Methods for integrating survey data and big non-survey data

Author: Elena Viti

Università di Pisa[1], Universität Trier[2]

Supervisors: Caterina Giusti[1], Ralph Münnich[2]

Co-supervisor: Johannes Straubinger[2]

Table of Contents

Abstract

The purpose of this thesis is to investigate methods for integrating data from different surveys. Data integration is a broad topic associated with many different statistical techniques. This thesis addresses the topic systematically, by first proposing integration techniques for probability samples, and then integration techniques between probability samples and big data. The term big data refers to non-probabilistic samples containing a large amount of variables and population elements. First, the determining conditions for choosing the technique, sample type and available information are presented. Then, a suitable statistical technique is selected and applied for survey integration. For a more in-depth analysis the selected statistical technique is applied to different models, in the case of integration between two probability samples, or to different types of big data, in the case of integration between a probability sample and big data. At the end, the estimates obtained in each case study is examined with particular regard to the relationship between the applied technique and different models or different types of big data.

1 Introduction

Data integration is a hot topic which is gaining a lot of interest during the last years in the field of statistics. This is explained by the growing availability of data in every sector and the need for answering challenges that finite population inference using probability sampling has always faced, (i) high-cost, and (ii) increasing response burden. Data integration means combining estimates from multiple surveys, a definition is given by Dalla Valle 2017: "Data integration is the process of combining heterogeneous data that originate from different sources, providing a unified view of this information".
The increasing amount of questions (coming mainly from politics but not only) that can't be addressed with a single existing survey has driven the development and growth of this field of statistics. Combining information from multiple surveys can be beneficial for both sampling and non-sampling errors [1] by using one survey to supply information that is lacking in another. In this way the resulting *combined* estimates are enhanced in estimating quantities and have a lower level of sampling errors.
To give the reader a chance to navigate the breadth of methodologies applicable in data integration [2], a systematic approach to classify statistical techniques on the basis of the type of sample considered is borrowed from Yang and J. K. Kim 2020:

1. integration between two probability samples
2. integration between a probability sample and big data.

Two different approaches will be explored for combining two probability samples - which is the simplest case - and one methodology will be applied for integrating a probability sample with big data [3].
The paper is organized as follows: first in section 2 the dataset used is presented. The type of sample and the information available in each sample, which constitute the basis for the application of the methodologies, will be selected from the dataset, so an in-depth knowledge of it is essential. Section 3 describes the approaches used for integrating probability samples. In subsection 3.1 the samples are combined to obtain more efficient estimator of the parameter of interest. The method is applied to different

[1] typically non sampling error are caused by missing data, coverage error and measurement or response error. While a survey can be planned to achieve a particular level of sampling error, it is more difficult to assess non-sampling error.

[2] suffice it to say that techniques such as *Statistical matching*, *Data harmonization* and *Imputation* can be associated to data integration.

[3] this last methodology could be also applied to combine a probability sample and a non-probability sample

type of information available in the samples and different sizes of the samples. In subsection 3.2 the second approach for integrating probability samples is described. The aim is creating a single synthetic dataset containing information available in both sample and then use it for estimation. The method is applied to different type of information available in the sample. Section 4 describes the statistical technique used for integrating probability sample and big data. A few words are spent on the general characteristics of big data to explain how was possible to obtain big data starting from the dataset presented in section 2. The method proposed is applied to different types of big data. In sections 3 and 4 simulations are implemented for every case and results are analyzed. To summarize, in section 5 general conclusion are drawn and future research are indicated.

2 Dataset

The dataset used for the analysis is the AMELIA dataset. AMELIA is a “synthetic but realistic dataset based on social science data” and “provides a realistic framework for open and reproducible research based on EU-SILC data. [...] AMELIA mimics real data, i.e. displays marginal distributions and basic interactions between variables of EU-SILC data (Burgard, Kolb, et al. 2017).

EU-SILC stands for European Union Statistics on Income and Living Conditions. The aim of the survey is "to collect timely and comparable cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions." [4] Information contained in EU-SILC includes: personal and household data, child care, type of housing, tenure and housing conditions, housing expenses and utilities, nonmonetary indicators of household deprivation, physical and social environment, personal and household level of income, level of education, health and access to health care and employment information. Some of the variables contained in AMELIA have a direct counterpart in EU-SILC. Table A1 shows the name variables used, the EU-SILC counterpart (if any), and, for categorical variables, a description of different levels. The levels are listed as used in this paper and do not necessarily correspond to those available in AMELIA, some variables have been re-categorized. The AMELIA dataset was developed within the scope of the AMELI (Advanced Methodolgy for European Laeken Indicators [5]) project, that concerns simulations for poverty measurements. AMELIA dataset is an answer to the difficulties in carrying out methodological research due to the lack of freely available adequate data.
The use of a synthetic dataset represents the ideal choice for this work, which follows a model-assisted approach.[6] Indeed, AMELIA dataset allows to leave out comparability issues that arise when implementing data integration and using real data. These issues that can arise are pointed out by Elliott, Raghunathan, and Schenker 2018:

- Differences in the type of respondents and/or source of responses’ information: consider for example the case of two face-to-face interviews; in one respondents provide information from their memory, and in the other one they answer by consulting some records available while providing information

[4] https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

[5] https://www.uni-trier.de/en/universitaet/fachbereiche-faecher/fachbereich-iv/faecher/volkswirtschaftslehre/professuren/wirtschafts-und-sozialstatistik/forschung-aktuell-1/surveystatisticsnet/ameli-1/about-ameli

[6] in a model assisted approach the aim is to improve inference and analysis by using correlated information, given an underlying theoretical or design model.

- Differing modes of data collection: like a face-to-face interview against a telephone interview
- Survey context: response error may differ according to whom is conducting the survey (a well-known National Statistical Institute against a reputed institution, but not so well-known)
- Differences in the survey design
- Differences in survey question (question wording or placement of the questions)

With AMELIA it is possible to evaluate and compare the procedure used without facing the above problems. The dataset [7] and its samples are freely available for the software R, the version of AMELIA used in this work is v0.2.3 [8]. The main properties of the dataset (as list in Burgard, Ertz, et al. 2020) are:

1. large population size (approximately 10 million observations of 33 variables on personal level and approximately 3.7 million observations of 27 variables on household level)
2. household structure available (both person-level and household-level variables are available)
3. regional structure available (4 regions, 11 provinces, 40 districts and 1592 cities)
4. maps of different regional structure (currently in preparation)
5. samples using different samples designs already drawn

Specifically, the sampling designs available are simple random sampling without replacement, stratified sampling, two-stage stratified sampling. In all the application presented in the next sections person-level variables are used.

[7] each variable is stored in a separated file.
[8] http://amelia.uni-trier.de/?page_id=121

3 Combining probability samples

Probability samples are representative of the target population since they are selected under known sampling design. The selection probability is known and inference is usually design-based [9]. The basic setup is the following: $\mathcal{U} = \{1,...,N\}$ is the set of N units for the finite population, N is the population size and it is known. $(x_i^T, y_i)^T$, with $i = 1,...,N$, is the realized value of random variables $(\mathrm{X}^T, Y)^T$ for unit i, X are auxiliary variables while Y is the target variable. The parameter of interest is $\mu_Y = N^{-1}\sum_{i=1}^{N} Y_i$. I_i is the sample indicator, if $I_i = 1$ the unit i is in the sample, if $I_i = 0$ the unit i is not selected in the sample. $\pi_i = P(I_i = 1 | i \in \mathcal{U})$ is the first-order inclusion probability, while $d_i = \pi_i^{-1}$ is the design weight. The sample size is $n = \sum_{i=1}^{N} I_i$. In this work simple random sampling without replacement is used (SRSWOR), therefore the inclusion probability is $\pi_i = \frac{n}{N}$. When the sampling design is known the first-order inclusion probability and design weights are also known, so Horvitz-Thompson estimator can be applied. The Horvitz-Thompson estimator or π estimator is an unbiased estimator for $t_y = \sum_{i=1}^{N} Y_i$ and it is defined as:

$$\hat{t}_{y,\pi} = \sum_{i=1}^{N} \frac{I_i}{\pi_i} \cdot y_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i \cdot y_i. \tag{1}$$

For estimating the mean of the parameter of interest the π estimator is divided by N $\hat{\mu}_Y = N^{-1}\sum_{i=1}^{n} d_i y_i$. The variance of the HT (Horvitz-Thompson) estimator for SRSWOR can be estimated using:

$$\hat{V}(\hat{t}_{y,\pi}) = N^2\Big(\frac{1}{n} - \frac{1}{N}\Big)S_{y,s}^2 \tag{2}$$

where $S_{y,s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ wiht $\overline{y} = \sum_{i=1}^{n} y_i$. To estimate the variance of $\hat{\mu}_Y$ it is sufficient to divide equation 2 by N^2.

If populations totals of auxiliary variables X are known it is possible to use them to correct the π estimator for Y. This process is called *calibration* or *inverse regression*, that is, the use of known data in the population about independent variables in order to estimate values of the dependent variable. During the calibration process, weights are assigned to the units in the sample so that the total X estimated using the sample matches the one observed in the population:

$$\hat{t}_X = t_X \tag{3}$$

[9] design-based inference means that the statistical model is known and the focus is on the experiment.

Equation 3 is the *calibration property* and belongs to all calibration estimators. When auxiliary information is available calibration estimators can be used to reduce the variance of the estimates by using the relationship between the target variable and the auxiliary variables. The effect of calibration becomes negative when there are too many marginal constraints to which calibrate information in the sample, in this case the variation of weights becomes very high and the minimization problem is an empty space. Calibration estimators minimize the distance between the original design weights d_i and the corrected weights. The estimator considered here is the generalized regression estimator (GREG) that minimizes the quadratic distance function between the weights (there could be other choices). The GREG estimator is:

$$\hat{t}_{y,GREG} = \hat{t}_{y,\pi} + \hat{B}(t_x - \hat{t}_{x,\pi}) \tag{4}$$

where t_x is the known population totals of X and $\hat{B}$ is the regression coefficient estimated from the sample. GREG estimator is a design-based estimator, but its efficiency depends on the ability of the model to describe the data. Indeed, the computation of GREG variance is based on the variance of the residuals $(y_i - \hat{y}_i)$, where $\hat{y}_i$ are the predicted values. Therefore, the higher the fitting of the linear working model the lower the variance of GREG estimator (or the higher its accuracy). On the other side, if the model underlying the GREG is not appropriate for the target variable, a too large variation of weights may increase the variance with respect to the HT estimator (GREG estimator can originate negative weights). To avoid an increase of the variance of the estimates, the variability of weights must be related with target variable (WILLENBORG, SCHOLTUS, and VAN DELDEN n.d.). $\hat{t}_{GREG}$ is asymptotically unbiased for t_y, this characteristics makes the GREG relatively robust to model choice (Hedlin et al. 2001).

When combing probability samples two different approaches can be identified basing on the level of information to be combined: a macro approach and a micro approach. The techniques that can be used for each approach depend on the type of missingness [10] and information available.

3.1 Macro approach

The aim of the macro approach is to "obtain summary information – point and variance estimates – from multiple data sources and combine those to obtain more efficient estimator of the parameters of interest, such as population means or totals (Yang and J. K. Kim 2020)". In other words, information from two different data sources are jointly

[10] the manner in which data are missing from a sample of a population.

pooled to obtain better estimates of the parameter of interest than if information from a single source is used.

The application for the macro approach consider a non-monotone missingness (see table 1). In this first application two probability samples A and B are involved, A and B are drawn from the population with simple random sampling without replacement (SRSWOR), so the first order inclusion probability is equal to $\pi_i = n/N$. With respect to the technique that will be presented, no particular disadvantage or advantage is mentioned in applying one sampling design rather than another.

Table 1: Non-monotone missingness for the macro-approach

	d	Z	X_1	X_2	Y_1	Y_2
Sample A	✓	✓	✓		✓	
Sample B	✓	✓		✓		✓

Looking at table 1, sample A and sample B have some auxiliary variables Z in common (called *common variables*), while X_1 and X_2 (*control variables*) are observed only in sample A and B respectively. For Z population totals are unknown, for X_1 and X_2 population totals are known. The target variable Y_1 is observed only in sample A, while the target variable Y_2 is observed only in sample B.

Renssen and Nieuwenbroek 1997, Merkouris 2004 and Merkouris 2010 address the problem of combining data from two independent probability samples to estimate μ_{Y_1} and μ_{Y_2}. The idea behind these papers is to obtain the estimation for the population totals of the common variables Z by pooling both surveys and then use them, together with the known population totals of the control variables, to improve the estimates of μ_{Y_1} and μ_{Y_2}. In other words, estimation of the population totals of the common variables by pooling both surveys makes possible to use them as additional regressors. Following this procedure it is obtained what Renssen and Nieuwenbroek 1997 call *adjusted general regressor estimator*. The weights of such regressor estimator are *reproductive* with respect to the control variables and *consistent* with respect to the common variables.

