Belgian Science Policy Office
belspo

TAKE

Reducing poverty through improving
the take-up of social policies

Manon Bolland

Generating the TAKEMOD input dataset – Cleaning, Merging and Imputations

TECHNICAL REPORT

REPORT NR. 4/2022

DECEMBER 2022

Universiteit Antwerpen
CSB | Centrum voor Sociaal Beleid Herman Deleeck

LIÈGE université
ESPRIst
Études et évaluations

Federal Planning Bureau
Economic analyses and forecasts

Federal Public Service
Social Security

http://takeproject.wordpress.com

Abstract

In this paper, we describe the different steps and procedures we followed to build the input dataset that was used in TAKEMOD to estimate the non-take-up rates of different means-tested social benefits granted to vulnerable people in Belgium. The input dataset includes 4,986 observations: 1,909 TAKE survey's main respondents, 860 other TAKE survey's respondents (main respondent's household members who participated to the survey), and 2,217 household members who did not participate to the survey. We describe the procedure for matching the survey data with administrative data. Furthermore, we had to impute the missing values of three survey variables for the simulations: 'personal monthly disposable income', 'personal savings', and 'amount of social benefits received'. For this purpose, we used the technique of multiple imputation by chained equation (MICE).

Keywords: TAKEMOD input dataset, merging TAKE survey data, missing data, multiple imputations, MICE

Manon Bolland (Université de Liège), manon.bolland@uliege.be

Suggested citation:

Bolland, M. (2022), *Generating the TAKEMOD input dataset – Cleaning, Merging and Imputations*, TAKE Report Nr 4/2022, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp

1 Introduction

In this technical report, we describe the different steps and procedures we followed to build the input dataset that was used in TAKEMOD to estimate the non-take-up rates of different means-tested social benefits (social assistance, income guarantee for the elderly, the increased reimbursement, and the heating allowance). This input dataset includes variables constructed from administrative sources and variables constructed on the basis of information available in the TAKE survey. It contains a total of 4986 observations, which are the main respondents of the TAKE survey and the other members of their household. These households were constructed on the basis of the information mentioned by the main respondents in the TAKE survey, which means that they reflect the real composition of the household rather than the official composition available in the administrative sources. Indeed, the actual composition of the household is necessary in order to determine the eligibility for certain social benefits.

First, we describe in detail the different steps of data merging we had to perform in order to obtain the final database. Then, we explain the imputation procedure used in order to impute the missing data of three variables used in the simulation (personal disposable income, personal savings, and amount of social assistances benefits received).

This paper is part of the basic methodological documentation of the TAKE Survey, alongside the following documents[1]:

- The TAKE questionnaires
- A report on the development of the TAKE questionnaire (Janssens et al., 2022).
- A report on the TAKE sample design and its implementation (Goedemé, 2022).
- The fieldwork report (Vergauwen et al., 2022).
- A report on the microsimulation models available for the TAKE data (Janssens and Derboven, 2022).

Key findings of the TAKE project are available in the Final report of the TAKE project (Goedemé et al., 2022).

2 Data Merging

2.1 Dataset with all survey participants

The raw data from the TAKE survey are divided into two datasets: a dataset that contains the information about the main respondents as well as some household-level information, and a dataset that contains the information about the other household members who agreed to participate to the survey. The merging steps we followed are the following:

1. *Merging of the main respondent dataset with the other participants dataset, based on variable 'selectpID'*

   The combined dataset includes 2778 observations: 1909 main respondents + 869 other participants. There were 10 people who completed both the main questionnaire and the individual questionnaire. As in these cases the individual questionnaire did not

[1] All these documents are available on the TAKE website: https://takeproject.wordpress.com/.

provide any additional information, this was dropped from the dataset. Furthermore, the gecodeerd_insz is missing for 128 observations: 9 main respondents and 119 other respondents. Consequently, we created and assigned identifiers to those 128 persons. In particular, we created identifiers starting from value 80000 for the main respondents and identifiers starting from value 90000 for the other participants. The dataset obtained at this step contains only survey data for both the main respondents and the other participants.

2. *Merging of dataset from step 1 with the national register file for year 2019, based on variable 'gecodeerd_insz'*

   At the second stage, we merged the dataset obtained in step 1 with the national register data for year 2019 based on the variable 'gecodeerd_insz' in order to include the administrative data of all survey respondents in the dataset. From the 2778 observations, 139 could not be matched with the administrative file. We therefore only have survey data for these people in the dataset. These 139 observations include:

   - 17 main respondents
     - 9 individuals with missing gecodeerd_insz
     - 6 individuals who died in 2019 according to administrative data (we found them in the administrative file that includes all the people who died in 2019).
     - 2 individuals who are not included in administrative file for 2019
   - 122 other participants
     - 119 individuals with missing gecodeerd_insz
     - 3 individuals who died in 2019 according to administrative data (we found them in the administrative file that includes all the people who died in 2019).

