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_12_Generalised Regression Estimator

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 Generalized regression estimator is an estimator applied on a sample of units selected according to a complex survey design with survey weights d, used when auxiliary information on some auxiliary variables is known at domain (for all domains) or unit level (for all units of a population). In particular, the Generalised regression estimator is a model assisted estimator designed to improve the accuracy with respect to other estimators by means of auxiliary information. Furthermore, the GREG estimator guarantees the coherence between sampling estimates and known totals of the auxiliary variables. It is defined as a particular form of the calibration estimator where the objective is to determine those transformed weights w that are the most similar to the initial survey weights d according to a Euclidean distance, among those systems of weights d that are able to estimate exactly the totals of the auxiliary variables on the whole population. The improvement in efficiency is ensured when a linear relationship between the target variable and auxiliary information holds true.

2. The related scenarios

2.1. The typical usage is for direct tabulation, i.e. modify the weights in order to take into account the available auxiliary information. Although auxiliary micro data can also be available, the typical situation in which GREG estimators are applied is when a sample is complemented with totals of some auxiliary variables, which corresponds to the configuration number 7 described in the Deliverable 1*.*

2.2. The typical statistical task that refers to GREG estimators is Multisource estimation at aggregated level, that refers to the sub-processes 5.6 “Calculate weights”, 5.7 “Calculate aggregates”.

2.3. This estimator is defined in the context of the calibration estimators. A specific GREG estimator is the ratio estimator, when there is only one covariate and the target variable variance given the auxiliary one is σ2 times the observed value of the auxiliary variable. If a linear relationship does not hold, it is possible to include generalized linear models (e.g. Lehtonen and Veijanen, 1998)). Nonparametric estimators can also be adapted (e.g. Breidt and Opsomer, 2000 and Montanari and Ranalli, 2005).

3. Description of the method

The GREG estimator (Cassel et al, 1976, Särndal et al, 1992) can be defined consistently in many different forms. At first, it can be seen as the difference between the traditional Horvitz-Thompson estimator of the target variable Y and a combination between the known totals of the auxiliary variables X and the corresponding Horvitz-Thompson estimators of the same totals, where the coefficients to be used in the combination are estimates of the regression coefficients of Y on X (in order to avoid the risks of homoscedasticity, X can be suitable transformed if necessary). Alternatively, if auxiliary variables are known for all units in the population, the predicted values of Y for each unit in the population can be estimated and the GREG estimator can assume the form of the total of the predicted values on all units in the population plus the sum of the differences between the observed and predicted values of Y on the samples units. Finally, the GREG estimator can be seen also as a sum of the sampled Y values, with weights given by the product of the initial weights d with correction factors that do not depend on the target variable Y.

4. Examples

A very detailed example on the application of the GREG estimator, consistent with the present document that aims at describing different data integration methodologies involving registers and archives, is available in Memobust (2014). This example refers to estimation problems relative to the Small and Medium-sized Enterprises (SME) sample survey, carried out annually by sending a postal questionnaire with the purpose of investigating profit-and-loss account of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulation n. 58/97 (Eurostat, 2003) and n. 295/2008.

Auxiliary variables are known from the Italian Statistical Business Register: the selected auxiliary variables, available also on the sample, are the Average number of employees in the year t-1 and the Number of enterprises.

5. Input data (characteristics, requirements for applicability)

The input data consists of:

1. a sample drawn according to a complex survey design, where the observed values corresponds to a target variable and a set of auxiliary variables.
2. known totals on the population of interest or on domains of interest for the auxiliary variables in order to constrain the GREG estimator to estimate exactly these totals; the method applies also if auxiliary variables are known at the unit level for all the units of the population of interest.

Missing data should be avoided, in case the sample is affected by missing data it would be preferable to impute it at first. Furthermore, missing data are not allowed on totals.

6. Output data (characteristics, requirements)

A sample with the same size as the one in the input, but transformed weights that are able to reproduce the known auxiliary variables totals.

