Estimation methods for the integration of administrative sources

Task 3: Comprehensive identification and enumeration of the possible estimation methods that could be used for the cases identified in task 2

Contract number:	Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
Responsible person at Commission:	Fabrice Gras Eurostat – Unit B1
Subject:	**Deliverable D3**
Date of first version:	11.01.2017
Version:	V2
Date of updated version:	25.01.2017
Written by :	Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang
Sogeti Luxembourg S.A.	Laurent Jacquet (project manager)
	Sanja Vujackov

Table of contents:

1 List of methods

Under the task 3, we provide a list of statistical integration methods grouped by statistical tasks.

Statistical tasks and methods

I. Data editing and imputation

- Most of the methods usually applied in the classical statistical context can be used in the setting of use and integration of administrative sources, as for instance automatic methods for error localization, outlier detection, imputation (see Memobust module: Editing administrative data). The methods are described in Memobust (2014a, 2014b, 2014c, 2014d, 2014e, 2014f, 2014g). More extensive descriptions and discussions can be found in De Waal et al., (2011).
- There are editing methods developed specifically for data obtained through an integration process. A description of the problem, the related context and the methods is reported in Appendix A, 'Micro-integration'.

II. Creation of joint statistical micro-data

- Data linkage: Identification of the set of unique units residing in multiple datasets.
  A description of the problem, the related context and a list of methods are given in Appendix B, 'Probabilistic Record Linkage'.
- Statistical matching: Inference of joint distribution based on marginal observations through the creation of a synthetic data set at micro level (micro objective). A description of the problem, the related context and a list of methods is given in Appendix C, 'Statistical Matching'.
- Depending on the confidentiality constraints and the context, linkage activity could include activities related to the building of synthetic and anonymised identifier through the use of hashing techniques (Smith and Shlomo, 2014). Hashing algorithms are used to anonymise persons identifying information such as names, dates of birth and addresses. The hash function, which converts a field into a condensed representation of fixed value, is a one-way process that is irreversible – once the hashing algorithm is applied it is not possible to get back to the original information without significant effort and the use of tools that are not available in the research environment. Hashing is mainly referred to the IT and Cryptography disciplines, no statistical methods are known at the moment.

III. Alignment of statistical data

IIIa. Alignment of units: harmonisation of relevant units, creation of target statistical units

No general statistical methods are available, while ad-hoc deterministic methods are generally used. For instance for business statistics, statistical units are derived from legal units using certain deterministic derivation rules that take account of ownership relations (Council Regulation 696/93; Memobust module "Derivation of Statistical Units").

Next, administrative units can be harmonised to the statistical units, by using data on the relations between the administrative units, the legal units and the statistical units.

IIIb. Alignment of measurements: harmonisation of relevant variables, derivation of target statistical variable

Until now, the alignment of measurements is generally performed by means of ad-hoc procedures. An example occurs at Statistics Netherlands with administrative data on wage components, hours worked and social benefit data (WS data) obtained from a government authority. Statistics Netherlands uses those data to publish monthly hourly wages of employees. The target output on hourly wages is derived such that the outcome is coherent across different NACE codes and such that monthly changes in hourly wages are not affected by incidental and additional salary payments. More precisely, the statistical target variable wage is computed from its administrative components as:

the fiscal salary (the administrative wage concept)
+ pension contribution
+ employers contribution to unemployment benefit
+ employer contribution to an employee savings regulation
- additional salary payments such as holiday allowances
- health insurance premium
- charges for private use of a business car.

Some of the above individual components are in turn derived from other variables in the WS data set.

In recent papers, a more general approach based on statistical modeling is emerging. It is based on *latent variable models*. The problem, the related context and the list of methods is in Appendix D, 'Variable harmonisation based on latent variable models'.

IV. Multisource estimation at aggregated level

IVa. Population size estimation: multiple lists with imperfect coverage of target population

The problem of estimating the population size by integrating data sources is described in Appendix E, 'Multiple-list models for population size estimation'. In the Appendix, a list of methods for dealing with different assumptions in this context is reported.

IVb. Univalent estimation: numerical and statistical consistent estimation of common variables

The problem of obtaining univalent estimates at cross-sectional level, the related context and the list of methods is in Appendix F, 'Statistical methods for achieving univalent estimates for cross-sectional data'.

The problem of obtaining univalent estimates at longitudinal level, the related context and the list of methods is in Appendix G, 'Macro-integration with a time component'.

IVc. Coherent estimation: aggregates that relate to each other in terms of accounting equations.

The problem, the context, and the methods are listed in Appendix F 'Statistical methods for achieving univalent estimates for cross-sectional data' in the sub-section 'Macro-Integration'.

References

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, New Jersey.

Memobust (2014a) Module Statistica Data editing - Deductive Editing.
https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014b) Module Statistical Data Editing - Selective Editing.
https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014c) Module Statistical Data Editing - Automatic Editing.
https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014d) Module Statistical Data Editing - Manual Editing.
https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014e) Module Statistical Data Editing - Macro Editing.
https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014f) Module Statistical Data Editing – Editing Administrative Data.
https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014g) Module Statistical Data Editing – Editing for Longitudinal Data.

https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

D. Smith and N. Shlomo, Privacy Preserving Record Linkage, Data Without Boundaries Deliverable D11.2, Report 2014-01, CMIST Working Paper (2014).
http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf

Appendix

Appendix A: Micro-integration

1. The statement of the problem

Integration of data sources at micro level may give rise to composite records that consist of a combination of values obtained from different sources: for instance an administrative data set combined with values obtained from a survey for the same units (obtained by record linkage), an integration of several surveys with non-overlapping units, in which case a unit from one source is matched with a similar (but not identical) unit from another source. In addition, records with values obtained from different sources can also arise as a consequence of item non-response and subsequent imputation in which case the two sources are the directly observed values versus the values generated by the imputation method.

In all these cases the composition of a record by combining information obtained from different sources may lead to consistency problems because the information is conflicting in the sense that edit rules that involve variables obtained from the different sources are violated. The purpose of reconciling conflicting micro-data (micro-integration) is to solve the consistency problems by making slight changes or adjustments to some of the variables involved. An illustrative example taken from Memobust (2014a) is provided.

Let us suppose in a business program we take the *total turnover* from an administrative data source and in a survey we measure the *main source of turnover* and *other sources of turnover.* When an integrated record is obtained, the balance constraint that should be fulfilled is

total turnover= main source of turnover + other sources of turnover

It can happen that the value of the total turnover observed in the admin data is different from the sum of the two components observed in a survey. In this case, data need to be changed in order to satisfy the balance edit.

2. The related scenarios

GSBPM

The GSBPM sub-phases essentially related to micro-integration are sub-phase 5.1 (Integrate) and sub-phase 5.4 (Impute). Sub-phase 5.3 is partly connected to micro-integration for the aspect concerning the definition of the edits.

Data configuration based on Komuso classification

Micro-integration is concerned with all the cases where the integrated record is composed of variables derived from different data sources, that is Data configuration 1, 3, 4 and 5 (see Section 'Usages in terms of data configurations' in Deliverable 1). In data configuration 1, micro-integration is generally applied to the values observed in the different data sources and pertaining to the same unit. In data configuration 3, micro-integration methods are applied to imputed values.

Data configuration 4 and 5 are characterised by having observed the same variables for the same unit in different data sources. In this case micro-integration methods will be applied to the unique value predicted in a single unit by taking into account the different observed values in the data sources.

3. The list of potential estimation methods

(details for the selected methods see T.5)

- Prorating,
- Minimum Adjustment Methods,
- Generalised Ratio Adjustments.

The methods are described in Pannekoek and Zhang (2015), Memobust (2014b), (2014c), and (2014d).

4. References

Memobust (2014a). Module Micro-Fusion - Reconciling conflicting microdata. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014b) Module Micro-Fusion – Prorating. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014c) Module Micro-Fusion - Minimum Adjustment Methods. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014d) Module Micro-Fusion - Generalised Ratio Adjustments. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Pannekoek, J. and Zhang, L.-C. (2015) Optimal adjustments for inconsistency in imputed data. Survey Methodology, vol. 41, pp. 127-144.

