Income Imputation strategies for the European Health Interview Survey in Luxembourg

Mauro Baldacchini[1], Maria Ruiz-Castell[2] & Gwenaëlle Le Coroller[1]

[1] *Luxembourg Institute of Health, Competence Center for Methodology and Statistics, Luxembourg - baldacchinimauro@gmail.com, Gwenaelle.LeCoroller@lih.lu*

[2] *Luxembourg Institute of Health, Department of Precision Health, Luxembourg - Maria.Ruiz@lih.lu*

Résumé. Lors la troisième vague de l'European Health Interview Survey (EHIS) menée au Luxembourg en 2019, 24.4% des réponses sur le revenu des ménages avaient des valeurs manquantes (combinaison de "Je ne souhaite pas répondre" et/ou aucune réponse), rendant cette variable inadaptée à l'analyse, incitant ainsi le besoin d'un processus d'imputation approprié (Lee et Huber 2021). Cependant, l'absence d'informations supplémentaires sur le revenu et les actifs des ménages dans le questionnaire d'EHIS rendait le processus d'estimation du revenu particulièrement difficile. Le revenu des ménages était exprimé en classes allant de "Moins de 1 000 €" à "Plus de 12 500 €". Ceci, combiné au fait que les données au Luxembourg ont été collectées à l'aide d'une enquête auto-administrée (sans enquêteur), a entraîné la présence de différentes erreurs et imprécisions dans les réponses. Avoir une estimation impartiale des informations sur le revenu est crucial pour développer des résultats destinés à l'élaboration de politiques efficaces pour un groupe déterminé de la population. Le but de ce travail est de trouver un moyen réalisable d'imputer ces données à l'aide de celles recueillies au sein d'EHIS. Pour l'analyse, il a été supposé que les informations sur le revenu étaient manquantes de manière non aléatoire (MNAR), c'est-à-dire qu'elles sont manquantes en raison du revenu lui-même, en plus de caractéristiques d'autres variables (Kim et al. 2007). Le travail est divisé en différentes étapes. La première étape a consisté à mener une revue de littérature sur les méthodes d'imputation du revenu, en mettant particulièrement l'accent sur les imputations de données MNAR. Dans un second temps, après un examen initial des variables, les caractéristiques de données manquantes ont été analysées, ainsi que les incohérences entre les réponses. En troisième lieu, des méthodologies pertinentes ont été appliquées à des données simulées, et leur efficacité a été évaluée en termes de Kappa pondéré de Cohen. Enfin, nous avons sélectionné la méthode Random Forest comme la meilleure pour nos données, et nous avons comparé les proportions obtenues grâce à l'imputation avec celles issues des cas complets. Les conclusions de ce travail seront bénéfiques pour d'autres enquêtes dans lesquelles le revenu peut servir de variable cruciale pour l'analyse, ainsi que pour les futures vagues d'EHIS, notamment celle prévue pour 2025, où la variable du revenu des ménages devra être imputée si le taux de valeurs manquantes dépasse 5%.

Mots-clés.EHIS, MNAR (Non manquant au hasard), Revue, Impartialité, Estimateur, Missing Data (données manquantes), Eurostat, Enquête

Abstract. In the third wave of the European Health Interview Survey (EHIS) conducted in Luxembourg in 2019, 24.4% of responses lacked household income data (merge of "I do not wish to answer" and/or missing value), making that variable unsuitable for analysis, thus prompting the need for an imputation process (Lee and Huber 2021). However, the absence of additional information related to income and household assets within the EHIS questionnaire

rendered the income estimation process particularly challenging. Household income was expressed in classes that ranged from "Less than €1,000" to "More than €12,500". This, combined with the fact that the data in Luxembourg were collected using a self-administered survey (without interviewer), leads to the presence of different errors and imprecisions on the answers. Having an unbiased estimation of the income information is crucial for developing policy-making aimed results that can be effective for a determined stratum of the population. The purpose of the present work was to find a feasible way of imputing this information using data gathered from the EHIS. For the analysis, income missing information was assumed to be Missing Not At Random (MNAR), i.e. that income information is missing due to income itself, in addition to other respondent's characteristics (Kim et al. 2007).The work is divided into different steps. The first one consisted in conducting a literature review on the methods to impute income, focusing especially on MNAR data imputations. Secondly, after an initial examination of the variables, missing data patterns were analyzed, as well as inconsistencies among answers. Thirdly, relevant methodologies were applied to simulated data, and their effectiveness was evaluated in terms of Weighted Cohen's Kappa. Finally, we selected Random Forest as the best method for our data, and we compared the proportions obtained from the imputation with those obtained from the complete cases. Results of this work will be beneficial for other surveys in which income may serve as crucial variable for the analysis, as well as for future waves of the EHIS, such as the one scheduled for 2025, where the household income variable will be imputed when the rate of missing values exceeds 5%.