7. Tools that implement the method

1. CALMAR (Deville, Särndal and Sautory 1993)
2. CLAN (Statistics Sweden, Anderson and Nordberg, 1994)
3. BASCULA (The Netherlands, Bethlehem, 1996)
4. GES (StatCan, Estevao, Hidiroglou and Särndal, 1995)
5. GENESEES (ISTAT, Falorsi and Falorsi, 1997)
6. Survey, an R package downloadable from the CRAN (Lumley, 2004, 2010, 2017)
7. Sampling, an R package downloadable from the CRAN (Tillé, Matei, 2016)
8. REgenesees (ISTAT, Zardetto, 2015), an R package downloadable from the JoinUP: https://joinup.ec.europa.eu/software/regenesees/release/release150#download-links.

8. Appraisal

1.1. The GREG estimator is nearly design unbiased, anyway it may suffer bias for small sample sizes.

1.2. If a linear model between the target variable and the auxiliary variables holds, the GREG estimator's variance is less than the one of the Horvitz-Thompson estimator. Anyway the opposite relation can occur if the linear model does not occur.

1.3. The GREG estimator can give negative weights.

9. References

Andersson, C. and Nordberg, L. (1994): A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys - Theory and a Software Implementation. Journal of Official Statistics, 10, 395-405

Bethlehem, J. (1996) Bascula 2.0 reference manual, Voorburg: Statistics Netherlands

Breidt, F. J. and Opsomer, J. D. (2000), Local Polynomial Regression Estimators in Survey Sampling. The Annals of Statistics 28, 1026–1053.

Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1976), Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. Biometrika 63, 615–620.

Deliverable 1. Identification of the main types of usages of administrative sources.

Deville, J. C., Särndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling, Journal of the American Statistical Association, 88, 1013–1020

Estevao, V., Hidiroglou, M.A. and Särndal, C. E. (1995) Methodological Principles for a Generalized Estimation System at Statistics Canada, Journal of Official Statistics, 11, 181–204

Falorsi, P.D. and Falorsi, S. (1997) The Italian Generalised Package for Weighting Persons and Families: Some Experimental Results with Different Non-Response Models. Statistics in Transition, 3, 357–381.

Lehtonen, R. and Veijanen, A. (1998), Logistic Generalized Regression Estimators. Survey Methodology 24, 51–55.

Lumley, T. (2004) Analysis of Complex Survey Samples. Journal of Statistical Software, 9, 1–19.

Lumley, T. (2010) Complex Surveys: A Guide to Analysis Using R. New York: John Wiley & Sons.

Lumley, T. (2017) Package “survey”, http://r-survey.r-forge.r-project.org/survey/

Memobust (2014). Method: Generalized regression estimator, in Memobust Handbook on Methodology of Modern Business Statistics. https://ec.europa.eu/eurostat/cros/content/memobust_en

Montanari, G. E. and Ranalli, M. G. (2005), Nonparametric Model Calibration Estimation in Survey Sampling. Journal of the American Statistical Association 100, 1429–1442.

Särndal, C.-E. (2007) The Calibration Approach in Survey Theory and Practice, Survey Methodology, 33, 99–119.

Särndal, C.-E., Swensson, B., and Wretman J. (1989) The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527–537

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

Tillé, Y., Matei, A. (2016). Package “sampling”, https://CRAN.R-project.org/package=sampling

Zardetto, D. (2015) ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys, Journal of Official Statistics, 31, 177–203

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_13_14_ Estimates with Model-Based Methods - EBLUP Area Level for Small Area 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

