Estimation methods for the integration of administrative sources

Task 5b: Review of estimation methods identified in Task 3 – a report containing technical summary sheet for each identified estimation/statistical method

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 D5b**
Date of first version:	30.03.2017
Version:	V1
Date of updated version:	-
Written by :	Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang, Pim Ouwehand
Sogeti Luxembourg S.A.	Laurent Jacquet (project manager)
	Sanja Vujackov

Method 1: T5_21_ State space models

Deliverable D5b gathers the presentation of the methods, the contextual framework as well as the conditions of applications, the pros and cons, a possible example of use in NSIs, and the related software.

LIST OF ESTIMATION METHODS

I. Data editing and imputation:

1. Deductive editing
2. Selective editing
3. Automatic editing
4. Manual editing
5. Macro-editing
6. Deductive Imputation
7. Model-Based Imputation
8. Donor Imputation
9. Imputation for Longitudinal Data (Little and Su Method)
10. Imputation under Edit Constraints
11. Outliers /extreme values detection
12. Generalised Regression Estimator
13. Estimates with Model-Based Methods
14. EBLUP Area Level for Small Area Estimation (Fay-Herriot) Method
15. Small Area Estimation Methods for Time Series Data

II. Creation of joint statistical micro data:

16. Data Fusion at Micro Level (relevant choice of Statistical Matching Methods)
17. Matching of Object Characteristics (Unweighted & Weighted Matching)
18. Probabilistic Record Linkage
19. Reconciling Conflicting Micro-data: Prorating, Minimum Adjustment Methods, Generalised Ratio Adjustments
20. Data hashing & anonymisation techniques

III. Alignment of statistical data:

21. State space models (estimation of unobserved variable and possible application to the alignment of statistical data)
22. Temporal Disaggregation/benchmarking methods: Denton's Method & Chow-Lin Method

IV. Multisource estimation at aggregated level:

23. RAS
24. Stone's Method
25. Harmonisation based on latent variable models
26. Multiple-list models for population size estimation
27. Statistical methods for achieving univalent estimates for cross-sectional data
    - 27.1. Repeated weighting
    - 27.2. Mass imputation
    - 27.3. Repeated imputation
    - 27.4. Macro-integration

1. Purpose of the method

A large part of the statistics produced by national statistical institutes involve time series data. For users of these statistics, it is important that those time series are consistent with each other. However, since different data sources may have been used and many methodological steps are taken in constructing the statistics, this is not always the case. Therefore, often some specific effort is needed in order ensure consistency is some way.

The problem can arise in different forms, and can require different kinds of consistency. A first situation is the case were a target variable is measured at different frequencies, e.g. at quarterly and annual frequency, or at monthly and quarterly frequency. It is then important to ensure temporal consistency, that is, to make sure that the time series observations at the higher frequency are consistent with the time series at the lower frequency by satisfying a certain relation in time between the two. For example, the average monthly growth rate of a certain industry should be equal to the quarterly growth rate for the three months belonging to the same quarter, or the quarterly total turnover should be equal to the sum of the three monthly turnovers. This problem is called the benchmarking problem.

Another situation occurs where there are multiple time series (at the same frequency) that have some sort of relation that should be satisfied. For example turnover for a certain industry is made up of the turnover of certain subindustries, and so the time series of the subindustries add up to the time series of the industry as a whole. When some sort of methodology is applied to each of the series, such as calendar adjustments or seasonal adjustment, this additivity is lost. This problem can be called contemporaneous consistency.

A third situation occurs when data is collected at different frequencies and is combined (aligned) to make statistics at the desired frequency. For example, companies report their turnover to the tax authorities on a monthly or quarterly basis. When only the monthly data from the tax register is used to estimate total monthly turnover for a certain industry, this is likely to be subject to a large amount of selectivity. By using both monthly and quarterly data a better estimate can be obtained.

Although many different solutions exist for the above problems, here we focus on time series methods, in particular on the opportunities state space models offer.