Keywords. EHIS, MNAR, Missing Not At Random, Review, Unbiasedness, Estimator, Missing Data, Eurostat

1 Background and problem description

The European Health Interview Survey (EHIS) is a cross-sectional population-based survey conducted by the statistical office of the European Commission (Eurostat). The aim is to measure the health status of the EU citizens and use of health care services, as well as limitations in accessing them on a harmonized basis and with a high degree of comparability among Member States (MS). The target population is composed of all citizens of the EU MS aged 15 or over, living in private households.

In Luxembourg, the EHIS was conducted in 2014 and 2019. In both waves, some problems regarding the household income question occurred. More specifically, in 2014 and 2019, the rate of Non Responses[1] to the question "*What is your household's total net monthly income?*" were 28.9% and 24.4%, respectively. These significant proportions made the variable unsuitable for analysis purposes, making the imputation of missing values crucial (Lee and Huber 2021). This is particularly relevant given the potential impact of income on individual's health status.

The process of imputing household monthly income in the EHIS presents several challenges due to various factors:

[1]This encompasses both individuals who left the question unanswered and those who selected the "*I do not wish to answer*" option.

- **Absence of additional information**: given the health-oriented nature of the EHIS, it lacks additional questions related to income[2] that could aid in estimating household monthly income. Moreover, the absence of general information about respondents and the fact that collected data are anonymous, make it unfeasible to use additional information from other surveys or administrative sources, as has been done in some other MS (Eurostat 2022);

- **Absence of an interviewer**: in Luxembourg, the survey adopts a self-administered format without any interviewer involvement. This absence of direct interaction may lead to a notable degree of inconsistencies among answers;

- **Answer format**: the question about household income adopts a single-choice close-ended format. Respondents are asked to choose from predefined intervals of income values, each with different ranges. These intervals can vary by increments of €500, €1,000, €2,500, and more. For example, the highest category, "More than €12,500", could include values such as €13,000 or even €60,000. Such response formatting amplifies precision-related concerns.

The purpose of this study is to find a viable solution to address this issue and make income a usable variable within the context of the EHIS or other health-related surveys.

2 Methodological phases

The research is divided into three parts, each representing different phases of the work:

2.1 Literature Review and Research

The initial step involves a comprehensive review of existing literature on the imputation of household income in health surveys. Moreover, the review was focused on identifying relevant methodologies, techniques, and best practices for handling MNAR data in health-related or general surveys. This involved examining studies, articles, and publications from reputable sources to gain insights into the current state-of-the-art approaches. The different steps that lead to the selection of the final papers for this study are represented in Figure 1.

We then extracted information from the papers and synthetized it to understand what has already been done in the literature to impute MNAR values. Among the reviewed papers, only 20% focused on imputation methods specifically for categorical data. Furthermore, fewer than 10% addressed income imputation, and in these studies, income was treated as a continuous variable.

[2]There are no questions regarding first and an eventual second source of income, or benefits from the Government.

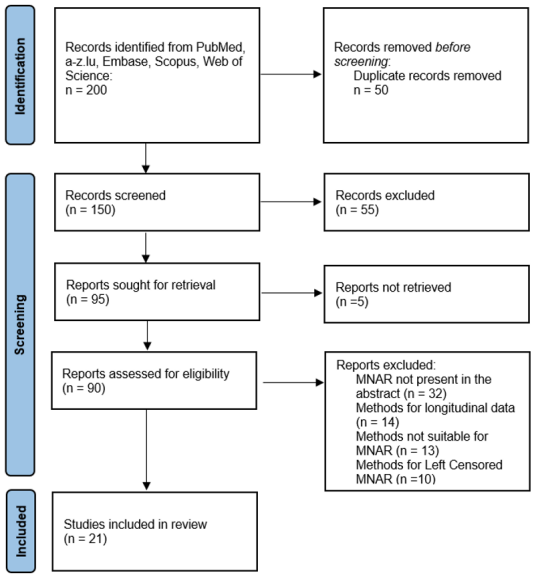

Figure 1: PRISMA diagram for the literature review.