3. *Merging of dataset from step 2 with the national register file for year 2018, based on variable 'gecodeerd_insz'*

   In order to find the administrative data of those people who have no missing gecodeerd_insz and who could not be matched with the national register file of year 2019, we merged the dataset from step 2 with the national register file for year 2018, again based on the variable 'gecodeerd_insz'. The 9 people who died in 2019 could be matched with the national register file of year 2018. At this stage, 130 observations are still not matched with their administrative data:

   - 11 main respondents
   - 119 other participants

   The 2 main respondents with no missing gecodeerd_insz, who were also not included in the national register file for year 2018 could however be found in the national register file for year 2017.

   The only individuals that could not be matched with the administrative file are those who have a missing gecodeerd_insz. This is not surprising since so far we have used the variable 'gecodeerd_insz' to perform the matching.

4. *Merging of dataset from step 3 with the national register file for year 2019, based on 3 variables: the reference person of the household, the age and the gender of the individual*

   Since people without a gecodeerd_insz in our dataset cannot be found directly (based on their personal identifier) in the administrative file, we tried to merge them with the national register file by using the information about the reference person of the household in which these persons are living, their age and their gender (the survey and the national register file both contain information on these three components).

   In the survey, each respondent was asked to report his or her age. However, the value given in the survey does not always correspond to the age in the national register file. In most cases, the difference between the two values is only one year. Therefore, in the administrative file, we created three versions of the age variable used for the merging: version 1 is simply keeping the age reported in the national register; version 2 is equal to the age from the national register file plus 1 year; version 3 is equal to the age from the national register file minus 1 year.

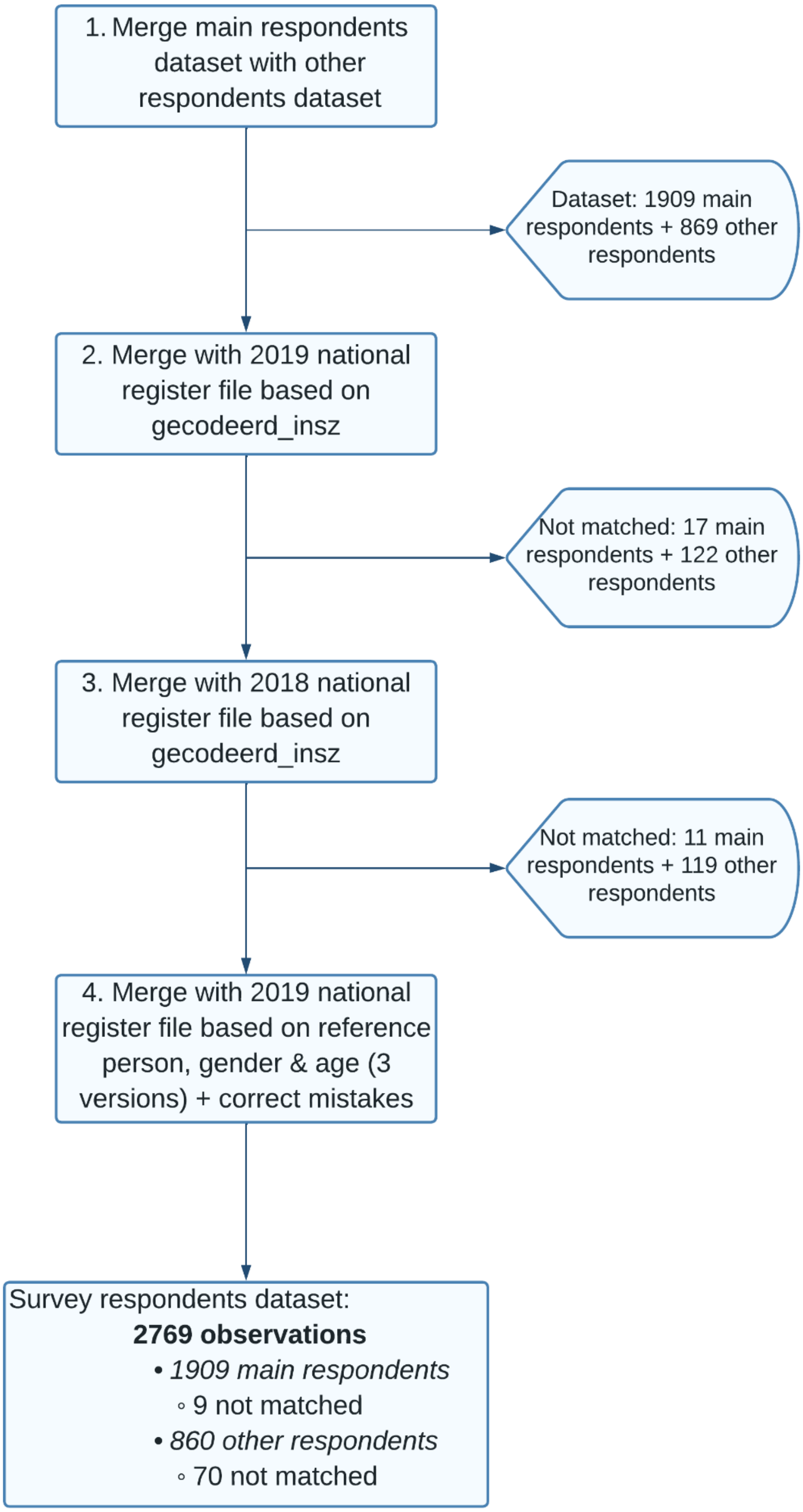
   By using version 1 of age for the merging (as well as the reference person variable and gender), we were able to match 34 additional other respondents with the administrative data. At this stage, from the 2778 observations, 94 are still not matched with the national register file (9 main respondents & 85 other participants). We looked at these 85 other participants in more details and noticed that for 9 of them the main respondents did not mention them to be part of the household (they did not report any information regarding these persons) and that no information was included in the administrative file. We assumed that this was an error and deleted these 9 observations from our dataset. We now have a dataset with 2769 observations, among which 85 observations with no administrative data (9 main respondents & 76 other participants).

