the ERWLF. The E-income register cover all “normal” payments of wage laborers in Denmark. The construction of the E-income register is solid and there is no reason to believe that the E-income register will cover more or less payments in near future. Instead of trying to make ERWLF cover unreported work or other undercovered labor, the ERWLF is simply defined as ‘The number of wage laborers receiving payment through the E-income register’. Hence by definition ERWLF does not suffer from neither undercoverage nor overcoverage. Solving the coverage issue by definition of the statistics can be considered as a too simplistic solution. On the other hand quantification of coverage problems on ERWLF is very difficult if not impossible to estimate. Due to taxes all (legal) wage payments in Denmark are payed through the E-income register, so it does make perfectly good sense, to define the number of persons in ERWLF as the total number of persons receiving wages through the E-income register.

Accuracy simulations

The ERWLF statistics results in total number of employees and number of employees by NACE sections. When all data are gathered the total number of employees is fixed. We will perform simulations that can assess the impact from NACE misclassifications on output estimates in the ERWLF. We will also perform sensitivity analysis investigating the effect of misclassification rates being higher or lower than the rates estimated by Statistics Denmark.

Table 2. Number of wage earner by activity section in Denmark in March 2016.

Activity section	Description	Number of wage earners
A	Agriculture, forestry and fishery	38,268
B	Mining and oil industry	4,394
C	Industry	291,214
D	Energy	10,231
E	Water and renovation	11,139
F	Construction	144,467
G	Trade	408,681
H	Transport	136,453
I	Hotels and restaurants	98,619
J	Information and communication	100,471
K	Finance and insurance	78,691
L	Real estate and rent	36,811
M	Knowledge based services	140,754
N	Travelling, cleaning and other operations services	141,088
O	Public administration, defense and police	138,963
P	Education	230,469
Q	Healt and social security	491,261
R	Culture and leisure	49,482
S	Other services	59,430
X	Unknow activity	319
Total		2,611,205

Proportion of enterprises with wrong activity codes

Once all data are available the total number of wage earners is fixed since by definition the total number of wage earners equals the total number of persons in the E-income register. Wrong NACE classifications on the other hand will lead to wrong figures on number of wage earners by activity section. With known fractions of wrongly coded enterprises in the BR it is possible to simulate 'likely distributions' of wage earners by activity section.

Accuracy simulations

Data from March 2016 are used to simulate distributions of wage earners in Denmark by activity section. There were 189,306 enterprises paying salaries to 2,611,206 wage earners. The enterprises are grouped at a 20 category level with number of wage earners ranging from 4,394 to 491,261 in each category (Table 2).

Simulation model:

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

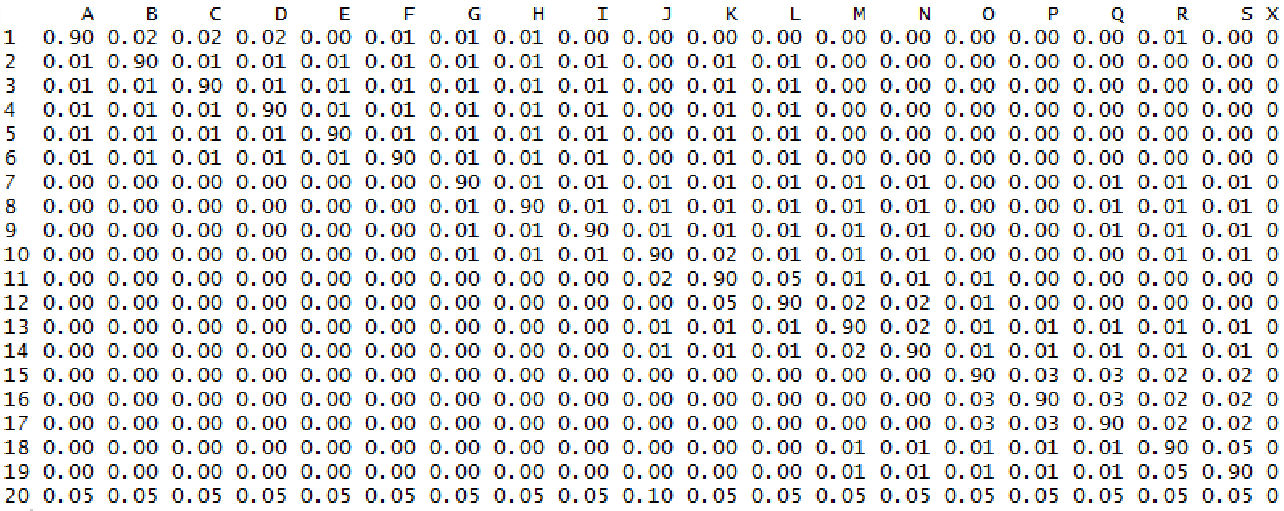
Statistics Denmark has conducted three so called recoding projects where the aim was to correct NACE codes. Two surveys with 3,000 enterprises in 2006 and 2009, and one survey with 50,000 enterprises in 2007. Even though no surveys have conducted with the same purpose since 2009, the impression from 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. The estimated proportions of misclassifications within size group of business and NACE aggregation level can be found in Table 3.