Appendix B: Probabilistic record linkage

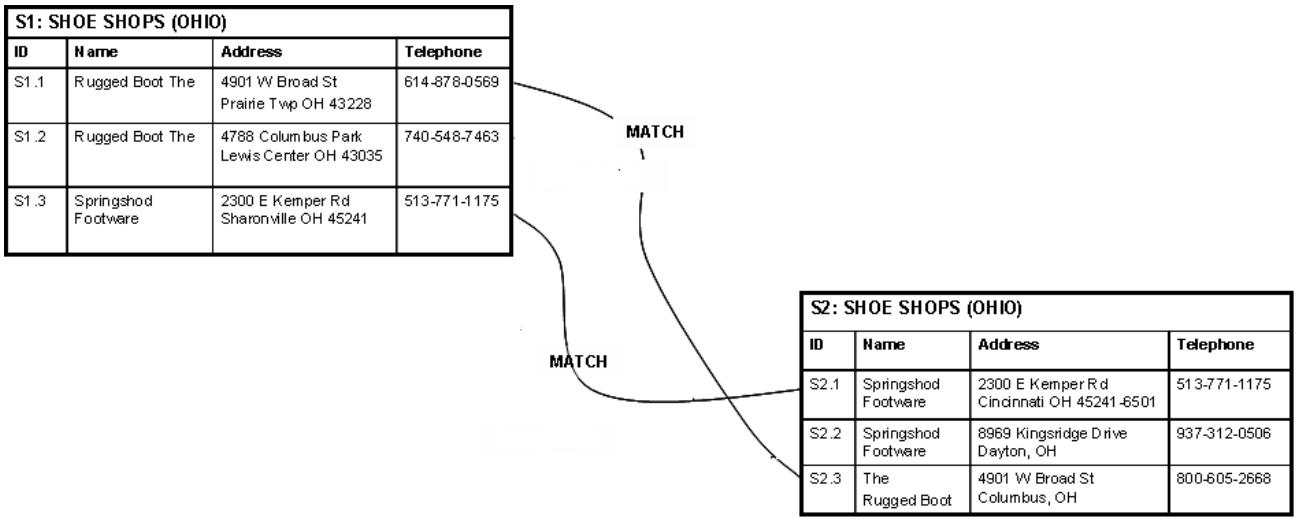
1. The statement of the problem

The integration of data at micro level with the main purpose of accurately recognize the same real world entity at individual micro level, even when differently stored in sources of various type, is known as record linkage. At the end, the record linkage process creates an integrated data set still composed by unit at micro level. Record linkage is also referred to as object identification, record matching, entity matching, entity resolution, reference reconciliation. In case of the National Statistical Institutes (NSIs) the joint use of statistical and administrative sources is a product of a rationalization of all the available sources to reduce costs, response burden and, most of all, to enrich the information collected in order to produce high quality statistics.

An important distinction between record linkage and statistical matching needs to be made, the former regards the "fusion" of sources composed mainly of the same units, partially or completely overlapping, e.g., in the case of integration of administrative registers and sample surveys, while the latter concerns the integration of different units, e.g. derived from different sample surveys.

The record linkage problem requires statistical estimation methods when unique identifiers are not available for all the units in the sources or the unique keys are affected by errors or a unique key can be derived from a combination of error-prone variables. For instances, the following figure, taken from Fortini et al. (2006), illustrates the integration of two data sets A and B, connecting records belonging to the same unit, on the basis of some common variables (name, address, telephone). It is possible that some agreement is not perfect (as in the telephone of the first record of the left data set and third record of the right data set), but the records still belong to the same unit.

Figure 1. Example of record linkage



The presence of errors in the unit identifiers and the use of statistical estimation methods for linking data files, may introduce linkage errors in the integrated data. This point needs to be taken into consideration in the subsequent analyses, in fact standard statistical estimation approach can produce inaccurate results in case of linkage errors. Statistical methods in order to evaluate and adjust for linkage errors are available.

It is the case to mention the deterministic record linkage approach that individuates links if and only if there is a full agreement of unique or common identifiers - the matching variables - or if it satisfied some a priori defined specific criteria. All the couples that present the same values on the selected matching variables or satisfy the a priori chosen criteria are assigned certainly to the set of matches. On the basis of this approach, the uncertainty in the matching procedure between two different databases is minimized but the linkage rate could be very low; the linkage quality can be assessed only by means of accurate and expensive clerical reviews or by means of re-linkage procedures.

Compared with the deterministic approach, the probabilistic one can solve problems caused by bad quality data and can be helpful when variables differently spelled, swapped or misreported are stored in the two data files; the attention in this section is mainly devoted to the probabilistic record linkage approach which is formally based on statistical methodology and allows also to evaluate the linkage errors, calculating the likelihood of the correct match. Generally, the deterministic and the probabilistic approaches can be adopted jointly in a two phase process: firstly the deterministic method is performed, choosing the high quality variables, then the probabilistic approach is adopted on the residuals which are the units not linked in the first step, the joint use of the two techniques depends on the aims of the whole linkage project (ESSnet ISAD, 2011).

Record linkage procedures may be applied to all the usages (both direct and indirect) delineated in Deliverable 1, that involves more than one source and requires the input and output data at micro level when error-free unique unit identifiers are lacking in the sources.

The usage of record linkage estimation techniques may be subject to restrictions, according to the privacy laws, but also may be used to assess both the risk of disclosure of the confidential micro-data file and the efficiency of the applied protection method (Shlomo and Skinner, 2010).

2. The related scenarios

GSBPM

The GSBPM identifies a sub-phase 5.1 called Integrate data within process 5. The record linkage estimation techniques belong to this sub-phase. However, the application of these estimation methods involves other sub-phases (for instances for coding and harmonizing the identifying variables stored in the different sources) and at the same time, other sub-phase may require linkage methods (for instances, the sub-phase 6.4 Applying disclosure control may use linkage for assessing the risk of disclosure). Moreover, if the results of the integration process may be affected by linkage errors, the propagation of these ones should be taken into account in all the sub-phases that make use of such integrated data.

Data configuration based on Komuso classification

Linkage estimation procedures may be required in the first and most basic data configuration 1 (see Section 'Usages in terms of data configurations' in Deliverable 1), called the "split-variable" case, concerning multiple cross-sectional data sources covering the target population where the different data sets contain different target variables and common variables able jointly to identify the units. This linkage step is also needed in the basic data configurations 4 and 5; the former is characterised by multiple cross-sectional data sources covering the target population with

overlapping units between the different data sources, the latter is characterised by under coverage of the target population in the different integrated sources.

In an ideal case, the data are error-free and the linkage process doesn't introduce extra errors in integrated data, to produce output statistics the data can simply be "added". More likely, statistical linkage procedures, able to dealt with error-prone data, are required; in addition, to produce the output statistics some adjustment and more sophisticated techniques are needed in order to avoid inaccurate estimates given by linkage errors.

3. The list of potential estimation methods

(details for the selected methods see T.5)

Standard approach

The standard statistical methods for probabilistic record linkage are by Fellegi and Sunter (1969) and Jaro (1989). They propose to define the record linkage problem as a classification one, where the matching status is unknown and needs to be estimated by means of a decision/classification rule.

Bayesian approach

In alternative, the record linkage problem can be faced in a Bayesian framework, (Fortini et al, 2001, Tancredi and Liseo, 2011) where prior distribution of the linkage probabilities and the number of matches and posterior distributions are simulated using MCMC. More recently, Steorts et al (2014) propose a different Bayesian approach, based on a parametric model for categorical data that addresses matching different files, detecting simultaneously duplicate records within the lists.The pattern of matches and non-matches was represented as a bipartite graph, in which records are directly linked to the true but latent individuals which they represent while they are only indirectly linked to other records..