   We then used version 2 of age (age from national register + 1 year) for the merging. Doing this, we could match some extra 6 other respondents to the administrative file. The dataset has now 79 observations that can still not be matched, 9 main respondents and 70 other participants.

   After this, we tried to merge the remaining observations with the national register by using version 3 of age (again, together with the reference person and gender), however, no observations could be matched.

Figure 1 shows a summary of the different steps that were followed to build our combined dataset for the survey participants. After these different steps, we obtain a database containing the survey data of 2769 participants: 1909 main respondents and 860 other respondents. Most of these people could be found in the national register file and therefore matched with their administrative data. In total, 79 observations, 9 main respondents and 70 other household members (all with a missing gecodeerd_insz), could not be merged with the administrative file. Consequently, we only have survey data for these people in our dataset.

Figure 1 - Steps for merging survey & administrative data for the survey participants


1. Merge main respondents dataset with other respondents dataset
Dataset: 1909 main respondents + 869 other respondents
2. Merge with 2019 national register file based on gecodeerd_insz
Not matched: 17 main respondents + 122 other respondents
3. Merge with 2018 national register file based on gecodeerd_insz
Not matched: 11 main respondents + 119 other respondents
4. Merge with 2019 national register file based on reference person, gender & age (3 versions) + correct mistakes
Survey respondents dataset:
2769 observations
• 1909 main respondents
◦ 9 not matched
• 860 other respondents
◦ 70 not matched

2.2 Dataset with all household members

In order to run the TAKEMOD microsimulation model, a dataset with one line for each household member is required. Therefore, the input dataset had to be expanded by adding data on household members without an interview, and most notably children (who were not invited for an interview). The main questionnaire included questions to create a partial 'household grid', i.e., we asked about how many people were living with them, their age, their gender and the relationship between the main respondent and every other household member. Based on this information, we created one additional row for each household member mentioned by the main respondent (adults and children) who did not participate to the survey. By doing this, we obtain a dataset of 4,995 observations, including 1,909 main respondents, 860 other household member participants, and 2,226 household members who did not participate to the survey.

The objective was then to add the administrative data of the additional household members to our new dataset. We again performed several rounds of merging with the national register file of year 2019 based on the reference person, the age and the gender. We used 5 different versions of the administrative age variable: age, age + 1, age – 1, age + 2, age – 2. From the 2,226 additional household members, 380 could not be merged with the administrative file. Furthermore, an extra person who did respond to the individual questionnaire could be retrieved based on version 4 of the administrative age variable (age + 2).

In a final effort to find the administrative data for all the individuals in our dataset, we attempted to search for the 458 individuals without personal identifier (9 main respondents, 69 other respondents, and 380 other household members) in the national register file. When the survey data on the total number of persons living in the household corresponded to the total number of persons mentioned in the administrative file, we compared the survey information on age and gender of all persons in the household with the information from the administrative file and tried to match these two data sources. We noticed that for some observations, the age difference between the one given in the survey and the one from the administrative file was larger than 2 years. Moreover, the gender reported by the main respondent for the other persons living with him or her did not always correspond with the gender reported in the national register (especially for children). Finally, we noticed that in some rare case, the main respondents included themselves in the information given about the other household members. We therefore deleted 9 observations among the additional household members (and only kept the already existing line for these main respondents).

The final input dataset for the simulation contains *4,986 observations*:

- *1,909 survey main respondents*
  - 1,900 were matched with administrative data
  - 9 not matched with administrative data

- *860 other survey respondents* (who are part of the main respondent's household)
  - 791 were matched with administrative data
  - 69 not matched with administrative data

- *2,217 other household members* (who did not participate to the survey)
  - 2075 were matched with administrative data

    - 142 not matched with administrative data

3 Imputing missing data

The TAKEMOD model requires complete data on all household members for key variables that are required for simulating their eligibility status. These variables include the individual's total monthly disposable income, the individual's personal savings, and the individual's monthly amount of social benefits received. While unit non-response at the household level is treated by applying a non-response correction to the survey weights, individual non-response (when a household member did not participate in the survey) and item-non-response (when a respondent refuses to respond or indicates not knowing the correct answer) were dealt with through imputation. With imputation, the missing data are replaced with an estimated value based on other available information. In what follows, we first describe the number of missing values for three key variables of interest. Thereafter, we describe the imputation procedure that we applied.