In the model-based approach, the estimation is obtained on the assumption of a model and the resulting estimators are the best in terms of model Mean Square Error, namely in a model-based approach the aim of the estimation step is the Best Linear Unbiased Predictor (Royall, 1970, Vaillant et al., 2000). In official statistics, the model-based estimators are particularly exploited in the so-called Small Area Estimation, that is when the sample size is not large enough, at least for some target domains, to obtain direct sample-based estimates with sufficient variability to be published (see EUROSTAT, 2013, for some examples of threshold on reliability of the estimators). National Statistical Office surveys are usually planned at a higher level, hence, whenever more detailed information is required, the sample size may be not large enough to guarantee release of direct estimates and in some cases, smaller domains may happen to be without sample units. Small area methods increase the reliability of estimation by "borrowing strength" from a set of areas in a larger domain for which the direct estimator is reliable. This means that information from other areas is used and/or additional information from different sources is exploited. The area level EBLUP described herewith is a linear combination of the area (domain) direct estimator and a predicted component based on a linear mixed model. The model relates the parameter of interest to known auxiliary variables for each of the domains that constitute the partition of the whole population. An effect to account for (within) domain homogeneity is included in the model. A large development in terms of methods and software for Small Area Estimation, as well as real applications in official statistics has been produced in recent years (see Rao, 2003, EURAREA project, ESSnet SAE, BLUE-ETS project).

2. The related scenarios

2.1. 1.1 Usages and Komuso Data Configurations (deliverable 1): Usage: Indirect estimation, the Area Level Small Area Estimation method applies to all the Komuso Data Configurations where data are available at least at aggregated (area) level but the sample size at these domains are not enough large to obtain sample-based estimates with sufficient variability to be published, e.g. configurations 1, 3, 7.

2.2. Statistical tasks (deliverable 2): Estimation.

3. Description of the method

In the module "Weighteing and Estimation" – Theme "EBLUP area Level for SAE" of the ESSnet Memobust Handbook a short but very effective description of the method is given. According to this reference, the EBLUP area level, also referred to as Fay-Herriot model (Fay and Herriot 1979), is based on a linear mixed model which formulates the relationship between the parameter of interest and auxiliary area level information. Let θ_d be the parameter to be estimated for each domain d. A linear relationship between θ_d and a set of covariates whose values are known for each domain of interest is assumed. In details

$$\theta_d = X^T_d \beta + u_d, \qquad (1)$$

where X_d is the vector of covariates for domain d and the u_d (d=1,...,D) are domain effects assumed to be distributed with mean zero and variance σ^2_u. The random effects account for the extra variability not explained by the auxiliary variables in the model.

Beside the model on the parameters, let us specify the sampling model. A design unbiased direct estimators $\hat{\theta}_d$ is supposed to be available (but not necessarily for all the domains), that is

$$\hat{\theta}_d = \theta_d + e_d \qquad (2)$$

where the e_d is are the sampling errors associated with the direct estimators, for which $E(e_d\,|\theta_d)$=0, i.e. , the direct estimator is assumed to be unbiased, and $V(e_d\,|\theta_d) = \varphi_d$, where the variances φ_d are

supposed to be known.
Combining equations (1) and (2) a linear mixed model is obtained. The model is formulated as follows:

$$\hat{\theta}_d = X_d^T \beta + u_d + e_d \qquad (3)$$

Normality for e and u is commonly assumed for estimation of the Mean Square Error (MSE), but this assumption is not necessary for estimating the parameter. On the basis of model (3) the empirical best linear unbiased estimator (EBLUP) is

$$\hat{\theta}_d^{EBLUP\_AREA} = \gamma_d \hat{\theta}_d + (1-\gamma_d) X_d^T \hat{\beta} \qquad (4)$$

where $\gamma_d = \hat{\sigma}_u^2 \;/(\hat{\sigma}_u^2 + \varphi_d)$ is the weight of the direct estimator and $\hat{\beta}$ is the weighted least square (WLS) estimator of the regression coefficient vector β , where the weights for estimating β are provided by a diagonal matrix whose generic element is given by $\hat{\sigma}_u^2 + \varphi_d$.

For more details on model specification, methods for estimation of $\hat{\sigma}_u^2$ see Rao (2003, pp. 115-120). Details on the estimation of the MSE are given in Rao (2003, pp. 103 and 128-130).

4. Examples

Small area methods are traditionally used to income and poverty estimation (Bell, 2009; Pratesi,2016), for instance in Italy, using EU-SILC survey data and Population census data poverty estimation at regional and sub-regional level were provided (Giusti et al 2009). Other applications are related to the estimation of number of employees at sub-regional level by gender (Falorsi e solari et).