2. The related scenarios

2.1. Usages and Komuso Data Configurations (deliverable 1):

Alignment; The methods apply to Configuration 8 where longitudinal data are available.

2.2. Statistical tasks:

Temporal measurement alignment.

2.3. Possible competing methods or related methods could be quoted in this part:

Alternative, related time series methods for temporal consistency are Denton (Denton, 1971) and Chow-Lin (Chow and Lin, 1971). Also other time series and non-time series methods exist.

3. Description of the method

State space models (SSM)

State space models (Harvey, 1989; Durbin en Koopman, 2012) offer a versatile environment for modelling many time series problems. This class of models describes time series as a sum of several (usually unobserved) components, where each of these components follows a stochastic model. A time series is typically assumed to consist of a level, a slope, and a seasonal component. Also cyclical components and explanatory variables can be included in the model. Each of the components is modelled explicitly and thus, after estimation, makes it clear how much each of the components contributes to the dynamics of the time series.

Estimating a SSM is not always easy, since it contains many unknowns: i) the unobserved state variables, and ii) the variances of the disturbance terms of each of the time series components, also called the hyperparameters. Estimating such a model therefore consists of two steps. First, via a Maximum Likelihood procedure the hyperparameters are estimated. Next, the Kalman filter is applied in order to recursively estimate the state variables. When all observations are treated, a backward recursion is applied that smoothes the estimates.

SSM and their estimation method have several advantages. First of all, by selecting only certain time series components or restricting their hyperparameters, many special cases arise. Secondly, the estimation methods can deal with missing observations. For this reason, they be applied in multivariate situations were some series start or end earlier than other series, causing so called 'ragged edges'. Another advantage, is that they can be easily extended to deal with multiple frequencies, as is required for solving some of the above problems.

Temporal consistency

This problem, also referred to as benchmarking, can be translated into a state space framework. In Durbin and Quenneville (1997) a model is developed for a situation with a monthly series of observations which is obtained from sample surveys and is thus subject to survey errors. Also, a series of annual values is available, which is considered more accurate than the survey observations. Two solutions are presented. The first is a two-stage method in which first fit a state space model is fitted to the monthly data alone and then combined with the benchmark series. In the second solution a single series is constructed from the monthly and annual values together and a SSM is fitted in a single stage.

Contemporaneous consistency

Often several series are expected to satisfy some sort of additive relationship. In Bikker et al. (2016) a model is presented for Gross Domestic Product (GDP) and its breakdown in underlying categories or domains. This system is consistent up to the moment seasonal adjustment is applied. After adjusting all of the series separately, additive is lost. A multivariate state space model is proposed in which total GDP and its breakdown in underlying domains are modelled in one model. Restrictions are imposed that the sum over the different time series components for the domains are equal to the corresponding values for the total GDP. In the proposed procedure this approach is applied as a pre-treatment to remove outliers, level shifts, seasonal breaks and calendar effects, while obeying the aforementioned consistency restrictions. Subsequently, X-13ARIMA-SEATS is used for seasonal adjustment. This reduces inconsistencies remarkably.

Remaining inconsistencies due to seasonal and calendar adjustment are removed with a benchmarking (Denton) procedure.

Combining time series

This concerns integrating several time series, possibly at different frequencies, in order to measure a single underlying phenomenon. For this, in a multivariate SSM, two or more series are modelled simultaneously, which allows to model cross-sectional dependency between these series.

In Van den Brakel en Krieg (2016), a model is presented for estimating a monthly time series of turnover with incomplete register data. Since enterprises can choose to declare VAT on a monthly, quarterly or annual basis, it is a challenge to compile short term business statistics on a monthly frequency from this data source. A SSM is proposed that combines time series on a monthly and a quarterly frequency with the purpose to estimate short term business statistics on a monthly frequency. For this, two time series are used: one with all information from both monthly and quarterly VAT declarations, and one with only the monthly VAT declarations. These series are then assumed to share a correlated trend. By modelling both series simultaneously, the estimate of the underlying trend improves.

