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_6_Deductive editing

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

Detecting and treating errors in a deductive manner. Deductive editing is the phase where methods for detecting and treating errors with a structural cause that occurs frequently in responding units (systematic errors) are used.

2. The related scenarios

2.1. The method applies to a single data set composed of microdata

The data set can be the result of a combination of several data sets. It mainly refers to data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing.

2.2. Statistical tasks: Data Editing and Imputation, measurement alignment.

3. Description of the method

Deductive editing is the phase where methods for detecting and treating errors with a structural cause that occurs frequently in responding units (systematic errors) are used. A systematic error is commonly defined as an error with a structural cause that occurs frequently between responding units (e.g., reporting financial amounts in units instead of the requested thousands of units).

A separate deductive method is needed for each type of systematic error. The exact form of the deductive method varies per type of error; there is no standard formula. The difficulty with using this method lies mainly in determining which systematic errors are present in the data. Sometimes, such an investigation can bring systematic errors to light that have arisen due to a shortcoming in the questionnaire design or a bug in the processing procedure. In that case, the questionnaire and/or the procedure should be adapted.

In administrative data, especially when more sources are used, deductive editing has an important role in the production process. Variables collected in the administrative sources may have similar definitions but they may have structural gaps given to the convenience of declaring some information in an item rather than in another one, for instance, declaring something either in a cost or in an investment item. The first step in an E&I process should be to look for systematic errors in the observed values, also in the case the definition of variables is almost the same with respect to the corresponding statistical target variable. Hence, deductive editing is substantially the same as the one carried out in a classical data editing process, in fact the detection of systematic errors implies the involvement of subject matter experts, and the error treatment, that is usually completely automated, is not affected by the large dimension of administrative databases.

4. Examples

4.1. Correction rules for subject-matter related errors

Subject-matter related errors can often be detected and treated by means of deterministic checking rules. The general form of a correction rule is as follows:

if (condition) then (correction).

Here, condition indicates a combination of values in a record that is not allowed. Subsequently correction describes the adjustment that is made to the record to resolve the inconsistency.

An example of a correction rule is:

```
if ( Number of Temporary Employees > 0 and Costs of Temporary Employees = 0 )
  then Number of Temporary Employees := 0.
```

This rule detects an inconsistency that occurs when a business reports to have employed temporary staff without reporting associated costs. In this example, the inconsistency is treated deductively by making the number of temporary employees equal to zero.

We notice that the correction rule operates under the assumption that the variable Costs of Temporary Employees is reported more accurately than the variable Number of Temporary Employees. Making such assumptions in a valid way generally requires subject-matter knowledge and knowledge of the data collection process.

4.2. The unit of measurement error

Business surveys usually contain instructions to the reporter that all financial amounts must be rounded to thousands of euros (dollars, pounds, etc.), that all quantities must be rounded to thousands of units, etcetera. Some respondents ignore these instructions and, consequently, report values that are a factor 1000 larger than they actually mean. Traditional methods for detecting unit of measurement errors usually work by comparing one or more reported amounts with reference values. For instance: in the Dutch Short Term Statistics, thousand-errors are detected as follows (Hoogland et al., 2011). The total turnover indicated by the respondent for period t, say x_t , is compared to the turnover from the most recent period for which a statement from the respondent is available, up to a maximum of six previous periods. The stated turnover for this earlier period must also not be equal to zero. A thousand-error is detected in x_t if the following applies:

$|x_t| > 300 \times |x_{t-i}|$, for some i=1,…,6.

If no data from the respondent from an earlier period are available, then the median of the turnover from the previous period in the stratum of the respondent is used instead. The stratification is based on economic activity and number of employees. A thousand-error is detected in t x if the following applies:

$|x_t| > 100$ x stratum median(x_{t-1}) -1.

If a thousand-error is detected by either formula, then it is resolved by dividing the total turnover and all the sub-items by 1000.

5. Input data (characteristics, requirements for applicability)

A data set containing unedited microdata. Missing values are allowed.

6. Output data (characteristics, requirements)

A data set containing edited microdata with respect to the systematic errors treated. It is an updated version of the first input data set.

7. Tools that implement the method

The R package deducorrect, which can be downloaded for free at http://cran.r-project.org,
contains an implementation of deductive editing methods for several generic errors:

- sign errors and interchanged values;
- simple typing errors;
- rounding errors (very small inconsistencies with respect to equality constraints).

