CEPS
INSTEAD

WORKING PAPERS

Income missing values imputation: EVS 1999 and 2008

Francesco SARRACINO

Working Paper No 2011-05
January 2011

L'European Values Study (EVS) est une enquête réalisée au Luxembourg en 2008 auprès d'un échantillon représentatif de la population résidante composé de 1610 individus âgés de 18 ans ou plus.

Au niveau national, cette enquête fait partie du projet de recherche VALCOS (Valeurs et Cohésion sociale), cofinancé par le FNR dans le cadre du programme VIVRE. Au niveau international, elle est partie intégrante d'une enquête réalisée dans 45 pays européens qui a pour objectif d'identifier et d'expliquer en Europe les dynamiques de changements de valeurs, et d'explorer les valeurs morales et sociales qui sous-tendent les institutions sociales et politiques européennes (www.europeanvaluesstudy.eu).

Plus d'infos : http://valcos.ceps.lu.

Fonds National de la Recherche Luxembourg

Income missing values imputation: EVS 1999 and 2008*

Francesco Sarracino†
Population et emploi, CEPS/Instead, Luxembourg

January 2011

Abstract

Missing data is a very frequent obstacle in many social science studies. The absence of values on one or more variables can significantly affect statistical analyses by reducing their precision and by introducing selection biases. Being unable to account for these aspects may result in severe mis-representation of the phenomenon under analysis. For this reason several approaches have been proposed to impute missing values. In present work I will adopt multiple imputation to impute income missing data for Luxembourg in the European Values Study data-set of 1999 and 2008.

Keywords: multiple imputation, missing data, income, EVS, cross-section.

JEL classification codes: C01, C11, C31.

*This research is part of the VALCOS project supported by the Luxembourg 'Fonds National de la Recherche' (contract FNR/VIVRE/06/01/09) and by core funding for CEPS/INSTEAD from the Ministry of Higher Education and Research of Luxembourg.

†Francesco Sarracino is supported by an AFR grant (contract PDR-09-075) by the National Research Fund, Luxembourg cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND). The author would like to thank Nizamul Islam and the VALCOS team for their precious advices and comments on previous versions of present work. The author remains the only responsible for the content of present article.

1 Introduction

The aim of present work is to impute missing values of income in EVS data on Luxembourg. The main feature of the present work is to enable other users to run proper statistical analysis using standard methods at their disposal on a complete data-set.

Missingness is a well-known problem to scientists and as such several approaches have been proposed to deal with it. In present work I will adopt multiple imputation, a modern technique allowing to overcome many of the limitations of more traditional methods.

Many studies in social science research are affected by data missingness. This can be due to survey design reasons explicitly skipping questions for specific samples of the population or, more problematically, to respondants not being able or willing to answer specific questions.

Independently from its causes, data missingness represents a challenge for researchers. The absence of values on one or more variables can significantly affect statistical analyses by reducing their precision and by introducing selection biases. For example, some respondants can refuse to answer questions for specific reasons. Being unable to account for these reasons may result in dramatic mis-representation of the phenomenon under analysis.

These issues are less relevant when the missingness arises because of survey design reasons. In such cases the literature agrees that analysis can be run by simply ignoring missing data (listwise method). In all other cases, and particularly when percentage of missing values is high, understanding the reasons behind the non-response becomes fundamental for running reliable analysis.

Information on income is usually difficult to obtain. People are usually reluctant to declare their own revenues and it is certainly possible that those refusing to answer are systematically different from the responding ones. Therefore, ignoring the missingness and running a standard listwise analysis can lead to significant distortion of results because of neglecting some characteristics of the population. Finally, ignoring missing cases usually strongly reduces the size of the analyzed sample.

1.1 Measuring social cohesion

Economic data are widely used in several research domains either as an outcome or as explanatory variables. This is the case of the VALCOS[1] project. The main aim of this project is to measure social cohesion starting from the individual level.

Many different definitions of social cohesion have been provided so far and it is difficult to find a generally accepted one. The literature on this topic has been previously influenced by the academic debate developed in sociology and social psychology (Berger-Schmitt, 2002, Gough and Olofsson, 1999, Lockwood, 1999) and, more recently, by a political debate in which economic and social dimensions gained a new relevance (Osberg, 2003).

Social cohesion is generally regarded as a composite concept and various approaches, both at macro and micro level, have been proposed for its measurement. On a macro level, several social indicators are adopted by institutions such as Eurostat (2009) and OECD (2009). On a micro level, social cohesion measurements point at some relevant domains of social life. For example, Jenson (1998) considers five dimensions of social cohesion: 1. affiliation/isolation; 2. insertion/exclusion; 3. participation/passivity; 4. acceptance/rejection; 5. legitimacy/illegitimacy. Bernard (1999) considers three domains of social cohesion (economic, political and socio-cultural) and distinguishes for each domain a formal and a substantial character. The formal character of a domain refers to individuals' attitudes whereas the substantial character refers to individuals' behaviours. More recently, Chan et al. (2006) present a two dimension measurement each characterized by a subjective (people's state of mind) and an objective (behavioural manifestations) component.

In this framework, some more recent measurement methods have been proposed by Rajulton et al. (2007) and Dickes et al. (2008, 2009). Both methods rely on an individual based exploratory and confirmatory factor analysis to create factor scores for the different dimensions of social cohesion as defined by Jenson (1998) and Bernard (1999).

Using data from EVS[2] 1999 and, more recently, 2008, Dickes et al. (2009) develop an index of social cohesion starting from several individual level variables and test it in 33 countries. The data-base allows them to perform a micro analysis of the main dimensions of social

[1] Valeurs et Cohésion sociale, `http:\valcos.ceps.lu`
[2] `http://www.europeanvaluesstudy.eu`

cohesion. Indeed EVS includes a large number of both subjective and objective items allowing to observe attidues and behaviour related to social relations, participation, trust and confidence in institutions at various levels of social reality. In this way the authors propose a "bottom-up" conceptualization of social cohesion in which individual attitudes and behaviours allow to define a society as cohesive. In particular, their aim is to gain information on the way individuals relate to supra-individual phenomena such as social relations and interactions, involvement and confidence in organizations and institutions.

Given its complex and multifaceted nature, social cohesion measurement involves a lot of information from various domains. Among these, economic dimensions, and particularly income, appear to be a natural element to account for income inequality.

1.2 Data source

European Values Study (EVS) data have been collected in four waves from 1981 to 2008 every 9 years. Data on Luxembourg are available only in 1999 and 2008.

Unfortunately, EVS is a poor data-set for what concern income. The use of this variable is constrained by 3 main aspects:

- it is collected in ranges (see tab.1a and 1b). Hence, the variable takes discrete values;
- ranges differ across waves;
- the percentage of missing values for income is particularly high in 1999 (see tab. 1a and 1b).