A total of 94 distinct imputation methods were identified across the papers. The methods that emerged as both widely used and highly effective were:

1. Random Forest;
2. k-Nearest Neighbors;
3. Multivariate Imputation by Chained Equations (MICE).

The papers utilized a variety of evaluation metrics for assessing the performances of the imputation algorithms. For continuous missing data, the most commonly employed metrics were the Root Mean Square Error (RMSE), its variants, and Bias. For categorical missing data, Sensitivity, Specificity, and the Macro F1 score were the predominant evaluation criteria.

2.2 Analysis of the EHIS Income Data

The second phase includes an examination of the EHIS income data, both to gather useful information needed for imputation and to check for inconsistencies among the responses. To do that we performed two different analysis on the income data: 1) A first analysis to describe the main characteristics of three groups of individuals: i) those who provided income information, ii) those who left the question unanswered (NA) and iii) those who selected the option "I do not wish to answer" (IDNWTA). 2) A second analysis describing only the group of respondents that selected the lowest income category (less than €1,000). These two analysis aim to analyze the income variable from two different perspectives: the first one is needed to search possible patterns among individuals without information on income (those who did not answer and those who answer IDNWTA), while the second one aims to check the reliability and quality of the answer to the income question. More details on these two analysis are reported in the following sections. It is important to notice that the Luxembourg version of the EHIS has 4,504 observations that responded to 117 questions in 2019 and, 4,004 observations that responded to 104 questions in 2014. As mentioned before, approximately

24% of the 4,504 observations (in 2019) lacked income information (which corresponds to 1,097 individuals).

2.2.1 Characteristics of the three income groups

The first analysis conducted focused on examining various socio-demographic variables across three groups of individuals: 1) those who left the question unanswered (NA), 2) those who selected the "I do not wish to answer" (IDNWTA) option, and 3) those who provided their income. This approach allowed us to identify specific patterns that could help explain the reasons behind missing income data.

For example, the light red and blue curves in Figure 2, particularly the difference between the first two quartiles (22.0 years and 35.7 years), clearly suggest that different factors may be responsible for the missing values. Younger individuals might have left the household income question blank due to a lack of knowledge, while older individuals, who know their income but prefer not to disclose it, may have opted for the "I do not wish to answer" option.

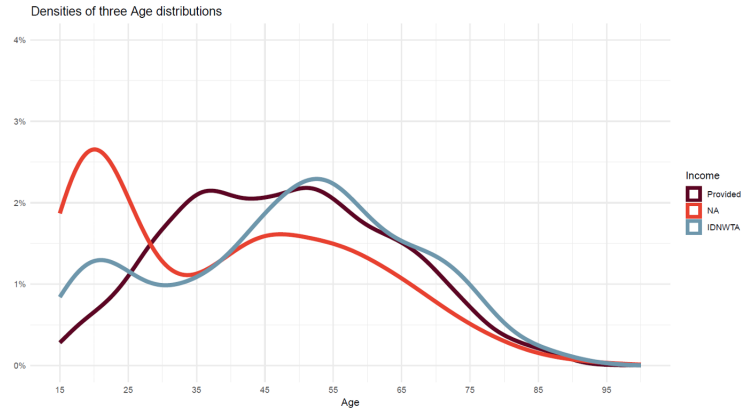

Figure 2: Densities of the age distributions in the three income groups.

The Kruskal-Wallis test revealed significant differences among the three groups (p-value $<$ 0.001). Post-hoc comparisons using Dunn's test with the Benjamini-Hochberg adjustment showed that the ages of individuals with missing data significantly differ from both those who chose not to respond (p-value adjusted $<$ 0.001) and those with available income (p-value adjusted $<$ 0.001).