3.1 Personal monthly disposable income

Information on the individual's monthly disposable income was available in the survey for both the main respondents and the other household members who responded to the individual questionnaire. Table 1 below shows the missing data of this variable for the main respondents, the other survey respondents, and the other household members. This information is obviously missing for all those who did not take part to the survey.

Table 1 - Missing data of individual's monthly disposable income

	Missing	Total	%Missing
Main respondents	238	1909	12,47
Other respondents	40	860	4,65
Non-participants	2217	2217	100
Total	2495	4986	50

We assumed that children below the age of 16 years old had no personal income and therefore assigned a value of 0 to those who had a missing value for this variable. The missing data for personal disposable income after this first imputation are presented in table 2.

Table 2 - Missing data of individual's monthly disposable income after imputing a 0 value to children < 16 years old

	Missing	Total	%Missing
Main respondents	238	1909	12,47
Other respondents	39	860	4,53
Non-participants	989	2217	44,61
Total	1266	4986	25,39

3.2 Personal savings

Main respondents and other respondents were both asked questions related to their personal savings in the survey. Missing data for this variable are shown in table 3.

Table 3 - Missing data of individual's personal savings

	Missing	Total	%Missing
Main respondents	160	1909	8,38
Other respondents	160	860	18,60
Non-participants	2217	2217	100
Total	2537	4986	50,88

As for personal income, we assumed that children below the age of 16 years old had no personal savings and therefore assigned a value of 0 to those who had a missing value for this variable. The result obtained after performing this are summarized in table 4.

Table 4 - Missing data of individual's personal savings after imputing a 0 value to children < 16 years old

	Missing	Total	%Missing
Main respondents	160	1909	8,38
Other respondents	159	860	18,49
Non-participants	989	2217	44,61
Total	1308	4986	26,23

3.3 Amount of social benefits received

In the survey, the main respondent was asked to indicate who in their household received social benefits. After this question, the main respondent was asked to indicate the amount of social benefits received for only one recipient in the household, giving priority to himself/herself if he/she received any, otherwise to his/her partner. If he/she and his/her partner were not receiving social benefits, then he/she was asked to indicate the amount received by the person in his/her household about whom he/she has the most information. The first step was to assign the amount mentioned by the main respondent to the relevant recipient in the household. The non-recipients in the household were assigned a value of 0 for this variable. As shown in table 5, we could identify 551 social benefit recipients based on the information given by the main respondent. Moreover, we were able to match the amount reported to 418 recipients, which gives us a total of 133 missing values for the survey variable about the amount of social benefits received.

Table 5 - Missing data of social benefits

	Missing	Total	%Missing
Non-recipients	0	4435	0
Recipients	133	551	24,14
Total	133	4986	2,67

3.4 Multiple imputation by chained equation (MICE)

As mentioned above, we assumed that children below the age of 16 years old did not have any personal income and personal savings and therefore imputed the value of 0 to these children for these variables. However, the variable ‘personal disposable income’ still has 1,266 missing values, and this number reaches 1,308 for the variable ‘personal savings’. In addition, we also have 133 observations with a missing value our third variable of interest: amount of social benefits received. We decided to apply the method of multiple imputation by chained equation (MICE). This method imputes multivariate missing values on a variable-by-variable bases (Van Buuren, 2018). The main steps of the process are the following (Azur et al. (2011):

- **Step 1** - Simple imputation (e.g., imputing the mean) is performed for every missing data point of the variables included in the imputation model (including both the variables of interest, and the other covariates).

- **Step 2** - The mean imputations for one variable (e.g. *personal income*) are set back to missing.

- **Step 3** – Observed data of *personal income* are regressed on the other variables included in the imputation model. In our particular case, we included 16 variables in the imputation model, meaning that when *personal income* is the dependent variable in the regression, the other 15 variables (which also include the other two variables of interest) are used as explanatory variables.

- **Step 4** - The missing values for *personal income* are then replaced with predictions (imputations) from the regression model. *Personal income* is subsequently used as an independent variable in the regression models for other variables where both the observed and these imputed values of this variable will be used.

- **Step 5** - Steps 2 - 4 are then repeated for each variable that has missing data. This constitutes what is called a ‘cycle’ and results in all missing values being replaced with predictions from regressions that reflect the relationships observed in the data.

- **Step 6** - Steps 2 through 4 are repeated for a number of cycles, with the imputations being updated at each cycle. The idea is that by the end of the cycles the distribution of the parameters governing the imputations (e.g., the coefficients in the regression models) should have converged in the sense of becoming stable.

- **Step 7** – We keep the imputed values for personal disposable income, personal savings and social benefits received, and delete again the imputed values of the other covariates.

There are two major approaches for handling multivariate missing data: joint model (JM) imputation and multiple imputation by chained equations (MICE). The MICE method has numerous advantages compared to the JM technique (Azur et al. (2011); White et al. (2011); Wulff (2017)). As described above, with MICE, a series of regression models is run in which each variable with missing data is modelled conditionally on the other variables in the data, meaning that each variable with missing data can be modelled separately (Azur et al. (2011)). By contrast, the JM method requires a specified joint model for the complete data, however formulating the joint distribution of the data may be difficult with large numbers of variables and different levels of measurement. MICE offers more flexibility as it can handle different types of variables (continuous, binary, ordered categorical and unordered categorical) and incorporates restrictions, bounds and survey skip patterns(White et al. (2011); Wulff (2017)). The method is therefore very practical when working with large datasets, where missing values often occur in several variables. Finally, MICE is easy to implement and available in a lot of common software programs, including STATA, S-Plus, R, IVEware, and SPSS (Azur et al. (2011); White et al. (2011)).