Dealing with the impact of linkage error

The impact of linkage error on the subsequent analyses and the related adjustments are largely studied in recent years (e.g. Chambers 2009).

4. The ongoing research fields

In official statistics, some research activities are devoted to solve recent issues related to the use of record linkage techniques. One of the active research fields is the privacy preserving record linkage (PPRL), related to the use of pseudo-anonymized data, so as to preserve data from privacy issues and be able to link data provided in a more protective framework; see Fienberg, (2010) and Christen (2012) and for the most comprehensive survey Vatsalan et al, (2013).

Another topic is the simultaneous linkage of more than two sources, the so called multiple record linkage (Sadinle et al.). The longitudinal nature of some linkage process, that involves updated version of administrative data has to be taken into account as well. In these cases, the statistical estimation methods are sometimes confused with IT solutions as incremental record linkage Gruenheid et al (2014), and parallel record linkage Christen et al (2002).

5. References

Chambers R. (2009). Regression Analysis Of Probability-Linked Data. Official Statistics Research Series 4.

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Heidelberg: Springer, Chapter 8: Privacy Aspects of Data Matching.

Christen P., Hegland M., Roberts S., Nielsen O. M., Churches T., Lim K. (2002). Parallel computing techniques for high-performance probabilistic record linkage available at http://datamining.anu.edu.au/linkage.html

ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (ISAD), Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data, 2011, available at https://ec.europa.eu/eurostat/cros/content/deliverables-wp2di-integration-applications-isad-wp2_en

Fellegi I P and Sunter A B (1969) A theory for record linkage, Journal of the American Statistical Association, Vols. 64: 1183-1210.

Fienberg S. Hall R. (2010) Privacy-preserving record linkage, in: Privacy in Statistical Databases, Springer Lecture Notes in Computer Science, vol. 6344, Corfu, Greece, 2010, pp. 269–283.

Fortini, M., Scannapieco, M., Tosco, L., and Tuoto, T., 2006. Towards an Open Source Toolkit for Building Record Linkage Workflows. Proceedings SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS'06), Chicago, USA, 2006.

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001), On Bayesian record linkage, Research In Official Statistics, 4, 185–191.

Gruenheid A. Dong X. L., Srivastava D. (2014) Incremental Record Linkage. Proc VLDB Endowment 7 (9):697-708.

Jaro M.A. (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 84, 414–420.

Sadinle M., Hall R., and Fienberg S. Approaches to Multiple Record Linkage available at https://www.cs.cmu.edu/~rjhall/ISIpaperfinal.pdf

Shlomo, N., & Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey micro-data. Annals of Applied Statistics, 4(3), 1291-1310. DOI: 10.1214/09-AOAS317. Publication link: fbb6d0a6-32fd-41c0-9aac-6db410555bab

Steorts, Hall, Fienberg (2014) SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication, a new approach in the Bayesian framework, Journal of Machine learning research.

Tancredi, Liseo (2011) Hierarchical Bayesian approach to record Linkage and population size problems. The Annals of Applied Statistics, 2011, Vol. 5, No. 2B, 1553—1585.

Vatsalan, Christen, Verykios (2013) A taxonomy of privacy-preserving record inkage techniques Information Systems 38 (2013) 946–969.

Appendix C: Statistical matching

1. The statement of the problem

A statistical matching problem occurs when:

1. (*objective*) knowledge on a (possibly conditional) joint distribution function of a couple of random variables, or on some of the parameters representing the (conditional) joint relationship between the two random variables, or on a data set representative of the (conditional) joint distribution of the two variables, is requested;
2. (*available data*) the two random variables of interest are not available jointly in any datasets, but it is possible to use two random samples, each one observing one of the two random variables, both observing all the conditioning variables. Furthermore, the two samples should be representative of the same population, but the overlap between the two observed sets of units is either null or consists of subsets of units that are not representative of the population itself.

In item 1. it is possible to recognize the two most important outputs that statistical matching can produce: either macro (estimation of a joint distribution or of a relationship parameter) or micro (a data set with complete observations on the variables of interest, i.e. both the couple of random variables and the conditioning variables) .

Note that even if a unique personal identifier had been available in the two samples, it would have not been useful because the data sets to integrate do not overlap. Identification is not requested for the statistical matching problem. This statement should be considered as a warning: a statistical matching problem does not correspond to the identification of the most similar unit to the one to integrate (as stated in Okner, 1972, p. 327), hence adaptations of the record linkage methods for statistical matching purposes via the use of imputation procedures should be generally avoided (see Sims, 1972).

In fact, the main source of error in a statistical matching problem is not the linkage error, as in record linkage, but the possible introduction of specific relationship models between the variables never jointly observed by means of the statistical matching method itself. In a micro approach, the data set resulting from such an activity becomes representative of the imposed model, possibly very different from the truth. The typical imposed model in this framework is the statistical independence between the never jointly observed variables given the conditional ones. This is, for instance, the imposed model by any imputation procedure filling in the missing variables in the two observed samples by a function of only the conditioning variables.

In order to avoid a result that can be completely misleading, the new approaches to statistical matching aim at introducing a new form of uncertainty that goes beyond sample variability: uncertainty due to lack of joint knowledge on the couple of variables of interest.

2. The related scenarios

Statistical matching naturally conforms to the GSBPM phase "Process". Specifically it can be one of the methods to use in the sub-process 5.1 "Integrate". Anyway, given its nature in discovering information on joint variables never jointly observed, it can correspond also to the sub-phase "Derive new variables" (as a micro objective) or directly "Calculate aggregates" if the objective is macro.

According to what represented in Deliverable 1, this problem corresponds to a basic data configuration 3. There can be slight modifications to this basic data configuration:

1. when there exists a third sample of complete data (used in Renssen, 1998, and Singh 1993);
2. when knowledge on some of the relationship parameters is available (e.g. Paass, 1986).

3. The list of potential estimation methods

(details for the selected methods see T.5)

Micro-objective methods (imputation)

The most popular statistical matching methods have a micro objective and are based on imputation techniques. They are mainly based on cold deck methods (see Singh et al., 1993, and references therein). The resulting data set can be used for inferential purposes on the joint (conditional) distribution of the pair of variables of interest only if the conditional independence assumption between the pair of random variables of interest given the conditional ones holds. This assumption seldom applies, with the remarkable exception of the case one of the pair of variables of interest is highly correlated with one of the conditional variables. An example of this last approach is in Donatiello et al (2016).

Methods for complex survey designs

Statistical matching in the context of samples drawn according to complex survey designs have been defined in Rubin (1986, with the so called concatenated sample) and Renssen (1998, with the incomplete and synthetic two-way stratification).

Methods for uncertainty evaluation

Given the lack of joint information on the pair of variables of interest, the joint distribution given the conditional variables is non-identifiable. In other words, the traditional estimators of the joint distribution or its parameters (either based on the likelihood principle or in a Bayesian setting) select a group of equally plausible estimates (for maximum likelihood estimators, the so called likelihood ridge; in a Bayesian setting, the set of equally maximum a posteriori parameters, noting that as Rubin (1974) says the non-identifiable parameters prior distribution cannot be updated by the observed likelihood). This set of equally plausible estimates correspond to all those joint distributions that are compatible (or that can co-exist) with the maximized identifiable distributions, namely the marginal distribution of the conditioning variables and the two conditional univariate distributions of the pair of random variables of interest given the conditional ones.

This set of estimates is studied in D'Orazio et al (2006a, 2006b) in a maximum likelihood context. In a non-proper Bayesian case, this set of estimates is studied in Rubin (1986) and extended to a proper Bayesian context in Raessler (2002).

Other references discussing statistical matching and related methods are Memobust (2014), Memobust (2014a) and the documents produced by the *ESSnet on Integration of Surveys and Administrative Data (https://ec.europa.eu/eurostat/cros/content/isad-finished_en)* and by the CENEX project on Data Integration (https://ec.europa.eu/eurostat/cros/content/data-integration-finished_en).