Harvey and Chung (2000) treat a similar situation, in which the objective is to estimate the underlying change in employment. This series is measured at a quarterly frequency. Another measure of employment (based on administrative sources) is measured at a monthly frequency. By using the latter as an auxiliary variable, a bivariate model is constructed in which the two series are considered to follow similar underlying trends. In this way monthly estimates of the change in employment can be obtained.

4. Examples

See Section 3.

5. Input data (characteristics, requirements for applicability)

Multiple time series that are to be aligned. This can be one ore more series plus a benchmark series, or multiple series plus a restriction on these time series.

6. Output data (characteristics, requirements)

Multiple time series that are consistent according to some criterion.

7. Tools that implement the method

A powerful tool for programming state space models is the Ox package (Doornik, 2009), accompanied by Ssfpack (Koopman et al., 2008), containing routines for carrying out computations involving the statistical analysis of time series models in state space form. Another tool is the statistical programming language R (R Core Team, 2017), accompanied by several packages for modelling state space models (e.g., Helske, 2017). In both software tools, tailor-made programs and scripts are used.

8. Appraisal

State space models offer good opportunities for modelling a broad spectrum of time series problems. However, the optimal model may vary from situation to situation and requires expert knowledge. Problems can arise when these models are estimated. Especially in complex (multivariate) settings, this can lead to estimation problems. Such as instability of (hyper)parameter estimates, or large computation times.

Therefore, when applying these models some experience is required. These potential drawbacks may also hamper the application of these models in a more automatic fashion.

It should be realised that when performing one of the methods that yield temporal or contemporaneous consistency, some of the time series are altered. Although consistency is achieved, this may be at the cost of a lower quality of those individual series. For this reason, for example Eurostat (2015) advises not benchmark seasonally adjusted series to the unadjusted data or the calendar adjusted data, unless strong users' requirements justify the benchmarking.

9. References

Bikker, R., J. van den Brakel, S. Krieg, P. Ouwehand, R. van der Stegen (2016), Consistent Multivariate Seasonal Adjustment for Gross Domestic Product and its Breakdown in Expenditures, Statistics Netherlands report (available upon request).

Brakel, J. van den, en S. Krieg (2016), Estimating monthly turn- over with incomplete register data, Statistics Netherlands report (available upon request).

Chow, G. and Lin, A. (1971), Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. The Review of Economics and Statistics 53, 372–375.

Denton, F. (1971). Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization, Journal of the American Statistical Association, 66, 99-102.

Doornik, J.A. (2009). An Object-oriented Matrix Programming Language Ox 6. London: Timberlake Consultants Press.

Durbin, J. and S. J. Koopman (2012), Time series analysis by state space methods, Oxford University Press.

Durbin, J. and B. Quenneville (1997), Benchmarking by state space models, International Statistical Review, 65, 1, 23-48.

Eurostat (2015), ESS Guidelines on seasonal adjustment, 2015 edition, European Union, Luxembourg

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

Harvey, A. and C. Chung (2000), Estimating the underlying change in employment in the UK, Journal of the Royal Statistical Society Series A, 162, 3, 303-339.

Helske, J. (2016), KFAS: Exponential Family State Space Models in R, http://cran.r-project.org/package=KFAS.

Koopman, S.J., N. Shephard, and J.A. Doornik (2008). SsfPack 3.0: Statistical algorithms for models in state space form. London: Timberlake Consultants Press.

R Core Team (2017), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org

Estimation methods for the integration of administrative sources

Task 5b: Review of estimation methods identified in Task 3 – a report containing technical summary sheet for each identified estimation/statistical method

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 D5b**
Date of first version:	14.03.2017
Version:	V1
Date of updated version:	-
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

Method 1: T5_22_ Benchmarking methods / temporal disaggregation

Deliverable D5b gathers the presentation of the methods, the contextual framework as well as the conditions of applications, the pros and cons, a possible example of use in NSIs, and the related software.