One limitation of the MICE procedure is that it lacks theoretical justifications. However, this does not seem to be an issue in practice (White et al., 2011; Wulff, 2017). Another drawback is that choosing an appropriate imputation model is a difficult task (Wulff, 2017). The number of variables to include in the model, the appropriate functional form for the continuous variables, as well as the appropriate type of model are all difficult choices. When working with datasets that contain many variables, we may end up with large and complex imputations models (White et al., 2011). In practice, fitting these kinds of complex models may be impossible due to convergence problems. Moreover, it may be too computationally intensive for the software. Finally, MICE is based on the assumption that missing data are Missing At Random (MAR), which means that the probability of a value being missing depends on observed values and not on unobserved values (Azur et al., 2011). The problem is that this assumption cannot be tested without additional information about the process that generated the missing data (see Rhoads, 2012). However, as explained by Collins, Schafer & Kam, (2001), unless causes of missingness being strongly correlated with outcomes, the consequences of falsely assuming MAR are minor. White et al. (2011) explain that one must be very cautious when imputing missing data of variables that contain more than 30-50% of missing values because this amplifies the consequences of any departures from the MAR assumption and any misspecifications in the imputation models.

Most experts recommend making the imputation model as general as possible, by incorporating variables that are highly correlated with responses or explanatory as well as variables that explain the mechanism leading to missing data (Hardt et al., 2012).

Most experts recommend making the imputation model as general as possible, by incorporating variables that are highly correlated with responses or explanatory as well as variables that explain the mechanism leading to missing data (Hardt et al., 2012). To impute the missing values of the three variables of interest (personal monthly disposable income, personal savings, and amount of social benefits received), we constructed an imputation model which includes 16 variables in total (the three variables of interest and 13 additional predictors). We did not include too many predictors in our model to avoid convergence problems and too heavy computations that could cause the software to fail. When working with datasets that contain hundreds of variables or more, Van Buuren (2018) explained that using all variables as predictors is not optimal and instead recommends selecting no more than 15 to 25 variables. The missing values (absolute numbers and percentages) of the 16 variables included in our models are shown in table 6. The percentage of missing data is below 30% for all of them.

Table 6 - Missing data of variables imputed using MICE

Variable	Source	Type	Missing	Total	% Missing
Personal monthly disposable income	Survey	Continuous	1266	4986	25,39
Personal savings	Survey	Continuous	1308	4986	26,23
Amount of social benefits received	Survey	Continuous	133	4986	2,67
Age	Survey	Continuous	29	4986	0,58
Gender	Survey	Binary	84	4986	1,68
Household type	Survey	Unordered categorical	0	4986	0,00
Number of children in hh	Survey	Ordered categorical	5	4986	0,10
Number of adults in hh	Survey	Ordered categorical	5	4986	0,10
Marital status	Administrative	Unordered categorical	564	4986	11,31
Global (joint) net Taxable income	Administrative	Continuous	218	4986	4,37
Personal net taxable income	Administrative	Continuous	218	4986	4,37
Income from property	Administrative	Continuous	218	4986	4,37
Amount paid on private transfers	Administrative	Continuous	218	4986	4,37
Amount of social benefits received	Administrative	Continuous	218	4986	9,75
Household disposable income	Survey	Continuous	193	4986	3,87
Difference between household and personal disposable income	Survey	Continuous	1359	4986	27,26

Note: The administrative data are treated as regular missing values for respondents that could not be linked to the administrative dataset.

The predictors include both administrative and survey variables. In particular, we selected income variables which could potentially be correlated with our variables of interest, as well as some common demographic variables. We looked at the correlation matrix and found that the variable ‘personal monthly disposable income’ was significantly (at 5% level or better) correlated with all the variables from table 6 (the other two variables of interest included), except ‘number of adults’. This predictor was however significantly correlated with the variable of interest ‘amount of social benefits’, which on the other hand was not significantly correlate with ‘number of children’, ‘personal net taxable income’ (administrative variable), ‘amount paid on private transfers’ and ‘personal savings’. Finally, ‘personal savings’ was only significantly correlated with 5 variables. We tried to make the imputation model more general by including other variables from the TAKE survey or administrative sources that were correlated with either the variables of interest or with the other predictors but the inclusion of these variables in the imputation model led to convergence problems. We therefore decided to include only these 16 variables in the imputation model.