This technique is well suited to split questionnaires survey design where instead of having a long questionnaire there are two shorter questionnaires, 1 and 2. The first part of both forms contains questions regarding the common variables, then one part of the remaining questions is assigned to form 1 and the other part to form 2. This allows to decrease the respondent burden and to increase the response rate (trivially because the questionnaires are shorter). It can be claimed that the reduced number of observations entails a loss of precision respect to the case where X_1 and X_2 are observed together. This is true, however the loss of precision can be limited if the common vari-

ables are highly correlated with the target variables.

3.1.1 Methodology

Going into the details of the methodology (found in Renssen and Nieuwenbroek 1997 and Merkouris 2004) the first thing to do is to estimate the unknown population totals of the common variables:

$$\hat{t}_z = P \cdot \hat{t}_{z_A,GREG} + Q \cdot \hat{t}_{z_B,GREG} \tag{5}$$

where $\hat{t}_{z_A,GREG}$ and $\hat{t}_{z_B,GREG}$ are the general regression estimators for t_z in sample A and B respectively. P and Q are two square matrices such that $P + Q = I$. In the literature two different choices for P and Q are proposed:

1. **proportional choice**: takes into account the difference in sample sizes;

$$\begin{aligned} P &= (n_A + n_B)^{-1} \cdot n_A \\ Q &= (n_A + n_B)^{-1} \cdot n_B. \end{aligned} \tag{6}$$

2. **optimal choice**: that takes into account the difference in samples sizes, the use of control variables and the efficiency of the design;

$$\begin{aligned} P &= V(\hat{t}_{z_B,GREG}) \cdot [V(\hat{t}_{z_A,GREG}) + V(\hat{t}_{z_B,GREG})]^{-1} \\ Q &= V(\hat{t}_{z_A,GREG}) \cdot [V(\hat{t}_{z_A,GREG}) + V(\hat{t}_{z_B,GREG})]^{-1}. \end{aligned} \tag{7}$$

Once the totals for Z are obtained it is possible to compute the adjusted general regression estimator, that is like the general regression estimator plus an adjustment term:

$$\hat{t}_{y,AR} = \hat{t}_{y,\pi} + \hat{B}(t_x - \hat{t}_{x,\pi}) + \hat{D}(\hat{t}_z - \hat{t}_{z,\pi}) \tag{8}$$

where $\hat{t}_{y,\pi} + \hat{B}^t(t_x - \hat{t}_{x,\pi})$ is the general regression estimator $\hat{t}_{y,GREG}$ and $\hat{D}(\hat{t}_z - \hat{t}_{z,\pi})$ is the adjustment term. To estimate the mean $\hat{\mu}_Y$ it is sufficient to divide $\hat{t}_{y,AR}$ by N.

3.1.2 Simulation

The methodology explained above is applied to all the cases presented in table 2. Residential status (RES) is a categorical variable that indicates whether the individual currently lives in the house (level 1) or if he/she is temporarily absent (level 2). Person with highest income in the household (PWHI) is a categorical variable with two

categories: 1 if the individual is the person with highest income in the household, 2 otherwise. Basic activity status (BAS) is a categorical variable available in AMELIA with 4 levels: level 1 = at work, level 2 = unemployed, level 3 = in retirement or early retirement or has given up business, level 4 = other inactive person. In this work the variable has been re-categorized in three levels by grouping level 2 and level 3 together. The choice to group in this way was made on the basis of the correlation between basic activity status and the target variables. For example, through a graph showing the relationship between the BAS and personal income it is possible to note how for levels 2 and 3 of basic activity status the average personal income does not change; for this reason the two categories have been merged into a single one. Age (AGE) is the age of the individual and in AMELIA is censored at 80. Age has been categorize in 3 levels based, again, on the relationship with the target variables: levels 1 = [0-20], levels 2 = [21-60], levels 3 = [61-80]. Self-employment (SEM) is a categorical variable with value 1 if individual is self-employed and value 2 if the individual is not self-employed. Age and self-employment are the control variables, their totals are available in the population and used for computing $\hat{t}_{z_A,GREG}$, $\hat{t}_{z_B,GREG}$ and $\hat{t}_{y,AR}$.
The choice to study the cases presented in table 2 is inspired by the simulation study in Renssen and Nieuwenbroek 1997. The idea is to compare the behaviour of the general regression estimator, the adjusted general regression estimator with optimal choice, and the adjusted general regression estimator with proportional choice, when there are different sample sizes and different number of control variables.

The target population is all people of AMELIA dataset. Sample A and sample B are two probability samples drawn from the same population N.
The simulation is implemented as follows:

1. 1000 samples A and B are randomly drawn from the population with SRSWOR. In each simulation the seed is set to i (i indicates number of the simulation and goes from 1 to 1000).

2. for each of the 1000 samples A the totals of the common variables Z are estimated. The same is done for each of the 1000 samples B. To estimate $\hat{t}_{z_A,GREG}$ and $\hat{t}_{z_B,GREG}$ the known total of the common variables is used in the calibration.

3. using the estimated totals from sample B and sample A it is possible to compute the matrices P and Q for the optimal case (to compute P and Q in the proportional case it is sufficient to know the sample size). The matrices P and Q are used to compute the totals of the common variables $\hat{t}_z$. For each of the common variables two versions, using P and Q proportional or optimal, of the totals are obtained (1000 for proportional choice and 1000 for optimal choice, for each case).

Table 2: Cases analyzed in the macro approach

	Case 1	
	Sample A	Sample B
Sample size	6000	2000
Target variables	Personal income	Equivalised disposable income
Common variables	Residential status, Person with highest income in the household, Basic activity status	
Control variables	Age Self-employment	Age
	Case 2	
	Sample A	Sample B
Sample size	3000	3000
Target variables	Personal income	Equivalised disposable income
Common variables	Residential status, Person with highest income in the household, Basic activity status	
Control variables	Age Self-employment	Age
	Case 3	
	Sample A	Sample B
Sample size	6000	2000
Target variables	Personal income	Equivalised disposable income
Common variables	Residential status, Person with highest income in the household, Basic activity status	
Control variables	Age	Age

4. the totals of the common variables estimated by pooling both surveys are used, together with the control variables X_1 for sample A and X_2 for sample B, to estimate the totals of the target variables, Y_1 for sample A and Y_2 for sample B.

To compute GREG estimators and adjusted GREG estimators the function calibrate in R package *survey* (Lumley 2020) is used.
Figure 2 represents a multiple violin plot showing the distributions of GREG and adjusted GREG estimators in all cases in table 2 for personal income.

Violins in figure 2 represent distributions of the estimators in the 1000 simulations performed. The wider part of the violin represent values assumed by the estimator

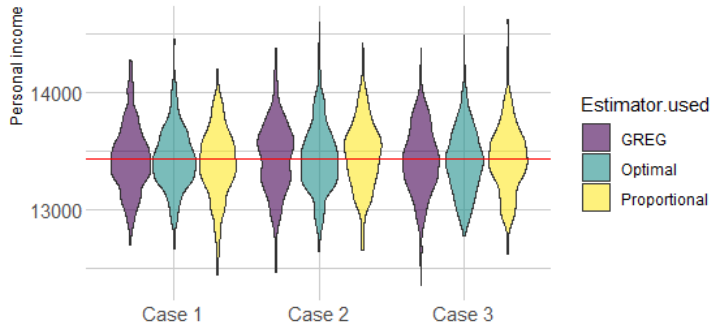


Figure 2: Violin plot of income for different models in macro approach. Red line is the true value of mean personal income in the population

with higher probability. The red horizontal line is the true value of the mean personal income observed in the population. The fact that this value is fairly in the middle of all violins is therefore a good sign; it means that by running a simulation there is a high probability of obtaining an estimated personal income close to its true value in the population. In contrast, the narrower and more elongated parts of the violins represent values assumed by the estimator with less probability. When a violin has a particularly elongated shape, such as the GREG estimator in case 3, it means that in worst-case scenarios we observe values that are particularly far from the true value in the population. In general, the GREG estimator and adjusted GREG estimators are more compact and centered around μ_Y in the case of personal income and more elongated when equivalised disposable income is estimated (see figure A1 in the appendix). To test whether there was a gain in terms of variance reduction, for each case and for both samples, the variance of the Horvitz-Thompson estimator, the variance of the GREG estimator and the two GREG adjusted estimators are computed. The variance of the HT estimator is taken as reference and set to 100. The results are reported in table 3 (a similar comparison is made by Renssen and Nieuwenbroek 1997). The first notable thing is that all estimators show an improvement in terms of variance over the Horvitz-Thompson estimator (they are lower than 100). This means that the weights obtained as results of the calibration process are correlated with the target variables and the efficiency of the estimates improves (this is more evident for personal income than for equivalised disposable income). In case 1 and case 2 samples A and B differs for the number of control variables. Sample A uses 2 control variables (age and self-employment, 5 categories in total) and sample B uses 1 (age, 3 categories in total). When the number of control variables is higher the GREG estimator performs better

	Sample A Personal income	**Sample B** Equivalised disposable income
Case 1		
GREG	93.38	99.93
Adjusted GREG		
1) optimal	83.30	98.78
2) proportional	83.30	98.78
Case 2		
GREG	93.33	99.92
Adjusted GREG		
1) optimal	83.12	98.79
2) proportional	83.12	98.79
Case 3		
GREG	96.55	99.93
Adjusted GREG		
1) optimal	86.39	98.81
2) proportional	86.39	98.81

Table 3: Estimated variances of general regression estimator (GREG) and adjusted general regression estimator (adjusted GREG) for different choices of P and Q matrices relative to the corresponding estimated variance of the Horvitz-Thompson estimator

in terms of variance, 93.38 against 99.93 and 93.33 against 99.92. When the control variables are the same for the two samples, as in case 3, the GREG estimator performs better in sample A, 96.55 against 99.93. When the GREG estimator is calculated using more control variables, it is normal to expect it to perform better, as can be seen from cases 1 and 2. The fact, however, that in case 3, where the common variables are the same, the GREG performs better again for sample A is evidence of other factors affecting its accuracy. These factors are, the sample size - in cases 1 and 3 sample A has a larger sample size than sample B - and the correlation between the target variables and the control variables - the correlation is stronger with personal income than with equivalised disposable income.

Comparing the results of the GREG estimator with those of the adjusted GREG we see that the latter performs better in all three cases, both with the optimal and proportional choice. The difference in the performance is more evident in sample A than sample B. In all cases in sample A the adjusted GREGs have a variance about ten percentage points lower than that of the GREG, while in the case of sample B this distance is reduced to one percentage point.

When common variables are used the factors that influence the variance reduction are: the difference in sample sizes between the two samples, the choice of P and Q matrices, the partial correlation between target and control variables and the partial correlation

between common and control variables (Renssen and Nieuwenbroek 1997). In cases 1 and 2 we expect to see a difference between proportional and adjusted GREGs due to the different number of control variables. In particular we would expect the optimal choice to outperform the proportional choice since the first takes into account the difference in sample sizes and in the use of control variables, while the second one takes into account only the difference in sample sizes. To be more precise; in case 1 optimal choice consider that sample A has two control variables and sample B just one and that sample A is bigger than sample B, while proportional choice only takes into account that sample A is bigger than B. However, contrary to expectations, for cases 1 and 2, in both samples A and B, the two different types of P and Q matrices give the same results (83.30-83.30, 98.78-98.78, 83.12-83.12, 98.79-98.79). The same occurs in case 3 (86.39-86.39, 98.81-98.81), where, however, unlike the other two cases, this result was expected since the number of control variables used and sample sizes are the same in the two samples. The differences between optimal and proportional choice *within* the same sample can be attributed to the fact that the number of control variables used in the two sample is almost the same (5 categories, 3 of AGE and 2 of SEM, versus 3 categories of AGE) and can't reflect a difference in the variance of the adjusted GREGs. Instead, the differences *between* sample A and sample B concern the partial correlation between target and control variables and the partial correlation between common and control variables, which in both cases is greater with personal income than with equivalised disposable income.
To try to bring to light numerical differences between the variances of proportional and optimal choice a fourth simulation was conducted using three more common variables (9 categories added in total) for sample A than for case 1 (see table 4). The added common variables are marital status (3 categories), household size (4 categories) and sex (2 categories).