LIST OF ESTIMATION METHODS

I. Data editing and imputation:

1. Deductive editing
2. Selective editing
3. Automatic editing
4. Manual editing
5. Macro-editing
6. Deductive Imputation
7. Model-Based Imputation
8. Donor Imputation
9. Imputation for Longitudinal Data (Little and Su Method)
10. Imputation under Edit Constraints
11. Outliers /extreme values detection
12. Generalised Regression Estimator
13. Estimates with Model-Based Methods
14. EBLUP Area Level for Small Area Estimation (Fay-Herriot) Method
15. Small Area Estimation Methods for Time Series Data

II. Creation of joint statistical micro data:

16. Data Fusion at Micro Level (relevant choice of Statistical Matching Methods)
17. Matching of Object Characteristics (Unweighted & Weighted Matching)
18. Probabilistic Record Linkage
19. Reconciling Conflicting Micro-data: Prorating, Minimum Adjustment Methods, Generalised Ratio Adjustments
20. Data hashing & anonymisation techniques

III. Alignment of statistical data:

21. State space models (estimation of unobserved variable and possible application to the alignment of statistical data)
22. Temporal Disaggregation/benchmarking methods: Denton's Method & Chow-Lin Method

IV. Multisource estimation at aggregated level:

23. RAS
24. Stone's Method
25. Harmonisation based on latent variable models
26. Multiple-list models for population size estimation
27. Statistical methods for achieving univalent estimates for cross-sectional data
    - 27.1. Repeated weighting
    - 27.2. Mass imputation
    - 27.3. Repeated imputation
    - 27.4. Macro-integration

1. Purpose of the method

Estimates of time series on the same phenomenon, but that are published with different frequencies, might confuse users. For instance one might have a monthly indicator of the business cycle based on a small survey followed by quarterly figures with (partly) the same variables that are based on a larger set of administrative data. It might turn out that the quarterly results derived from the monthly survey do not coincide with the later published quarterly figures based on the administrative data. To avoid the issue of publishing two figures on the same phenomenon, benchmarking methods have been developed.

Temporal disaggregation is very similar to benchmarking, but the starting point is slightly different. Temporal disaggregation is used for the situation that one has a low frequency time series for a target variable and one aims to publish it at a higher frequency, but there are no data available for this target variable at the higher 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.

2. The related scenarios

2.1. The high frequency series may be based on sample survey data that are benchmarked upon a low frequency series of (more complete) administrative data (indirect estimation – creation of population benchmarks). Also low frequency survey data may be disaggregated with administrative data containing a variable indicative of the high frequency changes of the target variable (indirect estimation – use administrative data in a predictive setting). It may also be possible that a low frequency series of administrative data (used in direct estimation) is disaggregated with an indicative survey variable.

2.2. Statistical tasks: "Univalent estimation".

2.3. An alternative method is the use of state-space models and cubic splines.

3. Description of the method

Methods for benchmarking and temporal disaggregation are so similar that they are treated here in one description. We will subsequently treat methods of pro-rating, Chow and Lin's (1971) method, Denton-based methods and the growth rate preservation method. The general idea of all of those methods is that some difference between the original and the adjusted figures is minimized subject to equality or inequality constraints.

Pro-rating

Pro-rating is a method that adjusts the level estimates (for benchmarking or temporal disaggregation) with the same relative factor. Unfortunately, pro-rating leads to the so-called step-problem. The step-problem implies that the adjusted (benchmarked) series may have considerable adjustments in the transition from one low-frequency period to the next.

Several more advanced methods are available. These methods can be divided into a statistical modelling approach and a purely numerical approach. Statistical models explicitly take any supplementary information about the underlying error mechanism and stochastic properties of the series into account, while numerical approaches have the advantage of being simple, robust and applicable for large-scale applications (see e.g. Bloem *et al.*, 2001).

Chow and Lin.

A seminal paper for the statistical modelling approach is Chow and Lin (1971). Chow and Lin (1971) minimizes the adjustments to the original level estimates for each of the time periods by using a regression approach. In this regression approach, the high frequency target variable is estimated as a linear regression function of a number of high frequency indicator variables subject to the constraint that the estimated high frequency levels of the target variable add up to the low frequency level estimates. Likewise to pro-rating, the Chow and Lin's method also leads to the already mentioned step-problem. Therefore, pro-rating and Chow and Lin's method are generally not considered to be suitable methods when the changes in the original time series are to be retained.