The quality of the data to impute impose some restrictions on the choice of the imputation technique:

1. provided that income variable is an ordered categorical variable in which each category corresponds to a given income interVal, the imputation method has to preserve the original scaling;
2. since income ranges change in the two waves, income imputation has to be run for each of the two waves separately;

income category	obs (%)
11	0.55
12	1.12
13	2.54
14	2.89
15	3.44
16	3.91
17	4.00
18	5.25
19	3.04
20	3.51
21	2.41
22	2.09
23	2.91
24	3.05
25	1.59
26	1.93
27	1.48
28	0.86
29	1.81
30	1.10
32	0.07
33	0.51
34	3.03
don’t know	10.17
don’t reply	36.74
Total	(1211)
	100.00

(a) 1999

income category	obs (%)
b	0.22
c	0.21
d	0.40
e	3.62
f	5.45
g	8.18
h	11.48
i	15.71
j	13.23
k	9.95
l	5.23
m	5.34
n	3.00
don’t know	6.74
don’t reply	11.24
Total	(1610)
	100.00

(b) 2008

Table 1: Net household income rankings

3. no upper bounds are available for the highest income category, forcing us to truncate the right side of the distribution tail.

2 Background of imputation techniques

2.1 Missing data mechanisms

A preliminary step in dealing with missing data is to understand the reasons causing the missingness. In other words the researcher should figure out the mechanisms generating the absence of the data. In the literature, three broad classes of mechanisms for missingness have been identified. Each of these classes has distinct implications for the analysis leading to different methodologies (Schafer, 1997, Little and Rubin, 2002).

The first mechanism is usually labeled as *missing completely at random* (MCAR). In this case data are randomly missing. This might be due, for example, to the fact that a page of the questionnaire was missing or because a data processing error happened or, simplier, because of a change in the data collection procedure. In all these cases, the reason for missingness is completely independent from the respondant (Schafer and Graham, 2002, Streiner, 2002).

Data are said to be *missing at random* (MAR) if, given the observed data, the missingness mechanism does not depend on any unobserved data. That is to say, if the probability of a missing observation does not depend on the respondant's score on the variable, after controlling for other variables in the study. These "controls" represent the mechanisms for explaining missing values. MAR means that data are conditionally randomly missing (Acock, 2005).

The last case is represented by data being *missing not at random* (MNAR). This case is generally referred to as the residual one. That is to say, it applies when the other two cases don't. In this case the missingness mechanism depends on the unobserved data, even after taking into account all the information in the observed data (Schafer and Graham, 2002).

Identifying the pattern of missingness is fundamental for at least two reasons. The first one is *representativeness* of the sample. When data are MNAR, the sample does not correctly mirror the population it is supposed to represent (Schafer and Graham, 2002). In these cases ignoring missing data would lead to biased and non-representative estimates. The second reason concerns imputation techniques. In many cases, data imputation methods assume data to be at least MAR. Hence, it is fundamental to understand which mechanism applies in order

to adopt a proper imputation method (Little and Rubin, 2002).

A further relevant aspect is represented by percentages of missingness. In general, small amount of missing values are considered less problematic and can be addressed with simplier data imputation methods (Schafer, 1997). Unfortunately, there is no consensus in the literature on how much "small" is: Tabachnick and Fidell (1983) consider small a percentage of missingness ranging between 0 and 5%, while Little and Rubin (2002) extend this qualification to cases with less than 20% of missing values.

Finally, the problem of data missingness matters depending on whether the relevant variable is an outcome or rather an explanatory variable (Saunders et al., 2006). This is particularly relevant in case of regression analysis. Pigott (2001) shows that coefficients are less biased when large missing data affect the independent variables rather than the dependent one.

2.2 Traditional techniques

Depending on the quality of the data at hand, various techniques for dealing with missing data are available.

Listwise deletion This is the most common solution to deal with missing data. Basically, listwise deletion excludes missing observations from the analysis. That is why it is sometimes called also *case deletion*. If the data are MCAR, then listwise deletion is a reasonable approach. Indeed, even if it results in a smaller sample and in higher standard errors, coefficients are still reliable. In other words, adopting this approach when data are MCAR may result in higher risk of a Type II error. Vice versa, in cases when data are not MCAR, listwise deletion can significantly bias results. In such cases its use is strongly discouraged.

Pairwise deletion In order to reduce problems linked with loss of observations, pairwise deletion uses all available information from pairs of variables regardless of whether respondants answered other variables. This method minimizes the number of dropped variables due to missing data, but it generates a new problem: potentially every couple of variables can involve different subsamples of participants virtually making any regression analysis impossible.

Mean substitution Mean substitution is a very simple and straighforward strategy to tackle with missing data. It simply substitutes mean of the total population (or of specific subgroups) for missing values. Unfortunately, the simplicity of this method has significant drawbacks. The first one is that it requires missing data to be MCAR. Secondly, the estimate of the standard deviation and the variance are downward biased. Nonetheless, this method can be considered a cheap and acceptable solution in case of very small percentages of missing data.

Hotdecking A widely adopted method, in particular when data-set are meant to be widely available, is *hotdecking*. This method is very intuitive. Let's assume we have two observations, A and B. A has missing values for some variables, while B has complete information. In that case, A's missing values are replaced with B's information provided that A and B are similar enough. Hotdecking procedes as follows: it first identifies a set of variables which are highly correlated with the variable with missing data; observations are then sorted by one of these highly correlated variables; finally, missing values are replaced by the value that appears for the preceding participant. Basically, similarity among observations is guaranteed by closeness of observations based on highly correlated variables. Obviously, this is also the main weak point of hotdecking. Indeed, for the imputation to be reliable, the variable used to sort observations has to be really highly correlated.

Regression imputation This method simply retrieves missing data using predictions from a regression model. Basically, a set of "explanatory variables" highly correlated with the variables with missing data is selected. These variables are then employed in a regression model taking the variables with missing information as dependent variable. Coefficients are estimated using listwise method and applied to predict missing values for incomplete cases.

Regression imputation is an appealing method, but it is subject to some limitations. First of all, imputed data are predicted using other observed data resulting in smaller variance and deflated standard errors (Allison, 2001). Secondly, coefficients are estimated assuming the existence of a linear relationship among variables, but this may well not be the case biasing imputations and estimates. Finally, good imputations require good predictors and these are not always available.

All these methods are quite easy to implement and have been widely used in the past.

Unfortunately, it is now clear that they have severe downsides unless very specialized circumstances apply (Acock, 2005). These conditions mainly refer to the mechanism of missingness: some of these methods work relatively well when data are MCAR. Unfortunately, this hipothesis hardly applies. In such situations, traditional methods can yield to unpredictable biases, increasing Type II errors and/or underestimating correlations and coefficients.

This is why a new set of approaches has been developed and recently integrated in the largest part of available statistical software. These new solutions are usually grouped into two categories: maximum likelihood solutions and multiple imputation (Howell, 2009).

In the remaining part of present section I am going to briefly outline the main characteristics of these two families of methods.

2.3 Modern techniques

Full information maximum likelihood approaches This family of imputation techniques does not impute missing values, but uses all the available information to provide a maximum likelihood estimation. This approach basically follows the algorithm developed by Little and Rubin (2002). The main disadvantage of this method is that models usually include only those variables that have an explicit role in the analytical model, while omitting other variables that can be mechanisms for missingness. On the other side, structural equation modelling and multilevel software packages provide many ways of working with missing values making this method easier to implement.