Another interesting difference between these three groups of individuals, regards the education levels: there is a statistically significant difference between the three groups of individuals (Chi-Squared test p-value $<$ 0.001). Table 2.2.1 shows the different level of education according to their mechanism of answering to income question. The disparity in education levels between respondents who left the question unanswered and those who selected "I do not wish to answer" might be helpful for the income imputation by suggesting that the two types of missing income individuals must be treated differently. This because the individuals with a secondary education who left the question unanswered are expected to have lower income values than the ones who deliberately decided to not respond.

Education Levels	Income: NA	Income: IDNWTA	Income: provided
Primary	27.3	19.5	17.5
Secondary	**35.1**	28.3	31.2
Tertiary	28.6	**48.7**	**48.2**
No Information	9.0	3.5	3.1

Table 1: Education level distribution, divided by answer to the income question. Values in percentage. Highest value for each column marked in bold. IDNWTA = I do not wish to answer.

The analysis of these variables, provides a basis for adopting **two different approaches for income imputation**: one for individuals in the NA group following a MCAR or MAR mechanism, and one for individuals in the IDNWTA group following a MNAR mechanism. The analysis performed on all the other socio-demographical variables, indicates that there are differences between the three groups, suggesting that a tailored imputation approach may be justified. To determine whether the missing values in the NA group follow a MCAR or MAR pattern, we conducted Little's MCAR test on the dataset, excluding the IDNWTA units, and rejected the null hypothesis that the missing pattern resembles MCAR behavior (Little's MCAR test p-value < 0.0005). This indicates that the data cannot be considered MCAR. Additionally, the analysis performed in this section, revealed that respondents who skipped the question shared common characteristics, notably being predominantly younger individuals, females and with lower education. This, combined with the MCAR test result, led us to classify the NA group of missing values as MAR.

2.2.2 Characteristics of the group of respondents in the lowest income category

The second type of analysis focused on respondents who selected the lowest income class, declaring a monthly household net income below €1,000 (hereafter referred as Low-Income Individuals, or LII). The objectives of this analysis were to identify the demographic and socioeconomic characteristics of LII and assess the accuracy and reliability of the collected data.

The analysis revealed that the LII group comprises 73 individuals, representing just 1.60% of the total sample. An examination of their characteristics showed that a significant portion of LII were young, with 25% being under 20 years of age and 50% being younger than 27 years old. More concerning, however, is that 58% of LII reported living in households with three or more members. This raises potential issues regarding the accuracy of their reported income, as it is challenging for a household of this size to rely on less than €1,000 per month in Luxembourg. This is supported from results shown in Figure 3, where, individuals living in households composed of more than two people were younger than those living in households with one or two people (excluding the single individual in a seven-person household). This trend is particularly noticeable in three-person households, where the third quartile age is 22 years. This pattern suggests that some respondents, especially younger ones, may have misunderstood the question and reported their personal monthly income instead of their household one.

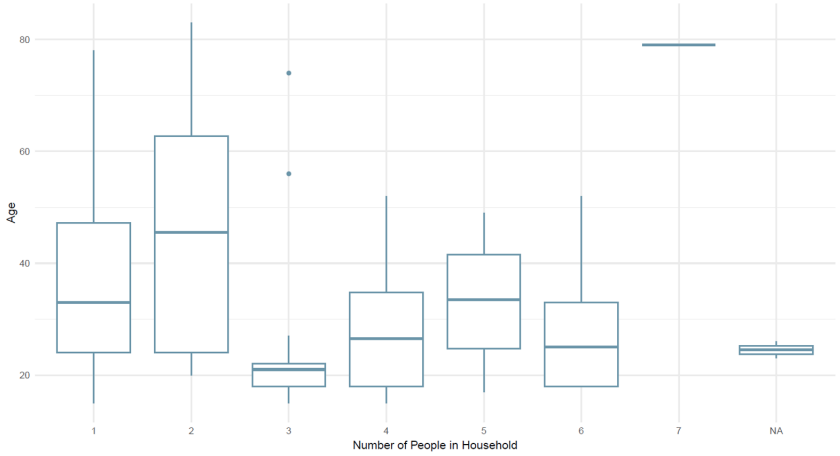

Figure 3: Age distribution by household size.

