**Quality evaluation of register-based statistics1**

Bart F.M. Bakker, Statistics Netherlands and VU University, bfm.bakker@cbs.nl

**Abstract**

*Due to the increased availability of administrative registers, more and more register-based official statistics are published. To decide whether a register-based statistic is of sufficient quality to replace an existing survey-based statistic it is important to evaluate this quality. In this paper, we present methods to assess this quality, using recently developed quality frameworks that have tried to mimic the Total Survey Error approach for register data. Main indicators are under and over coverage, measurement error and linkage error. However, these quality frameworks do not present the methods to measure these indicators. We present capture-recapture methods (CRC) to estimate under coverage of already linked registers. Over coverage can be determined by searching for units that do not belong to the target population. Duplicates can be identified by linking the records in the combined registers to each other. Measurement error can be estimated by Structural Equation Models (SEM, for numerical variables) and Latent Class Analysis (LCA, for categorical variables) with a measurement component if another source that measures the same concept is linked to the register data. Linkage error can be estimated using probabilistic linkage methods. However, none of these methods is error free in itself. Linkage error also could have impact on the outcomes of CRC, SEM and LCA. The paper shows the interdependencies between these methods.*

**Keywords:** Total survey error, representation error, measurement error, linkage error

**1. Introduction**

More and more statistical offices make use of register data for producing their official statistics. Driving force is that the use of register data is much cheaper than primary data collection with interviewing. Important to find out is whether the quality of the register-based statistics is sufficient to replace the primary data collection by registers. A customary method to evaluate this is to assume that the outcome from the existing primary data collection is the gold standard. If the register-based results differ substantially from this gold standard, the quality is assessed as too low. This method is rather naïve because all sources contain error and cannot be used as a gold standard (Bakker, 2012; Hand, 2018; Scholtus, 2018). A framework for the errors in register-based statistics has been developed by Bakker and Daas (2012) and Zhang (2012). As in the total survey error approach (Groves et al. 2007), they have defined possible error sources for each step in the process of register-based statistics.

1 Bart F.M. Bakker is head of the Methodology Department of Statistics Netherlands in The Hague and professor in Methodology for Register Data for the Social Sciences at VU University Amsterdam. The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author thanks Sander Scholtus, Peter-Paul de Wolf and Daniel Oberski for their critical review of this paper. This paper is a thoroughly revised version of Bakker (2017).

The process of register-based statistics typically is that a combination of registers are used to produce a list of the target population on a reference date. We call the result of this step a base register and these can exist for all kinds of units (persons, companies, regions, etc.) (Wallgren & Wallgren, 2014). The quality of the base register is a decisive factor in the quality of register-based statistics and therefore needs special attention. In particular, if part of the target population is not covered by the base register, there is a high risk of biased results. After the base register has been created, the other necessary registers and surveys are linked to it, the data are edited, the target variables are deduced including the classifications used for these variables and a solution is applied to the missing values in these variables (e.g. introducing a category unknown or using imputation). In the processing of the data, the representation error of the linked registers and surveys can be evaluated using information from the base register. The last step is the estimation of the outcomes. In the case of fully register-based statistics, this is merely simple counting. However, if a combination of registers and surveys is used, some estimation procedure has to be applied (e.g. weighting, consistent repeated weighting, macro-integration techniques). All steps can contribute to the total survey error.

Bakker and Daas (2012) and Zhang (2012) distinguish between representation and measurement error. A dataset contains representation error if the data do not represent the target population. It contains measurement error if a variable measures something else than the target variable. An inevitable step in the process of register-based statistics is that different registers, and sometimes surveys are linked. However, methods how to apply this total survey error approach for administrative data are lacking. In this paper, I discuss the different sources of representation and measurement error based on this framework, the methods to estimate the size of these errors and, if available, the methods to adjust the outcomes for these errors.

**2. Representation error**

In the first step of the production of register-based statistics, the target population is deduced from a base register in which ideally all units are covered (Wallgren & Wallgren, 2014). There should be a base register for different kinds of units: persons, companies, jobs, regions, etc. These base registers should be longitudinal of nature, and each unit should have a starting and, if applicable, ending date. Therefore, it is possible to derive a list of the units of the target population for each reference date or period. In countries that have a lot of registers at their disposal, the production of base registers is sometimes an easy step. For instance, those countries that have a well-functioning population register, the base register for persons can be similar to this population register. The representation error in such a base register consists of three kinds of error: units are in the base register while they do not belong to the target population (over coverage), units who do belong to the target population are missing from the base register (under coverage) and there are duplicates in the base register. Under and over coverage and duplicates in the base register reduce the quality of the register-based statistics, because it is assumed to describe the target population entirely correctly. Linking other registers and surveys to the target population deduced from the base register leads to under and over coverage of the register-based statistics. Therefore, it is very important to evaluate the quality of the base registers.