Expectation Maximization algorithm This method, shortly labelled EM algorithm, uses maximum likelihood to impute a single new data set. This method estimates the parameters of the data model on the basis of the observed data. Successively, it predicts missing data on the basis of those parameters. Up to this point, the EM algorithm is very similar to the traditional *regression imputation* technique. The main difference is that it iterates the two steps several times using the newly obtained completed data-set at each iteration. Every new iteration injects a degree of random error to reflect uncertainty of imputation. Values are imputed iteratively until successive iterations are sufficiently similar.

This method results in a significant improvement over traditional approaches. Nonetheless, it still produces a single imputation thus underestimating standard errors and overestimating

the level of precision.

Multiple imputation A second family of imputation techniques is the so-called multiple imputation (MI). In this case missing values are replaced by m>1 simulated versions of the variable with missing data, where $3 < m < 10$. The idea behind MI is to repeat the imputation process more than once, producing multiple "completed" data-sets. Basically, this method creates a small number (m) of completed matrices in which the missing values have been replaced by plausible values. The variability among the m imputations reflects the uncertainty about the hypothetically observed, but unknown, values. In this way MI allows for unbiased standard errors (Acock, 2005, Schafer, 1999b).

One of the major problems with MI is its implementation. This method requires three computationally intensive steps: 1. generating imputed values on the basis of existing data. Usually this step is performed using EM algorithm; 2. adding an error component to the predicted values of the variable with missing values. The error component is randomly drawn from the Bayesian posterior distribution at hand. Each time we impute data, we will obtain a slightly different result. This step is repeated several times until the process stabilizes; 3. several m complete data-sets are generated. Little and Rubin (2002) showed that, thanks to randomness inherent in the algorithm, three to five data sets are sufficient to account for uncertainty in the estimates (Allison, 2001, Schafer, 1997).

Under quite general conditions, it has been shown that:

- if the complete data model leads to valid inferences in the absence of non-response;
- if the imputation procedure is proper with respect to the non-response mechanism;

then MI yields valid inferences (van Buuren et al., 1999)

Each of the simulated complete datasets is then analysed by standard methods. The results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Intuitively, the validity of the method hinges on how the m imputations are generated (Little and Rubin, 2002). As I will show in section 3, the Bayesian theorem will allow us to get "proper" imputations.

Recent development of Monte Carlo (MC) simulation procedures made MI considerably easier to perform allowing for the development of some ad-hoc statistical software. Currently,

many of the most widely used softwares (such as Stata, SPSS, SAS and R) include a MI package. Schafer's NORM program[3] is known to be one of the first and most complete softwares to perform MI with data augmentation (a MC procedure) in S-plus (Schafer, 1999a).

3 Rubin's rule and the Bayes's theorem

The imputation problem is then how to impute a vector (**Y**) of missing values for a given variable, where $\mathbf{Y} \sim N(\mu, \psi), i = 1, ..., n$ and its parameters $\theta = (\mu, \psi)$ is unknown. For ease of explanation, let us assume the existence of a variable $\mathbf{Y}$ with n observations $\mathbf{Y} = (y_1, ..., y_n)$. A fraction of $\mathbf{Y}$ is observed $Y_{obs} = (y_1, ..., y_a)$ and the residual part is missing $Y_{mis} = (y_{a+1}, ..., y_n)$.

The Bayesian theorem offers an ideal framework to impute the missing part. In that case, we could re-write the theorem as follows (Rubin, 1987):

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta \tag{1}$$

where θ has a prior distribution and Y_{mis} is ignorably missing.

Given Y_{obs}, we can compute $\overline{y}_{obs}$ and σ^2_{obs}, that is to say we can retrieve both the mean value and the variance for the observed cases. It is important to stress that these parameters are different from the ones we had if a complete data-set was at hand. I will indicate the original and unknown parameter as μ for the average and ψ for the variance.

Under a standard non informative prior $P(\theta) \propto \psi^{-1}$, we can get θ as follows:

1. randomly drawing $\psi|Y_{obs} \sim (a-1)\sigma^2_{obs}/\chi^2_{a-1}$;
2. randomly drawing $\mu|\psi, Y_{obs} \sim N(\overline{y}_{obs}, a^{-1}\psi)$;
3. based on μ and ψ, we can generate Y_{mis} using eq.1.

Hence, Y_{mis} can be computed by: 1. simulating a random draw of θ from its observed-data posterior $\theta^* \sim P(\theta|Y_{obs})$ and 2. randomly drawing missing values from their conditional predictive distribution $Y^*_{mis} \sim P(Y_{mis}|Y_{obs}, \theta^*)$ (Schafer, 1999b). In order to ensure convergency of parameters and in the (frequent) case of non standard distribution of θ, this procedure

[3] `http://www.stat.psu.edu/~jls/misoftwa.html`

is repeated thousands of times using MC procedure. Iterations stop when a desired stationary distribution is reached. Many tools are available to assess whether stationarity has been reached or not. In general, given the increased computational possibilities of modern computers, it is sufficient to set the program to run more than one thousand iterations to be sure that stationarity has been reached (Schafer, 1999b).

4 Imputing income data for Luxembourg: EVS 1999 and 2008

Discussion in section 2 made it clear that the first step to impute missing data is identifying mechanisms and patterns of missingness. This analysis is aimed at assessing whether missing data can be considered at least MAR and, eventually, to point out which are the main characteristics of non respondants. These aspects will turn to be very relevant to define the imputation model.

4.1 Patterns of missingness

The first aspect arising by looking at descriptive statistics in tab. 1a and tab. 1b is that the number of missing data is higher in 1999 (46.9%) than in 2008 (18%). In the first case, 36.7% of respondants refuse to reply to the question and a further 10.17% declare not to know their own income. In the second one, the fraction of the sample not replying is 11.24%, while a further 6.74% does not know its own income.

When compared with data on sex, cross-tabulations in tab. 2 report that non respondants are approximately equally distributed across sex with women reporting slightly higher non respond rates than men.

Table 2: Frequency distribution of missing data about net household income across men and women. The weighted absolute number of cases is reported in parentheses. Information about categories with less than 30 observations are considered unreliable and are not commented in the text.

	1999				2008			
sex	observed	don't know	don't reply	Total	observed	don't know	don't reply	Total
men	54.28	9.99	35.73	100.00	84.86	5.88	9.25	100.00
	(323.74)	(59.61)	(213.08)	(596.43)	(695.71)	(48.24)	(75.87)	(819.81)
women	51.94	10.33	37.73	100.00	79.07	7.63	13.30	100.00
	(319.19)	(63.50)	(231.88)	(614.57)	(624.78)	(60.31)	(105.09)	(790.19)
Total	53.09	10.17	36.74	100.00	82.02	6.74	11.24	100.00
	(642.93)	(123.11)	(444.96)	(1211.00)	(1320.49)	(108.55)	(180.96)	(1610.00)

In both waves non respondants appear to be housewives, retired people, student and civil servants. Among these, in 1999 9% of housewives declare not to know their net household income and a further 44% refused to provide an answer to the qeustion. Similarly, 45% of students declared to ignore their own income. For the remaining categories, people mainly refused to answer. This is the case for workers (33%) and civil servants (50.8%) (see tab. 3).