In STATA, we applied MICE by using the command *mi impute chained.* With this command, we need to list the variables that need to be imputed along with the univariate method used for imputing the missing values of each single variable. There are 9 methods available: *regress*, *PMM*, *truncreg*, *intreg*, *logit ologit*, *mlogit*, *poisson*, *nbreg*. In our case, we used the following 3 parametric methods: logistic regressions (*logit*) for binary variables, ordered logistic regressions (*ologit*) for ordered categorical variables, and multinomial logistic regressions (*mlogit*) for unordered categorical variables. To impute missing values of a continuous variable, either predictive mean matching (PMM) or normal linear regression can be used. However, White et al. (2011) recommend using predictive mean matching (PMM) to impute missing values of continuous variables that are non-normally distributed. This is a semiparametric method that combines both parametric and non-parametric techniques. The procedure works as follows (see Bailey et al., 2020; White et al., 2011). Assume that *y* is a variable we want to impute which consists of n_{obs} number of missing values (or recipients) and n_{mis} number of non-missing values (or donors). At the parametric stage, PMM uses the normal linear regression to obtain the predicted means, $\hat{y}_{\mathrm{i}}$, for the n_{obs} elements of y and the posterior predicted means, y_h^*, for the n_{mis} elements of *y*. For each $h = 1, \dots, n_{mis}$, it then finds a set of k donors that minimizes the distance between $\hat{y}_{\mathrm{i}}$ and y_h^*, $|\hat{y}i - y_h^*|$ $(i = 1, \dots, n_{obs})$. At the nonparametric stage, PMM randomly selects a single donor from this set of k donors and uses the observed value from this donor as the imputed value for recipient *h*. As the imputed values of a variable are sampled from the observed values of that variable, the distribution of the observed values is preserved in the missing part of the data (White et al., 2011). The advantage of this method is that it is less sensitive to model misspecification, such deviations from normality, non-linear associations, and heteroscedastic residuals. For these reasons and given that our continuous variables are highly skewed, we applied PMM as imputation method for our continuous variables, setting the size of the matching set equal to k=10 as recommended in the literature (Morris et al., 2014). For practical reasons, we decide to perform only one set of imputations[2].

Finally, we performed some imputation diagnostics to help determine whether the imputations for the three variables of interest are reasonable. Some recommended diagnostics in the literature include graphical comparisons of the observed and imputed data

[2] See table A1 in Appendix for coefficients estimates and standard errors from the univariate models obtained at the final imputation cycle for the three variables of interest.

(Nguyen et al., 2017). Furthermore, Stuart et al. (2009) also proposed numerical diagnostics, which consists of comparing the means and variances of observed and imputed values, to identify variables of concern. They suggest marking variables with the following features: 1) the absolute difference in means between the observed and imputed values is greater than 2 standard deviations 2) the ratio of variances of the observed and imputed values that is less than 0.5 or greater than 2. The descriptive statistics of the observed and imputed data for the variables 'personal monthly disposable income', 'personal savings', and 'amount of social benefits received' are presented in Table 7. In the last 3 columns of table 7, we also respectively display the absolute difference in means between the observed and imputed data, the 2 standard deviations threshold for this first numerical test proposed by Stuart et al. (2009), and the ratios of variances of the observed and imputed values. We also found it interesting to compare the results obtained with MICE when using PMM (semiparametric method) as imputation method for the continuous variables with the results obtained with MICE when using simple linear regressions (fully parametric method) as imputation method. The results show that the absolute differences in the means are acceptable for all three variables when using PMM as imputation method, while this difference is greater than 2 standard deviations for the variable "amount of social benefits received" when using simple linear regressions. Concerning the variance ratio criterion, none of the variables raise any concern for both imputation methods. When looking at the imputed data of 'amount of social benefits' obtained with MICE when using PMM as imputation method, we found that 80 observations receive a value of 0. These imputed values were however not appropriate since the main respondents of the TAKE survey mentioned that these persons were social benefits recipients. We therefore used the option *conditional* in STATA and specified that the missing values must be imputed based only on observed data in the sample of benefits recipients (who have a positive amount of social benefits) which consists of 418 people. The results obtained are reported in table 7 (see row 'PMM with condition on social benefits'). The variance ratio criterion is satisfied but the difference in means criterion not. However, as mentioned by Stuart et al. (2009), differences between observed and imputed values do not necessarily imply a problem, given that the characteristics of respondents and non-respondents can differ. As it is more logical for beneficiaries to receive a positive amount of social benefits, we kept the results obtained with the adjusted PMM imputation model.

Table 7 – Descriptive statistics of observed and imputed data & numeric imputation diagnostics (absolute difference in means criterion & ratios of variances criterion)

		Observed data			Imputed data			$\|(1) - (3)\|$	$2 * (4)$	$\frac{(2)^2}{(4)^2}$
		N	**Mean (1)**	**SD (2)**	**N**	**Mean (3)**	**SD (4)**			
PMM	*Personal disp. income*	3720	817.59	1024.38	1266	911.96	1136.38	94.37	2272.76	0.81
	Personal savings	3678	103.61	687.57	1308	90.82	583.75	12.79	1167.5	1.39
	Amount social ben. received	4857	87.18	300.13	129	216.64	420.01	129.46	840.02	0.51
PMM with condition on social benefits	*Amount social ben. received*	4857	87.18	300.13	129	955.74	386.92	868.56	773.84	0.60
Linear regression	*Personal disp. income*	3720	817.59	1024.38	1266	840.35	1036.91	22.76	2073.82	0.98
	Personal savings	3678	103.61	687.57	1308	104.91	701.01	1.3	1402.02	0.96
	Amount social ben. received	4857	87.18	300.13	129	926.21	373.04	839.03	746.08	0.65