Also for this additional case, the variance has been calculated for each estimator (GREG and the two adjusted GREGs) using the Horvitz-Thompson estimator variance value as 100. Results are shown in the second part of table 4. Despite the increase of control variables in sample A the proportional estimator still performs as well as the optimal (to find a difference between the two one has to look to the second decimal place). To see greater differences between the two matrices P and Q used one can:

- further increase the common variables for sample A
- increase the difference between sample A and sample B using two different sampling designs. The proportional choice takes into account the efficiency of the

Table 4: Additional case in the macro approach

	Case 4	
	Sample A	Sample B
Sample size	6000	2000
Target variables	Personal income	Equivalised disposable income
Common variables	Residential status, Person with highest income in the household, Basic activity status	
Control variables	Age, Self-employment, Marital status, Household size, Sex	Age

	Sample A	**Sample B**
	Personal income	Equivalised disposable income
Case 4		
GREG	92.56	99.93
Adjusted GREG		
1) optimal	82.80	98.79
2) proportional	82.81	98.78

sampling design and may perform better than the proportional choice, which does not take this into account

Other statistics about the simulation, relative bias and relative root mean squared error (RMSE), are reported in the appendix in table A2.

3.2 Micro approach

The aim of the micro approach is to "create a single synthetic dataset that contains all available information for all data sources (Yang and J. K. Kim 2020)". In other words, the goal is to collect information from two or more different sources and combine them into a single, more easily consultable dataset. The synthetic dataset created can then be used to estimate parameters of interest[11].

The application for the micro approach consider a monotone missingness (see table 5).

Sample A and sample B are collected from two independent surveys referring to the same target population. From the first survey sample A, which contains information

[11] again, the ultimate scope is the creation of the synthetic dataset itself, not the estimation of the parameter, contrary to what happens for the macro approach.

Table 5: Monotone missingness considered for the micro-approach

	d	X	Y
Sample A	✓	✓	✓
Sample B	✓	✓	

on both X and Y, is obtained; from the second survey sample B, which contains information only on auxiliary variables X, is obtained. As in the macro approach, there is no particular advantage or disadvantage in applying one sampling design rather than another; so both sample A and sample B are drawn from the target population with SRSWOR, the first-order inclusion probability is $\pi = n/N$. The target population is again all people in AMELIA dataset.

The technique applied is mass imputation (or synthetic data imputation); this methodology allows to create imputed values for variables not observed in a survey by using information contained in other surveys. In mass imputation an independent sample is used as training dataset and imputation is applied to *all the units* in the other sample (differently from usual imputation where an imputation model is developed to impute missing values in a dataset). Legg and Fuller 2009 and J. K. Kim and Rao 2012 develop synthetic imputation approach to combine different sources. The idea behind this papers is to use sample A to fit a working model relating the variable of interest Y to the auxiliary variables X. The model fitted is then used to predict the target variables Y associated with the values of auxiliary variables X in sample B. Then a projection estimator for μ_Y is obtained from design weights of B and synthetic (imputed) values of Y.

This technique is suitable in cases where sample B is much larger than sample A and collect information on Y is relatively expensive, so measuring Y in sample B would require a large amount of money. In this paper sample A and B are two samples drawn from the same target population, also called nonnested two-phase sampling. The same methodology can be applied to traditional two-phase sampling, or double sampling, where a large first-phase sample is selected and then a much smaller second-phase subsample where also Y is observed. Sample A and B could be also selected from different frames as long as X is comparable in the two surveys.

3.2.1 Methodology

Going into the details, in this technique sample A is used as training sample for predicting Y in sample B. The steps are the following:

1. fit a working model $E(Y|X) = m(X;\beta_0)$ basing on the data $\{(x_i;y_i) : i \in A\}$ from sample A

2. predict the values of Y associated with X in sample B, $\tilde{y}_i = m(x_i; \hat{\beta})$ for $i \in B$. $\tilde{y}_i$ are called predicted values or synthetic values.

3. impute predicted values $\tilde{y}_i$ to sample B, thus obtaining a single synthetic dataset

4. compute the projection estimator

$$\hat{Y}_p = N^{-1} \sum_{i \in B} d_{i,B} \tilde{y} \tag{9}$$

The projection estimator is asymptotically design unbiased if

$$\sum_{i \in A} d_{i,A} \{y_i - m(x_i; \hat{\beta})\} = 0 \tag{10}$$

The second term of the difference $\{y_i - m(x_i; \hat{\beta})\}$ represents predicted values of Y when the working model is applied to values of X in sample A. If condition 10 is satisfied then the bias correct-corrected projection estimator

$$\hat{Y}_{p,bc} = N^{-1} \sum_{i \in B} d_{i,B} \tilde{y} + N^{-1} \sum_{i \in A} d_{i,A} \{y_i - m(x_i; \hat{\beta})\} \tag{11}$$

coincides with equation 9. An advantage of this procedure, as pointed out by J. K. Kim and Rao 2012 is that the same design weight $d_{i,B}$ is used even when different working models generate the synthetic values. The variance of $\hat{Y}_p$ is equal to:

$$V(\hat{Y}_p) = V\Big[N^{-1} \sum_{i \in B} d_{i,B} \tilde{y}\Big] + V\Big[N^{-1} \sum_{i \in A} d_{i,A} \{y_i - m(x_i; \hat{\beta})\}\Big] \tag{12}$$

The first term in 12 is the variance due to sampling in survey B, the second term is the variance due to sampling in survey A. Excluding the second term in 12 can lead to underestimation of the variance even if 10 is satisfied. When some populations totals of X are known first calibration weights, for sample A and sample B, are obtained such that $\sum_{i \in A} d_{i,A} x_i = t_x$ and $\sum_{i \in B} d_{i,B} x_i = t_x$. Then these new weights are used in equations 10 and 9 respectively.

3.2.2 Simulation

The methodology presented is well-suited to the case where sample B is much larger of sample A, as mentioned in paragraph 3.2, therefore for the simulation study sample A has a size of $n_A = 1500$, and sample B $n_B = 15000$. Both samples are drawn with SRSWOR from the target population, all people in AMELIA dataset. The target variable

Y is personal income. Different cases regarding the auxiliary variables observed in the sample and the known population totals are considered, these cases are presented in table 6. In the first case sample A and B contain 6 auxiliary variables and for 3 of them totals are known in the population, in the second case 4 auxiliary variables are available and no totals are known in the population, in the last case two auxiliary variables are observed in the samples and total is available in the population for on of them. Age has been categorized in 3 levels; level 1 = [0-20] years, level 2 = [21-60] years, level 3 = [61-80] years. Basic activity status has three levels (in AMELIA it is available with four levels); level 1 = at work, level 2 = unemployed or in retirement or early retirement or has given up business, level 4 = other inactive person. The categorization of AGE and BAS is based on their relationship with income. Marital status (MST) is a categorical variable available in AMELIA with five categories; level 1 = 1: never married, level 2 = married, level 3 = separated, level 4 = widowed, level 5 = divorced. In this work levels 3, 4 and 5 has been grouped together (again basing on the relationship with personal income). Re-categorizing variables allows also to reduce the numbers of marginal constraints in the calibration process. Residential status (RES) has value 1 if the person is currently living in the household and value 2 otherwise. Self-employment (SEM) is 1 if the person is self-employed and 2 if not. Person with highest income in the household (PWHI) is 1 if the individual has the highest income in the household, 2 if not.

Table 6: Cases analyzed in the micro approach

Case 1	
Auxiliary variables	**Population total is known**
Age	Yes
Basic activity status	Yes
Marital status	No
Residential status	No
Self-employment	No
Person with highest income in the household	Yes
Case 2	
Auxiliary variables	**Population total is known**
Age	No
Basic activity status	No
Self-employment	No
Person with highest income in the household	No
Case 3	
Auxiliary variables	**Population total is known**
Age	Yes
Self-employment	No

The simulation is implemented as follows:

1. 1000 samples A and B are randomly drawn from the population with SRSWOR. In each simulation the seed is set to i (i indicates number of the simulation and goes from 1 to 1000).

2. sample A and sample B are calibrated to the known totals of the population (if there are some) and calibrated weights are obtained.

3. for each case in table 6 a working model is fitted. To fit the model repeated k-fold cross validation is used. Sample A is divided into k-subsets, in our case k=10. When the algorithm is run, it will be trained on 90% of the model and tested on 10%, each run of the algorithm a different 10% is leave out from the training. The procedure is repeated 3 times. Dividing sample A in train sample and test sample allows to see how the algorithm performs on data that the model was not trained on; this matters since the aim is to use the model to make predictions on unseen data (sample B). In this step R packages *caret* (Kuhn et al. 2020) and *klaR* (Weihs et al. 2005) are used.

4. predicted values of personal income are computed for sample A to check condition in equation 10 (these values are also used for the computation of the variance, see equation 12)

5. predicted values of personal income are computed for sample B.

6. projection estimator and its variance are computed for each case.

Figure 3 shows a violin plot combined with a box plot for each case. The violin describes the distribution of the data, while the box plots give information on the interquartile range, the median, the maximum and minimum values, and the outliers. The red horizontal line represents the true value of income.

The graph shows that the projector estimator distributions are extremely similar in all three cases. All distributions are centered with respect to the true value μ_Y, that crosses the median in each of the three cases. Wider sections of the violin represent a higher probability that the elements of the sample will take on that value. Therefore, in all three cases there is a high probability that the projector estimator estimates exactly or it is very close to μ_Y. In case number 3 the violin is slightly more stretched than in the other two. This means that when the worst-case scenario occurs, that is, when the sample elements take on an extreme value, the estimator $\hat{Y}_p$ is more further away from μ_Y in case 3 than in cases 1 and 2. Case 1 and Case 3 have the same percentage of auxiliary variables with known total in the population and auxiliary variables with unknown total. Specifically, in case 1 the known totals are 3 out of 6 auxiliary variables

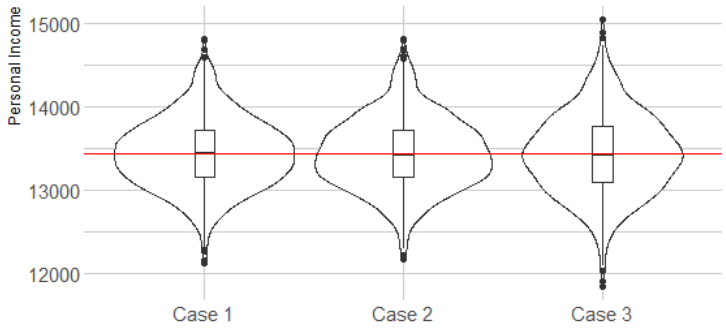


Figure 3: Violin and box plots of personal income for different models in micro approach. Red line is the true value of mean personal income in the population

and in case two the known total is 1 out of 2 auxiliary variables. Known totals in the population are used to calibrate $d_{i,B}$ in equation 9. In case 2, on the other hand, there are 4 auxiliary variables and 0 known totals. From figure 3 it seems that what affects the distribution of $\hat{Y}_p$ is the total number of auxiliary variables, the more the better [12], and not whether their totals are known or not.