However, several extensions of Chow and Lin are available that do consider the step-problem. These are regression models with autoregressive error terms. Examples are the: Fernandez random walk model (Fernandez, 1981), the Litterman random walk Markov model (Litterman, 1983) and ARIMA methods (Hillmer and Trabelsi, 1987).

Denton-based methods

A seminal paper for the purely numerical approach is Denton (1971). The most commonly applied variants of the Denton method minimize the squared differences between adjustments of two subsequent time periods, also referred to as *first differences*. For instance, let x_t be the original value of survey variable x in month $t = 1, \dots, M$ and let $\hat{x}_t$ be its reconciled value. Further, let y_k be the value of the same variable, but now observed in an administrative source on a quarterly basis, that is, we have quarter $k = 1, \dots, K$, with $K = M/3$. We can now for instance minimise:

$$\min_{\hat{x}_t} \sum_{t=2}^{M} \left(\hat{x}_t - x_t - (\hat{x}_{t-1} - x_{t-1})\right)^2, \tag{1}$$

subject to

$$\sum_{t=3(k-1)+1}^{3k} \hat{x}_t = y_k, \qquad k = 1, \dots, K \tag{2}$$

This minimisation is also referred to as the additive differences. A variation to this is minimizing the relative differences:

$$\min_{\hat{x}_t} \sum_{t=2}^{M} \left(\frac{\hat{x}_t}{x_t} - \frac{\hat{x}_{t-1}}{x_{t-1}}\right)^2, \tag{3}$$

subject to the constraint in (2). Equation (3), which is proposed by Cholette (1984), is often known as the Denton method, which is actually a slight modification of the original method (Denton, 1971).

Growth rate preservation method (GRP)

The Denton-based methods are also known as movement preservation methods (MP). Another numerical method, the so-called GRP-method was introduced by Causey and Trager (1981). This method minimizes the squared differences between the original period-to-period changes and the corresponding adjusted growth rate:

$$\min_{\hat{x}_t} \sum_{t=2}^{M} \left(\frac{\hat{x}_t}{\hat{x}_{t-1}} - \frac{x_t}{x_{t-1}}\right)^2, \tag{4}$$

subject to the constraint in (2).

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.

So far, we described the univariate case. One might also have the situation with several time-series that are benchmarked simultaneously, with constraints between the time series. 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. (2013) and Bikker and Buijtenhek (2006). Stone's method is explained in a separate description.

It is also interesting to note that there are several extensions or new optimisation functions, see for instance Bloem *et al.* (2001) and Dagum and Cholette (2006) for an overview of methods. Fortier and Quenneville (2007) developed a generalisation function of which a number of benchmarking optimisation functions are special cases.

4. Examples

As mentioned in Task 3, quarterly business survey data can be benchmarked upon yearly business survey data (Fortier and Quenneville, 2007). At Statistics Netherlands, multivariate times series of quarterly accounts are benchmarked upon the yearly accounts that become available after the four quarters of the year have been completed.

5. Input data (characteristics, requirements for applicability)

The input data for benchmarking and temporal disaggregation are macro data.

6. Output data (characteristics, requirements)

The output data of benchmarking and temporal disaggregation are macro data.

7. Tools that implement the method

Currently tailor-made programs and scripts are used.

8. Appraisal

A strong point of benchmarking through macro-integration type methods is that you can explicitly control which aspect of the original time-series (e.g. levels, absolute or relative changes) one aims to preserve as well as possible.

A good point is also that some of these benchmarking methods can reconcile time series while simultaneously accounting for cross-sectional restrictions.

An advantage of benchmarking methods over structural time-series is that the mathematical structure of benchmarking is somewhat more straightforward.

