Method: Macro-integration, including RAS and Stone's method

1. Purpose of the method

Different estimates for the same phenomenon could lead to confusion among users of these figures. Many other NSIs, such as Statistics Netherlands, have therefore adopted a one-figure policy. According to this one-figure policy, estimates for the same phenomenon in different tables should be equal to each other, even if these estimates are based on different underlying data sources. We call this *univalency* and say that estimates should be *univalent*.

When using a mix of administrative data sources and surveys to base estimates upon, the one-figure policy becomes problematic as for different (combinations of) variables data on different units, e.g. different persons, may be available. This means that different estimates concerning the same variable may yield different results, if one does not take special precautions.

2. The related scenarios

2.1 Macro-integration imputation can be used for data sets composed of macrodata. It can be used in data configurations 5 and 6 (see Deliverable 1). Statistical usage is "Direct tabulation".
2.2 Statistical tasks: "Integrate data".
2.3 Alternative methods are repeated weighting, repeated imputation and mass imputation.

3. Description of the method

Macro-integration is the process of reconciling statistical figures on an aggregate level. These figures are usually in the form of multi-dimensional tabulations, obtained from different sources. When macro-integration is applied, only estimated figures on aggregated level are adjusted. The underlying microdata are not adjusted or even considered in this adjustment process. The main goal of macro-integration is to obtain a more accurate, consistent and complete set of estimates for the variables of interest. Several methods for macro-integration have been developed, such as the methods by Stone, Champernowne and Meade (1942), Denton (1971), Byron (1978), Sefton and Weale (1995) and Magnus, Van Tongeren and De Vos (2000).

The starting point of macro-integration is a set of estimates in tabular form. These can be quantitative tables, for instance a table of average income by region, age and gender, or frequency tables, for instance a cross-tabulation of age, gender, occupation and employment. If the estimated figures in these tables are based on different sources and (some of) the tables have margins in common, these margins are often conflicting.

When one wants to use macro-integration to reconcile the data, it is generally important that (an approximation or indication of) the variance of each entry in the tables to be reconciled is computed. The entries of the tables are then adjusted by means of a macro-integration

technique so all differences between tables are reconciled and the entries with the highest variance are adjusted the most.

In a macro-integration approach often a constrained optimization problem is constructed. This is, for instance, the case for the method by Stone, Champernowne and Meade (1942). A target function, a quadratic function of differences between the original and the adjusted values, is minimized, subject to the constraints that the adjusted common figures in different tables are equal to each other and additivity of the adjusted tables is maintained.

In adjusting the initial data the method of Stone, Champernowne and Meade (1942) uses information on the relative reliabilities of the initial data, in particular a covariance matrix. Data that are considered most reliable are modified least and vice versa. The method yields a set of fully reconciled data, with minimum variance. The solution of the method includes the reconciled data as well as its covariance matrix. Analytical expressions can be derived for both results.

In some macro-integration approaches inequality constraints can be imposed in the optimization problems. Analytical expressions for the solution are then no longer available and (co)variances become hard to estimate (see, e.g., Knottnerus 2016).

In the literature also Bayesian macro-integration methods have been proposed based on a truncated multivariate normal distribution (see Magnus, Van Tongeren and De Vos 2000, and Boonstra, De Blois and Linders 2011).

The RAS method (or iterative proportional fitting – IPF, see Memobust 2014 and Deming and Stephan 1940) is a simple and well-known method for data reconciliation. It is a special case of (generalized) calibration. In the original papers discussing this method (see, e.g., Bacharach 1970) a vector of row multipliers was designated by *r*, a table of inter-industry transactions in coefficient form in a base year by *A*, and a vector of column multipliers by *s*. These three letters led to the name RAS.

RAS can be used for

- estimating contingency tables when the marginals are fully observed and incomplete information are available for the cells in the inner part of the table
- updating a given table to new marginal totals while preserving as much as possible the structure of the initial table.