Table 7 compare the relative bias and the relative RMSE of $\hat{Y}_p$ in the 3 different cases and with the estimator $\hat{\mu}_{Y,A}$ obtained from sample A. In case 1 and 3 $\hat{\mu}_{Y,A}$ is a GREG estimator using as known totals AGE, BAS and PWHI in case 1 and AGE in case 3. In case 2 $\hat{\mu}_{Y,A}$ is a π estimator since there are no known totals in the population. The

	Projection estimator		**GREG or π estimator**	
	Relative bias	Relative RMSE	Relative bias	Relative RMSE
Case				
1	0.000572	0.000989	0.000926	0.001028
2	0.000451	0.001011	0.000556	0.001247
3	-0.000404	0.001148	0.000003	0.001190

Table 7: Relative bias and relative root mean squared error (RMSE) in different cases of micro approach in comparison with the estimators obtained from sample A in each case

relative bias measures the relative difference from the true parameter:

$$\text{Relative bias} = \frac{\sum_{i=1}^{k}(\hat{\mu}_Y - \mu_Y)}{\mu_Y} \tag{13}$$

[12] obviously to have an improvement in the estimation of μ_Y the variables must be correlated with personal income

where $k = 1000$ is the number of simulations. When the relative bias is equal to 0 there is no difference between the estimator and the true value. Relative RMSE instead is a measure of the accuracy of the model

$$\text{Relative RMSE} = \sqrt{\frac{\frac{1}{k}\sum_{i=1}^{k}(\hat{\mu}_Y - \mu_Y)^2}{\sum_{i=1}^{k}\hat{\mu}_Y^2}}. \tag{14}$$

If relative RMSE=0 then the model has a perfect fit. In general, from table 7 we see that for each estimator, as the number of auxiliary variables increases, the relative bias worsens while the relative RMSE improves. In cases 1 and 2 the projection estimator performs better than the GREG and π estimator from sample A both in terms of relative bias and in terms of relative RMSE. In case 3 the projection estimator outperforms the GREG obtained from sample A in terms of relative RMSE (with lower magnitude compared to case 1 and 2) but not in terms of relative bias. Also in this case it appears that the determining factor, for the projection estimator to outperform the estimator obtained from sample A, is the number of auxiliary variables and not whether or not their totals are known in the population.
Analysis of the graph 3 and table 7 leads to the conclusion that the difference in the relative bias and relative RMSE between case 1 and 2 can be attributed mainly to the difference in the total number of auxiliary variables (6 versus 4) and in a smaller part to the difference in known totals. To test this hypothesis, a simulation was carried out for a fourth case (case 4) with the same 6 auxiliary variables as in case 1 but with all totals unknown. In case 4 the relative bias is 0.000457, and relative RMSE is 0.001009. Compared to case 1 in case 4 the lack of known totals is good news for the relative bias, which decreases due to fewer restrictions, and on the other hand is bad news for the accuracy of the model. The presence of known totals affects the projection estima-

Case	Projection estimator	GREG or π estimator
1	87.98	83.97
2	98.97	100
3	99.52	99.01
4	98.90	100

Table 8: Estimated variances of projection estimator relative to the corresponding estimated variance of the Horvitz-Thompson estimator. Value of 100 indicates that no totals were available and π or HT estimator is computed. The variances of GREG estimators are smaller than those of projection estimators because the latter consist of the sum of two variances; the variance due to sampling in survey B, and the variance due to sampling in survey A (see equation 12)

tor through $d_{i,B}$ (equation 9) and the variance through $d_{i,B}$ and $d_{i,A}$ (equation 12). Is looking at the variance that the greater difference emerges. Table 8 report the variance

of the projection estimators in all 4 cases using as reference value the variance of the π estimator. The table shows that the variance in case 1 is smaller than variance in case 4 (87.98 against 98.90).

4 Combining probability sample and big data

Big data is[13] a term used to indicate data sets that are too large or too complex to be dealt with by the available computer or storage power. In other words, big data is a non-probability sample with large amount of variables and population elements. Being a non probability sample implies that "it fails to represent the target population because of inherent selection biases (Yang and J. K. Kim 2020)". Today big data can be considered 1,024 terabytes of information, including billions or even trillions of records from millions of people[14]. Usually big data is defined through its characteristic, the so-called three Vs (Tam and Clarke 2015):

- volume: the number of data records, their attributes and linkages
- velocity: how fast data is produced and changed, and the speed at which they must be received, processed and understood
- variety: the diversity of data sources, formats, media and content.

Big data contains information of poor quality due to heterogeneity, selection bias and high dimensionality, for these reasons it is not suitable for production of statistics; but it can be used to increase the cost efficiency of produced estimates. The growing demand for more frequent and richer statistical estimates (nowadays the demand for surveys exceed the supply) led the scientific community to wonder if it was possible to solve problems tipically connected to big non-survey data, such as representativeness and measurement errors, by combining them with probability samples, which are much smaller in size but contain high-quality information. For their part, probabilistic samples suffer from high nonresponse rates and require expensive surveys to be conducted. The fact that big data, and non-probabilistic samples in general, are increasingly available provides unprecedented opportunities for research purposes. Where big data is lacking gold standard data sources (census and survey) have their strengths, and vice versa, with data integration it is possible to combine them to eliminate their weaknesses. Big data can come from a variety of different sources (listed by Tam and Clarke 2015 and Couper 2013):

1. administrative sources (administrative or private sector programs): e.g. electronic medical records, hospital visits, insurance records

[13] in the literature there is no uniformity in attributing the singular or plural to the term big data, in this paper the line of adopting the singular was chosen

[14] https://itchronicles.com/what-is-big-data/#:~:text=%E2%80%9CBig%20data%E2%80%9D%20is%20a%20term,records%20from%20millions%20of%20people.

2. transactional data: arising from the transaction between two entities, e.g. credit card transaction or online transaction
3. sensor network sources: e.g. satellite imaging, road sensors and climate sensors
4. tracking device sources: e.g. tracking data from mobile telephone and GPS (Global Positioning System)
5. opinion data sources: e.g. comment on social media
6. behavioral data sources : e.g. online searches and online page views
7. web panel survey : a survey utilizing samples from web panels

For the analysis that will be constructed in the following sections, it is necessary to derive big data from the AMELIA dataset described in section 2. For this reason the focus now will be on one of the sources listed above, web panel surveys. This source is the one from which (imaginatively) the big data used in subsection 4.2 are derived.
"A web panel – or online/internet panel" – is defined as "an access panel of people willing to respond to web questionnaires [...] an access panel is a sample database of potential respondents who declare that they will cooperate for future data collection if selected (Svensson and Sweden 2013)". In other words, web panel are like the sampling frames for web panel surveys. Participants in web panel surveys can be recruited in different ways. However the sample selected from the web panel survey is not a probability sample, even if the recruitment was done thought probability sampling, and so it is subject to selection bias. In this work self-selection as recruiting method is considered (the so-called self-selection surveys are surveys simply put on the web). Since there is no sampling design and everyone one can just complete the survey all selection probabilities are unknown and Horvitz-Thompson estimator can't be applied. As pointed out by Bethlehem 2016 self-selection web surveys may suffer from more problems: such as "people who participate more than once, respondents from outside the target population, and groups of people who together attempt to manipulate the outcomes of the survey actuating fraudulent or inattentive behavior". The risk of manipulation from professional survey takers that take part in many different web-panels is real. According to Couper 2013 "there is evidence that a relatively large number of surveys are completed by a relatively small number of active panelist, many of whom belong to several panels". Self-selection also causes the exclusion of certain subgroup of people; for example it is expected that some population groups are underrepresented, owing to the differential adoption and use of technologies and Internet of people. Indeed, "the main issue with self-selection is that responders differs from non-responders, estimating the effect only from responders might confound the effect and

the choice to respond (Dalla Valle 2017)". According to Bethlehem 2010 the bias resulting from self-selection will be diminished if the relationship between participation behavior and the target variable is reduced. Another problem in web panel surveys is undercoverage, since the target population of a survey is often wider to those having access to the internet. The bias risk may be most severe "for surveys on elderly, low-educated and ethnic minority groups, since they have lower Internet coverage" as pointed out by Svensson and Sweden 2013. After this considerations is this clear that big data needs to be supplemented with survey data to cover unrepresented segment of the population.
The following paragraph 4.1 will describe the methodology used to integrate big data and probability sample. In paragraph 4.2 it will be explained how big data was obtained from AMELIA and the simulation implemented.

4.1 Methodology applied

The scenario considered for combining big data with a probability sample is the one in table 9. Sample A is the probability sample (in fact design weight $d_i = \pi_i^{-1}$ is known) and sample B is the big data not representative of the target population. In both samples target and auxiliary variables are observed but in some cases target variable in sample B is measured with error (Y^*=target variable with measurement error).

Table 9: Scenario considered for combining big data with probability sample

	d	X	Y	Y*	Representative
Sample A	✓	✓	✓		Yes
Sample B		✓	✓	✓	No

The idea is to use the information contained in the big data (about X and Y or about X and Y^*) to improve the estimator $\hat{\mu}_Y$ obtained from sample A. To exploit such information the big data sample is considered as a new population to use in the calibration process and sample A is like a second-phase sample from the big data.Yang and Ding 2019 and J.-K. Kim and Tam 2021 use the incomplete finite population in big data sample for calibration weighting. Below is presented the methodology used by J.-K. Kim and Tam 2021, an advantage of this technique is that no missing-at-random assumptions are made.
The first step is to identify the subset of units in probability sample A that also belong to the big data operating an individual-level matching (J.-K. Kim and Tam 2021 also

propose an alternative when this is not possible). To conduct the matching, a variable δ_i is created for $i \in A$ such that:

$$\begin{cases} \delta_i = 1 & i \in B \\ \delta_i = 0 & \text{otherwise} \end{cases} \tag{15}$$

The true total of Y (t_Y) in the population can be written as:

$$t_Y = t_b + t_c = \sum_{i=1}^{N} \delta_i y_i + \sum_{i=1}^{N} (1 - \delta_i) y_i \tag{16}$$

Explained in words, the target population N is stratified in big data stratum (of size $n_B = \sum_{i=1}^{N} \delta_i$) and missing data stratum (of size $N - n_B$). From the big data stratum t_b is estimated while t_c can be estimated using the probability sample A. To estimate t_Y correctly, a regression data integration (RDI) estimator is proposed, such estimator is unbiased even in the case where Y is measured with error in sample B. Since in the N population all elements contained in the big data are associated with $\delta_i = 1$ (remember $n_B = \sum_{i=1}^{N} \delta_i$) and all elements not contained in the big sample are associated with $\delta_i = 0$, the idea behind RDI is using $\delta_i x_i$ for $i \in A$ as auxiliary variable and calibrating it to $\sum_{i=1}^{n_B} x_i$ (which act as a known total observed in the new population n_B). Therefore in calibration estimation $(1 - \delta_i, \delta_i, \delta_i x_i, \delta_i y_i)$ it used and known totals are $(\sum_{i=1}^{N}(1 - \delta_i) = N - n_b, \sum_{i=1}^{N} n_b, \sum_{i=1}^{n_B} x_i, \sum_{i=1}^{n_B} y_i = t_b)$. If for an auxiliary variable X the total is known in the population N and not just in the big data $(1 - \delta_i, \delta_i, x_i, \delta_i y_i)$ is used in the calibration and $(\sum_{i=1}^{N}(1 - \delta_i) = N - n_b, \sum_{i=1}^{N} n_b, \sum_{i=1}^{N} x_i = t_X, \sum_{i=1}^{n_B} y_i = t_b)$ are the known totals.

It can happen that in big data sample variable Y is measured with error. For example, thinking about web panel surveys, it may be the case that in answering certain questions approximate answers are given, or that to the same question asked in two different surveys done some time apart the same answer is given because of laziness (also because often the same people participate in different web panel surveys). It is intuitive to believe that the approximate answers concern a continuous variable such as income rather than a categorical variable in which detection usually takes place by putting a cross on a box. When in the big data sample we observe Y^* in the calibration it is used $(1 - \delta_i, \delta_i, x_i, \delta_i y_i^*)$ and $(\sum_{i=1}^{N}(1 - \delta_i) = N - n_b, \sum_{i=1}^{N} n_b, \sum_{i=1}^{N} x_i = t_X, \sum_{i=1}^{n_B} y_i^* = t_b^*)$ are the known totals. The assumption is that y_i^* can be obtained from sample A when $\delta_i = 1$ (this does not mean that target variable Y is measured in sample A with error).

J.-K. Kim and Tam 2021 offer also an alternative definition of δ_i when duplicates are observed in the big data sample. Duplicate measurements are a real big data problem.

Thinking of the case of the web panel survey, in which participation is on a voluntary basis, it is possible for a person to access the panel with two different email addresses and complete the questionnaire twice. This phenomenon is certainly connected to the the existence of professional survey takers interested in manipulating the result of web surveys. When duplicates are detected in big data, δ_i for $i \in A$ is defined as the number of times that the unit i appears in sample B. Also in this case $(1-\delta_i, \delta_i, x_i, \delta_i y_i)$ is used in the calibration, and $(\sum_{i=1}^{N}(1-\delta_i) = N - n_b, \sum_{i=1}^{N} n_b, \sum_{i=1}^{N} x_i = t_X, \sum_{i=1}^{n_B} y_i^* = t_b^*)$ are the known totals.

4.2 Simulation

The methodology presented in section 4.1 is applied here. The target population N is all the people in the AMELIA dataset. Sample A is SRSWOR drawn from the target population of size $n_A = 1500$. In sample A auxiliary variable X and target variable Y "personal income" measured always without error are observed. Regarding big data, different types of sample B are obtained from AMELIA dataset:

1. big data 1: all people with age under or equal to 40 years old and income above the first quantile of personal income. There is therefore a relationship between participation in the survey and the target variable personal income. The set of people represented has a sufficient income to guarantee access to the internet connection via any tool (computer or smartphone) and an age such that they can use this connection to complete web panel surveys, and therefore are aware of the their very existence. People are selected directly as a subset of the population N, in this way no sampling scheme is applied and a non-probability sample is obtained.

2. big data 2: big data 1 plus all people between 40 and 60 years old with income above the median. The subset of the population N for which people are expected to participate in the web panel survey is enlarged here. People older than 40 years (up to 60) may be aware of these surveys and use the internet to participate in them. In this case, the income threshold considered sufficient to access the subset is equal to the median of personal income. It is assumed that the older you are, the less you are familiar with technological tools; therefore if a person between 40 and 60 has the time to learn how to use the internet for purposes other than the most trivial ones and to learn about the existence of things like web panel surveys, it is assumed that his income allows him to acquire the technological tools and to have time to spend on the web.