For the LII we also analyzed their five main working status classes (employed (19.72%), unemployed (7.04%), retired (8.45%), students (43.66%) and people who fulfill domestic tasks (9.86%)) and described their characteristics (e.g. age, sex, number of people in the household, household type, educational level, migration status and citizenship). Through this analysis, individuals who might have provided unreliable information were identified. Specifically, those reporting a household of three or more members and a household type other than "other type of household" (N=26) were flagged as potentially unreliable (36% of the LII), since we thought that these kind of households could not rely on less than €1,000 per month in Luxembourg. However, we retained the data from individuals who reported living in an "other type of household", as this group may include students or young workers residing in shared accommodations. In these cases, it is possible that the individuals considered the number of people co-living in the apartment as the household size. Nevertheless, the income values for the flagged individuals were excluded and set as NA to prevent potential bias in the results of future analysis.

The findings underscore potential problems with income data obtained from EHIS. The difficulties encountered with LII may reflect broader issues affecting also respondents across various income classes (e.g. general misunderstanding of the question). However, it is impossible to check the reliability within the other income classes. To enhance the accuracy of income data in future EHIS waves, we recommend to clarify the income question, emphasizing that it refers to household income. This adjustment could help to reduce the risk of respondents providing their own income instead of household income.

2.3 Evaluation of selected methods

Following the initial literature review and the income analysis process, different methodologies for imputing household income were selected: Random Forest (RF), Extreme Gradient Boosting (XGBoost), k Nearest Neighbors (kNN), Support Vector Machines (SVM) and the Multiple Imputation by Chained Equations (MICE) algorithm. To determine which of the se-

lected algorithms performs best for imputing the data, we applied them to different simulated datasets. In this section, we will describe the steps involved in this simulation study.

2.3.1 Test dataset creation

In order to generate the datasets for the simulation, we used the complete case dataset from 2019 EHIS and artificially created missing values. This step involved selecting all individuals who had zero missing values in all socio-demographic variables: age, sex, birthplace, district of residence, citizenship, income, main working status, job status, type of occupation, part-time or full-time job, NACE code of the occupation in class, education level, language of the questionnaire, questionnaire type (online or paper version sent by mail), household type, number of people living in the household, number of people aged less than 13 years old living in the household, and migration status. The complete case dataset was composed by 2,954 people and missing values were created following the MCAR, MAR, and MNAR missing patterns in six different proportions: 5%, 10%, 15%, 20%, 25%, and 30%. This choice was made to identify which imputation method performs best with different proportions of missing data. In the end, we created 18 different datasets (six for every missing data mechanism) using a revised version of the missing data generation algorithms proposed in (Fouad et al. 2021).

MCAR algorithm: the algorithm for the MCAR generation is the easiest one to develop and apply. It generate a random number from one to the total number of complete cases and it will remove the income value of the randomly selected unit based on the required missing percentage. **MAR algorithm**: the algorithm for the MAR generation is less straightforward. In this case, a condition term, based on variables other than income, is needed to select the units at risk of losing their income value. If the condition is, for example, "Sex equals to female", this means that missing values will be generated between female respondents only. Since we noticed (with the analysis described in Section 2.2.1) that in 2019 EHIS data, there were more missing values among young people and women, we created the MAR missing values among women and individuals younger than 30. In this way we had a higher percentage of missing values among young females. **MNAR algorithm**: to generate MNAR missing values, we used the MAR algorithm with the income variable in the condition term. In this way, the algorithm generated missing income values based on the income variable itself, which is the concept behind the MNAR mechanism. In order to simulate a MNAR situation that follows the idea of the salary taboo (Tratchtman 1999), we created missing income values among individuals with income class lower than 6 and greater than 12, which represents the classes with the first and third quartile of income respectively.

2.3.2 Simulation results

After creating 18 different datasets, we applied the five selected algorithms with different set of hyperparameters to discover what were the best performing ones. As shown in Figure 4, Random Forest algorithm consistently performs better, particularly when using the default hyperparameters (mtry = 4). However, the version of the algorithm that utilizes CV for hy-

perparameter tuning (RF cv) yields comparable results in terms of weighted Kappa, and due to the generalizing property of the cross-validated method, this latter approach is preferred.

