**Confidence intervals for register-based statistics - A case study for the Austrian register-based labour market statistics**

Johannes Gussenbauer, Statistics Austria, Johannes.Gussenbauer@statistik.gv.at

Alexander Kowarik, Statistics Austria, Alexander.Kowarik@statistik.gv.at

Eliane Schwerer, Statistics Austria, Eliane.Schwerer@statistik.gv.at

Christoph Waldner, Statistics Austria, Christoph.Waldner@statistik.gv.at

**Abstract**

*Most of the time register-based statistics are published without quantitative measurements for the estimation error, e.g. a confidence interval. Based on an indicator proposed in the European project ‘Quality of multi-source statistics’ (KOMUSO), a resampling method is applied to the Austrian register-based labour market statistics with the goal to quantify the model error of the imputation process and the measurement error of the target variable in the administrative data. A survey data set is used to benchmark this measurement error and estimate transition probabilities to different states in the categorical target variable. A resample method is then applied to generate repeated estimates which are used to compute confidence intervals*

**Keywords:** Variance Estimation, Bootstrapping, Register Data

**1. Introduction**

In the classical sample theory the computation of confidence intervals or standard errors is established and well-known. In times of cost-saving the use of registers as direct input sources for statistics becomes more and more important. Since errors in administrative date differ from those in surveys, the big challenge arises to develop tools to measure quality in statistical figures similar to error estimates in sample theory. This challenge is an international one. In recent years many countries have developed different approaches to measure quality. Moreover multinational cooperation projects like ESSnet KOMUSO (“ESSnet on Quality of Multisource Statistics” 2017) are underway on these issues.

The presented article describes an approach to estimate confidence intervals or standard errors of register-based statistics by using only information on the classification and imputation errors. The used method is based on techniques presented in Delden, S., and J. (2016 where the authors present a methodology to estimate the variance and the bias of the total turnover of the industry “car trade”. We extend these methods by introducing an “imputation error” term to handle different error sources. Since the figures of the method can be biased, we attempt to correct them by an “inverse mapping” process. In contrast to Delden, S., and J. (2016) we propose an inverse mapping process which suffers less from a possible variance inflation of the bootstrap replicates. As a case study we apply the techniques to the final result of a register based statistics, the Austrian register-based labour market statistics (RBLMS). The results of this example will illustrate the method. The RBLMS is an annually multi-source statistics with characteristics of a population census. We focus on the variable highest level of education completed (EDU).

The paper is organized as follows. In section 2 we prepare the notation and the necessary theory on imputation, classification and their corresponding errors. In addition we address various difficulties when applying the proposed method. Section 3 is devoted to the results of the case study. The article is finished by some conclusions and outlook in the section 4.

**2. Methodology**

Consider a population , containing population units , , where each unit holds a set of characteristics . The characteristics in can take on values , or be unknown, e.g *missing values*. In the following we will denote as the value for characteristic of unit . Characteristics in a register are usually subject to classification errors. For instance information on main residence in the central population register does not always reflect the actual living reality. In this work we will only address classification errors and imputation of unknown values for characteristics as these are two of the main sources of errors. Other error sources like changes in population size, which might also have a significant effect, are not considered. Classification as well as imputation errors are assumed to be dependent on each other.

*2.1. Imputation Error*

Missing values usually occur in real life data sets and are in general challenging to deal with. The structure of missingness in data can be classified by three categories (see (Rubin 1976)).

Determining the type of missingness for a given data set is not always possible and in many cases only based on assumptions. For the work presented here we assume that the missing cells in the data are MAR. To be more precise, given that unit has a missing cell in characteristic , the true, but unobserved, value for can be estimated by a set of fully observed characteristics . The imputed value can thus be estimated via

(1)

*2.2. Classification Error*

We assume that the observed values for characteristic are subject to random classification errors that appear independently across units, meaning that each person has an unknown true and observed value for , and respectively, ,. The errors are assumed to follow a known or estimated transition matrix , with .

Applying the transition Matrix on the observed characteristics results in a new vector of characteristics , defined by

(2)

*2.3. Errors of population totals*

For characteristic we are interested in the population totals for values of ,

Population totals can easily be estimated through

where for unit , was imputed beforehand if it was unknown.

Estimating the confidence intervals or standard errors of this point estimate can be quite challenging since uncertainty stems from two different error sources which can be dependent on each other. Due to the complexity of these estimations it is reasonable to apply a bootstrap sampling procedure. First, the imputation procedure is applied on the data using equation (1) resulting in a vector of characteristics , . Afterwards the true but unobserved characteristics are estimated using the estimated transition matrix , thus resulting in a vector of characteristics . For the last step the transition probabilities are replaced by

The above procedure is applied times, resulting in a set of so called bootstrap replicates for large . With each bootstrap replicate one can estimate via

(3)

Using the bootstrap replicates it is straight forward to estimate standard errors or confidence intervals, see (Efron and Tibshirani 1986).