In order to have a more complete picture of the data, we compared the distribution of the three variables in the observed, imputed, and completed data. Figure 2 shows the distributions of personal monthly disposable income when PMM was used as method of imputation. The distribution of income in the imputed data is really similar to the one in the observed data. We compare these results to the one obtained when using simple linear regressions as imputation models for the continuous variables (both PMM and simple linear regressions are implemented within the MICE framework). The latter are shown on figure 3. Since personal income (as well as the other continuous variables used in the model) is not normally distributed, PMM appears to be a much better choice for imputing this variable. The imputations from figure 2 seem reasonable and are therefore included in the TAKEMOD input dataset.

Figure 2 - Distribution of personal monthly disposable income in the observed, imputed, and completed samples (using PMM as imputation method)

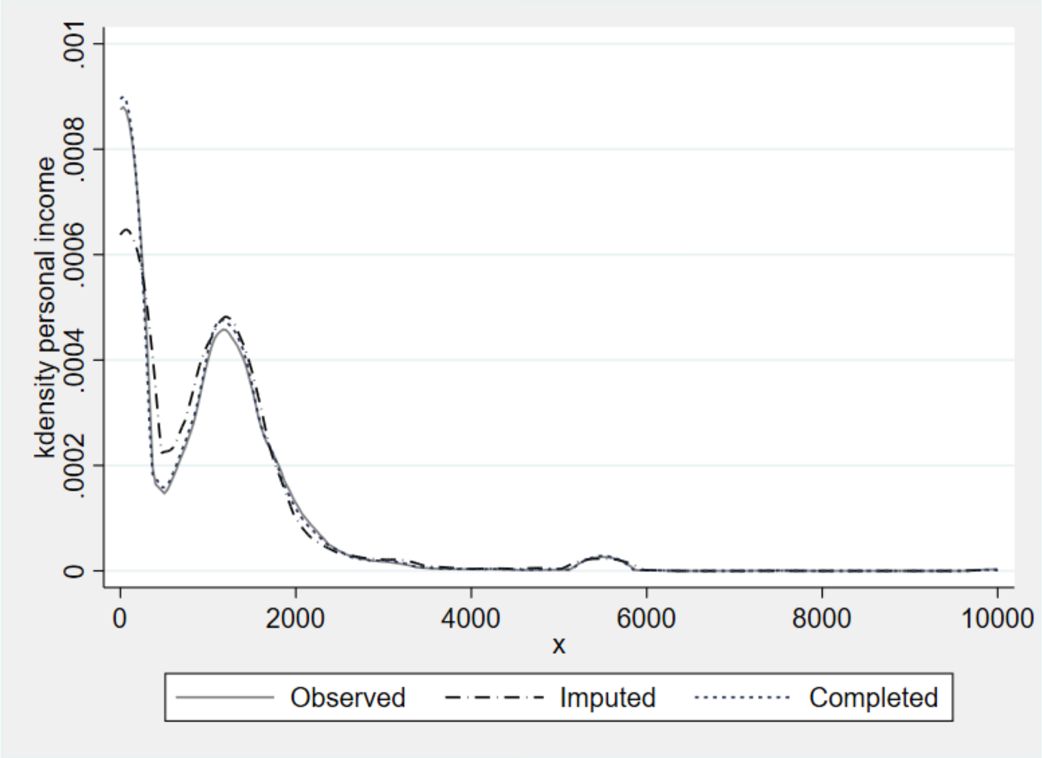
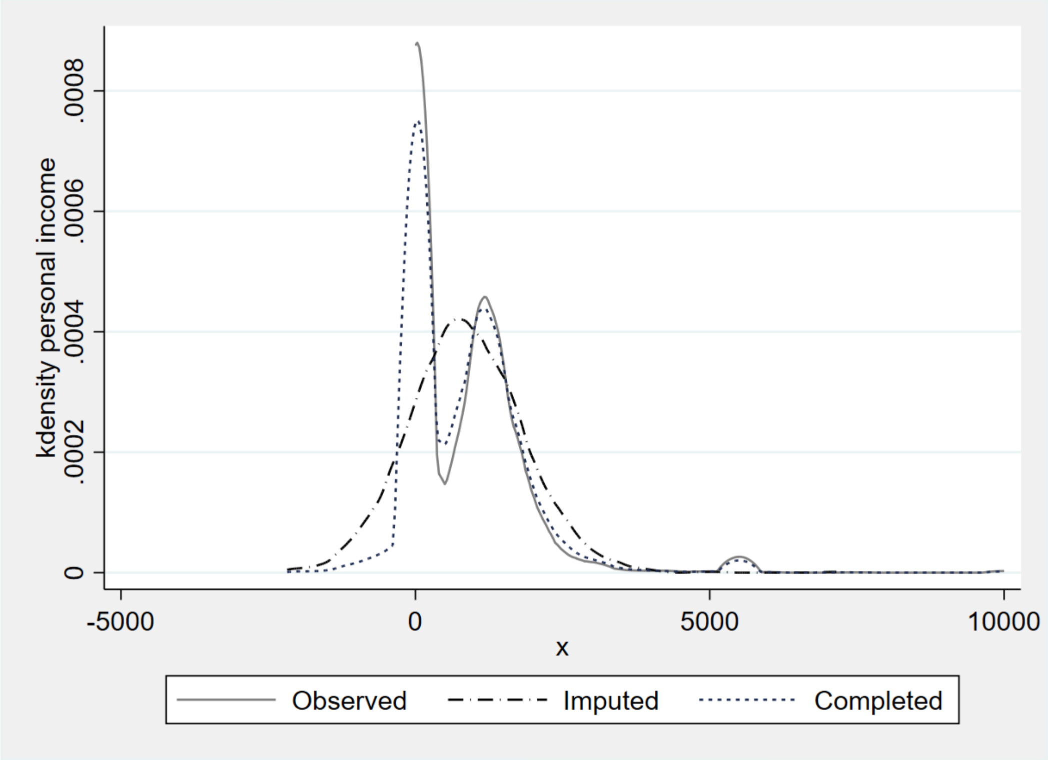