4. The ongoing research fields

As far as the definition of uncertainty and its use in order to select a distribution, see Conti et al (2012 and 2015). Note that the same authors have analysed uncertainty for distributions in a non-parametric framework (Conti et al, 2012) and for ordered categorical data (Marella et al, 2013).

The study of uncertainty when some of the common variables are misclassified is in Di Zio and Vantaggi (2017).

In a Bayesian multiple imputation approach, the extensions in Reiter (2013) are very important because they give an answer to the establishment of variance for parameters estimated in a statistical matching context.

Ahfock et al (2016) propose a Gibbs sampler to study uncertainty in high-dimensional statistical matching problems.

For samples drawn according to complex survey designs, Conti et al (2015) establish the asymptotic properties of estimators, given that an evaluation of the available solutions for finite sample sizes do not lead to a definite conclusion (D'Orazio et al, 2009).

Finally, D'Orazio et al (2016) look for strategies in selecting the conditional variables when the two samples have many common variables.

5. References

Ahfock D., Pyne S., Lee S.X., McLachlan G.J. (2016) Partial identification in the statistical matching problem. *Computational Statistics & Data Analysis*, Volume 104, December 2016, Pages 79–90

Conti P.L., Marella D. Scanu M. (2015). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, to appear (available on line)

D'Orazio M., Di Zio M., Scanu M. (2006a). *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8

D'Orazio M., Di Zio M., Scanu M. (2006b). Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints, *Journal of Official Statistics*, **22**, 137-157.

D'Orazio M., Di Zio M., Scanu M. (2009). Uncertainty intervals for non-identifiable parameters in statistical matching. ISI conference, Durban (South Africa).

D'Orazio M., Di Zio M., Scanu M. (2016). The use of uncertainty to choose matching variables in statistical matching. 8th International Conference on Soft Methods in Probability and Statistics (SMPS), Rome, September 12-14 2016.

Di Zio M., Vantaggi, B. (2017). Partial identification in statistical matching with misclassification. To appear in International Journal of Approximate Reasoning.

Donatiello D., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016) The role of the conditional independence assumption in statistically matching income and consumption. In press, Statistical Journal of the IAOS

Fosdick, B. K., De Yoreo, M. and Reiter, J. P. (2016), Categorical data fusion using auxiliary data, *Annals of Applied Statistics*, to appear.

Marella D. Conti. P.L., Scanu M. (2013) Uncertainty analysis for statistical matching of ordered categorical variables. *Computational Statistics and data Analysis*, **68**, 311, 325

Marella D. Conti. P.L., Scanu M. (2016) How far from identifiability? A systematic overview of the statistical matching problem in a non-parametric framework. *Communications in Statistics – Theory and Methods*, in press. DOI#10.1080/03610926.2015.1010005. Journal ISSN 1532-415X

Memobust (2014) Module Micro-Fusion - Statistical Matching. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Memobust (2014a) Module Micro-Fusion - Statistical Matching Methods. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Okner, B.A. (1972) Constructing a new data base from existing micro-data sets: the 1966 merge file. *Annals of Economic and Social Measurement* **1**(3), 325–342.

Paass, G. (1986) Statistical match: evaluation of existing procedures and improvements by using additional information. In G.H. Orcutt, J. Merz and H. Quinke (eds) *Microanalytic Simulation Models to Support Social and Financial Policy*, pp. 401–422. Amsterdam: Elsevier Science.

Raessler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Reiter, J. P. (2012), Bayesian finite population imputation for data fusion, *Statistica Sinica*, **22**, 795 - 811.

Renssen, R.H. (1998) Use of statistical matching techniques in calibration estimation. *Survey Methodology* **24**, 171–183.

Rubin, D.B. (1974) Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* **69**, 467–474.

Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* **4**, 87–94.

Sims, C.A. (1972) Comments on Okner (1972). *Annals of Economic and Social Measurement* **1**(3), 343–345.

Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology* **19**, 59–79.

Appendix D: Variable harmonisation based on latent variable models

1. The statement of the problem

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 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.

This problem is related to the problem of reconciling conflicting micro-data from different sources (see 'Micro-integration'). Reconciling conflicting micro-data is used to solve inconsistencies between *related* variables from different data sources (composite records), while variable harmonisation applies to situations where different versions of the *same* variable are available.

2. The related scenarios

GSBPM

Variable harmonisation can be considered part of the GSBPM phase 5 (Process) in the sub-processes 5.1 (Integrate) and 5.4 (Impute).

Data configuration based on Komuso classification

In principle, the problem of variable definitions that differ between data sources can arise in all data configurations that involve complementary combinations of micro-data (data configurations 2–5 in Section 'Usages in terms of data configurations' in Deliverable 1). To be able to apply variable harmonisation, overlapping observed values from at least two different data sources are required for at least part of the units. Therefore, strictly speaking the method applies only to data configurations 4 and 5. The method can also be used for data configurations 2 and 3 if the relation between the observed variables and target variables has been estimated previously.

3. The list of potential estimation methods

Traditionally, deterministic derivation rules have been used to derive the target variables from the observed variables in each input data source (Bakker, 2011). Such a rule can be completely based on subject-matter knowledge, or it can contain parameters that are estimated from a data set where values of the target variable are available. An example of the latter approach is provided by Van Delden et al. (2016) for VAT turnover. These authors used a robust linear regression model to estimate the intercept *a* and slope *b* of a derivation rule of the form

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

Here, the parameters *a* and *b* vary by type of economic activity. If there are no differences in definition, then $a = 0$ and $b = 1$. In this application, turnover values from an existing survey were used to approximate the target turnover values, i.e., the survey variable was used as a 'gold standard' measurement of the target variable.

In practice, all observed variables may contain measurement errors. Harmonisation methods that account for this fact 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. Once the model has been estimated, the unobserved values of the target variable can be predicted and used to impute a harmonised variable for each unit. *Measurement error models* for micro-data have a long tradition among researchers in the social and behavioral sciences (Lord and Novick, 1968; Saris and Andrews, 1991) and in econometrics (Durbin, 1954; Bound, Brown and Mathiowetz, 2001). Recently, some statistical institutes have investigated the possibility of using error models for variable harmonisation.

For categorical data, measurement models based on *latent class analysis* can be used (e.g., Hagenaars and McCutcheon, 2002). In this case, the model describes the probability of observing each category of each observed variable, given the category of the latent target variable. After estimating the model, the latent category can be predicted for each unit and used as the harmonised target variable. 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).

To model measurement errors in numerical data, so-called *structural equation models* can be used (e.g., Bollen, 1989). Bakker (2012) advocated the use of such a model with combined administrative and survey data. An application of structural equation modelling to variable harmonisation is discussed by Scholtus, Bakker and Van Delden (2015). Alternatively, 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 the assignment of units to subsets is not known a priori. 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.

4. The ongoing research fields

In principle, measurement error models are an attractive approach for variable harmonisation compared to traditional deterministic derivation rules, because the latter rely strongly on subjective decisions. A drawback of this approach is that the models that have been proposed so far 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.

5. 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.

Bound, J., C. Brown and N. Mathiowetz (2001), Measurement Error in Survey Data. In: Heckman & Leamer (eds.), *Handbook of Econometrics*, Volume 5, pp. 3705-3843, Amsterdam: Elsevier.

Durbin, J. (1954), Errors in Variables. *Review of the International Statistical Institute* **22**, 23-32.

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.

Lord, F.M. and M.R. Novick (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

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.

Saris, W.E. and F.M. Andrews (1991), Evaluation of Measurement Instruments Using a Structural Modeling Approach. In: Biemer, Groves, Lyberg, Mathiowetz and Sudman (eds.), *Measurement Errors in Surveys*, pp. 575-597, New York: John Wiley & Sons.

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.

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.