8. Appraisal

Deductive editing is very efficient and effective, but it should only be used to treat errors for which the error mechanism is known with sufficient reliability. Deductive adjustments based on invalid assumptions can produce biased estimators.

It may be difficult to maintain a large collection of deterministic correction rules over a long period of time. In particular, it becomes difficult to grasp the consequences of adding or removing a correction rule, or changing the order in which the rules are applied, when faced with a large collection of rules.

9. References

Hoogland, J., van der Loo, M., Pannekoek, J., and Scholtus, S. (2011), Data Editing: Detection and Correction of Errors. Methods Series Theme, Statistics Netherlands, The Hague.

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

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_2_Selective editing

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

Selective editing is a general approach to identify the records in a data set that contain potentially influential errors.

2. The related scenarios

2.1. The method applies to a single data set composed of microdata

The data set can be the result of a combination of several data sets. It mainly refers to data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing.

2.2. Statistical tasks: Data editing and imputation, measurement alignment.

2.3. Related methods: Macroediting, manual editing.

3. Description of the method

Observations are ranked according to the values of a score function expressing the impact of their potential errors on the target estimates. Units with a score above a given threshold are selected for editing. The score function is an instrument to prioritise observations according to the expected benefit of their correction on the target estimates. An example of score function is reported in the following. First a local score is computed for the unit *i* with respect to the variable Y_j as

$$S_{ij} = \frac{p_i w_i \,|\, y_{ij} - \tilde{y}_{ij} \,|}{\hat{T}_{Y_j}}$$

where p_i is the degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the *i*th unit, $\tilde{y}_{ij}$ is the corresponding prediction (e.g., an historical value) w_i is the sampling weight, and $\hat{T}_{Y\,j}$ is an estimate of the target parameter. Once the local scores for the variables of interest are computed, a global score to prioritise observations can be obtained by combining the local scores, as an example the sum of the local scores $GS_i^{(1)} = \sum_j S_{ij}$ can be taken. Once the observations have been ordered according to their global score, it is important to build a rule in order to determine the number of units to be reviewed. A first rule can be suggested by budget constraints. In this case, the first *n** observations are chosen such that the budget constraints are satisfied. A more complex approach is to select the subset of units such that the impact on the target estimates of the errors remaining in the unedited observations is negligible. Since the true values are unknown, this bias cannot be evaluated and an approximation is used. This approximation can be expressed in terms of the weighted differences between the raw values y_{ij} and the anticipated values $\tilde{y}_{ij}$ for the variable Y$_j$ in the units *i* not selected for interactive treatment. Let T_{Yj} be the target quantity related to the variable Y_j (for instance the total), the estimated bias is given by

$$EB_j(t) = \frac{|\sum_{i \notin E_t} w_i (y_{ij} - \tilde{y}_{ij})|}{\hat{T}_{Y_j}},$$

where w_i is the sampling weight of the i-th unit, $\hat{T}_{Y_j}$ is an estimate of the target quantity T_{Y_j}, and is the set of units to be selected. This set is composed of all the units having a global score *GS*> *t,* where *t* is a threshold values such that $EB_j(t)$ is below a predefined value. Details can be found in Memobust (2014) and de Waal et al., (2011).

A selective editing methods based on contamination models can be found in Di Zio et al., (2013). In this approach, the true (log)-data are considered realisations of a multivariate Gaussians distribution, while errors are supposed to act on a subset of data by inflating the variance of the true data distribution.

A probabilistic approach to selective editing.

Ilves and Laitila (2009) and Ilves (2010) propose a two-step procedure for selective editing. Their proposal is motivated by the fact that the non-selected observations may still be affected by errors resulting in a biased target parameter estimator. To obtain an unbiased estimator a sub-sample is drawn from the unedited observations (below threshold for global scores), follow-up activities with recontacts are carried through and the bias due to remaining errors is estimated. The estimated bias is used to make the target parameter estimator unbiased. Formulas for the variance and a variance estimator are derived by using a two-phase sampling approach.

4. Examples

Selective editing based on contamination model and integrated administrative data is applied to microdata on *gross investment*. Data on investments come from Istat Annual Survey on Economic and financial accounts of large enterprises. Two variables from administrative data are used as covariate in the model: the information on *expenditure for amortizable goods* reported in Value Added Tax declarations and a derived variable based on the assets at the end of the year minus assets at the beginning plus depreciation and revaluation that can be calculated from financial statements.