In 2008 the picture improves significantly. People with higher percentages of missing data are students (59%). Among these, 42% declare to ignore their income and a further 16.6% does not reply. A further 29% of missing data is attributable to unemployed people. In this case, missing cases are almost equally distributed between "don't know" and "don't reply". Finally, 20% of missing data is due to houseworkers and policy makers, respectively. In both categories, the majority of non respondants (14% on average) does not provide an answer (see tab. 3).

Table 4 informs that non respondants are mainly people with secondary education.

In 1999 49% of people with primary and secondary education declares to ignore its net household income. In 2008 the share of the population in the two categories is 22.5% and 13%, respectively (see tab. 4). Furthermore, in 2008 19% of non respondants have higher education. The main difference among these three groups is that, while people with primary and secondary education are almost equally distributed between not answering and not knowing how to answer to the question, 14.4% of people with higher education refused to provide an answer. In 1999 the vast majority of not answering is due to people not willing to answer to the question (see tab. 4).

For what concern the distribution of missing data across marital status, figures in tab. 5 inform that in 1999 missing data are approximately equally distributed among the 5 categories. In all these cases, the vast majority of non response is due to people refusing to answer the income question (see tab.5).

In 2008 the picture is slightly different: the main source of missing data are single (28%) and married people (16.6%). In the first case, the reason for missing data appears to be equally distributed between "don't know" (16.8%) and "don't reply" (10.3%). On the contrary, married people mainly refuse to provide an answer (12.8%) (see tab.5).

Finally, tab. 6 and tab.7 inform about the distribution across nationalities of non respondants.

Table 3: Frequency distribution of missing data about net household income by professions. The weighted absolute number of cases is reported in parentheses. Information about categories with less than 30 observations are considered unreliable and are not commented in the text.

	1999				2008			
isco classification	observed	don't know	don't reply	Total	observed	don't know	don't reply	Total
military professions	40.13	29.93	29.93	100.00	100.00	0.00	0.00	100.00
	(0.83)	(0.62)	(0.62)	(2.07)	(2.38)	(0.00)	(0.00)	(2.38)
policy-makers	53.53	2.02	44.45	100.00	79.88	4.46	15.66	100.00
	(10.61)	(0.40)	(8.81)	(19.82)	(51.01)	(2.85)	(10.00)	(63.86)
intellectual professions	56.90	8.26	34.84	100.00	84.57	2.73	12.70	100.00
	(77.42)	(11.23)	(47.40)	(136.05)	(122.27)	(3.94)	(18.36)	(144.57)
physic & technic professions	62.48	7.06	30.46	100.00	84.82	4.75	10.43	100.00
	(54.45)	(6.15)	(26.55)	(87.15)	(189.08)	(10.59)	(23.25)	(222.92)
civil servants	44.17	5.04	50.79	100.00	85.23	2.27	12.51	100.00
	(38.67)	(4.41)	(44.47)	(87.55)	(104.22)	(2.77)	(15.29)	(122.28)
traders merchants & vendors	60.45	6.76	32.79	100.00	92.97	3.57	3.46	100.00
	(42.57)	(4.76)	(23.10)	(70.43)	(86.92)	(3.34)	(3.24)	(93.49)
skilled workers	43.01	24.85	32.14	100.00	80.39	9.09	10.53	100.00
	(2.63)	(1.52)	(1.97)	(6.12)	(18.82)	(2.13)	(2.46)	(23.41)
artisanal workers	58.50	3.43	38.07	100.00	88.26	4.86	6.88	100.00
	(60.13)	(3.53)	(39.13)	(102.79)	(92.56)	(5.09)	(7.21)	(104.87)
factory workers	45.77	11.44	42.78	100.00	95.38	1.21	3.41	100.00
	(16.90)	(4.22)	(15.79)	(36.92)	(74.89)	(0.95)	(2.68)	(78.52)
unskilled workers	63.78	3.10	33.12	100.00	88.95	7.43	3.62	100.00
	(36.65)	(1.78)	(19.03)	(57.46)	(90.96)	(7.60)	(3.70)	(102.26)
retired	61.46	2.25	36.29	100.00	82.70	1.46	15.84	100.00
	(153.74)	(5.62)	(90.77)	(250.13)	(241.34)	(4.26)	(46.23)	(291.83)
houseworker	47.52	8.58	43.90	100.00	79.15	8.20	12.65	100.00
	(99.70)	(17.99)	(92.11)	(209.80)	(154.05)	(15.97)	(24.62)	(194.64)
student	30.86	45.56	23.58	100.00	40.94	42.41	16.65	100.00
	(39.34)	(58.08)	(30.06)	(127.48)	(39.44)	(40.86)	(16.04)	(96.35)
unemployed	58.78	19.82	21.40	100.00	70.68	11.55	17.78	100.00
	(8.27)	(2.79)	(3.01)	(14.07)	(31.27)	(5.11)	(7.86)	(44.24)
handicapped	32.54	0.00	67.46	100.00	87.30	12.70	0.00	100.00
	(1.03)	(0.00)	(2.14)	(3.17)	(21.30)	(3.10)	(0.00)	(24.40)
Total	53.09	10.17	36.74	100.00	82.02	6.74	11.24	100.00
	(642.93)	(123.11)	(444.96)	(1211.00)	(1320.49)	(108.55)	(180.96)	(1610.00)

Table 4: Frequency distribution of missing data about net household income by education level. The weighted absolute number of cases is reported in parentheses. Information about categories with less than 30 observations are considered unreliable and are not commented in the text.

	1999				2008			
education	observed	don't know	don't reply	Total	observed	don't know	don't reply	Total
primary	51.91	9.71	38.38	100.00	87.08	5.02	7.90	100.00
	(156.10)	(29.19)	(115.41)	(300.69)	(343.72)	(19.80)	(31.18)	(394.70)
vocational	54.15	6.22	39.64	100.00	83.77	6.69	9.54	100.00
	(179.01)	(20.55)	(131.05)	(330.61)	(272.07)	(21.72)	(30.99)	(324.78)
secondary	51.05	14.17	34.78	100.00	77.43	10.02	12.55	100.00
	(181.55)	(50.40)	(123.69)	(355.64)	(394.55)	(51.08)	(63.93)	(509.56)
higher	56.36	10.25	33.39	100.00	81.41	4.19	14.40	100.00
	(126.27)	(22.97)	(74.81)	(224.05)	(310.15)	(15.96)	(54.86)	(380.96)
Total	53.09	10.17	36.74	100.00	82.02	6.74	11.24	100.00
	(642.93)	(123.11)	(444.96)	(1211.00)	(1320.49)	(108.55)	(180.96)	(1610.00)

Table 5: Frequency distribution of missing data about net household income by marital status. The weighted absolute number of cases is reported in parentheses. Information about categories with less than 30 observations are considered unreliable and are not commented in the text.