Appendix E: Multiple-list models for population size estimation

1. The statement of the problem

The estimation of the unknown size of a target population is very important for official statistics. This is the situation when multiple sources (at least partially overlapping) are available but the combined data entail under coverage of the target population even in an ideal error-free state. In this case, the first statistical objective of the analysis is to estimate the unknown size of the target population collected in the different sources. The most common approach to face this task is the capture-recapture (CRC) method, which is originally been developed to estimate the size of animal populations (Fienberg, 1972; Bishop, Fienberg & Holland, 1975; IWGDMF, 1995). In case of two lists, the basic CRC method relies on the following assumptions (Wolters, 1986):

- the population is closed, so the population measured in both sources is the same;
- records from both sources can be linked without errors;
- the inclusion probability of being registered in the first source is independent of the inclusion probability in the second one;
- units have the same capture probabilities within each source (homogeneity probability assumption);
- over-count in both sources is negligible.

The violation of these assumptions can lead to serious bias in the CRC-estimation of the population size (e.g. Brown, Abott & Diamond, 2006; Van der Heijden et al., 2012; Baffour, Brown & Smith, 2013; Gerritse, et al., 2015a, 2016). CRC is in particular sensitive to violation of the assumption in the case of a low implied coverage, i.e. the second register overlaps greatly with the first register and adds relatively few new records to it (Brown, Abott & Diamond, 2006; Gerritse et al. 2016). So, several extensions of the CRC method were proposed in order to face problem connected to violation of the basic assumptions, we can divide them in two group: methods aiming at improving the CRC-method; alternatives methods.

While the class of CRC model was explicitly designed and developed to estimate the under-coverage, recently the estimation of the over-coverage has emerged as an important subject when studying population size estimation methods, in particular when the multiple lists are collected for administrative purposes. In fact, the risk that the administrative data contain units out-of-the target population, as well as duplicated units, is higher with respect to survey data collected for statistical purpose.

There are different approaches for measuring and/or integrating a measurement of the over-coverage into population size estimation, some of them dealt with different types of over-coverage separately, and therefore make separate adjustments to the population estimates (e.g. Statistics Canada, 2015 and ONS, 2012). Other methods have been developed in alternative to the CRC approach, both for the evaluation of the under-coverage and for the over-coverage.

2. The related scenarios

GSBPM

The population size estimation by means of CRC methods cannot be explicitly mapped into a single step of the GSBPM. For sure, the application of the CRC model requires a record linkage procedure, belonging to the sub-phase 5.1- Integrated data of the phase 5 - Process. Also methods and procedure relying to sub-phase 5.5- Derive new variable & statistical units can be involved in the population size estimation methods, due to the fact that one aims at profiling units that is not possible to observe in the considered sources. From the other hand, population size estimation implies to evaluate an aggregate measure (actually the population size) so the sub-phase 5.7 - Calculate aggregates can be involved, even if the procedure for obtaining the desired output is based on models and it is not the result of a simple aggregation.

Data configuration based on Komuso classification

The population size estimation methods are related to the basic data configuration 5 (see Section 'Usages in terms of data configurations' in Deliverable 1), characterised by a deviation from the basic data configuration 4, by which the combined data entail under coverage of the target population in addition, even when the data are in an ideal error-free state. Besides the under-coverage also the over-coverage, i.e., units not belonging to the target population that are in the data sources need to be considered. Record linkage techniques may be used to remove duplicated units, while model based approaches may be useful to classify whether or not units belong to the target population.

3. The list of potential estimation methods

(details for the selected methods see T.5)

The standard statistical method for evaluating the population size is based on the CRC model (also known as the Petersen or Lincoln-Petersen model). Extensions of this methodology that take into account the violation of the basic assumptions, aiming at improving the CRC model, include:

- methods connected to the correction for linkage errors (Ding and Fienberg, 1994; Di Consiglio and Tuoto, 2015);
- methods accounting for the over coverage (e.g. Zhang, 2015, Zhang and Dunne, 2017);
- methods dealing with partially overlapping populations (e.g. Zwane et al., 2004);
- methods relaxing the independence assumptions, considering more than two lists;
- methods dealing with heterogeneity in capture probabilities.

Alternatives methods with respect to the CRC model and its extensions have also been proposed in order to deal with overcoverage and partially overlapping populations; for instance, a latent class modelling approach is described in Di Cecco et al (2016).

Rasch models have been proposed in order to deal with dependencies between sources and heterogeneity of captures.

Finally, it is interesting to cite the Bayesian approaches to the estimation of the population size. Besides the Bayesian capture-recapture model (Ghosh and Norris 2005), in this field two main groups of methods can be identified: the former is mainly related to the record linkage topic and

its outcome in population size estimation (Steorts et al 2014, Tancredi and Liseo, 2011), the latter is connected to the use of Bayesian approaches in order to evaluate and projecting demographic stocks and flows in human population (Raftery et al 2012, Bryant and Graham, 2013).

Most of the cited approaches are also object of ongoing research in official statistics.

4. References

Baffour, B., J.J. Brown, P.W.F Smith, (2013). An investigation of triple system estimators in censuses. *Statistical Journal of the International Association for Official Statistics*, vol. 29, pp. 53-68

Bakker, B.F.M., & P. Daas, 2012, Some Methodological Issues of Register Based Research, *Statistica Neerlandica*, vol. 66, nr. 1, pp. 2-7

Bartolucci, F. and Forcina, A. (2001), Analysis of Capture-Recapture Data with a Rasch Type Model Allowing for Conditional Dependence and Multidimensionality, *Biometrics*, 57, 714–719

Bishop, Y., Fienberg, S., & Holland, P., (1975). *Discrete multivariate analysis, theory and practice* New York: McGraw-Hill.

Bryant and Graham, (2013). Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources. *Bayesian Analysis*, 8,3, pp.591—622

Brown, J.J., O. Abott & I.D. Diamond, (2006). Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society*. Series A (Statistics in Society), 169, 883-902

Di Cecco, D., Di Zio, M., Filipponi, D., Rocchetti, I. (2016). Estimating population size from multisource data with coverage and unit errors. *Proceedings of the ICES* V 20-23 June 2016 Geneva

Di Consiglio L., T. Tuoto, (2015). Coverage Evaluation on Probabilistically Linked Data, *Journal of Official Statistics*, vol. 31, nr. 3, 2015, pp. 415–429

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

Fienberg, S., 1972, The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, vol. 59, 409-439.

Gerritse, S. C., P.G.M. van der Heijden & B.F.M. Bakker, (2015a) Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models. *Journal of Official Statistics*, vol. 31, no. 3, pp. 357-379 http://dx.doi.org/10.1515/JOS-2015-0022

Gerritse, Susanna C. , Bart F. M. Bakker & Peter G. M. van der Heijden, (2015b). Different methods to complete datasets used for capture-recapture estimation, *Statistical Journal of the IAOS*, vol. 31, no. 4, pp. 613-627, 2015 (doi 10.3233/SJI-150938)

Gerritse, Susanna C. , Bart F. M. Bakker, Daan B. Zult & Peter G. M. van der Heijden, (2016). The effects of linkage errors and erroneous captures on the population size estimation (submitted)

Ghosh S.K., Norris J.L. (2004). Bayesian Capture-Recapture Analysis and Model Selection Allowing for Heterogeneity and Behavioral Effects, NCSU Institute of Statistics, Mimeo Series 2562, pp. 1-27

IWGDMF (International Working Group for Disease Monitoring and Forecasting), 1995, Capture-recapture and multiple record systems estimation. Part 1. History and theoretical development. *American Journal of Epidemiology*, 142, 1059-1068

ONS, 2012, 2011 Census: Over-count estimation and adjustment. Available at http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf

Raftery, A.E., Li. N., Ševčíková , H., Gerland, P. and Heilig, G.K. (2012). Bayesian probabilistic population projections for all countries. Proceedings of the National Academy of Sciences 109:13915-13921.