A third covariate has a peculiar nature. Istat may access the explanatory notes of corporations and limited companies in the form of non-standardized text files (one for each company) and in the form of an experimental dataset reporting the value of investment that is obtained using a software for automatic optical recognition from the non-standardized text files. Note that the variable on total investment reported in the dataset is exactly the target variable of the selective editing procedure. However, it cannot be used to produce SBS data or to automatically correct the data because of the errors due to the automatic optical recognition. For this reason, we use the value of total investment from the experimental database on the explanatory note as a third covariate in the selective editing procedure.

In Table 1, the number of selected units and impact on estimates for Manifacture industries are reported.

Table 1. Number of selected and/or edited observation, and differences between estimates computed on raw and edited data

ID	N.of Obs.	Selected	No-Edited	Edited	SBS Original Value*	Post-editing Value*	(B-A)/A %
					(A)	(B)	
M1	220	2	0	2	1.652.667	1.406.106	-15%
M2	227	1	0	1	579.315	588.957	2%
M3	36	0	0	0	73.309	73.309	0%
M4	64	2	1	1	500.077	283.803	-43%
M5	26	0	0	0	100.141	100.141	0%
M6	17	1	1	0	643.181	643.181	0%
M7	125	3	0	3	828.393	744.427	-10%
M8	97	1	0	1	826.628	831.095	1%

M9	154	0	0	0	575.771	575.771	0%
M10	121	0	0	0	642.383	642.383	0%
M11	122	1	0	1	889.292	876.648	-1%
M12	246	4	0	4	756.273	672.888	-11%
M13	90	3	1	2	374.818	357.631	-5%
M14	125	2	2	0	512.878	512.878	0%
M15	445	1	0	1	1.323.962	1.256.509	-5%
M16	128	3	0	3	1.041.013	1.332.121	28%
M17	53	2	2	0	804.111	804.111	0%
M18	117	0	0	0	301.780	301.780	0%

The selective editing procedure proved to be quite efficient: strong improvements in the results have been obtained selecting few units and the hit-rate was quite high. More details can be found in Di Zio et al. (2015).

5. Input data (characteristics, requirements for applicability)

Microdata with numerical variables.

6. Output data (characteristics, requirements).

A data set with observations flagged as affected by influential errors.

7. Tools that implement the method

R package SeleMix (Istat), Selekt (SSB).

8. Appraisal

A constraint for selective editing on administrative data derives from the difficulty of re-contacting units for this kind of data. This limitation is alleviated when multi-source data are used, in this case the availability of different values for the same observation is an important aspect that can help the statistician in understanding where the error is located and to recover a likely value.

9. References

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

Di Zio M., Guarnera U. (2013) Contamination Model for Selective Editing. Journal of Official Statistics, Vol. 26, n. 4, pp . 539-556

Di Zio M., Guarnera U., Iommi M., Regano A., (2015). Selective editing of business investments by using administrative data as auxiliary information. UNECE Work Session on Statistical Data Editing, 14-16 October, Budapest, Hungary.

Ilves, M. and Laitila, T. (2009), Probability-Sampling Approach to Editing. Austrian Journal of Statistics 38, 171–182.

Ilves M. (2010), Probabilistic Approach to Editing. Workshop on Survey Sampling Theory and Methodology Vilnius, Lithuania, August 23-27, 2010.

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

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_3_Automatic editing

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 goal of automatic editing is to accurately detect and treat errors and missing values in a data file in a fully automated manner, i.e., without human intervention.

2. The related scenarios

2.1. The method applies to a single data set composed of microdata

The data set can be the result of a combination of several data sets. It mainly refers to data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used.

2.2. Statistical tasks: Data editing and imputation.

3. Description of the method

Automatic editing is generally based on the Fellegi-Holt paradigm, which means that the smallest number of fields should be changed to a unit to be imputed consistently. The algorithms exploit edit rules (checking rules) that represent rules/constraints characterising the relationships among variables. An example of edit rule is that Turnover = Profit + Costs, or that the combination Gender = 'male' and Pregnant = 'yes' is not allowed. According to the Fellegi-Holt paradigm, one should minimise the number of observed values that have to be adjusted in order to satisfy all edit rules. This paradigm is often used in a generalised form, for which each variable is given a reliability weight w_i >0. A high value of w_i indicates that the variable x_i is expected to contain few errors. The generalised Fellegi-Holt paradigm now states that one should search for a subset of the variables E with the following two properties:

- The variables x_i ($i \in E$) can be imputed with values that, together with the observed values of the other variables in the record, satisfy all edit rules.
- Among all subsets that satisfy the first property, E has the smallest value of $\sum_{i \in E} w\_i$

In order to determine whether a set of variables can be imputed to satisfy all edits simultaneously, it is necessary to derive the so-called *implied edits* from the original set of edits. An implied edit is an edit rule that can be derived from the original edit rules by logical reasoning. For numerical data, the number of implied edits, that can be derived from even a single original edit, is actually infinite; e.g., if $x_1 > x_2$ is an edit rule, then so is $\square x_1 > \square x_2$ for any $\square$ 0 . Fortunately, for the purpose of solving the error localisation problem, it is not necessary to derive all possible implied edits, but only the so-called *essentially new implied edits* (see Fellegi-Holt, 1976). By adding the essentially new implied edits to the original set of edit rules, one obtains the *complete set of edits*. For a complete set of edit rules, it does hold that any subset of the variables which 'covers' all failed edit rules is a feasible solution to the error localisation problem. Hence, one can solve the error localisation problem for any given record in the following manner:

- Select all edits from the complete set of edits that are failed by the original record.

Find the smallest (weighted) subset of the variables with the property that each selected (original or implied) edit involves at least one of them.

4. Examples

4.1. Generic example

Let us consider four numerical variables with the following edit rules:

$x_1 - x_2 + x_3 + x_4 \geq 0$ (1), $-x_1 + 2x_2 - 3x_3 \geq 0$ (2)

It is possible to derive the essentially new implied edits from (1) and (2):

$$x_2 - 2x_3 + x_4 \geq 0 \quad (3)$$

$$x_1 - x_3 + 2x_4 \geq 0 \quad (4)$$

$$2x_1 - x_2 + 3x_4 \geq 0 \quad (5)$$

These five edit rules together constitute a complete set of edits. This means that we can now solve the error localisation problem for any record by solving an appropriate set-covering problem. Consider the record (x_1, x_2, x_3, x_4) = (3,4,6,1).

By checking the edit rules (1)–(5), it is seen that this record fails edits (2), (3), and (4). Thus, in order to solve the error localisation problem, we have to find the minimal subset of variables that 'covers' these three edit rules. By inspection, we see that the variable x_3 is involved in edit rules (2), (3), and (4). Thus, in this example, x_3 can be imputed to satisfy all the edit rules.

Since {x_3} is the only single-variable set with this property, changing the value of x_3 is in fact the optimal solution to the error localisation problem for this record. Note that the single-variable sets {x_1} and {x_2} cover the original failed edit (2), but not the implied failed edits (3) and (4). A consistent record can be obtained by imputing, for instance, the value x_3 = 1.

4.2. Practical example (same applies for incomplete administrative data or survey)

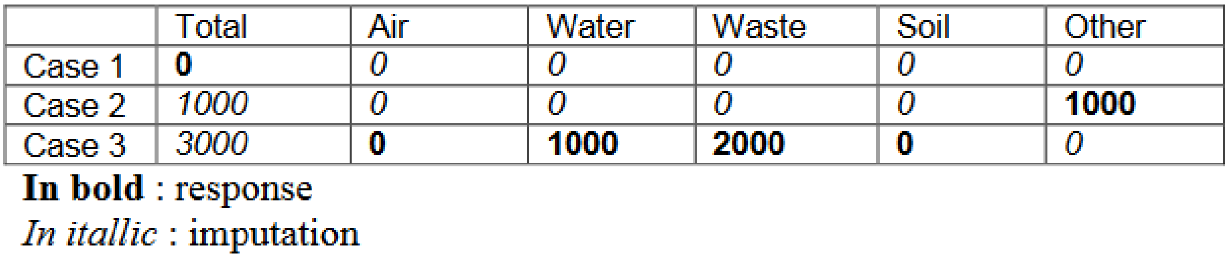
In a voluntary survey on current environmental expenditures by companies (by media: air, water, waste, soil, other), Statistics Belgium performed a deterministic imputation on the level of the databank following logical principles described below (Kestemont, 2004):

1-a company declaring zero for the total should have zero for the details;

2-a value in "other" is supposed to cover all remaining, not detailed, expenditures: a zero is attributed to all missing domains and the total is calculated accordingly;

3-a company that answered for all domains, but not for "others" is supposed to have classified all of its expenditures: other is then estimated to be zero, and the total is calculated accordingly.