RAS is an iterative scaling method. It multiplies each entry in one row or column by some factor, which is chosen in such a way that the sum of all entries in the row or column becomes equal to its target total. This operation is first applied to all rows of the table. As a consequence the table becomes consistent with all target row totals. Then, the columns are made consistent with their required totals. As a result consistency is achieved with the column totals, but the constraints on the row totals may be violated again. The rows and columns are adjusted in turn, until the algorithm converges to a table that is consistent with all required row and column totals. RAS can only be applied to nonnegative tables.

4. Examples

As mentioned in Task 4, macro-integration can be used to produce univalent estimated for the Population and Housing Census.

5. Input data (characteristics, requirements for applicability)

The input data for macro-integration are macrodata.

6. Output data (characteristics, requirements)

The output data of macro-integration are macrodata.

7. Tools that implement the method

Currently tailor-made programs and scripts are used. However, softwares for calibration can be used to implement the RAS method.

8. Appraisal

Macro-integration has an important advantage over other techniques that achieve univalency, such as repeated weighting and repeated imputation: macro-integration can reconcile all tables simultaneously instead of table by table, as long as the number of variables or constraints does not become too large. If tables are reconciled simultaneously, a better solution may be found.

Another strong point of the macro-integration approach is that some of the methods have been developed with longitudinal (numerical) data in mind instead of only cross-sectional data.

A drawback of macro-integration is that there is often no direct relation between the macro estimates and the underlying microdata.

9. References

Bacharach, M. (1970), *Biproportional Matrices & Input-Output Change*. Cambridge University Press, Cambridge.

Boonstra, H.J., C.J. De Blois and G.J. Linders (2011), Macro-Integration with Inequality Constraints an Application to the Integration of Transport and Trade Statistics. *Statistica Neerlandica 65*, pp. 407-431.

Byron, R.P. (1978), The Estimation of Large Social Account Matrices. *Journal of the Royal Statistical Society A 141*, pp. 359-367.

Deming, W. and F. Stephan (1940), On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known. *Annals of Mathematical Statistics 11*, pp. 427–444.

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

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

Magnus, J.R., J.W. van Tongeren and A.F. de Vos (2000), National Accounts Estimation using Indicator Ratios. *The Review of Income and Wealth 46*, pp. 329-350.

Memobust (2014), RAS. In: *Memobust Handbook on Methodology of Modern Business Statistics,* https://ec.europa.eu/eurostat/cros/content/memobust_en.

Memobust (2014), Stone's Method. In: *Memobust Handbook on Methodology of Modern Business Statistics,* https://ec.europa.eu/eurostat/cros/content/memobust_en.

Sefton, J. and M. Weale (1995), *Reconciliation of National Income and Expenditure*. Cambridge University Press, Cambridge, UK.

Stone, R., D.G. Champernowne and J.E. Meade (1942), The Precision of National Income Estimates. *Review of Economic Studies 9*, pp. 111-125.

1. Purpose of the method

When data from multiple sources are combined, differences in definition can occur between variables in different sources. In particular, variables in an administrative data source are defined according to the administrative purposes of the register owner. These definitions may differ from those of the target variables for statistical purposes. For example, a tax authority collects data of value-added tax (VAT) declarations from businesses which contain turnover values. Since the administrative purpose of these data is to levy taxes on turnover, the tax authorities will be interested only in the amount of turnover of each business that is derived from taxable economic activities. Depending on the specific tax regulations that apply, for some sectors these administrative turnover values will differ from the turnover values that a statistical institute needs: some economic activities that are relevant for economic statistics may be exempt from taxes, and vice versa (Rich and Burman, 2012; Van Delden et al., 2016). In the Netherlands, for instance, businesses that buy and sell second-hand cars may report their profit margin for VAT rather than their total turnover.

In case differences in variable definitions occur between data sources, these variables need to be harmonised during data integration. That is to say, for each unit in the integrated data set, the values of the target variable according to the desired definition need to be estimated from the observed values that are available.