Statistics Canada, (2015). Census Technical Report: Coverage. Available at https://www12.statcan.gc.ca/census-recensement/2011/ref/guides/98-303-x/index-eng.cfm

Steorts, Hall, Fienberg (2014) SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication, a new approach in the Bayesian framework, Journal of Machine learning research

Van der Heijden, P.G.M., J. Whittaker, M. Cruyff, B. Bakker & R. van der Vliet, (2012). People born in the Middle East but residing in the Netherlands: invariant population size estimates and the role of active and passive covariates, *Annals of Applied Statistics*, vol. 6, no. 3, pp. 831-852

Wolter, K.M. (1986). Some coverage error models for Census data, *Journal of the American Statistical Association*, vol. 81, pp. 338-346

Zhang, L.-C., (2015). On modelling register coverage errors. Journal of Official Statistics, vol. 31, nr. 3, pp. 381-396

Zhang, L.-C. & Dunne, J. (2017). Trimmed Dual System Estimation. In *Capture-Recapture Methods in Social and Medical Sciences (eds. D. Böhning, J. Bunge and P. v. d. Heijden),* Ch. 15, pp. 237-262.

Zwane, E. N., Van der Pal-de, B., Van der Heijden, P.G.M., (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations, *Statistics in Medicine*, 23, pp. 2267-2281

Appendix F: Statistical methods for achieving univalent estimates for cross-sectional data

1. The statement of the problem

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. That is, the published estimates in different tables are 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. For instance, if one uses a standard weighting approach to produce estimates, where one multiplies observed counts or values with surveys weights, one may get different estimates for different tables, as different units and hence different survey weights may be used to produce different tables.

In principle, these differences are merely caused by "noise" in the data, such as sampling errors. So, in a strictly statistical sense, different estimates concerning the same variables are to be expected and are not a problem. However, different estimates would violate the one-figure (univalency) policy and form a problem from this point of view.

The problem is described in more detail in De Waal (2016).

2. The related scenarios

GSBPM

Macro integration can be considered part of the GSBPM phase "Process" in the subprocess 5.1 "Integrate", taking into account that the subprocess 5.1 includes "…combining data from multiple sources, as part of the creation of integrated statistics such as national accounts".

Data configuration based on Komuso classification

Univalency problems between different tables can occur when aggregated data are to be integrated, either with other aggregated data or with micro-data. So, univalency problems between different tables can occur for basic data configurations 6 and 7 (see Section 'Usages in terms of data configurations' in Deliverable 1).

Univalency problems can also occur for longitudinal data, i.e. for basic data configuration 8. For longitudinal data benchmarking methods can be applied (see T3_MAcroIntegTimeComponent).

3. The list of potential estimation methods

When micro-data and aggregated data have to be estimated and reconciled, several methods are available, such as repeated weighting (RW), repeated imputation (RI), mass imputation and macro-integration (see De Waal 2016 for a discussion of the pros and cons of these methods).

Repeated weighting

In the RW approach (see, e.g., Kroese and Renssen 2000, Houbiers et al. 2003, and Houbiers 2004) a separate set of weights is assigned to sample units for each table of population totals to be estimated. In this approach the tables to be estimated are estimated sequentially and each table is estimated using as many sample units as possible in order to keep the sample variance as low as possible. The combined data from administrative data sources and surveys are divided into rectangular blocks. Such a block consists of a maximal set of variables for which data on the same units has been collected. The data blocks are chosen such that each table to be estimated is covered by at least one data block. Item non-response in a block is assumed to be treated beforehand by means of imputation.

How a table is estimated depends on the available data. Data from an available administrative data source covering the entire population can simply be counted or added. Data only available from surveys are weighted by means of regression weighting (see Särndal, Swensson and Wretman 1992). In that case weights must be assigned to all units in the block to be weighted. For a survey one usually starts with the inverse inclusion probabilities of the sample units (i.e. the design weights), corrected for response selectivity. These weights are then further adjusted by calibrating them to previously estimated totals. For a data block containing the overlap of two surveys, one usually begins with the product of the standard survey weights from each of the surveys as starting weight for each observed unit, and then corrects these starting weights by calibrating to totals known from administrative data sources and previously estimated totals.

When estimating a new table, all margins of this table that are known or have already been estimated for previous tables are kept fixed to their known or previously estimated values, i.e. the regression weighting is calibrated on these known or previously estimated values. This ensures that the margins of the new table are consistent with previous estimates.

Mass imputation

In the mass imputation approach, one imputes all variables for all population units for which no value was observed (see Whitridge, Bureau and Kovar 1990, Whitridge and Kovar 1990, Shlomo, De Waal and Pannekoek 2009). This leads to a rectangular data set with values for all variables and all population units. After imputation estimates of totals can be obtained by simply counting or adding the values of the corresponding variables.

The approach relies on the ability to capture all relevant variables and relevant relations between them in the imputation model, and to estimate the model parameters sufficiently accurately. Given that all relevant variables and all relevant relations among them can be captured accurately by the imputation model, the approach is very straightforward.

However, it is generally impossible to capture all relevant variables and relations in the imputation model, simply because there are not enough observations to estimate all model parameters accurately. This implies that many relations in the imputed data are spurious and do not reflect the relations in the population. Owing to the lack of degrees of freedom, the imputation model will have to ignore some important relations in the data (see also Kroese, Renssen and Trijssenaar 2000).

For rich data sets with many variables, especially if not all tables to be estimated are specified beforehand, it seems that mass imputation cannot be applied successfully. For data sets with a limited number of variables and where all tables to be estimated are specified beforehand mass imputation seems to a viable option, however.

Repeated imputation

RI is the equivalent of RW. The important difference is how estimates are produced: in the case of RW by means of a weighting method, in the case of RI by means of an imputation method. Like RW, RI is a sequential approach where tables are estimated one by one. For some variables in a table estimates may have already been produced while estimating a previous table. Similar to RW, these variables are then calibrated to the previously estimated totals.

A prerequisite for applying RI is an imputation method that succeeds in preserving the statistical aspects of the true data as well as possible, that is able to satisfy specified edit rules and that is able to preserve previously estimated totals. Such imputation methods have recently been developed by, for example, Coutinho, De Waal and Shlomo (2013), Pannekoek, Shlomo and De Waal (2013), and De Waal, Coutinho and Shlomo (forthcoming).

Thus far the RI approach has only been applied to small data sets, not on large data sets arising in practice. Evaluations on large data sets remain to be carried out. Software for the RI approach is not yet generally available.

Macro-integration

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 an aggregated level are adjusted. The underlying micro-data 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), Byron (1978), Sefton and Weale (1995) and Magnus, Van Tongeren and De Vos (2000).

Traditionally, macro-integration has mainly been applied in the area of macro-economics, in particular for compiling the National Accounts, for example to adjust supply and use tables to new margins (see Stone, Champernowne and Meade 1942). Also, applications in other areas have been studied at Statistics Netherlands, namely for the reconciliation of tables of Transport and Trade Statistics (see Boonstra, De Blois and Linders 2011), for the Census 2011 (see Mushkudiani, Daalmans and Pannekoek 2012), and for combining estimates of labour market variables (see Mushkudiani, Daalmans and Pannekoek 2012).

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.

In the macro-integration approach often a constrained optimization problem is constructed (see, e.g., Stone, Champernowne and Meade 1942, and Byron 1978). A target function, for instance a quadratic form 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. Inequality constraints can be imposed in these quadratic optimization problems.

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).

4. The ongoing research fields

At Statistics Netherlands mass imputation, RI and macro-integration are being examined as alternatives to RW for the Census 2021. In particular, it is examined to what extent macro-integration can solve some problems that may occur with RW, such as handling cells with zero observations in a survey that have to sum up to a known positive population total. Also, research is being carried out with respect to applying macro-integration efficiently for very large reconciliation problems. Research on mass imputation at Statistics Netherlands is focused on imputing educational level for all persons belonging to the Dutch population. The imputed data set obtained in this way is meant to be used for the Census 2021. Part of the research on mass imputation or macro-integration has been or is funded by the EU.

As far as we are aware no research is being done on RW outside Statistics Netherlands. Imputation methods that either satisfy edit rules or preserve known totals, which are needed for RI, have recently been proposed in the literature. Mass imputation and macro-integration are relatively broad research topics. As far as we are aware no research on these topics with a specific focus on official statistics has recently been carried out outside Statistics Netherlands.