The fact that big data 2 is bigger and contains more information than big data 1 does not mean that it is more representative, these are two non-probability samples and none is representative of the target population.

3. big data 3: big data 1 with duplicates. In particular;

   - 20% of rows of big data 1 without duplicates in big data 3
   - 40% of rows of big data 1 with one duplicate in big data 3
   - 20% of rows in big data 1 with two duplicates in big data 3
   - 15% of rows in big data 1 with three duplicates in big data 3
   - 5% of rows in big data 1 with four duplicates in big data 3

   Every time the rows with duplicates are randomly selected from big data 1. This means, for example, that every time the 20% of rows without duplicates change and never stay the same (the same goes for the other percentages). Since 1000 simulations are implemented, there will be 1000 different versions of big data 3, each time the duplicates rows, and the number of times they are duplicate, change.

4. big data 3: big data 2 with duplicates. In particular;

   - 20% of rows of big data 2 without duplicates in big data 4
   - 40% of rows of big data 2 with one duplicate in big data 4
   - 20% of rows in big data 2 with two duplicates in big data 4
   - 15% of rows in big data 2 with three duplicates in big data 4
   - 5% of rows in big data 2 with four duplicates in big data 4

   Again each time rows with duplicates are randomly selected from big data 2.

5. big data 5: big data 1 with measurement errors in personal income. It is assumed that 80% of personal income measurements are measured with error. In one case we assume a measurement error downwards

$$y_i^* = 0.7 \cdot y_i + e_i \tag{17}$$

   and in one case upwards

$$y_i^* = 1.3 \cdot y_i + e_i. \tag{18}$$

   Each time the rows containing Y^* are randomly extracted from big data 1. Therefore 2000 different versions of big data 5 are obtained; 1000 containing 80% of

observations (different each time) of personal income measured with error according to the equation 17 and 1000 containing each time 80% of observations (different each time) of personal income measured with error according to the equation 18. In equations 17 and 18 $e_i \sim \mathcal{N}(0, 0.5)$.

6. big data 6: big data 2 with measurement errors in personal income. Again, 80% of personal income measurements are measured with error. In one case we assume a measurement error downwards like in equation 17 and in one case upwards like in equation 18. Each time the rows containing Y^* are randomly extracted from big data 2. Therefore 2000 different versions of big data 6 are obtained.

A schematization of the variables used for each type of combination between probability sample A and big data is provided in the table 10.

big data 1 and 2	
Auxiliary variables	Population total is known
Sex	Yes
Region	Yes

big data 3 and 4 (duplicates)			
Case 1		Case 2	
Auxiliary variables	Population total is known	Auxiliary variables	Population total is known
Sex	Yes	Sex	Yes
Region	Yes	Region	Yes
Household size	No	Marital status	No

big data 5 and 6 (under and over measurement error)	
Auxiliary variables	Population total is known
Sex	Yes
Region	Yes

Table 10: Cases analyzed when combining probability sample and big data

For big data 1, big data 2, big data 5 and big data 6 we observe sex and region as auxiliary variables, and their totals are known in the population. In big data 3 and big data 4 there is an extra auxiliary variable, in one case household size and in another case marital status. Both household size and marital status have unknown totals, while sex and region still have known totals in the population.
Sex (SEX) is a categorical variable, assumes value 1 if the individual is male and 2 if the individual is female. Region (REG) is also categorical and indicates the regional identifier, the region in AMELIA population are four. Household size (HHS) has been

categorized in 4 levels according to the number of people living in the household; level 1 = 1 person, level 2 = 2 people, level 3 = 3 or 4 people, level 4 = from 5 to 16 people (considering all AMELIA population household size range from 1 to 16). Marital status (MST) has three levels; 1 = never married, 2 = married, 3 = widowed, separated or divorced. The reason for which household size and marital status are considered as auxiliary variables alternatively but not together, is to avoid risk of having too strongly correlated variables. Such risk is quite high when considering big data as a new population. If population totals are known adding auxiliary variables is not a problem, if they are unknown and we calibrate to totals in big data is more probable to have strongly correlated variables and the impossibility to apply function calibrate of R survey package (Lumley 2020).

The simulation is implemented as follows:

1. 1000 samples A are randomly drawn from the population with SRSWOR and the different types of big data are created. The seed is set to i and goes from 1 to 1000.

2. δ_i is created for sample A. In order to perform calibration, interactions between auxiliary variables and target variable with δ_i are constructed for each case in table 10.
   For big data 1, 2, 5 and 6 δ_i for $i \in A$ can assume value 1, if $i \in B$ or 0 otherwise. Interactions are constructed by multiplying $\delta_i \cdot x_i$ for $i \in A$ or $\delta_i \cdot y_i$. For big data 3 and 4 δ_i for $i \in A$ can assume value 0 if $i \notin B$, 1 if $i \in B$ 1 time, 2 if $i \in B$ 2 times, 3 if $i \in B$ 3 times and 4 if $i \in B$ 4 times. To calculate the interactions with the categorical auxiliary variables we construct as many interaction variables as the number of levels of X. For example consider the auxiliary variable marital status; marital status has three levels (1 = never married, 2 = married, 3 = widowed, separated or divorced) and so three interactions variables with δ_i are constructed. The first one is:
   $$\begin{cases} \delta_i \cdot x_i & x_i = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$
   the second one is
   $$\begin{cases} \delta_i \cdot x_i & x_i = 2 \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$
   and the third one is
   $$\begin{cases} \delta_i \cdot x_i & x_i = 3 \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

This way of proceeding is necessary to avoid miss-classification. Indeed, if one had simply multiply δ_i and x_i value of $\delta_i \cdot x_i = 6$ would have enclose cases in which $\delta_i = 2$ and $x_i = 3$ and cases in which $\delta_i = 3$ and $x_i = 2$.
For the continuous target variable personal income is sufficient to simply multiply $\delta_i \cdot y_i$.

3. calibration is applied using R package *survey* (Lumley 2020) and estimators for mean personal income are computed.

4.3 Results

In this section results obtained from the simulations will be analyzed. Figure 4 shows a comparison of the distributions of the personal income mean estimator in the case of big data 1 and big data 2 as new finite population. Blue and orange lines represent mean of personal income estimates when the new acting population is big data 1 and big data 2 respectively. Red line is the true value of income observed in the population N. The violin plots of big data 1 and big data 2 are far apart in figure 4. The estimates

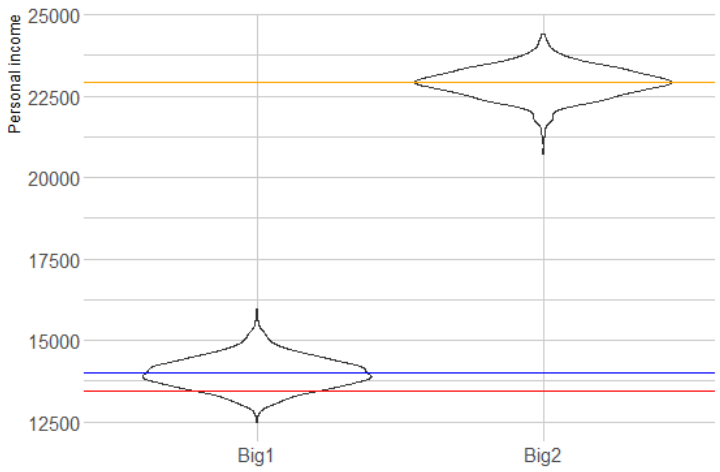


Figure 4: Violin plot of estimates of personal income for big data 1 and big data 2. Blue line is the mean of personal income estimates when big data 1 is acting as a new population. Orange line is the mean of personal income estimates when big data 2 is acting as a new population. Red line is the true value of mean personal income in the population N. The auxiliary variables used are sex and region, and the totals are known. The target variable is personal income and is observed in big data 1 and big data 2.

for big data 2 cluster around the value 22915 while the estimates for big data 1 cluster around the value 13994. Since the red line is the true value of personal income it

is possible to say that in the case of big data 1 the variables used are able to provide estimates fairly close to the true value of personal income while for big data 2 they are not (the estimates obtained are very far from the value observed in the population). This fact is interesting because the variables used to explain personal income are the same in the two cases. This means that when we expand the subset of people we imagine participating in the web panel survey to people aged 40-60 with income above the median, the situation becomes far more complicated. When we add this population segment, the variables sex and region are no longer good regressors for personal income. The estimates are not optimal in the case of big data 1 either, but if nothing else at least here the violin graph intersects the true value in the population.
When there are measurement errors the situation gets even worse (see figure A2 in the appendix). None of the violin plots of big data 5 and big data 6, either in the case of downward measurement error or upward measurement error, intersects the true value of personal income at its widest part. For big data 6 in the case of under measurement errors the violin graph is closer to the value in the population than in figure 4. This is not good news but the sum of two bad scenarios: auxiliary variables unable to explain the target variable and measurement errors.
Before illustrating how to proceed, the results of the simulation concerning big data 3 are shown in figure 5. The differences between big data 1 and big data 3 concern

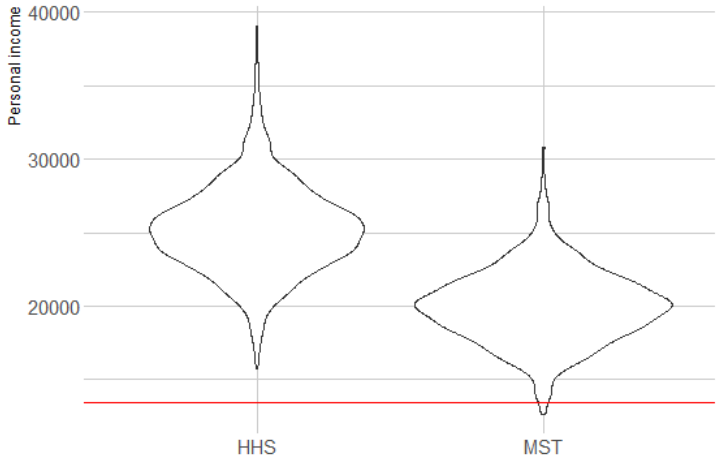


Figure 5: Violin plot of estimates of personal income for big data 3 using household size (left) and marital status (right). Red line is the true value of mean personal income in the population N. The others auxiliary variables used are sex and region, their totals are known. The target variable is personal income and is observed in big data 3 and big data 4.

the inclusion of duplicates and the use of two alternately observed auxiliary variables,

marital status and household size, observed in big data (which were not included in the previous analysis or the system would have been computationally singular), in addition to sex and region. In the case where household size is used as an additional auxiliary variable the violin plot does not intersect the true value observed in the population. In the case where marital status is used the intersection occurs at the lower end of the violin graph. This means that in the 1000 simulations carried out values very close to or equal to the average personal income in the population are obtained in extreme cases, that is, they occur very few times and are outliers in the distribution. Compared with figure 4, looking at the violin graph referring to big data 1, the ability to estimate the true value of the average personal income is much worse in figure 5. To understand whether this result can be attributed to (i) the addition of household size or marital status as auxiliary variable, (ii) the presence of duplicates in big data 3 or (iii) both of the previous, table 11 is created.