2. The related scenarios

2.1. The method can be used for "data validation". To be able to estimate a latent variable model for variable harmonisation, overlapping observed values from at least two different data sources are required for at least part of the units. Thus, strictly speaking, the method applies only to data configurations 4 and 5 (see Deliverable 1). If a latent variable model has been estimated previously and it may be assumed that the relation between the observed variables and target variables has remained stable, then the method can also be applied to data configurations 2 and 3 (see Deliverable 1).

2.2. Statistical tasks: alignment of measurements.

2.3. An alternative approach that is currently often used in practice is to apply ad hoc derivation rules to derive the target variables from the observed variables in each input data source (Bakker, 2011). These rules are formulated by experts, based on subject-matter knowledge. An example of this approach is mentioned in Section 1 of Deliverable 3 (the processing of administrative data on wages at Statistics Netherlands).

3. Description of the method

When microdata from different sources are to be combined, variable harmonisation is necessary if conceptual differences occur between the observed variables in different data sources. By explicitly modeling these conceptual differences, we may be able to correct for them and thus obtain harmonised measurements. It is assumed here that overlapping observed values from different sources are available for at least a subset of the units.

A relatively simple approach is obtained by treating the observed variable in one of the data sources as the 'gold standard' measurement of the target variable. The observed variables in the other data sources should then be corrected for any systematic deviations from this 'gold standard' observed variable. Using the overlapping units, for each data source one can estimate a model with the 'gold standard' variable as the dependent variable and the observed variable to be harmonised as the independent variable.

Van Delden et al. (2016) used this approach to investigate and possibly correct for conceptual differences between turnover as measured in administrative VAT data and turnover as defined for statistical purposes. These authors used a robust linear regression model of the following form

$$\text{Target turnover} = a + b \cdot \text{VAT Turnover} + \varepsilon,$$

where ε denotes a zero-mean random measurement error. The intercept a and slope b were allowed to vary by type of economic activity, since the definition of VAT turnover depends on industry-specific tax regulations. If there are no differences in definition, then $a = 0$ and $b = 1$ and no harmonisation step is needed. Otherwise, harmonised turnover values can be obtained from the estimated model as $\hat{a} + \hat{b} \cdot \text{VAT Turnover}$, for all units in the VAT data set. In this application, turnover values from an existing survey were linked to the VAT data and used as 'gold standard' measurements of the target turnover.

In practice, all observed variables may contain measurement errors. Models that do not account for this fact may lead to biased parameter estimates (Biemer, 2011). In particular, the estimated slope parameter $\hat{b}$ in the above example could be biased towards zero due to random measurement errors in VAT turnover (Van Delden and Scholtus, 2017) and the parameter estimates could also be biased due to systematic errors in survey turnover. More advanced harmonisation methods that can take measurement errors in all data sources into account are therefore of interest.

When an observed variable is available from multiple sources for at least some overlapping units, it is possible to model the measurement errors in each of these variables explicitly. The target variable itself is represented in such a model as an unobserved (*latent*) variable, of which the observed variables are error-prone measures. There are several types of latent variable models for different types of data.

Latent class models

For categorical data, measurement models based on latent class analysis can be used (e.g., Hagenaars and McCutcheon, 2002). The basic form of a latent class model that can be used to model errors in observed variables is given by

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{K} P(X = x) P(\boldsymbol{Y} = \boldsymbol{y} | X = x).$$

Here, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)'$ denotes a vector of the observed categorical variables in the different data sources and X denotes a latent class variable that represents the true value of the underlying target variable. Often, it is assumed that the classification errors in different observed variables are conditionally independent, given the true value of the target variable:

$$P(\boldsymbol{Y} = \boldsymbol{y} | X = x) = P(Y_1 = y_1 | X = x) P(Y_2 = y_2 | X = x) \cdots P\big(Y_p = y_p \big| X = x\big).$$

Note that, for instance, for units that in reality belong to the first category of the target variable ($X = 1$), the probability of misclassification on observed variable Y_j is given by $P\big(Y_j \neq 1 \big| X = 1\big) = 1 - P\big(Y_j = 1 \big| X = 1\big)$.