5. References

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.

Coutinho, W., T. de Waal and N. Shlomo (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. Journal of Official Statistics 29, pp. 299-321.

De Waal, T. (2016), Obtaining Numerically Consistent Estimates from a Mix of Administrative Data and Surveys. *Statistical Journal of the IAOS 32* pp. 231–243.

De Waal, T., W. Coutinho and N. Shlomo (forthcoming), Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions. To be published in Journal of Survey Statistics and Methodology.

Houbiers, M. (2004), Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics 20, pp. 55-75.

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

Kroese, A.H. and R.H. Renssen (2000), New Applications of Old Weighting Techniques; Constructing a Consistent Set of Estimates Based on Data from Different surveys. In: Proceedings of ICES II. American Statistical Association, Buffalo NY, pp. 831-840.

Kroese, B., R.H. Renssen and M. Trijssenaar (2000), Weighting or Imputation: Constructing a Consistent Set of Estimates Based on Data from Different Sources. Netherlands Official Statistics 15, pp. 23-31.

Magnus,J.T., 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.

Mushkudiani, N., J. Daalmans and J. Pannekoek (2012), Macro-Integration Techniques with Applications to Census Tables and Labour Market Statistics. Discussion paper, Statistics Netherlands.

Pannekoek, J., N. Shlomo and T. de Waal (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. Annals of Applied Statistics 7, pp. 1983-2006.

Särndal, C.E., B. Swensson and J. Wretman (1992), Model Assisted Survey Sampling. New York: Springer-Verlag.

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

Shlomo, N., T. de Waal and J. Pannekoek (2009), Mass Imputation for Building a Numerical Statistical Database. UN/ECE Work Session on Statistical Data Editing, Neuchâtel, Switzerland.

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

Whitridge, P., M. Bureau and J. Kovar (1990), Mass Imputation at Statistics Canada. In: Proceedings of the Annual Research Conference, U.S. Census Bureau, Washington D.C., pp. 666-675.

Whitridge, P. and J. Kovar (1990), Use of Mass Imputation to Estimate for Subsample Variables. In: Proceedings of the Business and Economic Statistics Section, American Statistical Association, pp. 132-137.

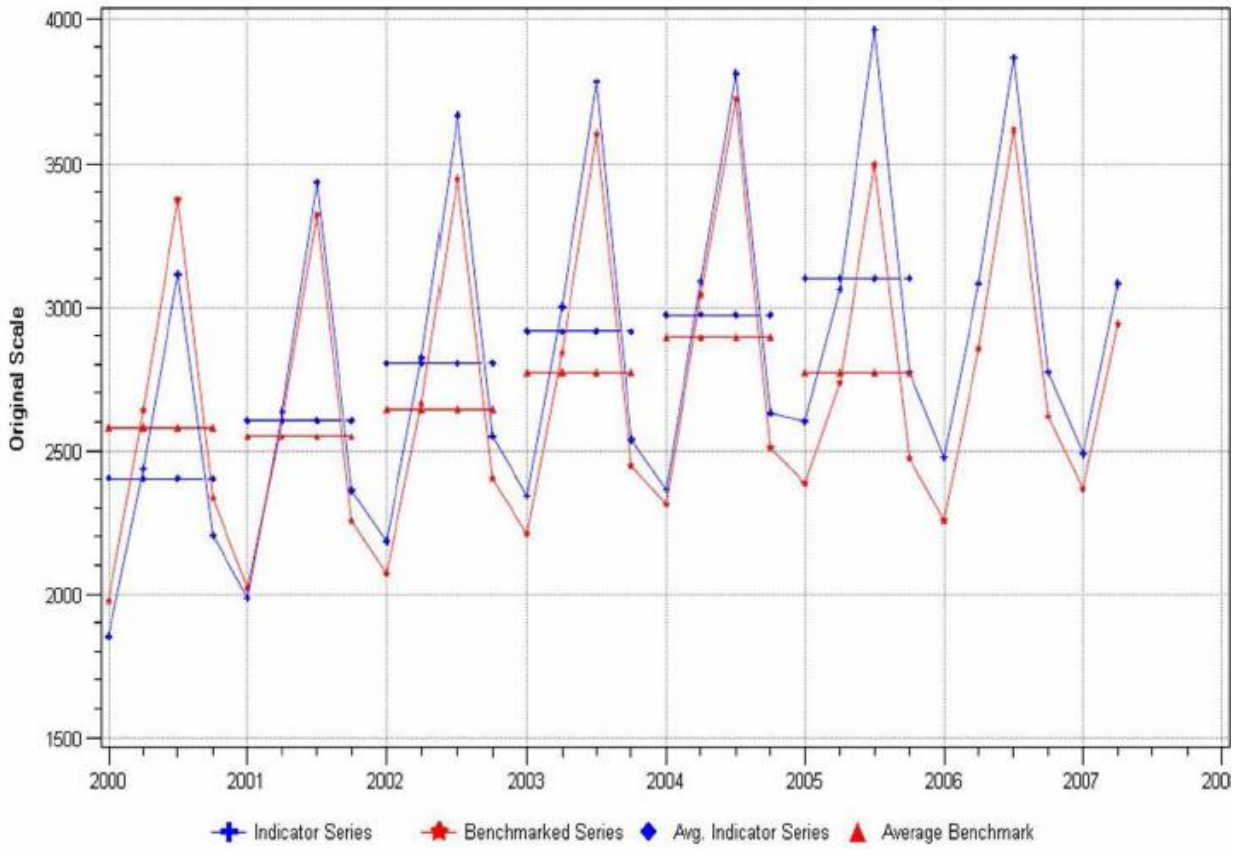
Appendix G: Macro-integration with a time component

1. The statement of the problem

The general problem behind macro-integration with a time component has a target variable with low frequency data (e.g. quarterly or yearly), and in addition one has a time series of the same variable with high frequency data (e.g. monthly data). The data of the lower frequency are usually based on a larger or more reliable set of data, and those data are often kept fixed. The high frequency data are then adjusted to them, where we would like to keep the adjustments (to be defined later) as small as possible. In the context of the combination of administrative data with survey data, an example is that a nearly complete set of administrative data are available on a quarterly basis whereas survey sample data on the same variable are available on a monthly basis.

An example of the problem at hand is given in Figure 1 taken from Fortier and Quenneville (2007). The example concerns on business surveys. The high frequency series concerns quarterly survey data that give information on the short-term movements, and are given as the blue lines in Figure 1, labelled as "indicator series". The low-frequency series concerns yearly figures that provide more reliable estimates, because more units are sampled and more variables (the editing process can be done more thoroughly). Those yearly values are given as red straight lines, labelled by "average benchmark"; they are computed as the yearly value divided by four. The reconciled values (after macro-integration) are given as the red lines and labelled as "benchmarked series".

Figure 1. Example of benchmarking (Fortier and Quenneville (2007)



2. The related scenarios

GSBPM

GSPBM has the sub-phase 5.7 called Calculate aggregates from the phase 5 Process. Benchmarking techniques belong to this sub-phase, although the sub-phase name calculate suggests that it concerns simply adding of outcomes. Benchmarking methods are estimation techniques that can take uncertainty of the estimates that are to be reconciled into account.

Data configuration based on Komuso classification

The problem of macro-integration of two time series is described in data configuration 8 (see Section 'Usages in terms of data configurations' in Deliverable 1). Figure 8 of data configuration 8 illustrates the situation where turnover data of enterprises are available on a monthly basis from a survey and data on a quarterly basis are available from the tax office (value added tax data) for the smaller units in combination with survey data for the largest units.

Data configuration 8 differs from that in data configuration 7 in that we have the time component to deal with. The important issue to deal with now is that one has to take account of the changes in the original time-series.

3. The list of potential estimation methods

(details for the selected methods see T.5)

For macro-integration with a time component there are three different types of methods: *benchmarking, temporal disaggregation* and *nowcasting*.