A limitation of the currently available methods occurs in the case of seasonally adjusted time-series. When benchmarked series are seasonally adjusted, it is possible to take care of the time-related restrictions between the high and the low-frequency series but it is not yet possible to simultaneously account for cross-sectional relations.

9. 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., 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).

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, Distribution and Extrapolation of Time Series by Related Series. Rev. Economics and Statistics 53, pp. 372-375.

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 movement 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

Fernandez, R.B. (1981), A methodological note on the estimation of time series. The Review of Economics and Statistics 63, pp. 471-476.

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.

Hillmer S.C. and A. Trabelsi (1987), Benchmarking of economic time series," Journal of the American Statistical Association 82, pp. 1064-1071.

Litterman, R.B. (1983), A random walk, markov model for the distribution of time series, Journal of Business and Statistics, 1, pp. 169-173.

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

Estimation methods for the integration of administrative sources

Task 5b: Review of estimation methods identified in Task 3 – a report containing technical summary sheet for each identified estimation/statistical method

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 D5b**
Date of first version:	14.03.2017
Version:	V1
Date of updated version:	-
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

Method 1: T5_26_ Multiple-list models for population size estimation

Deliverable D5b gathers the presentation of the methods, the contextual framework as well as the conditions of applications, the pros and cons, a possible example of use in NSIs, and the related software.

LIST OF ESTIMATION METHODS

I. Data editing and imputation:

1. Deductive editing
2. Selective editing
3. Automatic editing
4. Manual editing
5. Macro-editing
6. Deductive Imputation
7. Model-Based Imputation
8. Donor Imputation
9. Imputation for Longitudinal Data (Little and Su Method)
10. Imputation under Edit Constraints
11. Outliers /extreme values detection
12. Generalised Regression Estimator
13. Estimates with Model-Based Methods
14. EBLUP Area Level for Small Area Estimation (Fay-Herriot) Method
15. Small Area Estimation Methods for Time Series Data

II. Creation of joint statistical micro data:

16. Data Fusion at Micro Level (relevant choice of Statistical Matching Methods)
17. Matching of Object Characteristics (Unweighted & Weighted Matching)
18. Probabilistic Record Linkage
19. Reconciling Conflicting Micro-data: Prorating, Minimum Adjustment Methods, Generalised Ratio Adjustments
20. Data hashing & anonymisation techniques

III. Alignment of statistical data:

21. State space models (estimation of unobserved variable and possible application to the alignment of statistical data)
22. Temporal Disaggregation/benchmarking methods: Denton's Method & Chow-Lin Method

IV. Multisource estimation at aggregated level:

23. RAS
24. Stone's Method
25. Harmonisation based on latent variable models
26. Multiple-list models for population size estimation
27. Statistical methods for achieving univalent estimates for cross-sectional data
    - 27.1. Repeated weighting
    - 27.2. Mass imputation
    - 27.3. Repeated imputation
    - 27.4. Macro-integration

1. Purpose of the method

The availability of the administrative sources in the recent year have been constantly increasing. In order to reduce costs and response burden, the National statistical institute are considering the possibility of using in a different manner the administrative data also for producing statistics exclusively based on them (Wallgren and Wallgren, 2007). The estimation of the unknown size of a target population is very important for official statistics and when is based on several administrative sources, the misalignment between the scope of the administrative data and the statistical ones need to be taken into account because produces methodological challenges. This is the situation, for example, 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). The violation of the approach 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), especially in 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.

2. The related scenarios

2.1. The Multiple-list models for population size estimation is related to indirect estimation for the specific usage "estimation where administrative and statistical data are used on an equal footing".

2.2. The Multiple-list models for population size estimation is related to task IV. a) Multisources estimation at aggregated level for population size estimation.

3. Description of the method

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 standard dual system estimation (Peterson, 1896; Sekar and Deming, 1949; Wolter, 1986) is a well-known model to evaluate the population total. Assuming that N, the total amount of the populatio is fixed but unknown, the problem is to estimate N. Suppose that we have made two attempts to count the entire population and have obtained two lists of identified individuals, namely the List A and the List B. After matching the two lists, a 2×2 table is set up, as in Table 1. Table 1 is commonly referred to as The Dual System Estimation Table.