*2.4. Imputation within register*

Imputation in this section refers to item imputation which is the insertion of artificial but plausible information into a data record with a missing value in this specific attribute.

The information concerning the highest level of education completed is taken from the Register of educational Attainment which is an in-house register of Statistics Austria. The register of educational attainment contains information on the formal education degrees of the Austrian population of the age of 15 years and older. It was first filled with data on the highest educational attainment from the population census 2001 and has been updated since with degrees achieved at Austrian schools and universities as well as with data reported by the economic chamber and the agricultural chambers (apprenticeship examination, master craftsman examination), and the Federal Ministry of Health (diploma of perfusionists). In addition, the Public Employment Service annually informs about the educational attainment of the recipients of benefits. The educational attainment of immigrants who came to Austria after 2001 is known if they acquired a degree in Austria afterwards, if an academic degree was recorded in the central population register, if a foreign graduation was recognized, or if a person is a benefit recipient from the Public Employment Service.

However, for about 5 percent of the Austrian population with reference day 31 October 2015 the information about the highest level of education is unknown (mainly immigrants). To close this gap different imputation models based on other information were developed. First of all, the educational attainment is imputed based on information on school or university enrollment, if applicable. Secondly, the imputation is predicated on information concerning the current activity status (e.g. status in employment, branch of the employment). Finally, if the person in neither enrolled on school/university nor an economically active person, the imputation model is based on demographic and other information which has an impact on the highest level of education.

*2.5. Modelling with sample data*

The transition matrix for the classification error will in most cases be unknown and thus needs to be estimated beforehand. Delden, S., and J. (2015) have proposed a set of assumptions for calculating the matrix of classification error probabilities when dealing with classification errors for industry codes in a business register. Delden, S., and J. (2016) later relaxed those assumptions and suggested a methodology for estimating the transition matrix using an audit sample. In this work we also focus on using a survey for estimating . The Austrian labour force survey was chosen as audit sample.

Let be the characteristic of interest from population and a sample from , where the characteristic was surveyed. Using the observed values , , as values for , makes it straight forward to calculate the classification error of on the subset . For all units in the classification error can be estimated using, for example, a multinomial regression model

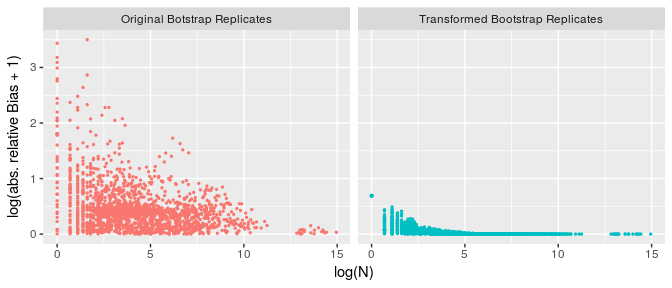
with as a set of independent variables taken from . Since was a sample it is reasonable to include the corresponding sampling weights into the model calculations. It should be noted, that using a statistical model is in practice only feasible if the sample is very large or the number of values in small. have proposed a methodology to estimate the transition matrix for small sampling size and large number of values in . In our case study however, we used only logistic regression models to estimate since the underlying sample was fairly large and the observed characteristics only held a small number of values. The set of modelling parameter was taken similar to the set of parameters used for imputation. Applying multinomial regression yields a transition matrix , dependent on the observed values of in the register and the auxiliary information in :

with . That is, the probability of changing to classification , given auxiliary variables and classification observed in the register. Using in equation (2) makes it straight forward to calculate the bootstrap replicates defined in equation (3). In addition to modelling the transition probabilities we also used a naive bootstrap as benchmarking approach, e.g a bootstrap over the whole population and afterwards aggregated over characteristic .

*2.6. Bias of bootstrap replicates*

(Delden, S., and J. 2015) have shown that the bootstrap replicates can be biased. This, for instance, will always occur if the expected number of units that change for characteristic , from to a different value is unequal the number of units that change from other values of to . Depending on the severity of the bias this can even lead to confidence intervals which do not enclose the original point estimate. We as well experienced bias in bootstrap replicates for our application, which was in some cases very severe. Figure 1 shows on the left hand side the distribution of relative bias when of the bootstrap replicates, for all results. The x-axis was transformed by the logarithm for better readability. It has to be noted that the magnitude of the bias results from the systemic differences regarding the education level asked in the Austrian LFS and the education level maintained in the register.