Table 3. Proportion of missclassified businesses by size of business in Full Time Employees (FTE) and aggregated level of business activities.

	Aggregation level of business activities	
Size of business (FTE)	Level 0 (10 cat.)	Level 1 (20 cat.)
Enterprises with 1 FTE or less	8%	10%
Enterprises with 2 to 10 FTE	6%	8%
Enterprises with more than 10 FTE	3%	4%

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 business is required. Each row in a confusion matrix will add to 1 and the values in the diagonal reflect the probability of correct coding within each activity group, derived as complement of the corresponding figure in Table 3. Table 4 shows the confusion matrix for smaller enterprises and level 1 NACE classification.

Table 4. Confusion matrix used for simulation between business sections.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	X
1	0.90	0.02	0.02	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0
2	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
3	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
4	0.01	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
5	0.01	0.01	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
6	0.01	0.01	0.01	0.01	0.01	0.90	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0
8	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0
9	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.90	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0
10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.90	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.90	0.05	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.90	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.90	0.02	0.01	0.01	0.01	0.01	0.01	0
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.90	0.01	0.01	0.01	0.01	0.01	0
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.03	0.03	0.02	0.02	0
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.90	0.03	0.02	0.02	0
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.90	0.02	0.02	0
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.90	0.05	0
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.05	0.90	0
20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.10	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0

The confusion matrix in Table 4 is constructed by letting the diagonal be the proportion of correct classified enterprises. The sum of each row equals 1 and corresponds to the probabilities for correct business sections. This transition matrix is used for simulating small enterprises on a 20 group level. Note the 0.9 in the diagonal corresponds to estimated proportion of smaller business that are classified correctly (Table 3) Hence the remaining probability is distributed in 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. Note that the last column (X = unknown activity) is zero. No enterprise correctly belongs to "unknown", hence the zeros.

A simulation study has been performed where the activity section is simulated by using the observed activity and confusion matrices as the one in Table 4. When all enterprises have a simulated activity code, simulated number of wage earners can be calculated in the same way as in the public figures. In the case where 0.9 is in the diagonal for enterprises with 1 or fewer FTE's, confusion matrices with 0.92 and 0.96 are used for enterprises with 2-10 FTE's and more than 10 FTE's respectively.

Fout! Verwijzingsbron niet gevonden. shows results from the simulations. The coefficient of variation varies between 0.9% and 13.2% with a relationship towards higher coefficient of variation for businesses with few wage earners (Figure 1). In the simulations all enterprises with unknown activity have been placed in other activity sections and hence the simulated number of wage earners with unknown activity always becomes zero.

Table 5. Results from simulation study of estimated number of wage earners by business section with basic proportion of correct classified enterprises equals 90% for small enterprises.

Activity section	Official figures	Std dev	Coeffcient of variation	5% quantile	95% quantile
A	38,268	1,246	3.3%	37,307	41,974
B	4,394	579	13.2%	2,488	4,490
C	291,214	3,562	1.2%	244,662	258,059
D	10,231	542	5.3%	8,043	10,050
E	11,139	467	4.2%	9,461	11,133
F	144,467	1,731	1.2%	128,678	135,190
G	408,681	3,484	0.9%	385,552	400,527
H	136,453	3,388	2.5%	120,790	133,705
I	98,619	2,076	2.1%	98,629	106,300
J	100,471	2,508	2.5%	93,355	103,324
K	78,691	2,376	3.0%	77,041	86,227
L	36,811	2,379	6.5%	54,884	64,748
M	140,754	2,786	2.0%	143,532	154,776
N	141,088	2,750	2.0%	132,085	142,471
O	138,963	3,479	2.5%	123,074	136,463
P	230,469	4,653	2.0%	219,074	238,078
Q	491,261	6,904	1.4%	467,298	495,908
R	49,482	3,254	6.6%	65,945	78,654
S	59,430	4,422	7.4%	98,701	116,425
X	319	0	0.0%	0	0

In order to study the sensitivity of the simulations confusion matrices with 0.85 and 0.95 correct classification proportions for small enterprises are also performed. Proportions of correct classified enterprises in other size groups are adjusted proportionally in the same directions. Results from the simulations can be seen in

Table 6. When percentage of wrong classified is doubled (correct changed from 90 to 95% for small enterprises), the standard deviation is reduced by just short of factor of square root of two in average. Seemingly the condition of a constant total of number of wage earners reduces the variation, compared with a the variation in a simple random sample.