	1999				2008			
marital status	observed	don't know	don't reply	Total	observed	don't know	don't reply	Total
married	56.78	4.14	39.08	100.00	83.37	3.78	12.85	100.00
	(389.67)	(28.42)	(268.17)	(686.26)	(783.30)	(35.55)	(120.75)	(939.60)
widowed	51.21	6.40	42.39	100.00	91.16	0.68	8.16	100.00
	(48.30)	(6.04)	(39.98)	(94.31)	(82.67)	(0.61)	(7.40)	(90.69)
divorced	53.79	0.00	46.21	100.00	90.99	2.17	6.84	100.00
	(29.99)	(0.00)	(25.77)	(55.76)	(130.33)	(3.10)	(9.80)	(143.23)
separated	36.86	10.32	52.82	100.00	94.64	2.92	2.45	100.00
	(3.13)	(0.88)	(4.49)	(8.50)	(27.31)	(0.84)	(0.71)	(28.86)
single	46.93	23.97	29.10	100.00	72.83	16.79	10.38	100.00
	(171.83)	(87.78)	(106.56)	(366.17)	(296.88)	(68.44)	(42.31)	(407.63)
Total	53.09	10.17	36.74	100.00	82.02	6.74	11.24	100.00
	(642.93)	(123.11)	(444.96)	(1211.00)	(1320.49)	(108.55)	(180.96)	(1610.00)

Table 6: Frequency distribution of missing data about net household income by nationality in 1999. The weighted absolute number of cases is reported in parentheses. Information about categories with less than 30 observations are considered unreliable and are not commented in the text.

nationality	observed	don't know	don't reply	Total
luxembourgish	53.76	9.72	36.52	100.00
	(430.98)	(77.92)	(292.73)	(801.63)
portuguese	43.15	10.02	46.83	100.00
	(63.95)	(14.85)	(69.41)	(148.21)
italian	55.03	12.71	32.26	100.00
	(38.37)	(8.86)	(22.49)	(69.73)
belgian	69.02	5.81	25.16	100.00
	(29.23)	(2.46)	(10.65)	(42.34)
french	53.28	11.38	35.34	100.00
	(27.55)	(5.88)	(18.27)	(51.70)
german	53.86	3.62	42.52	100.00
	(14.38)	(0.97)	(11.35)	(26.70)
dutch	58.95	0.00	41.05	100.00
	(10.55)	(0.00)	(7.35)	(17.90)
Other EU 15	72.14	27.86	0.00	100.00
	(15.48)	(5.98)	(0.00)	(21.46)
Central and eastern Europe	39.98	17.23	42.79	100.00
	(7.93)	(3.42)	(8.49)	(19.85)
North America	36.54	29.45	34.01	100.00
	(1.33)	(1.07)	(1.23)	(3.63)
Africa	49.38	26.66	23.95	100.00
	(3.17)	(1.71)	(1.54)	(6.42)
Middle East	0.00	0.00	100.00	100.00
	(0.00)	(0.00)	(1.44)	(1.44)
Total	53.09	10.17	36.74	100.00
	(642.93)	(123.11)	(444.96)	(1211.00)

In the first wave the main source of missing data are people from Portugal (56.8%) followed by French, Luxembourgish, Italian and Belgian people. In all these cases, missing data are mainly due to people not replying to the question on household income (see tab.6).

In 2008 the composition of non respondants slightly changes: missing values are mainly due to people from Italy (26%), Luxembourg (20%), Belgium (14.5%) and Portugal (12%) (see tab.7). Similar to previous cases, people mainly chose to refuse to reply to the income question.

Overall, these figures suggest that missing data have a specific pattern. Indeed, missingness is mainly due to people of both sexes with a middle level education, mainly workers or students, either single or married and from various nationalities.

4.2 Possible strategies

Given the quality of data, the number and the pattern of missing values, it appears clear that traditional methods can not be applied. The number of missing observations are large enough (in particular in the first wave) to considerably bias estimates if ignored. Similarly, cross

Table 7: Frequency distribution of missing data about net household income by nationality in 2008. The weighted absolute number of cases is reported in parentheses. Information about categories with less than 30 observations are considered unreliable and are not commented in the text.

nationality	observed	don't know	don't reply	Total
luxembourgish	79.88	7.37	12.75	100.00
	(808.47)	(74.58)	(129.07)	(1012.12)
portuguese	87.51	5.87	6.63	100.00
	(214.17)	(14.36)	(16.22)	(244.75)
italian	74.44	8.83	16.74	100.00
	(49.05)	(5.82)	(11.03)	(65.90)
belgian	85.50	4.39	10.11	100.00
	(46.31)	(2.38)	(5.47)	(54.16)
french	89.35	4.24	6.41	100.00
	(75.72)	(3.60)	(5.43)	(84.74)
german	84.64	4.00	11.36	100.00
	(30.54)	(1.44)	(4.10)	(36.08)
dutch	91.35	1.97	6.68	100.00
	(9.66)	(0.21)	(0.71)	(10.58)
Other EU 15	87.20	2.15	10.65	100.00
	(33.33)	(0.82)	(4.07)	(38.22)
Other western Europe	92.38	0.00	7.62	100.00
	(4.23)	(0.00)	(0.35)	(4.58)
Central and eastern Europe	86.21	9.41	4.38	100.00
	(33.80)	(3.69)	(1.72)	(39.20)
North America	100.00	0.00	0.00	100.00
	(3.86)	(0.00)	(0.00)	(3.86)
Latin America	61.54	38.46	0.00	100.00
	(1.36)	(0.85)	(0.00)	(2.21)
Africa	83.63	6.46	9.90	100.00
	(8.64)	(0.67)	(1.02)	(10.34)
Asia	41.46	4.51	54.03	100.00
	(1.36)	(0.15)	(1.77)	(3.28)
Total	82.02	6.74	11.24	100.00
	(1320.49)	(108.55)	(180.96)	(1610.00)

tabulations from section 4.1 inform that data can not be considered MCAR. In this situation, any of the traditional imputation techniques would lead to biased and possibly misleading imputation.

Nonetheless, it is still possible to identify a set of characteristics "explaining" the missingness. In other words, it is still possible to assume data at hand to be MAR, once all these characteristics are taken into account. Hence, the most reasonable solution in this framework is to impute data using MI technique.

4.3 The model: MI with Ordered logit

Multiple imputations can be implemented in various ways depending on the quality of the variable we want to impute. In present context, income is reported with a set of ordered categories and it is fundamental that the imputed values respect such scaling. Hence, given the **ordered nature** of the income variable, the best strategy is to use an **ordered logit model** with K ordered categories. In order to fill in Y_{mis}, MI with Ologit in Stata performs the following steps:

1. fit an ordered logistic model to (Y_{obs}, Z_{obs});
2. obtain the maximum likelihood estimates (MLE) of the parameters, $\hat{\theta}$ = ($\hat{\beta}$, $\hat{\gamma}$), and their σ^2;
3. simulate new parameters, θ*, from the large-sample normal approximation, $N(\hat{\theta}, \hat{\sigma^2})$, to its posterior distribution assuming a non-informative prior (Rubin's rule);
4. obtain one set of imputed values , Y^1_{mis} using an ordered logistic distribution as defined by (3).
5. the last two steps are repeated several times to obtain M sets of imputed values.