**Figure 1. Distribution of relative bias for bootstrap replicates before and after bias reduction.**



(Delden, S., and J. 2016) have shown that it is possible to construct unbiased bootstrap replicates by using the inverse transition matrix . However this approach can lead to an increase in estimated variance and widen estimated confidence intervals. In our case study we did not estimate the inverse of directly but rather constructed an average transition matrix , depending on and additional characteristics . For our results the characteristic of interest was the education level and results were aggregated per gender, nationality and gender by nationality. Given a set of values the element of the average transition matrix is then defined by

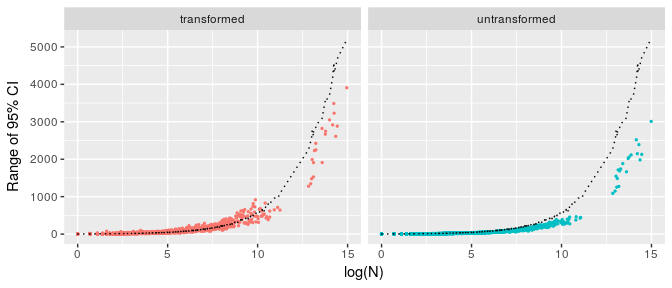
So for each set of values we average over the transition probability matrices of all units in population which take on the same values for characteristics . Calculating the inverse and applying it on the aggregates of each bootstrap replicate yields which are the unbiased estimates of :

where denotes the transposed matrix of . Figure 1 shows on the right hand side the relative bias when is applied after creating the bootstrap replicates.

*2.7. Variance inflation*

Although the bias can be greatly decreased, using can lead to an increase in estimated variance or widen confidence intervals. In Figure 2 the range of the confidence intervals using , left side of the panel, and confidence intervals of the not transformed bootstrap replicates, right side of the panel, are shown. The x-axis presents the number of observations, transformed by the logarithm, where for the not transformed case the mean over all bootstrap replicates was used. The dotted line represents the range of confidence intervals when using the naive bootstrap procedure instead of estimating transition probabilities. Looking at the right side of Figure 2 one can see that, disregarding the fact that the bootstrap replicates are biased, the confidence intervals are smaller when using transition probabilities matrices instead of the naive bootstrap procedures. The decrease in confidence intervals can be interpreted as the descriptive power of the model used to explain the differences between the educational level in the register and the Austrian LFS. Looking at the left hand side of 2 we see that the confidence intervals have increased and are in many cases larger then the ones we would receive using a naive bootstrap procedure. It has to be noted, that naive bootstrap replicates seems to already yield acceptable, although conservative, confidence intervals.

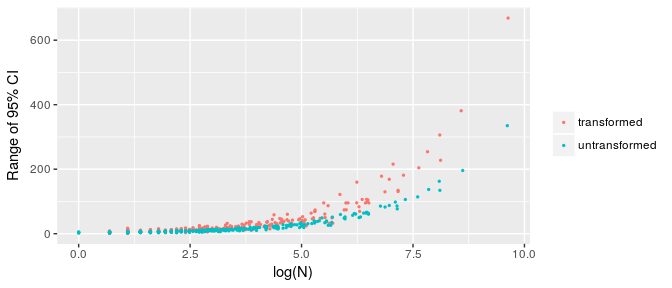
**Figure 2. Range of confidence intervals before (left side of panel) and after bias reduction (right side of panel), where the x-axis represents the logarithm of the expected totals. The dotted line shows the results achieved by using a naive bootstrap procedure.**



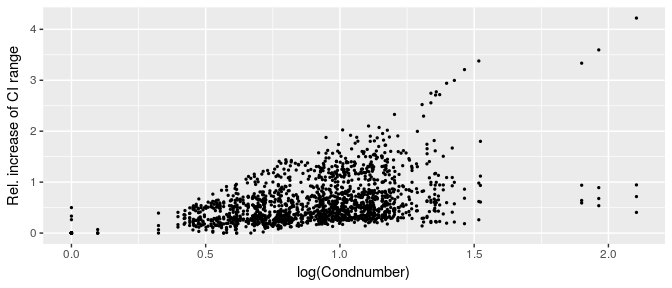
One of the root causes for the increase in confidence intervals is the fact that building the inverse of a matrix can be an ill-conditioned problem, e.g. the condition number is large. Given a matrix , the condition number is defined by

where is a norm. can indicate how much the noise in the data is magnified by solving a system of linear equations. In our setting this translates to how much the random noise, produced by our bootstrap replicates, is magnified by transforming the biased into the unbiased bootstrap replicates. The increase in confidence intervals can even occur if the initial bootstrap replicates are not biased to begin with. Figure 3 shows the range of the confidence intervals with and without transformation of bootstrap replicates for cases where the initial bootstrap replicates had little bias to begin with.