Deterministic imputation is illustrated below:

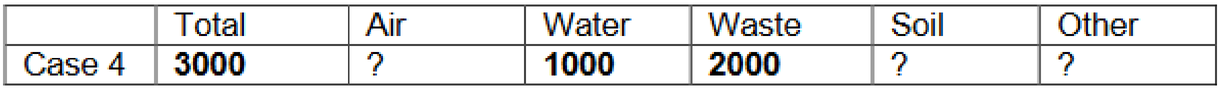


	Total	Air	Water	Waste	Soil	Other
Case 1	**0**	*0*	*0*	*0*	*0*	*0*
Case 2	*1000*	*0*	*0*	*0*	*0*	**1000**
Case 3	*3000*	**0**	**1000**	**2000**	**0**	*0*

In bold : response

In itallic : imputation

A remaining case is more problematic:



	Total	Air	Water	Waste	Soil	Other
Case 4	**3000**	?	**1000**	**2000**	?	?

In this frequent situation, a company answered something for several domains, and gave as total, the total of its detailed answers. This could be interpreted in 2 ways:

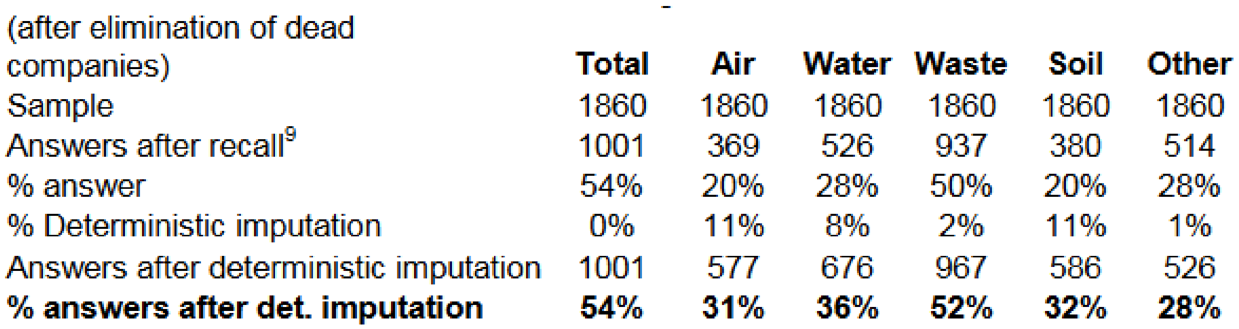
1-either the answer is fully correct and we should add zero to the remaining fields;

2-either the company declared the total of what it could identify, and we cannot affirm that there is no expenditure for other or undifferentiated domains.

In this latest situation, Statistics Belgium did NOT impute the missing values. This implies that the total answered could be considered “doubtful” (possibly underestimated) and is subject to post-editing (once we would have estimated detailed missing values with other methods).

As a summary, we can say that this phase of deterministic imputation consisted mainly to add missing zero’s. The answer rate before and after deterministic imputation is given in table 1.

Table 1: Answer rate after deterministic imputation



(after elimination of dead companies)	**Total**	**Air**	**Water**	**Waste**	**Soil**	**Other**
Sample	1860	1860	1860	1860	1860	1860
Answers after recall[9]	1001	369	526	937	380	514
% answer	54%	20%	28%	50%	20%	28%
% Deterministic imputation	0%	11%	8%	2%	11%	1%
Answers after deterministic imputation	1001	577	676	967	586	526
% answers after det. imputation	**54%**	**31%**	**36%**	**52%**	**32%**	**28%**

5. Input data (characteristics, requirements for applicability)

Microdata free of systematic errors.

6. Output data (characteristics, requirements).

Microdata not failing the edit rules and potentially free of errors.

7. Tools that implement the method

Banff (Statistics Canada), CherryPi (CBS), SCIA (Istat), SLICE (CBS), R package *editrules*.

8. Appraisal

In principle, if the focus is just on one data source, we are in the same situation as the one we would have in an E&I process of statistical survey data. If different data sources are integrated some additional problems may arise. A first issue to take into account is whether the data sources should be treated simultaneously as a unique data set after the integration process. This could be an interesting option, because the amount of information would increase, and an improvement in the E&I procedure is expected. In this case, edits simultaneously involving variables of the different data sources should be considered.