Table 6. Standard deviations for simulated number of employees within NACE 19 group classification with three different base probabilities of correct classification.

NACE 19 classification group	85% correct classified	90% correct classified	95% correct classified
A	1,457	1,246	884
B	666	579	407
C	4,108	3,562	2,479
D	684	542	433
E	598	467	363
F	2,042	1,731	1,280
G	4,492	3,484	2,652
H	4,013	3,388	2,639
I	2,384	2,076	1,364
J	3,072	2,508	1,904
K	2,910	2,376	1,775
L	3,028	2,379	1,751
M	3,417	2,786	2,046
N	3,215	2,750	1,929
O	4,084	3,479	2,427
P	5,737	4,653	3,343
Q	8,585	6,904	4,989
R	3,902	3,254	2,284
S	5,577	4,422	3,222
X	0	0	0

Comparison of frozen versions

The Business Register (BR) is a living register, where information continually is updated. That means that the same query in the BR today can give another result tomorrow, e.g. because of business closure or reclassification of NACE group. Published statistics are always based on frozen versions of the BR. This makes it possible to reconstruct published statistics, by using the same frozen version as the one used originally. A change in the preliminary frozen version of the same period is triggered once information about enterprises is corrected back in time in the BR.

Table 7. Number of enterprises that have changed activity compared to the preliminary version of the BR and number of employees in these enterprises. The first two lines are changes in frozen versions of the same period, while the next two lines are changes from the preliminary frozen version in the given quarter to the preliminary frozen version in the following quarter.

Year	2014				2015				2016	
Quarter	1	2	3	4	1	2	3	4	1	2
Changes from preliminary version to final frozen version										
Enterprises	3,479	7,656	3,984	2,001	3,645	3,652	2,478	3,757	4,789	3,029
Employees	5,443	9,421	5,930	5,018	7,045	8,417	6,575	7,061	15,507	3,745
Changes compared to next final frozen version										
Enterprises	18,387	23,648	17,297	20,862	15,888	16,620	18,143	24,656	19,591	19,318
Employees	22,950	18,818	14,654	19,954	17,737	17,737	17,601	22,189	27,099	13,748
Ratio between corrections from preliminary to final frozen and changes to next final frozen version										
Enterprises	18.9%	32.4%	23.0%	9.6%	22.9%	22.0%	13.7%	15.2%	24.4%	15.7%
Employees	23.7%	50.1%	40.5%	25.1%	39.7%	47.5%	37.4%	31.8%	57.2%	27.2%

The ERWLF uses the preliminary frozen version of the BR. In Table 7 the numbers of enterprises that have changed activity from the preliminary version to the last frozen version are shown. Below is the number of enterprises changing activity from one period to another activity the following period.

The BR is updated more than three months back, so an update in the final version of a frozen version compared to the preliminary version, does not always compare to a change from one quarter to another. The frozen versions of the BR are corrected approximately one year back, compared to the preliminary version.

On the other hand if an update from one period to another is not registered within three months, but within a year it will always result in a change from the preliminary version to the final version. Hence the large amount of changes from one period to another reflects that by far the most changes of the BR are made within a time period, so that the first preliminary version is correct. The ratios between corrections of frozen versions and changes between preliminary versions are between 10 and 32% of the enterprises and 24 and 57% of the employees. These figures reflect that a fairly large part of the changes in the BR are corrected in time for the quarterly preliminary version to be updated with changed activity codes.

