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

<table>
<thead>
<tr>
<th>Contract number:</th>
<th>Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responsible person at Commission:</td>
<td>Fabrice Gras Eurostat – Unit B1</td>
</tr>
<tr>
<td>Subject:</td>
<td>Deliverable D5b</td>
</tr>
<tr>
<td>Date of first version:</td>
<td>14.03.2017</td>
</tr>
<tr>
<td>Version:</td>
<td>V1</td>
</tr>
<tr>
<td>Date of updated version:</td>
<td>-</td>
</tr>
<tr>
<td>Written by:</td>
<td>Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang</td>
</tr>
<tr>
<td>Sogeti Luxembourg S.A.</td>
<td>Laurent Jacquet (project manager)</td>
</tr>
<tr>
<td></td>
<td>Sanja Vujackov</td>
</tr>
</tbody>
</table>

Method 1: T5_27.2_Mass imputation

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)

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

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

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

2. The related scenarios

2.1. Mass imputation can be used for a single data set composed of microdata. It can be used in data configurations 3 to 5 (see Deliverable 1). Statistical usage is “Direct tabulation”.

2.2. Statistical tasks: “Direct tabulation” and “Integrate data”.

2.3. With respect to the statistical task “Integrate data”, alternative methods are repeated weighting, repeated imputation and macro-integration.

3. Description of the method

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

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

4. Examples

At Statistics Netherlands, mass imputation is examined for the Educational Attainment. By mass imputing educational attainment, an estimate for the highest educational level of all people in the Dutch population will become available. Those estimates can subsequently be used to compile part of the Dutch Population and Housing Census. If results of mass imputation are satisfactory, this mass imputation approach is planned to be used for the 2021 Census.

5. Input data (characteristics, requirements for applicability)

Mass imputation uses microdata as input data.

6. Output data (characteristics, requirements)

The output of mass imputation consists of microdata.
7. Tools that implement the method

Tailor-made programs and scripts are used.

8. Appraisal

A fundamental problem with a mass-imputed data set is that it may be used for purposes for which it was never intended. Moreover, in most application of mass imputation it is hard to tell from the imputed data set itself that one is using it for unintended purposes.

An example is combining the amount of money spent per month on dog food, (which may be known from a Budget Survey), with whether or not people have a dog as pet (which may be known from a Survey on Living Conditions). Including these two variables – the amount of money spent per month on dog food and whether one has a dog as pet or not – in an imputation model is, except in very exceptional cases, not deemed important enough. Including information on their correlation in an imputation model is even more unlikely.

If these variables and information on their correlation are not included in the imputation model and one is not aware of this, one may decide to analyse and publish the relation between these variables. In this case one may come to the unjustified conclusion that many people who do not have a dog as pet spent money on dog food, and that conversely many people who do have a dog a pet do not buy dog food.

The situation is different for the Educational Attainment data, as here mass imputation is used for a clear purpose and all relevant variables can be included in the imputation model.

In order to alleviate the problem of not knowing from the imputed data set itself that one is using it for unintended purposes, one could consider releasing additional information about which (combinations of) variables are included in the imputation model and are hence controlled for. For instance, for a data set with three variables $x_1$, $x_2$ and $x_3$, one could use the notation \([x_1 \mid x_2 x_3]\) (Zhang 2017) to express that $x_1$ and the combination $x_2$ and $x_3$ are included in the imputation model. As the user now knows that, for example, the combination $x_1$ and $x_2$ is not included in the imputation model, he/she will also know that estimates involving the combination of $x_1$ and $x_2$ will not be valid.

9. References

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


Zhang, L.-C. (2017), Personal communication.