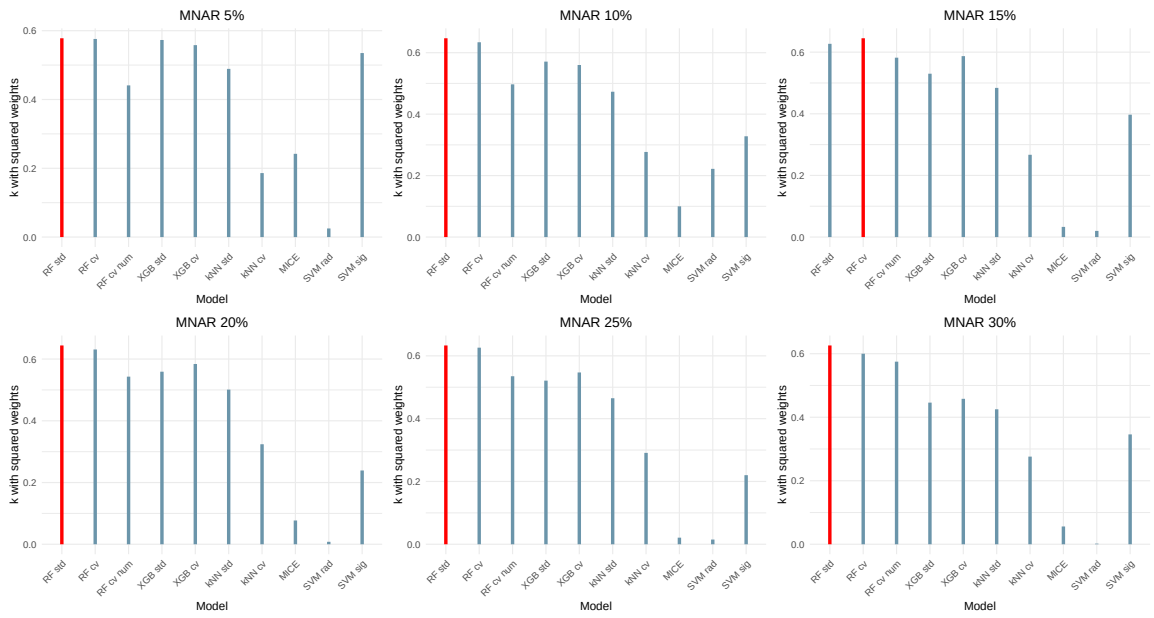

Figure 4: Imputation algorithm's performance on dataset with simulated MNAR values, by Weighted Cohen's K with squared weights. RF: Random Forest XGB: Extreme Gradient Boosting, kNN: k-Nearest Neighbors, MICE: Multivariate Imputation by Chained Equations, SVM: Support Vector Machines, num: income classes treated as numerical, cv: cross validation has been applied, std: no cross validation used, only default parameters.

Since the results for the MAR datasets were similar, we chose to use the Random Forest algorithm to impute both the IDNWTA and the classical NA missing income values.

3 Imputation of the real data

After having chosen the Random Forest for both the MAR and MNAR data, for the final imputation, we divided the real dataset into two separate ones: the first dataset included all data except the IDNWTA units, resulting in a dataset with 3,960 units, where the only missing income values were the NAs (i.e., the missing income values from people who skipped the question). The second dataset excluded the NAs, resulting in a dataset with 3,925 units, where the only missing income values were the IDNWTA ones.

We then implemented the Random Forest algorithm, using hyperparameters selected from the simulation study. Specifically, we set the number of features randomly selected at each split in the decision trees (*mtry*, the RF hyperparameter) to 14 for the MNAR dataset (which contained only the IDNWTA as missing income values) and 12 for the MAR dataset (which contained only the NAs as missing income values). This choice of hyperparameters was based on the fact that, for the income variable, the missing values represented 14.6% of the total units in the NAs dataset and 13.8% in the IDNWTA dataset. We derived the *mtry* values from the cross-validated Random Forests applied to the simulated MAR and MNAR datasets

with 15% of the values missing.