The under coverage of base registers can be estimated by means of capture-recapture methods (Bishop, Fienberg and Holland, 1975; Gerritse et al., 2015; Van der Heijden et al., 2018). This methods can be applied if besides the base register a second source is available (be it a register or a specially for that purpose designed survey). This second source is linked to the base register. One can generate a twoway table whether or not a unit is in one of the sources or in both sources. A log-linear model is applied to this table under the assumption that the two sources are independent. To get accurate outcomes from these models, several assumptions have to be met. In practice these assumptions are always violated to a certain extent, particularly when data are originally not collected for statistical purposes. Violation of the assumptions can lead to severely biased results (Brown et al., 2006; Gerritse, 2016)

However, there are a few guidelines how to diminish the impact of violation of the assumptions. The first assumption that has to be met is that inclusion of persons in the first data source is independent of inclusion in the second data source. This assumption can be relaxed in two ways: by linking a third data source or by adding covariates to the model. If both guidelines are followed, the assumption is relaxed to the assumption that the higher order interaction of the three inclusion probabilities is zero within each cell of the crosstable of the covariates. The second assumption is that the population is closed during the data collection period. This assumption can be easily met by using data of only one reference date. If that is not possible, then take data that is collected in a very short period. The third assumption is that the sources are perfectly linked. False negative and false positive links lead to biased estimates. Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015) have developed methods to adjust the outcomes of capture-recapture analysis for linkage error. An important condition to apply this method is that you use probabilistic linkage to link your sources. It is based on the idea that you make use of the estimated probabilities that linked pairs are correct links. These improvements of the capture-recapture method seem promising. However, these methods are restricted to the use of two sources and a model without covariates. De Wolf et al. (2018) generalized the models for linkage correction and Zult et al. (2018) try to generalize the method to the use of three or more sources and covariates.

To reduce over coverage, records from units who are in the base register while they do not belong to the population have to be removed. Of course, the success of this step depends on the possibilities of identifying those units. An example of this is the use of a population register for the target population ‘usual residents’. Important is then to determine the date that persons became a resident. In register data, it is usually possible to deduce the starting date that people are registered as living, working or following education in the country. However, identification of persons and determining the starting date of such events will certainly not be perfect because in practice the data quality on e.g. the persons who immigrate is not as high as we want it to be, often containing a relatively large number of missing values.

One should verify the possibility of duplicate records in the base register. Duplicate records can be detected on the basis of identifying variables. The amount of success of this step depends on the quality of these variables. If a lot of missing values appear in the identifying variables, we could expect that the remaining records contain to some extent duplicates. If individuals are identified by a personal identification number, the number of duplicates is usually small. However, if a personal identification number is missing or is not complete, address information is necessary to identify individuals. When registers of different quality and in particular with different levels of administrative delay are combined, individuals who move from one place to the other can be seen as multiple different persons. To identify those persons correctly one can use additional information on the history of removals. If that is not available, only expert guesses based on knowledge of the quality of the different sources are possible to get an idea how many duplicates remain. If that is substantial, then it would be advisable to give those records that are suspected to be a duplicate a weight of one half.

After the evaluation of the representation error in the base register, it can be used to evaluate the representation error of the sources that are linked to the base register. It is of course possible to do an evaluation similar to that of the base register, but that is usually too time consuming and too expensive. First, the target population can be derived from the base register by selecting those units that are present at the selected reference date or period. Records that link to the base register but not on the selected reference date or period can be handled as over coverage. The base register records that do not have a record linked form the other source, can be handled as under coverage. The representation error of the additional sources can be improved using traditional weighting techniques similar to non-response correction techniques (e.g. Bethlehem et al., 2011)

**3. Measurement error**

Register-based statistics contain measurement error if the register variables are conceptually different from the target variables, or if they are measured with a systematic or random measurement error. To determine the size of the measurement error, two methods are available for different levels of measurement. The first method is a Structural Equation Model (SEM) with a measurement component for continuous variables, with so-called multitrait-multimethod models (MTMM-models) as a special case, and the second method is Latent Class Analysis (LCA) for categorical variables. The basic idea is that you use (at least) two measurements of the same concept and conceive the association between each observed variable and the latent target variable as a measure of the quality of that observed variable. This association is known as the indicator validity of the observed variable (Bakker, 2012; Scholtus, 2018: p. 42). This can be done by linking a (small) survey collected by interviewing to the combined registers or by linking another register that contains the same variable(s).