A special but not infrequent case is when the same variable is observed in different data sources. For the sake of simplicity, let us suppose that there are only two data sets with the same variable. According to the Fellegi-Holt approach, we are assuming that with a high probability at least one of the two variables in turn is not affected by error. In the case that this assumption is not reliable, a different approach should be followed, for instance, a prediction conditionally on the observed values of the two variables can be obtained.

The method should be used for error localisation in microdata containing only random errors.

Any systematic errors that may occur in the original microdata have to be resolved beforehand, using deductive editing methods.

In general, it is not possible to construct a set of edit rules that always leads to the correct solution. Thus, the edited data may still contain some errors, although the edited records are consistent with the edit rules. For this reason, automatic editing should not be applied to crucial records, e.g., records belonging to very large businesses. In addition, the quality of automatic editing is lower for records that contain many errors. Both disadvantages can be circumvented by always using automatic editing in combination with a form of selective editing.

There are other methods that find the same solution of Fellegi-Holt, but they use different search algorithms: Algorithms based on vertex generation; Algorithms based on branch-and-bound; Algorithms based on cutting planes; Algorithms based on (mixed) integer programming, see de Waal et al. (2011).

9. References

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

Kestemont, B. (2004), Environmental expenditures by the Belgian industries in 2002, Statistics Belgium Working paper n°9, Direction générale Statistique et information économique, Brussels, http://statbel.fgov.be/fr/binaries/p009n009%5B1%5D_tcm326-34514.pdf

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

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_4_Manual editing

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

Records of microdata are checked for errors and, if necessary, adjusted by a human editor.

2. The related scenarios

2.1. The method applies to a single data set composed of microdata

The data set may result from the combination of several data sets. It mainly refers to the data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing.

2.2. Statistical tasks: Data editing and imputation.

2.3. Selective editing and macroediting are used to select units to be re-contacted.

3. Description of the method

In manual editing, records of microdata are checked for errors and, if necessary, adjusted by a human editor, using expert judgement or information gathered by a re-contact. Nowadays, the editor is usually supported by a computer program in identifying data items that require closer inspection. Moreover, the computer program enables the editor to change data items interactively, meaning that the automatic checks that identify inconsistent or suspicious values are immediately rerun whenever a value is changed. This modern form of manual editing is often referred to as 'interactive editing'.

Ideally, a person (editor) who performs manual editing should be an expert who has extensive knowledge of the survey subject, the survey population, and the kind of errors that are likely to occur in the survey data. If necessary, he or she may re-contact a respondent to check whether a suspicious value is correct, or to obtain a new value for a data item that was originally missing or incorrect. The editor may compare a survey unit's data to reference data, such as data on the same unit from a previous survey or from an external register, or data on similar units. Finally, he or she may have access to other sources of information, for instance through internet searches. More details are in Memobust (2014).

4. Examples

The statistics of road accidents are based on site records by the police. In case of death, the prosecution conduces an investigation. Statistics Belgium receives a copy of the "deaths 30 days" as reported by the prosecution, with the PV number and the characteristics (age, sex, date of death) of the victims. In both sources, several information might be missing or contradictory. It is mainly on the side of the police that information might be misleading, for example unknown of erroneous sex or age, or victim reported as "indemn" or "slightly wounded" when he/she finish to die from this accident. Sometimes, the investigation can conclude to a suicide, which make it to be deleted from the accident victims. Many other errors can be made by the policeman, including many "other" or "unknown" item responses when he typed his survey in the office after hours. An experimented staff compares all the available information, accident by accident, and possible additional data from the media. He deduces some corrections to be made on the basic administrative database. After 20 years editing, he is still very concerned by the high quality of each line of the statistics, because as he says, for any accident, he "still feels compassion even after so many years for every victim; you cannot avoid this compassion when you see that a father and a baby were injured in an accident". For example, when an item says "it was raining" and another says "the road was dry" and the location is on an open road (not a tunnel), other external or internal information might correct the wrong item. Other examples between many possible cases: hour of the accident 02:00, day, correction 14:00; brightness missing, weather sun, other information showing that there was no visibility problem, correction brightness clear, etc.

Most of the cases are unique and cannot be corrected automatically, only a human deduction considering various information and an overview of the circumstances of the accident can lead to editing. The expertise makes that the staff responsible for the editing “finds” inconsistencies very efficiently (the embedded home-made program also helps him to sort and identify possible errors). One of the most important outcome is a minimization of the “other”, “unknown” or “missing” items.