The table represents the distributions of the different levels of sex, region, marital

	REG = 1	**REG = 2**	**REG = 3**	**REG = 4**
Population N	24.6	26.7	22.3	26.4
Big data 1	23.5	25.8	22.7	28.0
Big data 2	24.4	26.8	22.8	26.0
	HHS = 1	**HHS = 2**	**HHS = 3**	**HHS = 4**
Population N	8.6	23.5	48.2	19.7
Big data 1	8.3	23.1	48.6	20.0
Big data 2	8.5	23.4	48.8	19.7
	MST=1	**MST=2**	**MST=3**	
Population N	41.4	47.0	11.6	
Big data 1	46.9	43.8	9.3	
Big data 2	37.7	53.4	11.9	
	SEX=1		**SEX=2**	
Population N	48.5		51.5	
Big data 1	50.2		49.8	
Big data 2	49.4		50.6	

Table 11: Percentage for different levels of region (REG), household size (HHS), sex (SEX) and marital status (MST) in target population N, big data 1 and big data 2

status and household size in the target population N, in big data 1 and big data 2. The distribution of these variables is the same in every population considered. Table 11 shows that there is no difference between observing one of the variables table in the big data or in the population. Taking sex as example, whether $\delta_i x_{i=1}$ for $i \in A$ is calibrated to the total $\sum_{i=1}^{n_B} x_{i=1}$ or whether $x_{i=1}$ for $i \in A$ is calibrated to the total $\sum_{i=1}^{N} x_{i=1}$ males always represent about 50% of the population. The sex variable, like the others in the table, is distributed approximately in the same way in all population. However

by looking at table 12 it is possible to see the average of personal income in big data 1 and big data 2 differs from that of population N. The chosen auxiliary variables reflect

	Population N	**Big data 1**	**Big data 2**
Mean of personal income	13435	18551	21669

Table 12: Mean of personal income in the population N, big data 1 and big data 2

a personal income distribution that remains the same in the pairs (i) target population N and big data 1 and (ii) target population N and big data 2. The estimates obtained are unsatisfactory for big data 1 (for big data 3 adding duplicates makes the situation much worse) and disastrous for big data 2. This is because there is much more difference between the distribution of the population subset considered for big data 2 and N than the distribution of the population subset considered for big data 1 and N. Anyhow, in both cases there is a need for variables that are able to capture the difference between N and the segments of the population once considered. This does not mean that sex, region, household size or marital status are not correlated with the target variable, but that they are correlated with personal income in the same way in big data stratum and missing data stratum. If we manage to find auxiliary variables capable of

	RES=1	**RES=2**	
Population N	98.3	1.7	
Big data 1	96.7	3.3	
Big data 2	97.6	2.4	
	SEM=1	**SEM=2**	
Population N	6.7	93.7	
Big data 1	8.3	91.8	
Big data 2	8.6	91.4	
	PWHI=1	**PWHI=2**	
Population N	37.7	62.2	
Big data 1	46.6	53.4	
Big data 2	51.5	48.5	
	BAS=1	**BAS=2**	**BAS=3**
Population N	42.3	22.4	35.3
Big data 1	63.8	8.2	28.0
Big data 2	65.1	12.7	22.2

Table 13: Percentage for different levels of residential status (RES), self-employed (SEM), person with highest income in the household (PWHI) and basic activity status (BAS) in target population N, big data 1 and big data 2

bringing to light the different characteristics of the two groups, big data stratum and missing data stratum, we could explain the difference in incomes and correct our estimate. The question arises spontaneously: do these variables exist or do the two groups

have identical characteristics except for personal income and age?
To investigate which variables could be useful for our purpose, the distribution of other categorical variables in population N and in the new big data 1 and big data 2 populations has been analyzed. The results of the analysis are in table 13. The variables analyzed are four: residential status, self-employed, person with highest income in the household and basic activity status. The results suggest using PWHI and BAS. There is a difference of almost 10 percentage points between the number of main income earners in the household in population N and in big data 1 (37.7 percent versus 46.7 percent), and a difference of more than 10 percentage points with big data 2 (37.7 percent versus 51.5 percent). Obviously such analyzes can only be done when one has the entire population available, which in reality does not happen. Where data is not available it is necessary to proceed with the logic. For example, it is quite intuitive to think that variables such as gender and region of residence are equally distributed among individuals who have different ages and income levels above a certain threshold.
The graphs in figure 6 show the distribution of estimates of mean personal income in the case of big data 1 and big data 2 when PWHI and BAS are entered as auxiliary variables with known totals in the population instead of region and gender. The problem

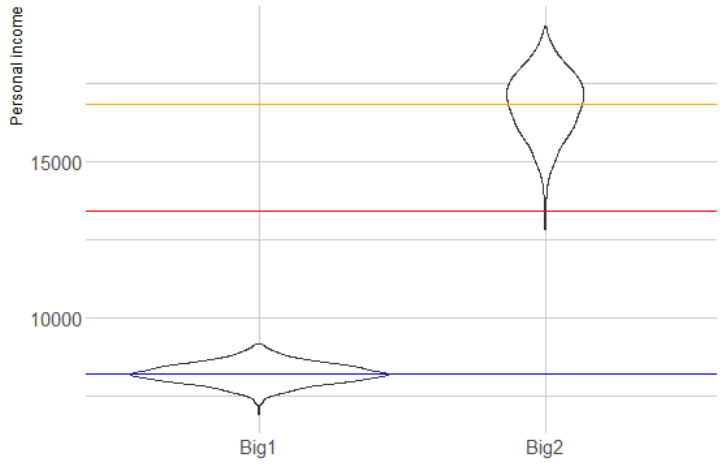


Figure 6: Violin plot of estimates of personal income for big data 1 and big data 2 using PWHI and BAS as auxiliary variables. Blue line is the mean of personal income estimates when big data 1 is acting as a new population. Orange line is the mean of personal income estimates when big data 2 is acting as a new population. Red line is the true value of mean personal income observed in the population N. The others auxiliary variables used are PWHI and BAS, their totals are known. The target variable is personal income and is observed in big data 1 and big data 2.

has not been solved. For big data 1 the situation has drastically worsened compared to figure 4. The violin graph is concentrated around the value 8216 and in none of the simulations we obtain values equal to or close to the parameter observed in the population (red line). For big data 2 the situation is certainly better than figure 4 but far from obtaining satisfactory results. The values of the distribution of the estimates are concentrated around the value 16817 and the true value of personal income in the population is an outlier. In the light of the results in figure 6, two things are evident:

- is not the solution to choose only variables that have different distribution between big data stratum and population N
- to explain personal income it is necessary to use different auxiliary variables for big data 1 and big data 2 (the distribution of personal income is very different in the two cases)

To try to understand which variables could help to obtain more precise estimates of income, several simulations were carried out trying different combinations of **X**. In doing this it is necessary to remember that when a total is known in the population, the GREG estimator could very well be applied using the sample A. Therefore, if in choosing auxiliary variables X it is assumed that their total is known, the precision of the GREG increases and the estimate of personal income obtained by integrating the probability sample and big data will have to be even more precise to be better than the one obtained using only the probability sample A. When there are no known totals in the population, the estimate obtained by integrating the two surveys must instead have a better performance than the π estimator that would be obtained from sample A. It is important to remember that this method of integration aims to improve the estimate obtained in sample A using the information contained in big data. In our case, the information provided by big data is the personal income of those who complete the web survey. One might consider including as few auxiliary variables as possible with known totals in the population to avoid simultaneously improving the accuracy of the GREG. Several objections can be raised to this: (i) the precision of the GREG estimator increases a lot when the first known totals are entered in the calibration and then marginally much less, therefore it is enough to have sex and region (whose totals are generally known in the population , think of censuses) to greatly improve the estimate (ii) inserting more than two variables to be calibrated to the totals observed in big data 1 or big data 2 makes it impossible to carry out the calibration itself because the variables are too correlated and the system is without solution (iii) even assuming to insert many auxiliary variables with known totals and compare the estimate obtained with the π estimator (rather than with the GREG estimator) this is still more efficient.

Several simulations have been conducted using different variables **X** for big data 1 and big data 2. Sometimes the calibration includes the interaction between δ_i and y_i=personal income, other times this is excluded. In the simulations, in addition to the already known variables, two new ones are inserted; PY010, the employee cash or near-cash income and PY020, the non-cash employee income. These variables are useful in explaining the distribution of income in the two strata of the population (without them the situation is worse) and can be inserted into the calibration as an interaction with δ_i without preventing the calibrate function from operating. It is now necessary to open a small parenthesis regarding the interaction between PY010 or PY020 and δ_i. Consider the non-cash employee income; first the value -1 (-1 is a value that no one in AMELIA population has for PY020) is assigned to $y_i = 0$, then the interaction with δ_i is constructed, and finally the value 0 is reassigned when $\delta_i \cdot y_i = 1$. This step is necessary because otherwise observations where $y_i = 0$ would generate zero value even if $\delta_i = 1$. In other words, without assigning the value -1 before performing the interaction, the case $\delta_i \cdot y_i = 0$ would not include only the cases in which $\delta_i = 0$, but also those in which $y_i = 0$, generating a miss-classification. This step is unnecessary when y_i=personal income because big data 1 and big data 2 include people with income levels above a certain threshold. So when $\delta_i = 1$ personal income is never zero. This prevented miss classification.
The conclusions to which the several simulations conducted led are the following:

1. including or not personal income in the calibration has no visible effect on the performance of the estimator;

2. for big data 1 and big data 2 some variables have been identified which are able to explain part of the different distribution of income in the two strata. These variables are PY010 and PY020 for big data 1, and PY020 for big data 2;

3. after the variables indicated in step two are entered into the calibration, the addition of other variables can slightly improve the situation or have drastic effects. However, not only is the relationship of the chosen variables with personal income important, but also the relationship between the auxiliary variables themselves. In other words, from table 12 we see that big data 1 overestimates the level of personal income compared to population N. The distribution of PY010 and PY020 manages to capture this phenomenon and bring the value observed in big data 1 closer to the real one. By entering PWHI the same phenomenon occurs and the final estimate of the income is lower than the real value (while actually it is overestimated). A similar effect is obtained when PY010 and PY020 are used simultaneously in big data 2.

In the light of the simulations conducted, the best situations resulted are; for big data

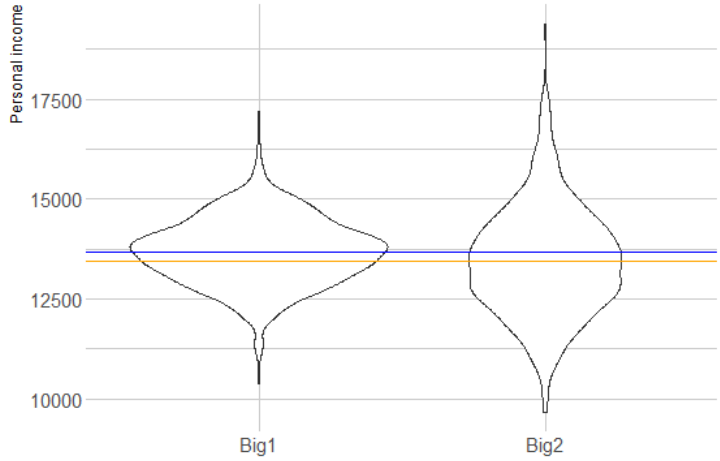


Figure 7: Violin plot of estimates of personal income using as auxiliary variables PY010, PY020, SEX and REG for big data 1 and PY020 and BAS as auxiliary variables for big data 2. Blue line is the mean of personal income estimates when big data 1 is acting as a new population. Orange line is the mean of personal income estimates when big data 2 is acting as a new population. Red line is the true value of mean personal income observed in the population N.

1 use the employee cash income (PY010), the non-cash employee income (PY020), sex (SEX) and region (REG); for big data 2 use non-cash employee income (PY020) and basic activity status (BAS). Employee cash income and non-cash employee income are calibrated to the total of big data 1 or big data 2, for this reason it was necessary to build the interaction with δ_i as described above. Instead, basic activity status and person with highest income in the household have known totals in the population.
Figure 7 shows the results obtained when these auxiliary variables are used.The situation is much better than figures 4 and 6. The red line (not very visible because covered by the orange one, which represents the average of the estimates obtained when big data 2 and the new acting population) represents the true value of personal income in the population. Both in the case of big data 1 and in the case of big data 2, the true value of personal income passes through the center of the violin graphs. This means that in the simulations carried out, the estimates obtained are very likely equal to or very close to the parameter observed in the population. The situation continues to be better for big data 1 because the violin is less elongated and the values are more concentrated around the red line, while for big data the shape is more elongated and values far from the red line are more frequent. Success is not repeated when considering the case with duplicates (figure 8). When duplicates are added the situation becomes similar to figure 4 again. For big data 1 compared to figure 4 the red and blue

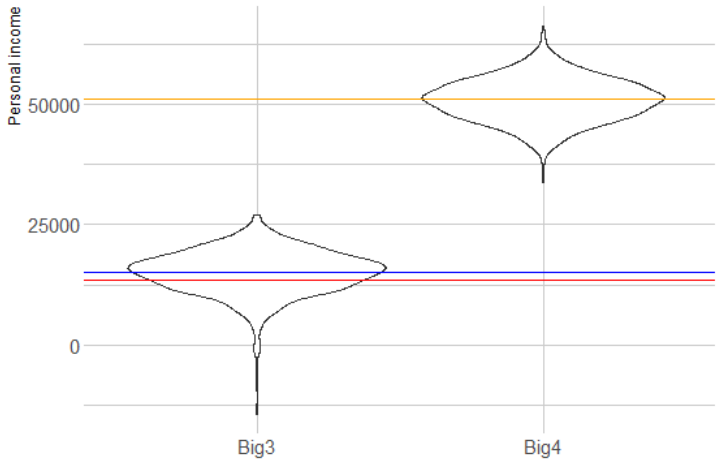


Figure 8: Violin plot of estimates of personal income using as auxiliary variables PY010, PY020, SEX and REG for big data 1 and PY010 and BAS as auxiliary variables for big data 2. Blue line is the mean of personal income estimates when big data 3 is acting as a new population. Orange line is the mean of personal income estimates when big data 4 is acting as a new population. Red line is the true value of mean personal income observed in the population N.