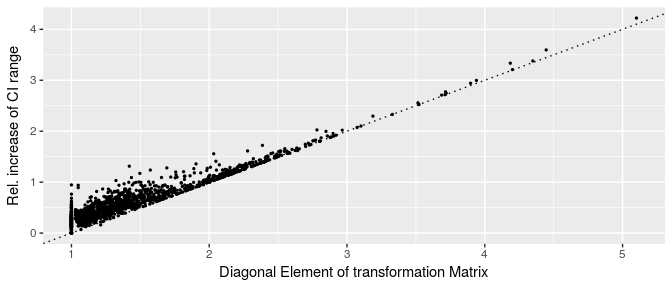
Since the condition number should indicate how much the noise of in the bootstrap replicates is magnified by applying , we would expect a high correlation between increase of confidence intervals and condition number of the matrix used. Figure 4 shows the relative increase of confidence intervals after transformation depending on the condition Number transformed by the logarithm. Altough the condition number does seem to indicate some increase in confidence intervals, this relation is not homogeneous throughout all the results. We also exhibit cases with high condition number but comparably little relative increase in confidence intervals.

***Figure 3. Increase of confidence intervals for estimates with little bias.*** 

**Figure 4. Relative increase of confidence interval range depending on the condition number**



Looking at the diagonal elements of shows a clearer picture. Figure 5 shows the relative increase of confidence interval depending on the diagonal elements of . It seems that there exists a more direct correlation between the diagonal elements of the inverse matrix and the relative increase of confidence interval than there is between the condition number in Figure 4. Therefore we would like to propose a method for constructing a transformation matrix, different to , with diagonal elements equal to 1, in order to limit the increase in confidence interval.

***Figure 5. Relative increase of confidence interval range depending on the diagonal element of the transformation Matrix***

*2.8. Constructing a transformation matrix*

Consider a characteristic , containing different values, and a transition probability matrix . For the population or a sub-part of the population , containing units, define the vector as the vector of aggregated totals for each value of ,

, where are some given values for . Applying on the vector yields the expected totals ,

The transformation matrix has to fulfill some criteria similar to , such as:

* The sum over each column must equal 1, so that the sum over all totals after transformation always equals .
* Applying the transition matrix on must yield the original totals .

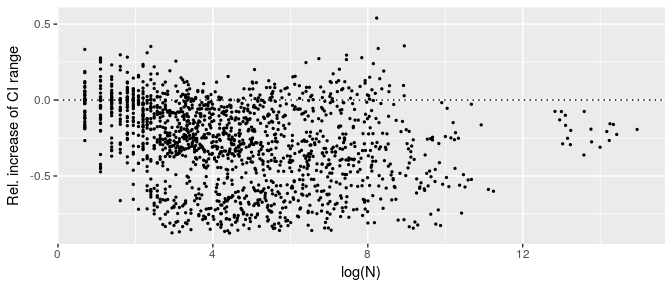
Considering these condition the transformation Matrix can be constructed by the following iterative procedure:

1. Set and initialize the transformation matrix as the -dimensional identity matrix
2. Estimate the bias and vector ,
3. Update the values in row of matrix by
4. If go to Step 7, otherwise
5. Calculate the column difference and vector ,
6. Update the values in column of matrix by
7. If
   * and
   * terminate, otherwise set and go to Step 2.

*2.9. Constructing a transformation matrix*

Applying this new transition matrix and comparing the results to the confidence intervals produced by using shows a clear reduction in confidence intervals. Figure 6 shows for each results the relative increase or decrease of confidence intervals when using . For observations above or below the dotted horizontal line the use of had an increasing or decreasing effect on confidence intervals, respectively. Although the decrease for the confidence interval was substantial in most of the cases, for some it had the opposite effect. However this effect appears only in 14.54 of the observations, leading to an average decrease in confidence interval of 27.96 . The following sections will compare the estimated confidence intervals for educational level by gender and/or nationality including the effect of the imputation scheme in more detail . For the transformation of the biased bootstrap replicates the transformation matrix, , was used.