We then merged the two datasets to obtain one complete dataset without any missing income value. In Figure 5 is presented the final income variable distribution (with values imputed for NA and IDNWTA) compared with the income variable without missing information (complete cases). We observed that the imputation gave more realistic results, especially by increasing the extreme classes that we thought to be underrepresented. Moreover the two distributions are significantly different (Chi-squared p-value < 0.0005), meaning that the imputation is worth to be applied.

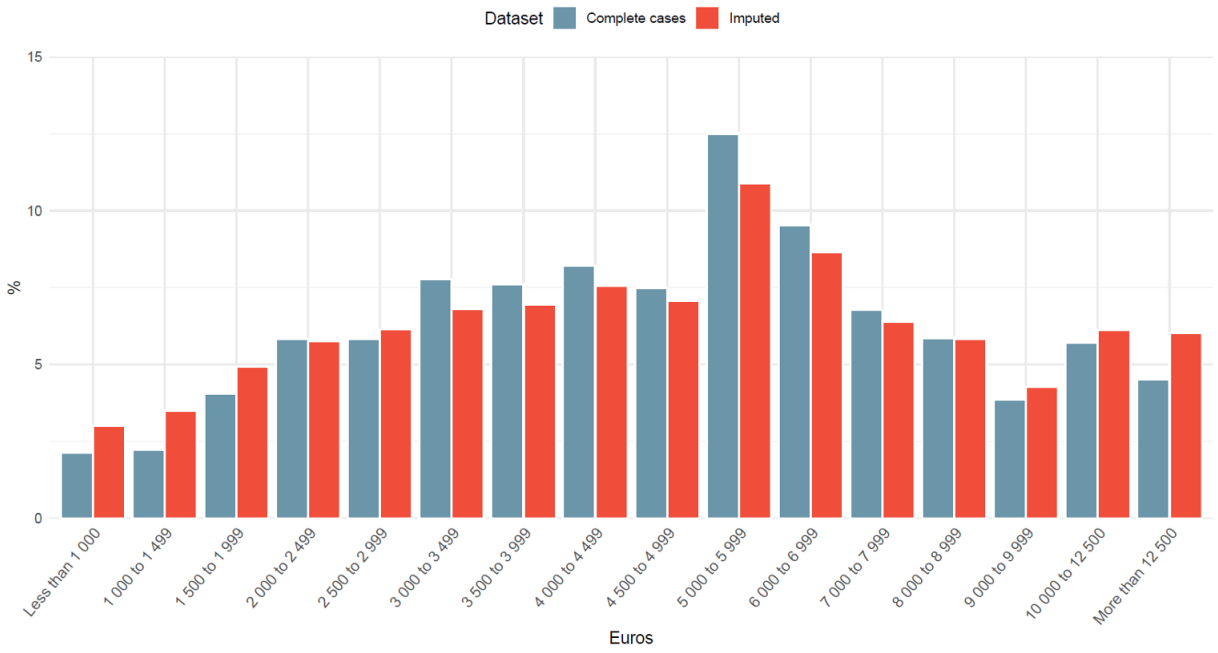

Figure 5: Distribution of the final imputed dataset and the complete cases one.

References

Eurostat (2022), *Quality report of the third wave of the European Health Interview Survey*, edited by Eurostat.

Fouad, K. M., M. M. Ismail, A. T. Azar, and M. M. Arafa. (2021). 'Advanced methods for missing values imputation based on similarity learning',*PeerJ Comput Sci*, 7: e619.

Jakobsson, U. and Westergren, A. (2005), Statistical methods for assessing agreement for ordinal data, *Scandinavian Journal of Caring Sciences*, 19, pp. 427-431.

Kim, S., Egerter, S., Cubbin, C., Takahashi, E. R., and Braveman, P. (2007), Potential implications of missing income data in population-based surveys: An example from a postpartum survey in California, *Public Health Reports*, 122, pp. 753-763.

Lee, J. H. and Huber, J. C., Jr. (2021), Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much?, *Iranian Journal of Public Health*, 50, pp. 1372-1380.

Little, R. J. A. (1988), A Test of Missing Completely at Random for Multivariate Data with Missing Values, *Journal of the American Statistical Association*, 83, pp. 1198-1202.

Trachtman, Richard. (1999). 'The Money Taboo: Its Effects in Everyday Life and in the Practice of Psychotherapy', *Clinical Social Work Journal*, 27: 275-88.