line are closer, it means that the mean of the estimates obtained when big data 1 is the new population is very close to the true value of personal income in the population. This is a positive note, on the other hand the fact that the lower part of the violin is very elongated means that during the simulations, if sample A is particularly bad, extreme values are obtained that are very far from the parameter in the population. For big data 2 the situation becomes disastrous again, a situation even worse than the one in figure 4 since the distribution is concentrated around the value 50912 (orange line). In this case, however, we know that the results are to be attributed to the presence of duplicates and not to the choice of auxiliary variables which are suitable to explain the target variable (see figure 7). The big data 2 violin plot in figure 8 suggests that the methodology used is not suitable in case of duplicates, while the big data 1 violin plot suggests the opposite. This difference is probably due to the fact that in big data 2 the duplicates present are in absolute value greater than those present in big data 1 (trivially because in big data 2 a larger segment of the population is considered).
Figure 9 shows results when measurement errors for PY010 and PY020 are present. PY020 is the variable measured with error in big data 5 and big data 6. The situation is certainly better than in figure A2 but the results are far from satisfactory. While in the case of duplicates the results were good for big data 3 and disastrous for big data 4, in this case the results are mediocre in all four cases. The distributions of the estimates for big data 5 are compact and not stretched but the personal income line

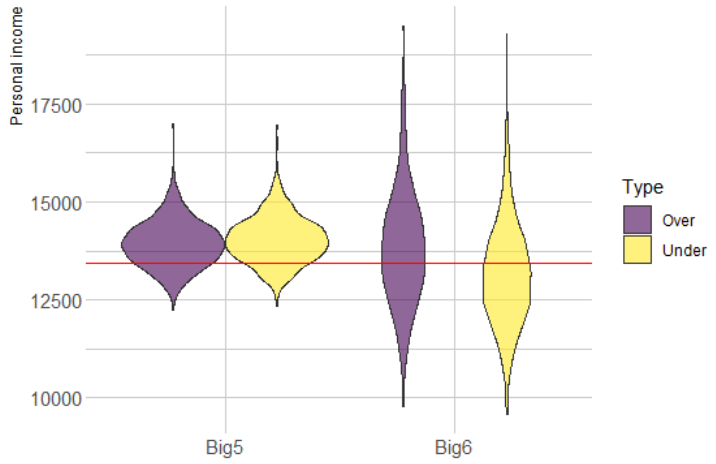


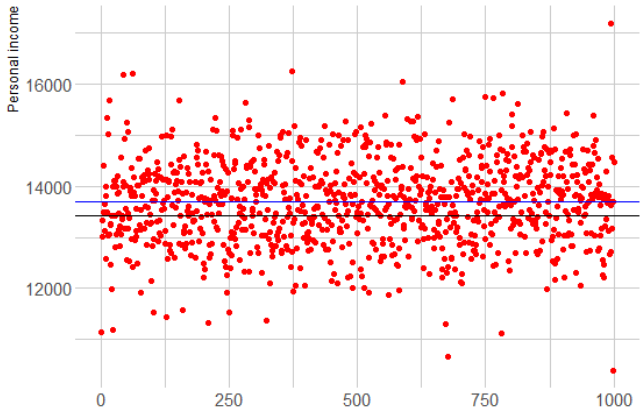
Figure 9: Violin plot of estimates of personal income using as auxiliary variables PY010, PY020, SEX and REG for big data 5 and PY010 and BAS as auxiliary variables for big data 6. PY020 is the variable measured with error in big data 5 and big data 6. Red line is the true value of mean personal income observed in the population N.

in the population does not cross them centrally, but is slightly shifted downwards. In the distributions of estimates for big data 6 the red line passes through the center but the distributions are much more elongated and consequently it is common to obtain estimates far from the true parameter in the population. In the case of big data 5 we note that both in the case of upward measurement errors and downward measurement errors, personal income is on average overestimated compared to the value observed in the population. Compared to figure A2 there are no differences between the distributions of over measurement errors and under measurement errors within big data 5 and big data 6. This is probably a consequence of the choice of better auxiliary variables.

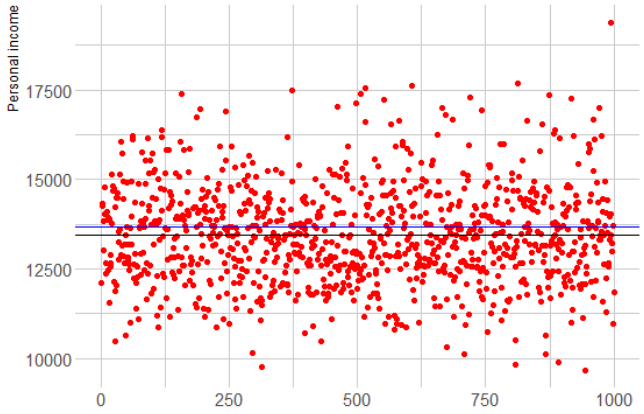
4.4 Monte Carlo effect and matching effect

In light of what emerged in Section 4.2 a new case study is considered for big data. Instead of selecting samples from N, the entire population is considered, so the sampling fraction $\frac{n}{N}$ is equal to 1. In this way it is possible to distinguish the matching effect, that is, the effect related to the accuracy of the method used, and the Monte Carlo effect, that is, the effect due to the variability of different samples A. The variables considered are employee cash income, non-cash employee income, sex and region for big data 1, 3 and 5 and non-cash employee income and basic activity status for big data 2, 4 and 6. When a sample containing 100% of the observations is considered, the variability due to sample A is eliminated; therefore there is no need to worry about whether A is a

good or bad sample. The only source of variation that remains is that due to the matching effect. If the estimates obtained for each type of big data B considered are close to the true value of the mean value of personal income in the population, it means that the integration method is suitable for estimating the parameter of interest and any departures from that value are to be attributed to the particular sample A (i.e., you were particularly unlucky) but not to the method applied, which remains suitable.
Figure 10 reports this analysis for big data 1 and big data 2. The black line represents



(a)

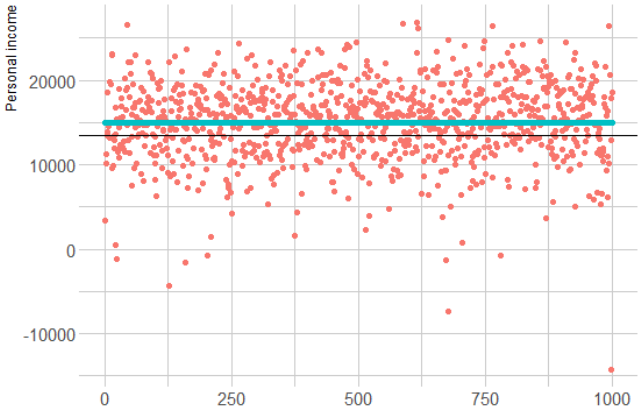


(b)

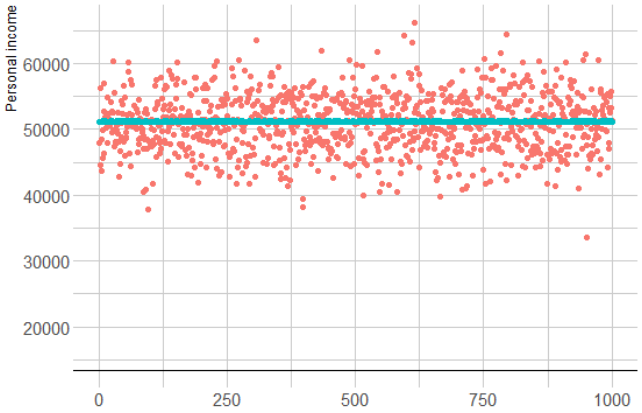
Figure 10: Results with a sample containing all observations for big data 1 (up) and big data 2 (down)

the true value of the parameter of interest in the population, the blue line is the estimate obtained when 100% of the observations are considered (the estimate does not change for the 1000 different types of sample A considered because big data 1 and big

data 2 do not change), and the red dots (they are 1000) report the estimates obtained for each sample A when the seed is equal to the number of the current simulation (1 to 1000).
For big data 3, 4, 5 and 6, there is no longer a blue line but 1000 light blue dots representing the results obtained when the Monte Carlo effect is equal to 0. In these cases there are 1000 points and not one line because big data 3,4,5 and 6 change every time sample A changes. Seeds are kept the same as in section 4.2. Figure 11 shows the results for big data 3 and 4. The appendix shows the results for big data 5 and big



(a)



(b)

Figure 11: Results with a sample containing all observations for big data 3 (up) and big data 4 (down)

data 6 (figure A3). All the graphs show that the variability introduced by the method used (matching effect) is almost null. The light blue dots are distributed along a line

that roughly corresponds to the mean of personal income estimates when that particular type of big data is considered to be the acting population. The values obtained when using 1000 different A samples are distributed around the line of light blue dots. What seems to be important is how to condition the position of the blue line of points, i.e. through the choice of good auxiliary variables. However, sometimes these variables are not sufficient to reach good estimates if there are duplicates or measurement errors. And this is not because there is much variability due to the matching effect, rather because the method used is not robust in the presence of too many duplicates or would require variables even more strongly correlated with personal income (like in the case of big data 2).

5 Summary

This thesis aimed to investigate methods for integrating different data sources. Specifically, two methodologies for integrating probabilistic samples with each other and one methodology for combining big data with a probabilistic sample were investigated. The first methodology followed a macro approach, where the goal is to combine multiple data sources to have more efficient estimates of the parameter of interest. The methodology applied here has an additional and not secondary purpose to that of obtaining more precise estimates of the variable of interest, that is to create consistency between the two samples with respect to the common variables. Regarding the first purpose, the gains are present only if the common variables are highly correlated with both target variables. This generates the problem of finding common variables able to explain two different phenomena. In this application the common variables were useful in explaining personal income, but not as much in explaining equivalised disposable income. However, although the gain in terms of parameter of interest estimation and variance were minimal, in the case of equivalised disposable income was still possible to obtain information on the unknown total of some variables using information from two different surveys. If our purpose had been only to estimate personal income and reduce the variance of the GREG estimator, we could still have used information from other surveys to estimate unknown totals in the population and obtained satisfactory results. What could not be shown is a significant difference in the use of optimal or proportional choice when the adjusted GREG is calculated. Nor does a significant difference emerge by greatly increasing the control variables in one sample compared to the other. However, the fact that a very small difference emerges suggests that if there were even more different conditions in sample selection (perhaps even with respect to sampling design) the optimal choice would have been able to capture them.

In contrast, the other technique for integrating probabilistic samples followed a micro approach. That is, the primary purpose is to create a synthetic dataset that summarizes information from different surveys into a single dataset. Although the primary purpose in this approach is not to obtain an estimator, it is evident that the synthetic dataset created is well suited for this. Regarding the accuracy of the model that is used to predict personal income values in the sample where they are missing, the more auxiliary variables are used, the better. Whether the totals of auxiliary variables are known is an additional positive factor, but less important than the number of variables themselves. For example, better six variables with unknown totals than five with a known total. Instead, evaluating the variance of the projection estimator, and no longer the

accuracy of the model, shows that it has substantial reductions when the weights are calibrated based on known totals. This method is effective and useful for avoiding measuring variables whose detection is costly in large samples.
Looking at the integration between probability samples and big data, the situation certainly becomes more complicated. The methodology involves using the information contained in the big data to improve the performance of the estimator obtained from sample A. To do this, the big data is considered as a new population in the calibration process. When big data is obtained from the AMELIA dataset, this is done by selecting age classes and income thresholds. The choice of ages and income levels was made considering that the big data used in this work were derived from a web panel survey, therefore, a relationship was established between the variable of interest and survey participation (as happens in web panel surveys). This relationship adds a lot of complication, finding variables that can explain the different income distribution in the two strata of the population becomes very difficult. Not only because none alone can explain the income distribution in the big data, but also because the positive contribution of two variables can become negative if both are included. Therefore, in order to apply this technique, it is necessary to have information about the type of variables that can be used in the calibration. This procedure becomes much easier when the variable of interest is not related to web survey participation. If for example big data had been created as a subset of the population including all females and no males, variables such as age, self-employed, basic activity status would have been able to describe the income distribution in big data. Eventually, even using an uncomfortable method of obtaining big data such as the one described here, it was possible to find variables that returned satisfactory estimates of personal income. This cannot be considered a complete victory since the π estimator still remain better, and thus it would not make sense to resort to integration techniques. Moreover, auxiliary variables that perform well in the simple case of big data and quite well in the case of measurements error, still fail in the case where duplicates are present. The proposed technique for dealing with duplicates does not seem adequate and it is advisable to remove them if they are detected. Part of the failure of this method can be attributed to the size of our big data, far from those needed to be defined as such. Perhaps with larger size it would be possible to include more variables in the calibration without running the risk of too much correlation. On the other hand, the proposed methodology is also applicable for non-probability samples. Therefore, even assuming that sample B obtained from the AMELIA dataset cannot be called big data, and that the number of observations in that sample is not comparable to the far greater number of observations that big data has, the applied method should still work since sample B, if not big data, is definitely

a non-probability sample.
By virtue of this last observation, we can conclude that although the analyses are based on synthetic data and simulations, the results provide useful insights into the integration of multiple data sources; specifically the integration of two probabilistic samples and the integration of a probabilistic sample and a non-probabilistic sample. Further developments could concern the use of different sampling schemes, available in the AMELIA dataset, to see if there is a change in the estimates obtained.