Formally, the model for imputation can be represented as following:

$$Pr(y_i = k|\mathbf{z}_i) = Pr(\gamma_{k-1} < \mathbf{z'}_i\beta + u \leq \gamma_k) \tag{2}$$

$$Pr(y_i = k|\mathbf{z}_i) = \frac{1}{1 + exp(-\gamma_k + \mathbf{z'}_i\beta)} - \frac{1}{1 + exp(-\gamma_{k-1} + \mathbf{z'}_i\beta)} \tag{3}$$

where $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{iq})$ records values of predictors of **y** for observation *i*, β is a vector of unknown regression coefficients and $\gamma = (\gamma_1, ..., \gamma_{k-1})$ are the unknown cutpoints with $\gamma_0 = -\infty$ and $\gamma_k = \infty$.

The model of equation 3 is repeated separately for each wave. This is due to the fact that the income categories differ across waves and running a separate model seemed the most reasonable way to respect the original scaling after imputation.

Tables 8 and 9 report the list of variables included in equation 3. Beside the list of variables used to explain the pattern of missingness, I include age of respondant, number of children and household size to account for unobserved heterogeneity. Furthermore, since imputed data are going to be adopted in various models, it is important to include among the predictors also a broad set of variables that can be object of future analysis by other researchers (Allison, 2001). Provided that present data will be mainly adopted by researchers for social cohesion analysis, I included also a set of variables to account for various dimensions of social cohesion, subjective well-being and weights. The income variable is used as dependent variable.

variable	mean	sd	min	max	obs	missing
age	41.18	16.64	15	86	1144	0
age (10 years classes)	3.135	1.728	0	7	1144	0
age (sesopi categories)	3.104	1.667	0	6	1144	0
higher level of education - valcos-sesopi classification (4 categories)	-	-	1	4	1144	0
higher level of education - isced classification (4 categories)	-	-	1	4	1144	0
isco socioprofessional classification (15 categories)	-	-	0	14	1144	0
isco socioprofessional classification (11 categories)	-	-	0	10	1144	0
marital status	-	-	1	5	1144	0
marital status (pacse recoded)	-	-	1	5	1144	0
employment status	-	-	1	9	1144	0
employment status: accounting for the detailed activity status	-	-	1	14	1144	0
employment status: accounting for the inactive status	-	-	1	7	1144	0
employment status (5 categories)	-	-	1	5	1144	0
employment status (active-retired-housekeeper-student)	-	-	1	8	1144	0
sex	-	-	1	2	1144	0
number of children	-	-	0	7	1144	0
nationality	-	-	1	8	1144	0
weight	1.012	0.408	0.220	3.650	1144	0
household composition (5 categories)	-	-	1	5	1144	0
socio-economic status of the respondent	-	-	1	4	1144	0
trust	-0.0897	0.980	-3.675	2.902	1144	0
solidarity	-0.0157	0.985	-2.817	3.046	1144	0
political participation	0.0707	1.054	-1.741	2.981	1144	0
socio-cultural participation	-0.0722	0.924	-0.995	5.819	1144	0
social relationships	0.00367	0.931	-1.958	3.084	1144	0
formal	-0.0164	0.986	-2.861	2.973	1144	0
substantial	0.000327	0.923	-2.032	3.329	1144	0
happiness	3.282	0.588	1	4	1144	0
life satisfaction	3.616	1.123	1	5	1144	0
income range index	-	-	1	24	603	0.473

Table 8: Descriptive statistics of the selected variables for the imputation - 1999. Non weighted data. Categorical variables have been recoded into dummies. Means and standard deviations for categorical variables have been omitted from the table.

variable	mean	sd	min	max	obs	missing
age	39.57	17.51	18	88	1605	0
age (10 years classes)	2.988	1.770	1	7	1605	0
age (sesopi categories)	2.950	1.692	1	6	1605	0
higher level of education - valcos-sesopi classification (4 categories)	-	-	1	4	1605	0
higher level of education - isced classification (4 categories)	-	-	1	4	1605	0
isco socioprofessional classification (15 categories)	-	-	0	14	1605	0
isco socioprofessional classification (11 categories)	-	-	0	10	1605	0
marital status	-	-	1	6	1605	0
marital status (pacse recoded)	-	-	1	5	1605	0
employment status	-	-	1	9	1605	0
employment status: accounting for the detailed activity status	-	-	1	14	1605	0
employment status: accounting for the inactive status	-	-	1	7	1605	0
employment status (5 categories)	-	-	1	5	1605	0
employment status (active-retired-housekeeper-student)	-	-	1	8	1605	0
sex	-	-	1	2	1605	0
number of children	-	-	0	7	1605	0
nationality	-	-	1	8	1605	0
weight	1.001	0.651	0.0205	2.904	1605	0
household composition (5 categories)	-	-	1	5	1605	0
socio-economic status of the respondent	-	-	1	4	1605	0
trust	0.0762	0.958	-3.572	2.627	1605	0
solidarity	-0.0957	0.977	-2.986	3.105	1605	0
political participation	0.0430	0.918	-1.766	2.855	1605	0
socio-cultural participation	-0.0250	0.986	-1.012	8.423	1605	0
social relationships	0.139	1.088	-2.136	3.980	1605	0
formal	-0.121	0.981	-3.052	2.869	1605	0
substantial	0.101	1.051	-2.157	5.023	1605	0
happiness	3.321	0.601	1	4	1605	0
life satisfaction	3.669	1.144	1	5	1605	0
income range index	-	-	1	13	1223	0.238

Table 9: Descriptive statistics of the selected variables for the imputation - 2008. Non weighted data. Categorical variables have been recoded into dummies. Means and standard deviations for categorical variables have been omitted from the table.

4.4 The code

Stata code to implement MI with an ordered logit model is quite straightforward. The first step is to define the data to be *wide*. This is required by Stata, but it should be clear that data can be defined in various ways. Given the data at hand, I chose the most conservative option.

```
mi set wide
```

The second step is to declare the variable to be imputed and the explanatory ones.

```
mi register imputed yindex
```

```
mi register regular ‘‘set of explanatory variables’’
```

Finally, it is possible to run MI with the following command:

```
mi impute ologit yindex = ‘‘set of explanatory variables’’,
add(10) rseed(47963) double noisily showstep
```

The structure of the command reflects usual commands in Stata. The option *add* tells Stata how many complete data-set to produce. Little and Rubin (2002) suggest that 3 to 5 imputed data-set should be a safe choice. Nonetheless, thanks to the increased computation speed of modern computers, I opted for a conservative choice generating 10 new complete data-set.

The option *rseed* allows to set a seed for the random number generator. This option is not mandatory, but it is highly recommended. It prevents Stata to produce different results because of different seeds. In this way, we are sure that every time we run the model, results will stay constant unless we explicitly change the model.

The last three options affect only the display of imputation process and of its results by showing each step and every intermediate output. The option *double* requires the 10 new imputed variables to be of double precision.