After estimating the latent class model, Bayes' rule can be used to compute corresponding estimates of the probability that a unit belongs to a particular latent class, given its observed values, i.e., $P(X = x | \boldsymbol{Y} = \boldsymbol{y})$. These estimated probabilities can be used to predict the latent (true) category and thus to impute a harmonised target variable for each unit.

Applications of latent class models to measurement errors in statistical data are discussed by Biemer (2011), Si and Reiter (2013), Pavlopoulos and Vermunt (2015), Boeschoten, Oberski and De Waal (forthcoming), and Oberski (2017). The precise form of the models used in these applications depends on particular features of the data that were analysed.

Structural equation models

To model measurement errors in numerical data, a so-called linear structural equation model with latent variables can be used (e.g., Bollen, 1989). For a single target variable with multiple observed variables, the model takes the form

$$Y_j = a_j + b_j\eta + \varepsilon_j,$$

where $Y_1, Y_2, \ldots, Y_p$ now denote observed numerical variables, η is a latent variable that represents the underlying true value, and ε_j denotes a zero-mean random measurement error. It is possible to include multiple target variables $\eta_1, \eta_2, \ldots, \eta_m$ in the same structural equation model, and in that case the model may also contain regression equations that describe the relations between different latent variables. In a general structural equation model, an observed variable may be related to several latent variables, but for the purpose of modeling measurement errors it will usually be natural to assume that each Y_j measures a single latent variable, as in the above equation.

Bakker (2012) advocated the use of structural equation models with combined administrative and survey data, to evaluate and compare the measurement quality of observed variables in different sources. Scholtus, Bakker and Van Delden (2015) described an application of structural equation modeling to variable harmonisation for VAT turnover. As discussed by these authors, to identify all parameters of the model, it is necessary to obtain error-free measurements of the target variable for a (small) random subsample of the original data, for instance by re-editing the data. This may be a drawback in practice. The audit sample can be avoided if it may be assumed that $a_j = 0$ and $b_j = 1$ for one of the observed versions of each latent variable in the model.

Finite mixture models

As an alternative approach for numerical data, so-called finite mixture models (e.g., McLachlan and Peel, 2000) can be used to handle situations where different error structures apply to different subsets of the population, and it is not known in advance to which subset each unit belongs. Finite mixture models for measurement errors in multiple data sources with overlapping units have been developed by Meijer, Rohwedder and Wansbeek (2012) and by Guarnera and Varriale (2015, 2016). The latter authors explicitly consider situations where the measurement errors are 'intermittent': part of the observed values are correct and the remaining values contain errors. This leads to a model of the form

$$Y_j = \eta + Z_j\varepsilon_j,$$

where Z_j is an indicator that is equal to 1 if the observed variable Y_j contains a measurement error and equal to 0 otherwise. Scholtus, Bakker and Robinson (2017) consider an extended version of this model which also contains intercept and slope parameters.

4. Examples

Two case studies that use a latent variable model for variable harmonisation can be found in Deliverable 4:

- Modeling measurement error in admin and survey variables on turnover *(example of structural equation modeling);*
- Estimating classification errors under edit-restrictions in combined register-survey data *(example of latent class analysis)*.

5. Input data (characteristics, requirements for applicability)

The input consists of microdata. At least for a subset of the units, multiple measurements of the same variable should be available, e.g., from different data sources. It is assumed that the estimated model parameters for this subset can be generalised to the other units.

6. Output data (characteristics, requirements)

The output consists of microdata.

7. Tools that implement the method

Latent class models can be estimated by the software *Latent Gold*.

Structural equation models can be estimated in the R environment for statistical computing by the package *lavaan*, as well as by several stand-alone software packages, including *LISREL* and *M-Plus*.

8. Appraisal

Compared to traditional subject-matter knowledge-based derivation rules, variable harmonisation methods based on latent variable models have the advantage of being objective and based on explicit assumptions that to some extent can be tested. For instance, fit measures are available for testing latent class models (Hagenaars and McCutcheon, 2002) and structural equation models (Bollen, 1989). However, the introduction of latent variables does make the modeling approach more complicated; for instance, model selection is less straightforward than for linear regression models and additional data may be required to identify all model parameters.