Benchmarking. The methods of *benchmarking* concerns the situation where the high and low frequency series concern the same variable. The idea is that some difference between the original and the adjusted figures is minimized subject to equality or inequality constraints. Maybe the most basic method to deal with benchmarking is a method that tries to minimize the adjustments to the original level estimates for each of the time periods. When each of the level estimates are adjusted with the same relative factor, this is referred to as *pro-rating*. Pro-rating leads to the so-called step-problem. The step-problem means that the adjusted (benchmarked) series may have a disproportional large adjustments in the transition from one low-frequency period to the next. To avoid step-problems methods have been developed that take account of the changes in the original time-series.

One well-known method that tries to preserve the changes in the original high-frequency series is the *growth rate preservation method* (GRP) by Causey and Trager (1981). That method minimizes the squared difference between the original period-to-period change and the corresponding adjusted growth rate. Another approach is the *movement preservation method* (MP) by Denton (1971). The most commonly applied variants of this method minimize the squared differences between adjustments of two subsequent time periods, also referred to as *first differences*. This can be further refined into minimizing additive or relative differences. A modification of the relative difference model, as proposed by Cholette (1984), is most popularly applied in practise (often called modified Denton). Mathematically, MP is easier to apply than GRP, because MP deals with a standard linearly constrained quadratic optimization problem, while GRP solves a more

difficult linearly constrained nonlinear problem (a ratio of two estimators) that can be efficiently solved by an interior-point-algorithm.

Note that benchmarking can also be applied to link two related time series, when there is a change of methods or revision in classifications between the two series (Fortier and Quenneville, 2007).

Temporal disaggregation. Methods for *temporal disaggregation* are very close to those of *benchmarking*. Temporal disaggregation is used for the situation that one has a low frequency series of a target variable that one wishes to disaggregate into a higher frequency series, but there are no data valuable for the target variable on this high frequency. Instead, for the disaggregation one uses data of other variable(s) that one considers to be indicative of the high-frequency changes of the target variable. An example of such a method is given by the seminal paper of Chow and Lin (1971). Chow and Lin (1971) minimizes the adjustments to the original level estimates for each of the time periods, so the disadvantage is that it leads to the already mentioned step-problem.

Nowcasting. An important problem with the estimation of the high-frequency series is that the benchmark of the low frequency series is not yet available when the preliminary estimates of the high frequency series are made: this is referred to as the timeliness issue. One of the means to cope with this timeliness issue is to pre-adjust the high-frequency series (e.g. Fortier and Quenneville, 2007). Another means to handle the timeliness issue is to use *nowcasting*. Nowcasting aims to provide an estimate for the present or recent past, based on a limited amount of data on the present. In particular, often information is used from sources that are available earlier, or with higher quality. Nowcasting can be done in many different ways, for instance by using an ARIMA model (Box and Jenkins, 1976) or by using a structural time-series model (STM) (Harvey, 1989). An example of the use of nowcasting using STM and benchmarking is given in Brakel and Krieg (2016).

4. The ongoing research fields

Several topics that relate to benchmarking receive attention in the field of official statistics. A first topic concerns research into new optimisation functions or extensions of existing functions for benchmarking, see for instance Bloem (2001) and Dagum and Cholette (2006) for an overview of methods. For instance, a comparison between the *growth rate preservation method* and the proportionate first differences can be found in Daalmans et al. (2016). Fortier and Quenneville (2007) developed a generalisation method of which a number of benchmarking optimisation functions are special cases. Sayal et al. (2017) introduce the use of wavelets as a means for benchmarking. Several authors have extended Denton's original method or univariate case with Stone's (1942) method of handling constraints between variables, e.g. Di Fonzo and Marini (2003 and 2005), Bikker et al. (2003) and Bikker and Buijtenhek (2006).

A second topic concerns seasonal adjustment. After seasonal adjustment the time series is no longer consistent with the yearly value before the adjustment. This can be solved by using a STM that incorporates a restriction on the yearly outcomes (Durbin and Quenneville, 1997), or by benchmarking the series afterwards (e.g. Dagum and Cholette, 2006). In addition, in the cross-sectional direction, comparing different aggregate levels, also inconsistencies may arise after seasonal adjustment. Research is being done to solve this by using a multi-variate STM (Bikker et al, forthcoming).

Finally we mention the problem that might occur that the reconciled values near the end of the time-series in the case of the proportionate Denton method tends to be over adjusted. This property of the Denton method has been described previously by Boot et al. (1967).

5. References

Bikker, R.P. and S. Buijtenhek (2006), Alignment of Quarterly Sector Accounts to Annual data. Statistics Netherlands, Voorburg, http://www.cbs.nl/NR/rdonlyres/D918B487-45C7-4C3C-ACD0-oE1C86E6CAFA/0/Benchmarking_QSA.pdf.

Bikker, R.P., Brakel, J. van den, Krieg, S., Ouwehand, P. and R. van der Stegen (forthcoming), Consistent multivariate seasonal adjustment for gross domestic product and its breakdown in expenditures

Bikker, R.P., J. Daalmans & N. Mushkudiani (2013), Benchmarking Large Accounting Frameworks: a Generalised Multivariate Model. Economic Systems Research 25, pp. 390-408.

Bloem, A., R. Dippelsman, and N. Mæhle, (2001), Quarterly National Accounts Manual: Concepts, Data Sources, and Compilation, (Washington, D.C. International Monetary Fund).

Boot, J. C. G., Feibes, W. and J. H. C. Lisman, 1967. Further Methods of Derivation of Quarterly Figures from Annual Data, Applied Statistics, 16, 65-75.

Box, G. E. P. and G.M. Jenkins (1976). Time Series Analysis: Forecasting and Control (2nd ed.). San Francisco, CA: Holden-Day.

Brakel, J. van den and S. Krieg, (2016). Estimating monthly turnover with incomplete register data. CBS report (available upon request).

Causey, B. and M.L. Trager (1981). Derivation of Solution to the Benchmarking Problem: Trend Revision. Unpublished research notes, U.S. Census Bureau, Washington D.C. Available as an appendix in Bozik and Otto (1988).

Cholette, P. (1984). Adjusting sub-annual series to yearly benchmarks, Survey Methodology, 10, 35-49.

Chow, G.C. and A. Lin (1971), Best Linear Unbiased Interpolation, and Extrapolation of Time Series by Related Series. Rev. Economics and Statistics 53, pp. 372-375.

Daalmans J., di Fonzo, T., Mushkudiani, N. and R.P. Bikker (2016). Removing the Gap between Annual and Sub- Annual Statistics based on Different Data Sources. Proceedings of the Fifth International Conference on Establishment Surveys, June 20-23, 2016, Geneva, Switserland: American Statistical Association.

Dagum, E.B. and Cholette, P. (2006). Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series Data. New York: Springer-Verlag, Lecture Notes in Statistics, volume186.

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

Di Fonzo, T. and M. Marini (2003), Benchmarking systems of seasonally adjusted time series according to Denton's moving preservation principle. University of Padova, Available at www.oecd.org/dataoecd/59/19/21778574.pdf.

Di Fonzo, T. and M. Marini (2005), Benchmarking a system of Time Series: Denton's movement preservation principle vs. data based procedure. University of Padova, Available at http://epp.eurostat.cec.eu.int/cache/ITY_PUBLIC/KSDT- 05-008/EN/KS-DT-05-008-EN.pdf

Durbin, J. and B. Quenneville (1997). "Benchmarking by state space models". International Statistical Review 65, 23–48.

Fortier, S. and B. Quenneville (2007). Theory and Application of Benchmarking in Business Surveys. . Proceedings of the Third International Conference on Establishment Surveys, June 18-21, 2007, Montreal, Quebec, Canada: American Statistical Association.

Harvey (1989), Forecasting, structural time series models and the Kalman filter, Cambridge University Press.

Sayal, H., Aston, J.A.D., Elliott, D and H. Ombao (2017). An introduction to applications of wavelet benchmarking with seasonal adjustment. Journal of the Royal Statistical Society Series A. (forthcoming)

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