5. Input data (characteristics, requirements for applicability)

Microdata with errors and missings on the checked records.

6. Output data (characteristics, requirements)

Microdata free of errors and without (or with less) missing on the checked records.

7. Tools that implement the method

8. Appraisal

If organised properly, manual/interactive editing is expected to yield high quality data. However, it is also time-consuming and labour-intensive. Therefore, it should only be applied to that part of the data which cannot be edited safely by any other means, i.e., some form of selective editing should be applied.

There are several potential problems associated with interactive editing. The most important of these are the risks of overediting and creative editing. Overediting occurs when “the share of resources and time dedicated to editing is not justified by the resulting improvements in data quality.” In creative editing: editors inventing their own, often highly subjective, editing procedures. Creative editing often involves complex adjustments of reported data items, done for the sole purpose of making the data consistent with a set of edit rules.

A problem when using integrated administrative data is that it is frequently not possible to re-contact the observed units, so one of the main advantages motivating interactive editing declines. However, interactive editing can be considered effective in order to understand error sources and possibly resolve errors in the short term, while in the long term it can contribute to the increase of the subject-matter expertise for the staff working on administrative data, increasing their knowledge of the characteristics and the contents of administrative data and gaining understanding of how the data can be used in a more suitable way (Wallgren and Wallgren, 2007).

9. References

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

Wallgren, A. and Wallgren, B. (2007), Register-based statistics – Administrative data for statistical purposes. John Wiley and Sons, ChichesterHarvey (1989), Forecasting, structural time series models and the Kalman filter, Cambridge University Press.

Method: Macroediting

1. Purpose of the method

Macro-editing (also known as output editing or selection at the macro level) is a general approach to identify the records in a data set that contain potentially influential errors. It can be used when all the data, or at least a substantial part thereof, have been collected.

2. The related scenarios

2.1 The method applies to data sets composed of microdata. It mainly refers to data configurations 1 to 5, and 7 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing, and Data validation/confrontation.

2.2 Statistical tasks: Data editing and Imputation

2.3 Selective editing has the same purpose but it selects units on a record-by-record basis, whereas macroediting selects units by considering all the data at once.

3. Description of the method

Macro-editing is a general approach to identify potentially influential errors in a data set for manual follow-up. It can be used when all the data, or at least a substantial part thereof, have been collected. In addition, the method is particularly effective when it is applied to data that contain only a limited number of large errors. Macro-editing may succeed in finding these errors by examining the data from a macro rather than a micro level perspective – in other words, looking at the whole data set instead of one record at a time. Macro-editing proceeds by computing aggregate values from a data set and systematically checking these aggregates for suspicious values and inconsistencies. The following types of checks are typically used:

- Internal consistency checks. An example is that, based on subject-matter knowledge, the fraction of total net 4 turnover from domestic sales may be expected to lie between certain bounds; i.e., a < X / X < b 1 for certain constants a and b.

- Comparisons with other statistics. It may be possible to compare aggregates to similar estimates from other data sources. If large differences occur, the corresponding aggregates are identified as suspicious.

- Comparisons with previously published statistics. In repeated surveys, one can compare current aggregates to a time series of previously published values.

It should be noted that in macro-editing all actual adjustments to the data take place at the micro level, not the macro level. Therefore, after one has found suspicious aggregates by any of the above means, the next step is to identify individual units that contribute to these aggregates and may require further editing. There are two generic approaches to do this:

the *aggregate method* proceeds by 'drilling down' from suspicious aggregate values to lower-level aggregates and, eventually, individual units; the *distribution method* examines the distribution of the microdata to identify outliers and other suspicious values.

3.1 Aggregate methods

The aggregate method starts by calculating estimates of aggregates at the highest level of publication based on the current data. If an aggregate is identified as suspicious, the next step is to zoom in on the cause of the suspicious value by examining the lower-level aggregates that contribute to the suspicious aggregate. In this way, macro-editing proceeds until the lowest level of aggregation is reached, i.e., the individual units. Finally, the units that have been identified as the most important contributors to a suspicious provisional publication figure are submitted to manual follow-up.