5. Input data (characteristics, requirements for applicability)

The input data set used by the method is macrodata referred to domain level with the target variables and the covariates.

6. Output data (characteristics, requirements)

The method provides estimates of the target variables at area level, as well as estimates of the related MSEs which result reduced with respect to the direct sample-based area level estimates.

7. Tools that implement the method

There are many codes providing Area Level Small area Estimation in SAS and SAE. A review is available in ESSnet SAE (2014)Work Package 4 “Software Tools” downloadable from
https://ec.europa.eu/eurostat/cros/content/sae-finished_en

8. Appraisal

8.1. The main advantage of the method results in its applicability for estimation when few or even no sample data are available for one or more domains of interest. Covariates are needed only at domain level. The method is useful to improve direct estimators if a set of covariates with a strong relationship with the variable of interest is available. The variances of the small area direct estimates has to be known. Usually a smoothed model for variance estimation is applied and variances are assumed to be known. This affects the MSE (see Bell, 1999).

8.2. Possible disadvantages of the method occur if the model is not correctly specified the estimator can be affected from bias. When adding up small domains estimates to a larger domain, it is not ensured that direct estimates at larger level are obtained. A simple way to ensure consistency is to ratio adjust the EBLUP area level estimator. Benchmarking can be also set as a constraint to obtain small area estimates. Symmetry of the distribution is required while in business survey skewness may be present. If transformation of variables does not suffice to reduce skewness advanced methods may be employed (Chandra and Chambers, 2007). Assumptions of normality with known variance might be untenable at small sample sizes. Model variance σ ^_u^2 can be estimated to be zero. This is an undesirable result. Hierarchical Bayesian methods are good alternatives and they always result in strictly positive variances, see, e.g., Bell (1999) and Buelens et al. (2012).

8.3. MSE and bias diagnostic of the resulting estimates (see the ESSnet/sae site https://ec.europa.eu/eurostat/cros/content/sae-finished_en) should be checked to evaluate the properties of the output data.

8.4. Alternative methods in Small Area estimation context are Synthetic Estimators, Composite Estimators, EBLUP Unit Leve (see https://ec.europa.eu/eurostat/cros/content/sae-finished_en for details about this methods), Benchmarking methods, Quantile regression Methods and its extension, Hierarchical Bayesian methods.

9. References

Bell, W. R. (1999), Accounting for uncertainty about variances in small area estimation. Bulletin of the International Statistical Institute. http://www.census.gov/did/www/saipe/publications/files/Bell99.pdf

Bell, W. R. (2009), The U.S. Census Bureau's small area income and poverty estimates program: a statistical review. http://cio.umh.es/data2/T1A%20William.R.Bell@census.gov.pdf

Buelens, B., van den Brakel, J., Boonstra, H. J., Smeets, M., and Blaess, V. (2012), Case study, report Statistics Netherlands. Essnet SAE WP5 report, 62–81.

Chandra, H. and Chambers, R. (2007), Small area estimation for skewed data. Small Area Estimation Conference, Pisa, Italy.

ESSnet SAE (2012), WP4 Final Report Deliverables of the project.
http://www.cros-portal.eu/sites/default/files//WP4report_0.pdf

EURAREA Consortium (2004), PROJECT REFERENCE VOLUME, Vol. 1.
http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatialanalysis-andmodelling /eurarea/index.html

Eurostat (2013), Handbook on precision requirements and variance estimation for ESS household surveys. Methodologies and Working papers.

Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association 74, 269–277.

Giusti, C., Pratesi, M. and Salvati, S. (2009) Small area methods in the estimation of poverty indicators: the case of Tuscany, Politica Economica, Vol.3, 369-380.

Memobust (2014), Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot) In: *Memobust Handbook on Methodology of Modern Business Statistics,* https://ec.europa.eu/eurostat/cros/content/memobust_en.

Rao, J. N. K. (2003), Small area estimation. John Wiley & Sons, Hoboken, New Jersey.

Pratesi, M. (ed) (2016), Analysis of Poverty Data by Small Area Estimation, New York, Wiley http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118815017.html