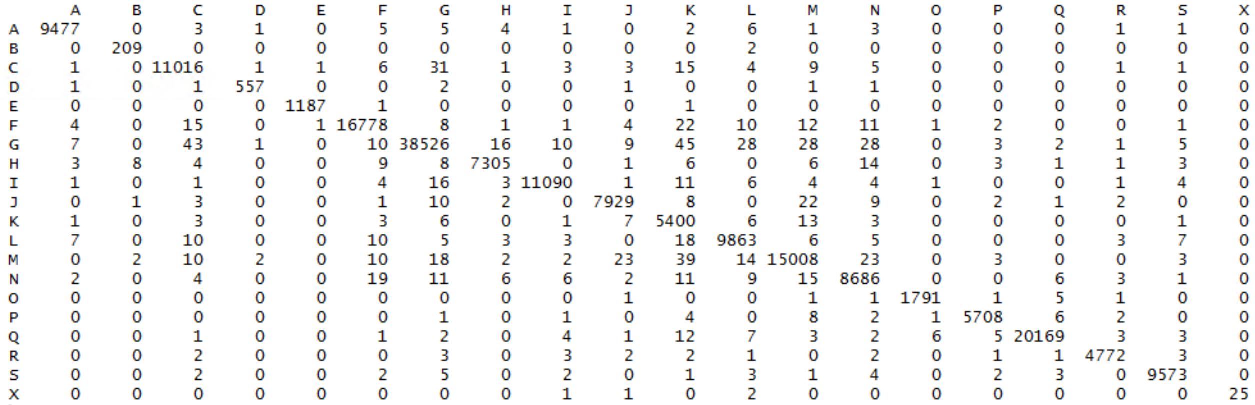
Conclusion

The total number of wage earners is not affected by wrong business classifications, but on activity section level official figures are affected by wrong NACE classifications. If there are estimates of the fractions of enterprises with wrong classifications we have demonstrated a method to calculated standard deviations and variation of coefficients of official figures as a function of wrong NACE classifications. Previous surveys conducted at Statistics Denmark have provided figures that can be used as input for these simulations.

Comparisons of frozen versions show that by far the most changes of activity the BR are made quickly after the change. By comparing changes of frozen versions over the same period with changes between different periods it is possible to conclude that most changes are made in time to be corrected already in the first preliminary version on a quarterly level.

Appendix. Other tables with changes in frozen versions.

Table 8: Number of enterprises in BRWLF split on NACE branches in first and last frozen version of BR.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	X
A	9477	0	3	1	0	5	5	4	1	0	2	6	1	3	0	0	0	1	1	0
B	0	209	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
C	1	0	11016	1	1	6	31	1	3	3	15	4	9	5	0	0	0	1	1	0
D	1	0	1	557	0	0	2	0	0	1	0	0	1	1	0	0	0	0	0	0
E	0	0	0	0	1187	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
F	4	0	15	0	1	16778	8	1	1	4	22	10	12	11	1	2	0	0	1	0
G	7	0	43	1	0	10	38526	16	10	9	45	28	28	28	0	3	2	1	5	0
H	3	8	4	0	0	9	8	7305	0	1	6	0	6	14	0	3	1	1	3	0
I	1	0	1	0	0	4	16	3	11090	1	11	6	4	4	1	0	0	1	4	0
J	0	1	3	0	0	1	10	2	0	7929	8	0	22	9	0	2	1	2	0	0
K	1	0	3	0	0	3	6	0	1	7	5400	6	13	3	0	0	0	0	1	0
L	7	0	10	0	0	10	5	3	3	0	18	9863	6	5	0	0	0	3	7	0
M	0	2	10	2	0	10	18	2	2	23	39	14	15008	23	0	3	0	0	3	0
N	2	0	4	0	0	19	11	6	6	2	11	9	15	8686	0	0	6	3	1	0
O	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1791	1	5	1	0	0
P	0	0	0	0	0	0	1	0	1	0	4	0	8	2	1	5708	6	2	0	0
Q	0	0	1	0	0	1	2	0	4	1	12	7	3	2	6	5	20169	3	3	0
R	0	0	2	0	0	0	3	0	3	2	2	1	0	2	0	1	1	4772	3	0
S	0	0	2	0	0	2	5	0	2	0	1	3	1	4	0	2	3	0	9573	0
X	0	0	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	25

Table 9. Comparisons of preliminary version of BR and final frozen version for 1st quarter 2016. Only enterprises with wage earners are counted.

	A	B	C	D	E	F	G	H	I	J
Wage earners first frozen version	38268	4394	291223	10231	11139	144479	408717	136453	98626	100471
Wage earners last frozen version	38292	4535	291173	10214	11140	144491	408922	135585	98659	100239
Enterprises first frozen version	9687	214	11252	571	1197	17049	39528	7536	11527	8132
Enterprises last frozen version	9681	223	11272	569	1197	17037	39423	7507	11508	8127
Change in number of wage earners	0.1%	3.2%	0.0%	-0.2%	0.0%	0.0%	0.1%	-0.6%	0.0%	-0.2%
Change in number of enterprises	-0.1%	4.2%	0.2%	-0.4%	0.0%	-0.1%	-0.3%	-0.4%	-0.2%	-0.1%
	K	L	M	N	O	P	Q	R	S	X
Wage earners first frozen version	78693	36811	140756	141089	138963	230469	491287	49482	59430	319
Wage earners last frozen version	79298	37080	140564	141366	139068	230395	491291	49534	59145	310
Enterprises first frozen version	5538	10029	15314	8926	1878	5830	20533	4833	9698	49
Enterprises last frozen version	5691	10050	15293	8948	1877	5827	20508	4832	9706	45
Change in number of wage earners	0.8%	0.7%	-0.1%	0.2%	0.1%	0.0%	0.0%	0.1%	-0.5%	-2.8%
Change in number of enterprises	2.8%	0.2%	-0.1%	0.2%	-0.1%	-0.1%	-0.1%	0.0%	0.1%	-8.2%