If one uses a SEM to estimate the size of the measurement error, one has to distinguish between the error in the estimated relationships between the variables, the indicator validity, and the intercept bias, i.e. an error in the observed means of the variables. Scholtus, Bakker and Van Delden (2015) introduce an additional test on the intercept bias, using similar models. However, in order to pin the “true level” a small audit sample is needed of flawless quality. Using the indicator validity and the intercept bias, it is possible to adjust the outcomes of the register data.

However, SEM can only be used if one has continuous conceptual variables. For instance educational attainment can be considered a continuous concept, but it can only be measured as a categorical variable (in years, or levels). In official statistics one finds also categorical variables. In that case errors are classification errors: an individual is wrongly classified to a category. For the determination of these classification errors one can apply LCA. Examples of this approach are Pavlopoulos and Vermunt (2015) and Pankowska et al. (2017) with applications to estimating permanent and temporary employment in the Netherlands. They use linked longitudinal categorical data from a register and the Labour Force Survey and apply a particular group of latent class models called Hidden Markov Models (HMMs). HMMs are applied to describe a turnover or transition in some characteristic assuming that it is driven by a process without memory and that it is measured with an error. These models probabilistically estimate the latent states at the individual level and provide also estimates for the distribution of these states as well as for the mobility between them. Pavlopoulos and Vermunt (2015) show that the latent transition rate from temporary to permanent employment in the Netherlands is less than half than the observed data from both sources suggest. LCA leads to a classification table which can be used to correct for the classification error.

A new development is the so-called generalized MTMM-model. In this model it is possible to combine categorical and numerical variables (Oberski et al., 2017). This is of course advantageous, because in official statistics one usually finds variables of both types in one table. The authors claim that for their models less assumptions are to be made; in the generalized MTMM-models one does not have to assume that the true values are normally distributed or that the errors are linear and homoscedastic.

**4. Linkage Error**

After the target population has been derived from the base register by choosing a reference date or period, the other registers are linked to the selected units of this base register. In the last step, for those who use survey data these data are linked too. Linkage error could then lead to both representation error (duplicate records, false negative links) and measurement error (false positive links). Moreover, the combination of register and survey data can lead to inconsistent outcomes.

False negatives exist if records that belong to the same unit were not linked. This results in missing values in variables in the combined dataset for those missed links. This can lead to severe quality problems. E.g. if one is interested in the question what is the proportion of people that belong to the working labour force at a certain reference date, one usually link an administrative source with information on jobs (the jobs base register) to the selected units from the base register of persons. However, false negatives lead easily to an underestimation of the number of individuals belonging to the workforce. It is very difficult to determine whether a job should be linked or not because the job register should describe the entire work force and also includes individuals not belonging to the population (cross border workers). If it is possible to identify the cross border workers, one can try to correct for the false negatives by weighting or imputation.

False positives exist if records that belong to different units were linked. This leads to a specific form of measurement error. The linkage process should focus on the minimisation of the false negatives and false positives. However, if a large number of false negatives remain after a simple deterministic linkage process in which all the identifying variables are identical, a form of probabilistic linkage is needed in which some deviation from total similarity is allowed. That leads inevitably also to some false positives. If the resulting file can be linked to a survey, one can apply LCA or SEM to estimate the quality of the estimations and to correct for measurement error as is described in section 3. New research presented at this conference concludes that HMM can account for the effects of false positives on estimates (Pankowska et al., 2018).

**5. Conclusion**

Register-based official statistics have become more and more popular, mainly because of the low costs. However, official statistics are very important for decision making and the quality should therefore be high. There are three interdependent errors in register-based statistics: under and over coverage, measurement error and linkage error.

Coverage error can be detected in two steps. In the first step base registers for the relevant units are created. After that the target population is derived from the base register by selecting the right units and the other registers and surveys are linked to the target population. In the second step the over and under coverage of the resulting records are evaluated using the base register information on the target population. Therefore, much effort should be put into the creation of the base registers, because coverage error in the base register leads subsequently to coverage error in the dataset after linkage. Under coverage of the base register can be estimated with capture-recapture methods. Duplicates can be estimated by linking the records in a (quasi) population register to each other. Over coverage can be solved by deleting the units that do not belong to the population, provided that one can identify those units.