3.2 The distribution method

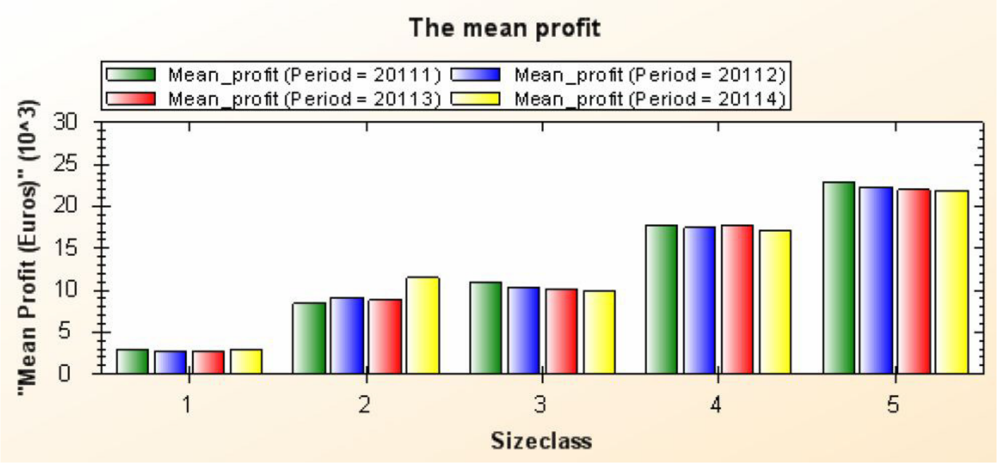
This method identifies observations that require further treatment by applying techniques for detecting outliers, i.e., observations that deviate from the distribution of the bulk of the data. Records are then prioritised for manual follow-up by ordering them on some measure of 'outlyingness'. Graphical displays can also be useful for detecting observations that deviate from the distribution of the bulk of the data.

In practice, the distribution method is often applied in conjunction with the aggregate method. Thus, the macro-analysis starts by identifying suspicious aggregates at the highest level and 'drills down' to suspicious aggregates at a lower level. Subsequently, the distribution method is applied to identify the records that are likely to contribute most to the total error in the identified low-level aggregates, more details can be found in Memobust (2014)

4. Examples

As an example taken from Hacking and Ossen (2012), Figure 1 shows a histogram that compares the mean value of profit across several reference periods and several size classes. It is seen that the mean profit in the last period for size class 2 is unusually high in comparison with previous periods and other size classes. This could be a reason to identify this aggregate as suspicious and drill down to the contributing units.

Figure 1. Example of a histogram for macro-editing.



Another example suggested by Statistics Belgium is the following. A given year, environmental investments by companies had almost doubled for the all country. By sorting the weighted microdata, it was discovered that one respondent (a public company responsible for water sanitation) with a weight 39 had invested in a major sewage treatment plant. The weight was given on the ground of the size of the company. An environmental expert statistician verified that only 5 major sewage treatment plants were constructed this year in the all country. The result of the survey gave the equivalent of 39 sewage treatment plants instead. Sewage plants are 100 % financed by the government and are thus not related to the size of the company. After discussion with the methodologist, it was decided to edit the item in order to reach an equivalent of 5 sewage treatment plants for the country for this strata. The item result error is linked to the generic calculation of a size and weight (depending on last available value added and employment), which is not stricto sensu appropriate for "low occurrence" items like investments.

5. Input data (characteristics, requirements for applicability)

Data with numerical variables. All (or nearly all) should be available.

6. Output data (characteristics, requirements)

A data set with observations flagged as affected by influential errors.

7. Tools that implement the method

MacroView (CBS) see Hacking and Ossen (2012).

8. Appraisal

In integrated administrative data macro-editing can be a useful tool to reveal important errors due to an incomparability of the sources in some estimation domain are present. For instance, it can happen that the definition of a variable is the same in two data sources. Nevertheless, for a specific economic sector some particular businesses could not provide the complete amount of the value in one source because of fiscal benefits typically allowed only for that segment of units. Macro-editing can be useful to isolate those critical situations that the subject matter expert may study and interpret in order to fix the problem wherever it is possible. Macro-editing can also reveal errors due to data linking or to the incomplete delivery of some sources, as anomalous aggregates may result from not enough covered domains from one time period to the subsequent one.

9. References

Hacking, W. and Ossen, S. (2012), User Manual MacroView. Report PMH-20121125-WHCG, Statistics Netherlands, Heerlen

Memobust (2014). Macro-Editing, in Memobust Handbook on Methodology of Modern Business Statistics. https://ec.europa.eu/eurostat/cros/content/memobust_en