References

Bethlehem, Jelke (2010). "Selection bias in web surveys". In: *International statistical review* 78.2, pp. 161–188.

— (2016). "Solving the nonresponse problem with sample matching?" In: *Social Science Computer Review* 34.1, pp. 59–77.

Burgard, Jan Pablo, Florian Ertz, et al. (2020). *AMELIA—Data description v0. 2.3. 1*.

Burgard, Jan Pablo, Jan-Philipp Kolb, et al. (2017). "Synthetic data for open and reproducible methodological research in social sciences and official statistics". In: *AStA Wirtschafts-und Sozialstatistisches Archiv* 11.3, pp. 233–244.

Couper, Mick P (2013). "Is the sky falling? New technology, changing media, and the future of surveys". In: *Survey Research Methods*. Vol. 7. 3, pp. 145–156.

Dalla Valle, Luciana (2017). "Data integration". In: *Wiley StatsRef: statistics reference online*. John Wiley & Sons.

Elliott, Michael R, Trivellore E Raghunathan, and Nathaniel Schenker (2018). "Combining estimates from multiple surveys". In: *Wiley StatsRef: Statistics Reference Online*, pp. 1–10.

Garnier, Simon et al. (2021). "Rvision-Colorblind-Friendly Color Maps for R". In: *R package version 0.6* 1, p. 2021.

Groves, Robert M (2011). "Three eras of survey research". In: *Public opinion quarterly* 75.5, pp. 861–871.

Hedlin, Dan et al. (2001). "Does the model matter for GREG estimation? A business survey example". In.

Hidiroglou, MA (2001). "Double sampling". In: *Survey methodology* 27.2, pp. 143–154.

Joshi, M and J Pustejovsky (2020). "Simhelpers: Helper functions for simulation studies". In: *URL: https://CRAN. R-project. org/package= simhelpers. R package version 0.1. 0*.

Kim, Jae Kwang and Jon NK Rao (2012). "Combining data from two independent surveys: a model-assisted approach". In: *Biometrika* 99.1, pp. 85–100.

Kim, Jae Kwang and Zhonglei Wang (2019). "Sampling techniques for big data analysis". In: *International Statistical Review* 87, S177–S191.

Kim, Jae-Kwang and Siu-Ming Tam (2021). "Data integration by combining big data and survey sample data for finite population inference". In: *International Statistical Review* 89.2, pp. 382–401.

Kuhn, Max et al. (2020). "caret: Classification and Regression Training. R package version 6.0-86". In: *Astrophysics Source Code Library: Cambridge, MA, USA*.

Legg, Jason C and Wayne A Fuller (2009). "Two-phase sampling". In: *Handbook of statistics*. Vol. 29. Elsevier, pp. 55–70.

Lumley, T (2020). *Package 'survey': analysis of complex survey samples, version 4.0*.

Merkouris, Takis (2004). "Combining independent regression estimators from multiple surveys". In: *Journal of the American Statistical Association* 99.468, pp. 1131–1139.

— (2010). "Combining information from multiple surveys by using regression for efficient small domain estimation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.1, pp. 27–48.

Renssen, Robbert H and Nico J Nieuwenbroek (1997). "Aligning estimates for common variables in two or more sample surveys". In: *Journal of the American Statistical Association* 92.437, pp. 368–374.

Schenker, Nathaniel and Trivellore E Raghunathan (2007). "Combining information from multiple surveys to enhance estimation of measures of health". In: *Statistics in medicine* 26.8, pp. 1802–1811.

Svensson, Jörgen (2014). "Web Panel Surveys-a challenge for official statistics". In: *Proceedings of Statistics Canada Symposium*.

Svensson, Jörgen and Statistics Sweden (2013). "Web panel surveys–can they be designed and used in a scientifically sound way?" In: *59th ISI World Statistics Congress*. Citeseer.

Tam, Siu-Ming and Frederic Clarke (2015). "Big data, official statistics and some initiatives by the Australian Bureau of Statistics". In: *International Statistical Review* 83.3, pp. 436–448.

Tam, Siu-Ming and Jae-Kwang Kim (2018). "Big Data ethics and selection-bias: An official statistician's perspective". In: *Statistical Journal of the IAOS* 34.4, pp. 577–588.

Weihs, Claus et al. (2005). "klaR analyzing German business cycles". In: *Data analysis and decision support*. Springer, pp. 335–343.

Wickham, Hadley (2016). "Data analysis". In: *ggplot2*. Springer, pp. 189–201.

WILLENBORG, Leon, Sander SCHOLTUS, and Arnout VAN DELDEN (n.d.). "Development and Structure of the Memobust Handbook". In: ().

Yang, Shu and Peng Ding (2019). "Combining multiple observational data sources to estimate causal effects". In: *Journal of the American Statistical Association*.

Yang, Shu and Jae Kwang Kim (2020). "Statistical data integration in survey sampling: A review". In: *Japanese Journal of Statistics and Data Science* 3.2, pp. 625–650.

Appendix

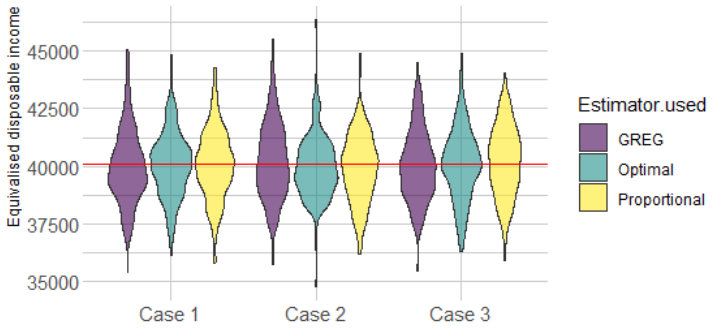


Figure A1: Violin plot of equivalised disposable income for different models in macro approach. Red line is the true value of mean personal income in the population

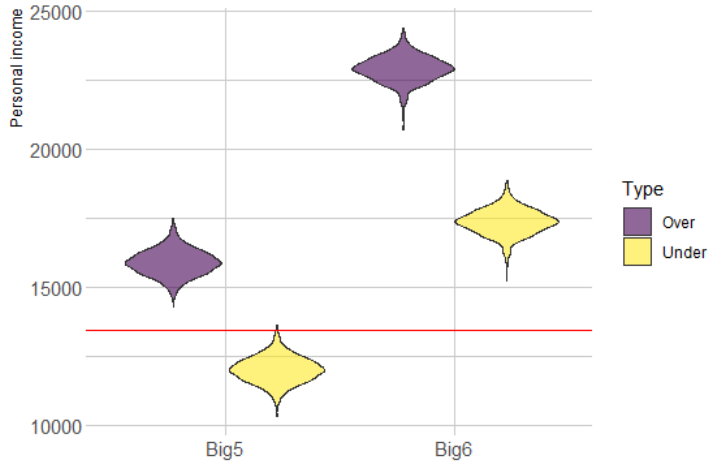


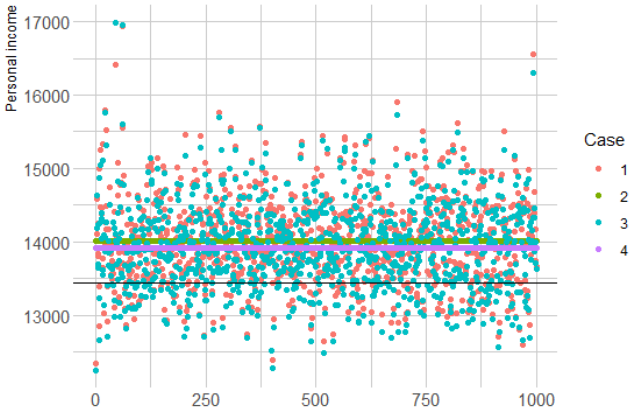
Figure A2: Violin plot of estimates of mean personal income for big data 5 and 6 using over and under measurements errors. Red line is the true value of mean personal income in the population N. The auxiliary variables used are sex and region, and the totals are known. The target variable is personal income and is observed in big data 1 and big data 2.

AMELIA	EU-SILC	Name	Notes
AGE	PB010- PB140	Age	1: [0-20] years old 2: [21-60] years old 3: [61-80] years old
BAS	RB210	Basic activity status	1: at work 2: unemployed, in retirement or early retirement or has given up business 3: other inactive person
EDI	HX090	Equivalised (disposable) household income	
HHS	HX040	Household size	1: 1 person in the household 2: 2 people in the household 3: 3 or 4 people in the household 4: [5-16] people in the household
INC	-	Personal income	
MST	PB190	Marital status	1: never married 2: married 3: separated, widowed or divorced
PWHI	-	Person with highest income in the household	1: person with highest income in household 2: not person with highest income in household
PY010	PY010	Employee Cash or near-cash income	
PY020	PY020	Non-Cash Employee income	
REG	-	Regional identifier	
RES	RB200	Residential status	1: currently living in the household 2: temporarily absent
SEM	-	Self-employment dummy	1: self-employed not self-employed
SEX	RB090	Sex	1: male 2: female

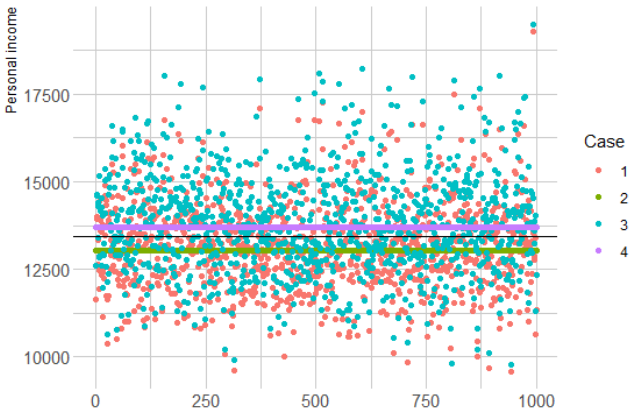
Table A1: Variables used

	Personal income		Equivalised disposable income	
	Relative bias	Relative RMSE	Relative bias	Relative RMSE
Case 1				
GREG	0.000593	0.000436	0.00130	0.000972
Optimal	0.000607	0.000441	0.00129	0.000961
Proportional	0.000607	0.000441	0.00129	0.000960
Case 2				
GREG	0.000808	0.000597	0.00102	0.000733
Optimal	0.000808	0.000597	0.00102	0.000733
Proportional	0.000808	0.000597	0.00102	0.000733
Case 3				
GREG	0.000603	0.000450	0.00130	0.000972
Optimal	0.000618	0.000459	0.00129	0.000959
Proportional	0.000618	0.000459	0.00129	0.000959

Table A2: Relative Bias and relative root mean squared error (RMSE) in different cases of macro approach, for mean of personal income and equivalised disposable income



(a)



(b)

Figure A3: Results with a sample containing all observations for big data 5 (up) and big data 6 (down). Case 1 = under measurement errors with sample A; Case 2 = under measurement errors when considering all observations; Case 3 = over measurement errors with sample A; Case 4 = over measurement errors when considering all observations. Black line is the true value of population mean personal income.

List of Figures

List of Tables

Lists of acronyms

Acronym	Explanation
AGE	Age
AMELI	Advanced Methodolgy for European Laeken Indicators
BAS	Basic activity status
EU-SILC	European Union Statistics on Income and Living Conditions
GREG	Generalised regression estimator
HHS	household size
HT	Horvitz-Thompson
MST	Marital status
PWHI	Person with highest income in the household
PY010	Employee Cash or near-cash income
PY020	Non-Cash Employee income
REG	Regional identifier
RES	Residential status
RMSE	Root mean squared error
SEM	Self-employed
SEX	Sex
SRSWOR	Simple random sampling without replacement