The entries in the table relate to: the number of people counted in the List A and in the List B, x_{11}; the number of people counted in the List A but not in the List B $x_{12} = N_1 - x_{11}$ and in the List B but not in the

List A, $x_{21} = N_2 - x_{11}$ where $x_{1+} = N_1$ and $x_{+1} = N_2$ are the population size reported in the List A and the List B, respectively; the number of people missed in both lists, x_{22}. Note that x_{22}, the marginal totals x_{+2}, x_{2+} and the population total N are unobservable, and therefore need to be estimated.

Table 1. Contingency table of the counts in the List A and in the List B

		LIST B		
		Present	*Absent*	*Total*
LIST A	*Present*	x_{11}	x_{12}	x_{1+}
	Absent	x_{21}	x_{22}	x_{2+}
	Total	x_{+1}	x_{+2}	N

Under the assumption of independent captures, the number of individuals in the contingency table follows the multinomial distribution.

An unbiased estimator of N, the well-known Petersen estimator, is given by

$$\hat{N} = N_1 \times N_2 / x_{11}.$$

Let p_{ab} be the probability of inclusion in the ab-th cell, with a, b = 1, 2, +.

The probability of being counted twice is the product of the marginal counting probabilities, $p_{11} = p_{1+}p_{+1}$, and the maximum likelihood estimators of the probabilities that a person will be counted in the Census and in the PES, p_{1+} and p_{+1}, respectively, are:

$$\hat{p}_{1+} = x_{11} / N_2 \quad , \quad \hat{p}_{+1} = x_{11} / N_1$$

Given these assumptions, these estimators are strongly consistent with asymptotic normal distribution (see, e.g., Alho, 1990). In fact, in many applications of the dual system methodology, it is not realistic to assume that both lists are based on a counting procedure that aims to count the entire population. Wolter (1986) provides a detailed description of an alternative model where the second list is based on a post-enumeration under-coverage sample survey. Moreover, several extensions and adjustments have been proposed in order to avoid biases due to any failure of these assumptions, i.e. under or over estimation of the real population total amount.

4. Examples

The following example is summarised from the paper Cibella et al. (2008c). It involves data from the 2001 Italian Population Census and its Post Enumeration Survey (PES). The main goal of the Census was to enumerate the resident population at the Census date, 21/10/2001. The PES instead had the objective of estimating the coverage rate of the Census; it was carried out on a sample of enumeration areas (called EA in the following), which are the smallest territorial level considered by the Census. The size of the PES's sample was about 70000 households and 180000 individuals while the variables stored in the files are name, surname, gender, date and place of birth, marital status, etc. Correspondingly, comparable amounts of households and people were selected from the Census database with respect to the same EAs. The PES was based on the replication of the Census process inside the sampled EAs and on the use of a capture-

recapture model (Wolter, 1986) for estimating the hidden amount of the population. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a record linkage between the two lists of people built up by the Census and the PES was performed. In this way the rate of coverage, consisting of the ratio between the people enumerated at the Census day and the hidden amount of the population, was obtained.

5. Input data (characteristics, requirements for applicability)

Different archives, say two lists A and B, with identified individuals, namely the List A and the List B.

6. Output data (characteristics, requirements)

The final output data is the N, estimate of the unknown population size, and the estimated capture probabilities $\hat{p}_{1+}$ and $\hat{p}_{+1}$. As intermediate result, one can consider the set of matched units n_{11} from the two lists A and B.

7. Tools that implement the method

None.

8. Appraisal

8.1. 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.

The violation of the basic assumptions of CRC method could cause serious bias in the estimation of the population size so, several extensions of the CRC method were proposed to overcome this problem.

8.2. How to evaluate the results of the method (measures of model fitting, accuracy measure, etc.). 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.

8.3. 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. Extensions of this methodology that take into account the violation of the basic assumptions, aiming at improving the CRC model (see T3 for details). 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.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).

9. 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

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.