**Figure 6. Relative reduction of confidence interval range when using instead of .**



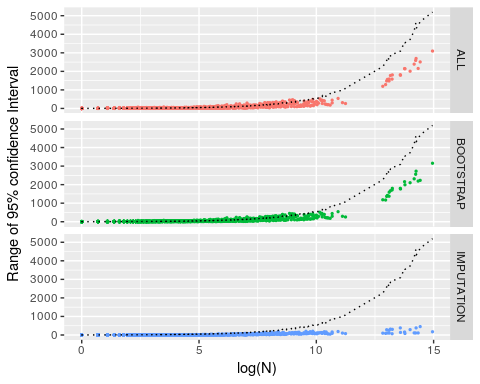
**3. Results**

The imputation and bootstrap schemes were applied onto register data, with repetitions, resulting in the following 3 types of confidence intervals:

* Confidence intervals resulting from imputation
* Confidence intervals resulting from bootstrapping
* Confidence intervals resulting from first imputing and afterwards bootstrapping.

Figure 7 shows the range of the resulting confidence intervals for each of the cases. The dotted line represents confidence intervals which would have been achieved by the naive bootstrap procedure. Looking at the panel for imputation the resulting confidence intervals are rather small as would be expected, since imputation was only applied on a comparably small number of observations. Interesting to see is that for large groups confidence intervals are always smaller compared when applying the estimated transition probability matrix. In contrast to small groups, where the confidence intervals from the naive bootstrap procedure seems to be more reliable. As described in the previous section this behaviour for small groups results from the bias reduction. We can conclude that for large groups the estimation of the transition probabilities does have a decreasing effect on confidence intervals. Even for biased bootstrap replicates the reduction in variance due to modelling the transition probabilities outweighs the possible increase in variance, due to transformation. For smaller groups a naive bootstrap procedure appears to be a more suitable method. This can also be supported by the fact the range of confidence intervals is bounded by 0 and therefore the possible decrease of variance due to modelling the transition probabilities is more narrow for smaller groups than larger. Concluding that in increase of variance due to bias reduction is more easily to outweigh the decrease due to modelling.

**Figure 7. Confidence interval resulting from different error sources. The dotted line shows confidence intervals resulting from using a naive bootstrap procedure.**



**4. Conclusion and Outlook**

Estimating confidence intervals for register data is in general a very challenging and quite new problem. Nevertheless it gains more and more importance since producing statistics based on register can lower costs and reduce respondent burden. Delden, S., and J. (2015) as well as Delden, S., and J. (2016) have already presented a methodology for estimating confidence intervals on business registers. In this work we presented an extension to their methodology by introducing, apart from classification errors, imputation errors as a second error source. Furthermore assumed the classification error to be depended on the imputation error. Regarding the treatment of biased bootstrap replicates and the possible resulting variance inflation we propose a specific transformation matrix, which is different from the inverse transition probability matrix. This transformation matrix does keep the composition of the data intact but, as our empirical results have shown, reduces the variance inflation when transforming the bootstrap replicates. However a more well-grounded theoretical background on the construction of such a matrix is needed. Interesting to see was the fact, that although classification errors can be modeled and used in order to calculate bootstrap replicates, applying a naive bootstrap procedure onto to register data does yield robust and conservative results. For the last Census 2011 Statistics Austria has developed a quality framework (Asamer et al. 2016) to assess the quality of register based statistics. With regard to Census 2021 the presented method will be used to enlarge the utilization of the quality indicators.

**5. References**

Asamer, E.M., F. Astleithner, P. Ćetković, S. Humer, M. Lenk, M. Moser, and H. Rechta (2016), Quality Assessment for Register-Based Statistics - Results for the Austrian Census 2011, *Austrian Journal of Statistics* 45 (2), The Austrian Statistical Society: 3–14. doi:[10.17713/ajs.v45i2.97](https://doi.org/10.17713/ajs.v45i2.97).

Delden, A. van, Scholtus S., and Burger J. (2016), Accuracy of Mixed-Source Statistics as Affected by Classification Errors, *Journal of Official Statistics* 32 (3), Statistics Sweden: 619–42. doi:[10.1515/jos-2016-0032](https://doi.org/10.1515/jos-2016-0032).

Efron, B., and R. Tibshirani (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, *Statist. Sci.* 1 (1), The Institute of Mathematical Statistics: 54–75. doi:[10.1214/ss/1177013815](https://doi.org/10.1214/ss/1177013815).

ESSnet on Quality of Multisource Statistics (2017), <https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en>.

Delden, A. van, Scholtus S., and Burger J. (2015) Sensitivity of Mixed-Source Statistics to Classification Errors, *Journal of Official Statistics* 31 (3): 489–506, <https://EconPapers.repec.org/RePEc:vrs:offsta:v:31:y:2015:i:3:p:489-506:n:9>.

Rubin, Donald B. (1976), Inference and Missing Data, *Biometrika* 63 (3), Oxford University Press: 581–92.