Tab. 10 provides an example of the output of the MI command. In this example, the data-set at hand includes 10 observations and four variables: age, happiness, respondant number and income (*y*). The last three respondants have missing values for the income variable. Imputation using MI produces a new dummy variable (*_mi_miss*) equal to 1 if the value is missing and 0 otherwise. Successively, if the option *add(10)* is specified, MI command generates 10

age	happy	obs	y	_mi_miss	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_10
41	4	1191	20	0	20	20	20	20	20	20	20	20	20	20
40	4	1196	20	0	20	20	20	20	20	20	20	20	20	20
49	3	1204	17	0	17	17	17	17	17	17	17	17	17	17
25	3	1205	23	0	23	23	23	23	23	23	23	23	23	23
53	4	1206	22	0	22	22	22	22	22	22	22	22	22	22
43	3	1207	20	0	20	20	20	20	20	20	20	20	20	20
31	4	1211	18	0	18	18	18	18	18	18	18	18	18	18
33	3	1	.	1	23	23	21	21	20	24	25	24	24	29
44	3	4	.	1	27	24	19	20	23	23	24	17	30	28
23	3	5	.	1	34	34	26	34	21	32	34	33	26	27

Table 10: An example of the data-set with multiple imputations.

new complete income variables (*y_1 . . . y_10*). Each of these new variables has the same income values as the original one for non-missing cases. In fact the first 7 observations of each income variable are the same. On the contrary, the remaining three values are imputed and changing across variables. This is meant to reflect uncertainty with respect to the original (but unobserved) values.

At this point it is possible to run statistical analysis including a complete variable for income.

The imputed income has been successively transformed into real euros 2005.

MI framework in Stata[4], or many other statistical software, provides a set of functions to deal with the imputed variable. For example, in case of a regression analysis, the software will first run a separate regression for each of the 10 imputed income variables. Successively, it will provide summary results for coefficients and standard errors.

Coefficients will be computed by averaging the 10 different coefficients:

$$\beta = \frac{1}{K}\sum_{k=1}^{K}\beta_k$$

while the variance of coefficients is computed as follows:

$$\sigma^2 = (1 + \frac{1}{K})\sigma_b^2 + \sigma_w^2$$

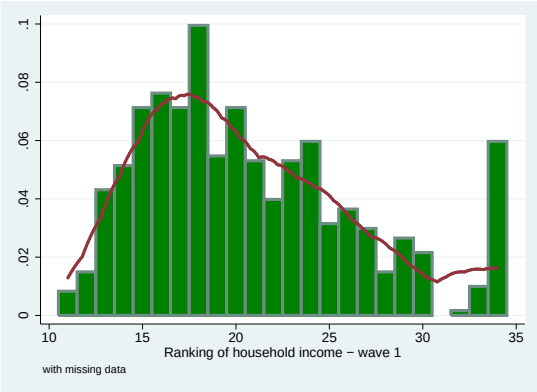
where $\sigma_w^2 = \frac{1}{K}\sum_{i=1}^{K}\sigma_k^2$ and $\sigma_b^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\beta_k - \beta)^2$

[4]The multiple imputation framework is available only in Stata 11 and newer versions.

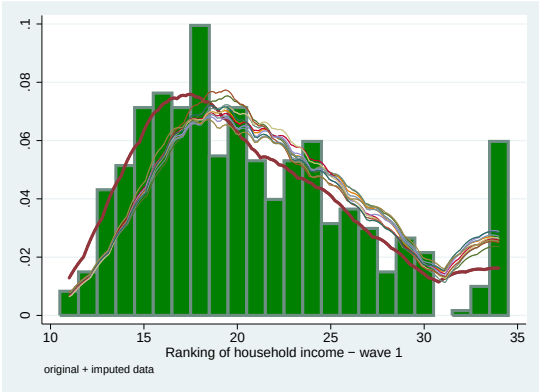
It is important to stress that the final user will not perform these steps manually. Stata's *MI* command offers a wide list of statistical analysis that will perform all the relevant steps automatically. All that is required from the user is to check that Stata is recognising the data-set at hand as containing imputed variables and to select the model of interest. The option "Multiple imputation" available under the menu "Statistics" in Stata provides an intuitive and graphic tool to perform all the required steps, i.e. 1. checking/defining a data-set to contain imputed data. It is also possible to perform other operations such as data reshaping and re-organization; 2. getting summary statistics; 3. choosing the relevant model; 4. getting final results.

5 Final remarks

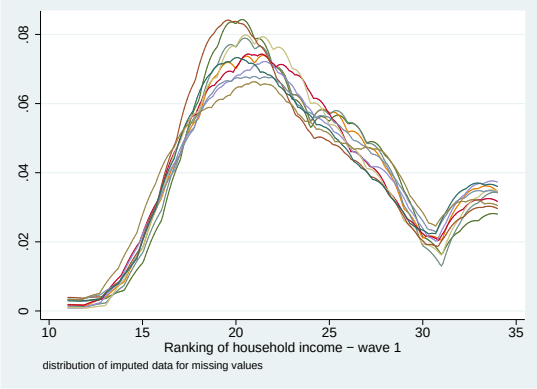
Once the missing values have been imputed, it is interesting to check how the imputed variables are distributed with respect to the observed variable. Figures 1 and 2 provide a graphic answer to this question for years 1999 and 2008 respectively.



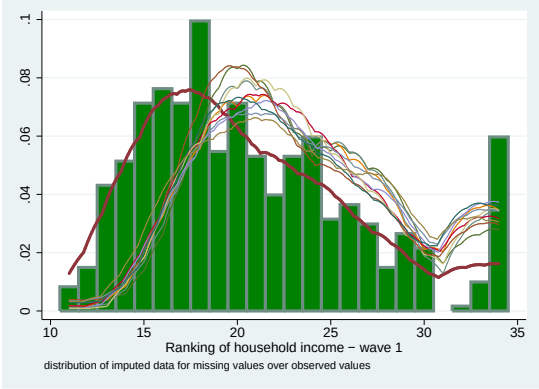
(a) *Density function and kdensity.*



(b) *Density function and kdensity for 10 complete variables.*



(c) *kdensity for imputed observations only.*



(d) *kdensity for imputed observations only over the distribution with missing values.*

Figure 1: Net household income distribution in 1999

Figure 1a shows how the original (and incomplete) income variable is distributed. The kernel density is added to make comparison with imputed variable easier. The distribution appear to be slightly skewed on the left reaching the maximum on the 18^{th} category. Nonetheless, the right tile appear quite heavy, probably because of the truncation of the last category.

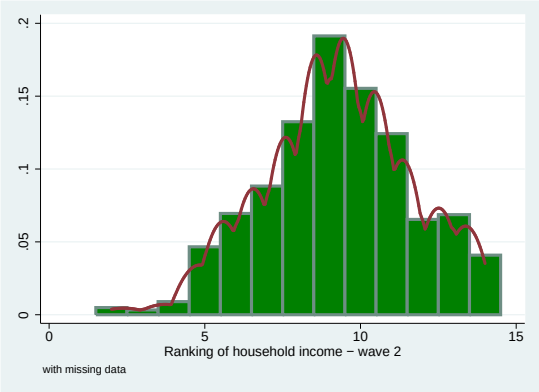
Figure 1b reports the same distribution of fig.1a adding kernel density for the 10 imputed variables. The first aspect arising from this chart is that the 10 variables are very similarly distributed. This is partly due to the fact that about 54% of its values are observed, while the remaining 46% are imputed values. As such they include some disturbance. Overall, the distribution of the new variables is right shifted with respect to the original variable, thus better approximating a normal distribution. Consistently, the right tile appears to be heavier.