Measurement error can be estimated by Structural Equation Models (for numerical variables) and Latent Class Analysis (for categorical variables) with a measurement component if another source is linked to the register data that measures the same concept. Linkage error can be estimated using probabilistic linkage methods. For each error source, it is possible to make some adjustments to the estimates. However, because the error sources are interdependent, it is to be seen whether it is possible to design a process that leads to the optimal outcomes. Moreover, more work has to be done on estimating accuracy and reliability measures of these outcomes. These are a few of the challenges that are still ahead (Hand, 2018).

**6. References**

Bakker, B.F.M. (2012), “Estimating the validity of administrative variables.” Statistica Neerlandica, 66, pp. 8-17.

Bakker, B.F.M. and Daas, P.J.H. (2012), Methodological Challenges of Register-Based Research, Statistica Neerlandica 66, pp. 2–7.

Bakker, B.F.M. (2017), Assessing the quality of register-based population censuses: processes and outputs, Paper ISI World Statistics Congress, Marrakech.

Bethlehem, J., Schouten, B., and Cobben, F. (2011), Handbook of Non-response in Household Surveys, John Wiley and sons, New York.

Bishop, Y., Fienberg, S. and Holland, P. (1975), Discrete Multivariate Analysis, Theory and Practice, McGraw-Hill, New York.

Brown, J., Abbott, O. and Diamond, I. (2006), Dependence in the 2011 One-Number Census Project, Journal of the Royal Statistical Society: Series A (Statistics in Society) 169, pp. 883–902.

De Wolf, P.-P., Van der Laan, J. and Zult, D. (2018), Connecting correction methods for linkage error in capture-recapture, Discussion Paper 2018, Statistics Netherlands, Den Haag / Heerlen.

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

Ding, Y. and Fienberg, S.E. (1994), Dual System Estimation of Census Undercount in the Presence of Matching Error, Survey Methodology, 20, pp. 149–158.

Gerritse, S.C., (2016), An application of population size estimation to official statistics. Sensitivity of model assumptions and the effect of implied coverage, Ph. Dissertation Utrecht University.

Gerritse, S.C., Van der Heijden, P.G.M. and Bakker, B.F.M. (2015), Population size estimation for violating parametric assumptions in loglinear models, Journal of Official Statistics, 31, pp. 357-379.

Groves, R. M., Fowler JR., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2007), Survey methodology, Wiley Interscience, New York.

Hand, D.J. (2018), Statistical challenges of administrative and transaction data, Journal of Royal Statistical Society A Part 3, 181, pp. 1-24.

Oberski, D.L., Kirchner, A., Eckman, S., Kreuter, F. (2017), Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models, Journal of the American Statistical Association, 112, pp. 1477-1489, DOI: 10.1080/01621459.2017.1302338

Pankowska, P., Bakker, B., Oberski, D.L. and Pavlopoulos, D. (2017), Reconciliation of two data sources by correction for measurement error: the feasibility of parameter re-use, Statistical Journal of the IAOS*,* vol. Preprint, pp. 1-13, DOI10.3233/SJI-170368

Pankowska, P., Bakker, B., Oberski, D.L. and Pavlopoulos, D. (2018), Use of latent class models in official statistics production, Paper European Conference on Quality in Official Statistics, 26-29 June 2018, Krakow

Pavlopoulos, D., and Vermunt, J. K. (2015), Measuring temporary employment. Do survey or register data tell the truth?, Survey Methodology, 41, pp. 197–214.

Scholtus, S., Bakker, B.F.M. and Van Delden, A. (2015), Modelling measurement error to estimate bias in administrative and survey variables. Discussion Paper 2015-17, Statistics Netherlands, Den Haag / Heerlen.

Scholtus, S. (2018), Editing and estimation of measurement errors in administrative and survey data, PhD Dissertation, VU University Amsterdam.

Van der Heijden, P.G.M., Smith, P.A., Cruyff, M., Bakker, B.F.M. (2018), An Overview of Population Size Estimation Where Linking Registers Results in Incomplete Covariates, With an Application to Mode Of Transport of Serious Road Casualties, Journal of Official Statistics*,* 34, pp. 239–263.

Wallgren, A., and Wallgren, B., 2014, Register-based statistics, Wiley, New York.

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

Zult, D., De Wolf, P.-P., Bakker, B.F.M. & Van der Heijden, P.G.M. (2018), Improvement of the capture-recapture model to estimate under coverage of administrative data sources, Paper European Conference on Quality in Official Statistics, 26-29 June 2018, Krakow.