A drawback of the models that have been proposed so far is that they rely on assumptions that may be violated in many practical situations (e.g., normally distributed data, independent errors between sources). At the moment, it is not clear how sensitive the estimated values of the harmonised variable are to (minor) violations of these assumptions. In addition, more realistic models should be developed.

In general, estimated relations between the latent variable and covariates will be valid if those covariates are included in the measurement error model. Boeschoten, Oberski and De Waal (forthcoming) noted that, for covariates not included in their latent class model, estimated relations with the latent variable may be biased. Since it is not always possible or desirable to include many covariates in the latent class model, research is currently being carried out to correct estimated relations between the latent variable and covariate not included in the latent class model, so these corrected relations are unbiased.

Kim, Berg and Park (2016) have recently proposed an imputation method based on statistical matching that incorporates measurement errors, which could also be used for variable harmonisation.

9. References

Bakker, B.F.M. (2011), Micro-Integration: State of the Art. In: State of the Art on Statistical Methodologies for Data Integration, report on WP1 of the ESS net on Data Integration.

Bakker, B.F.M. (2012), Estimating the Validity of Administrative Variables. Statistica Neerlandica 66, 8-17.

Biemer, P.P. (2011), Latent Class Analysis of Survey Error, Hoboken, NJ: John Wiley & Sons.

Boeschoten, L., D. Oberski and T. De Waal (forthcoming), Estimating Classification Error under Edit Restrictions in Combined Survey-Register Data. Journal of Official Statistics, conditionally accepted for publication.

Bollen, K.A. (1989), Structural Equations with Latent Variables. New York: John Wiley & Sons.

Guarnera, U. and R. Varriale (2015), Estimation and Editing for Data from Different Sources. An Approach Based on Latent Class Model. Working Paper No. 32, UN/ECE Work Session on Statistical Data Editing, Budapest.

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

Hagenaars, J.A. and A.L. McCutcheon (eds.) (2002), Applied Latent Class Analysis, New York: Cambridge University Press.

Kim, J.K., E. Berg and T. Park (2016), Statistical Matching using Fractional Imputation. Survey Methodology 42, 19-40.

McLachlan, G.J. and D. Peel (2000), Finite Mixture Models, New York: John Wiley & Sons.

Meijer, E., S. Rohwedder and T. Wansbeek (2012), Measurement Error in Earnings Data: Using a Mixture Model Approach to Combine Survey and Register Data. Journal of Business & Economic Statistics 30, 191–201.

Oberski, D. (2017), Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model. In: Biemer, De Leeuw, Eckman, Edwards, Kreuter, Lyberg, Tucker and West (eds.), Total Survey Error in Practice: Improving Quality in the Era of Big Data, New York: John Wiley & Sons.

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

Rich, S. and S. Burman (2012), Use of VAT and VIES Data for Validation in International Trade in Goods and Services. Paper presented at the European Conference on Quality in Statistics (Q2012), 29 May–1 June 2012, Athens, Greece.

Scholtus, S., B.F.M. Bakker and A. Van Delden (2015), Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. Discussion Paper 2015-17, Statistics Netherlands, The Hague.

Scholtus, S., B.F.M. Bakker and S. Robinson (2017), Evaluating the Quality of Business Survey Data before and after Automatic Editing. Working Paper, UN/ECE Work Session on Statistical Data Editing, The Hague.

Si, Y. and J.P. Reiter (2013), Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. Journal of Educational and Behavioral Statistics 38, 499-521.

Van Delden, A., R. Banning, A. De Boer and J. Pannekoek (2016), Analysing Correspondence between Administrative and Survey Data. Statistical Journal of the IAOS 32, 569-584.

Van Delden, A. and S. Scholtus (2017), Correspondence between Survey and Admin Data on Quarterly Turnover. Discussion Paper 2017-03, Statistics Netherlands, The Hague.