When considering the distribution of the imputed values only (fig. 1c, the effect of the random error in the imputation process becomes clearer. Curves are still following similar patterns across variables, but now distributions appear less concentrated.

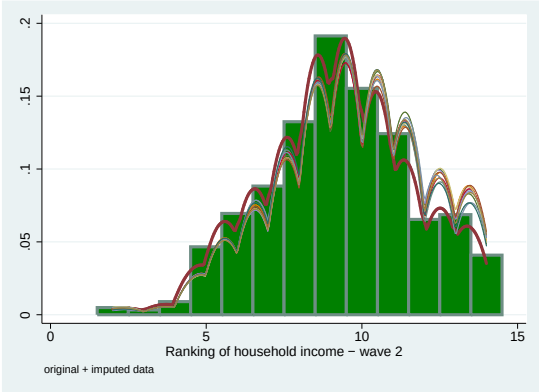
Finally, fig. 1d informs about the differences between imputed and observed values by superimposing the distribution of the original variable with the distribution of the imputed ones.

The income variable for 2008 is much more normally distributed (see fig. 2a than the one for 1999 (fig. 1a). A plausible explanation for this difference is in the number of non responses: there are less missing values in 2008 (18%) than in 1999 (46%). Nonetheless, also in this case the right tile of the distribution appears heavier than the left one, probably reflecting the effect of the truncation due to the absence of an upper limit for the last income category.

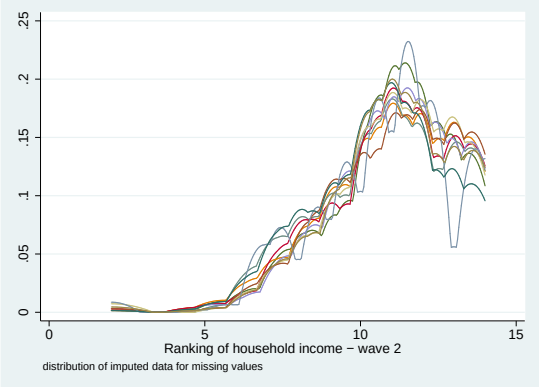
Comparing the distributions in fig. 2b we notice that the imputed values are rightward shifted with respect to the original distribution. In particular, for values below the average, new variables are slightly lower than the original ones, while for higher values, variables with imputed values are above the original variable. Also in this case the net effect of the imputation is approximating a normal distribution.



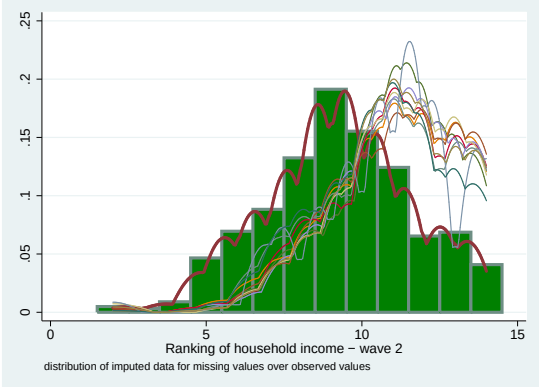
(a) *Density function and kdensity.*



(b) *Density function and kdensity for 10 complete variables.*



(c) *kdensity for imputed observations only.*



(d) *kdensity for imputed observations only over the distribution with missing values.*

Figure 2: Net household income distribution in 2008

References

A.C. Acock. Working with missing values. *Journal of marriage and family*, 67:1012 – 1028, November 2005.

P.D. Allison. Missing data. *SAGE university paper*, 136, 2001.

R. Berger-Schmitt. Considering social cohesion in quality of life assessments: concept and measurement. *Social indicators research*, 58:103 – 428, 2002.

P. Bernard. Social cohesion: a critique. *CPRN Discussion Paper*, F 09, 1999. Canadian Policy Research Networks.

J. Chan, H.P. To, and E. Chan. Reconsidering social cohesion: developing a definition and analytical framework for empirical research. *Social Indicators Research*, 75:273 –302, 2006.

P. Dickes, M. Valentova, and M. Borsenberger. Social cohesion: measurement based on the evs micro data. *Statistica applicata*, 20:1 –16, 2008.

P. Dickes, M. Valentova, and M. Borsenberger. Construct validation and application of a common measure of social cohesion in 33 European countries. *Social Indicators Research*, December 2009. published online first, [http://dx.doi.org/10.1007/s11205-009-9551-5].

Eurostat. Structural indicators: social cohesion, 2009. URL `http://epp.eurostat.ec.europa.eu/portal/page/portal/structural_indicato`

I. Gough and G. Olofsson. *Capitalism and social cohesion: essays on exclusion and integration*. Palgrave Macmillan, New York, October 1999.

D.C. Howell. Treatment of missing data, July 2009. URL `TOFILLIN`.

J. Jenson. Mapping social cohesion: the stat of Canadian research. *CPRN Discussion Paper*, F 03, 1998. Canadian Policy Research Networks.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, New York, 2nd edition, 2002.

D. Lockwood. Civic integration and social cohesion. In I. Gough and G. Olofsson, editors, *Capitalism and social cohesion: essays on exclusion and integration*, pages 63 – 84. Palgrave Macmillan, New York, 1999.

OECD. Society at a glance 2009 - OECD Social Indicators. Technical report, 2009.

L. Osberg, editor. *The economic implications of social cohesion*. University of Toronto Press, 2003.

T.D. Pigott. A review of methods for missing data. *Educational research and evaluation*, 7: 353 – 383, 2001.

F. Rajulton, Z. Ravanera, and R. Beaujot. Measuring social cohesion: an experiment using the Canadian national survey of giving, volunteerin and participating. *Social Indicators Research*, 80:461 – 492, 2007.

D. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley, New York, 1987.

J.A. Saunders, N. Morrow-Howell, E. Spitznagel, P. Doré, E.K. Proctor, and R. Pescarino. Imputing missing data: a comparison of methods for social work researchers. *Social work research*, 30(1):19 – 31, March 2006.

J. Schafer. Norm: multiple imputation of incomplete multivariate data under a normal model. Technical report, 1999a. URL `http://www.stat.psu.edu/~jls/misoftwa.html`. version 2.

J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, CRC Press Company, 1997.

J.L. Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8:3–15, 1999b.

J.L. Schafer and J.W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7:147 – 177, 2002.

D.L. Streiner. The case of the missing data: methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry*, 47:68 – 75, 2002.

B.G. Tabachnick and L.S. Fidell. *Using multivariate statistics*, chapter Cleaning up your act: screening data prior to analysis, pages 68 – 81. Harper & Row, New York, 1983.

S. van Buuren, C.H. Boshuizen, and D.L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 1:681 – 694, 1999.

CEPS
INSTEAD

B.P. 48
L-4501 Differdange
Tél.: +352 58.58.55-801
www.ceps.lu