Figure 3 - Distribution of personal monthly disposable income in the observed, imputed, and completed samples (using linear regression as imputation method)



We performed the same comparison for the variable 'personal savings'. Figure 4 and figure 5 show the distribution of this variable in the observed, imputed and completed data for respectively the case where the PMM method is applied to impute missing values of continuous variables and the case where linear regressions are used. Again, since personal savings (as well as the other continuous variables used in the model) is not normally distributed, the results based on the PMM method are much more appropriate than those based on the linear regressions. As with personal income, the minimum observed value for personal savings is 0, while some of the imputed values were negative in the case of imputations based on a linear regression. Therefore, we also decided to use the results of personal savings obtained with PMM in the final input dataset.

Finally, we compared the results of the amount of social benefits received obtained with PMM and the one obtained when using linear regressions instead. The three distributions obtained with PMM are plotted in figure 6 and those obtained when using linear regressions are plotted in figure 7. Again, the results obtained with PMM are more reasonable than those obtained when using simple linear regressions. On figure 8, we also show the results obtained when specifying that the missing values of the amount of social benefits must be imputed based on the sample of benefit recipients only.

Figure 4 - Distribution of personal savings in the observed, imputed, and completed samples (using PMM as imputation method)

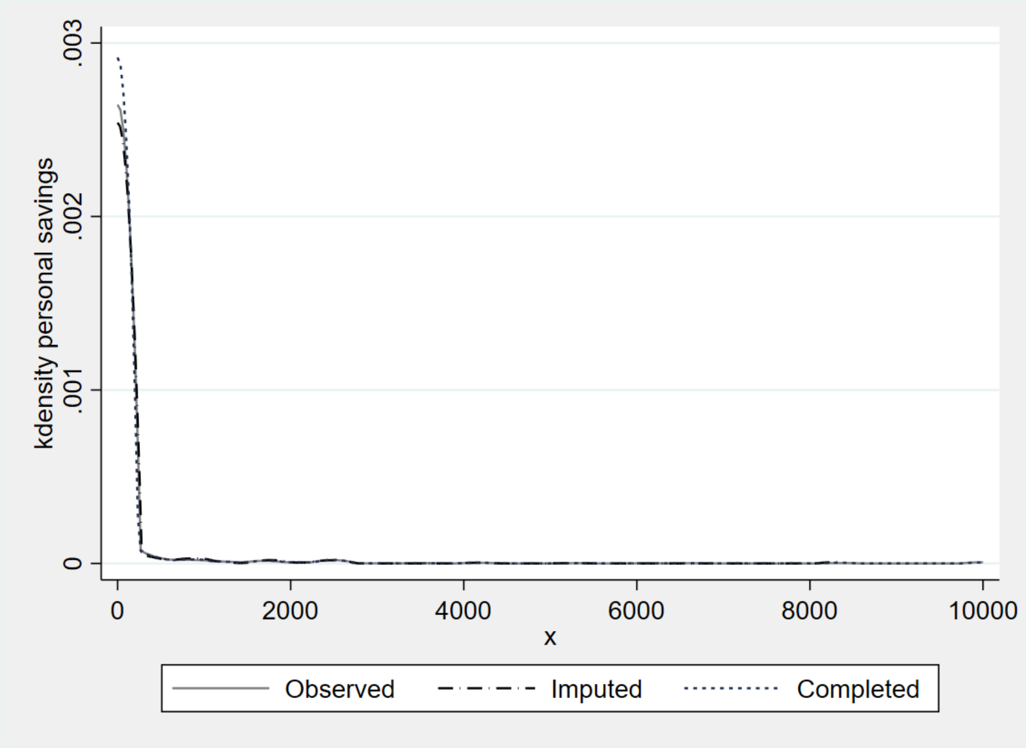


Figure 5 - Distribution of personal savings in the observed, imputed, and completed samples (using linear regression as imputation method)

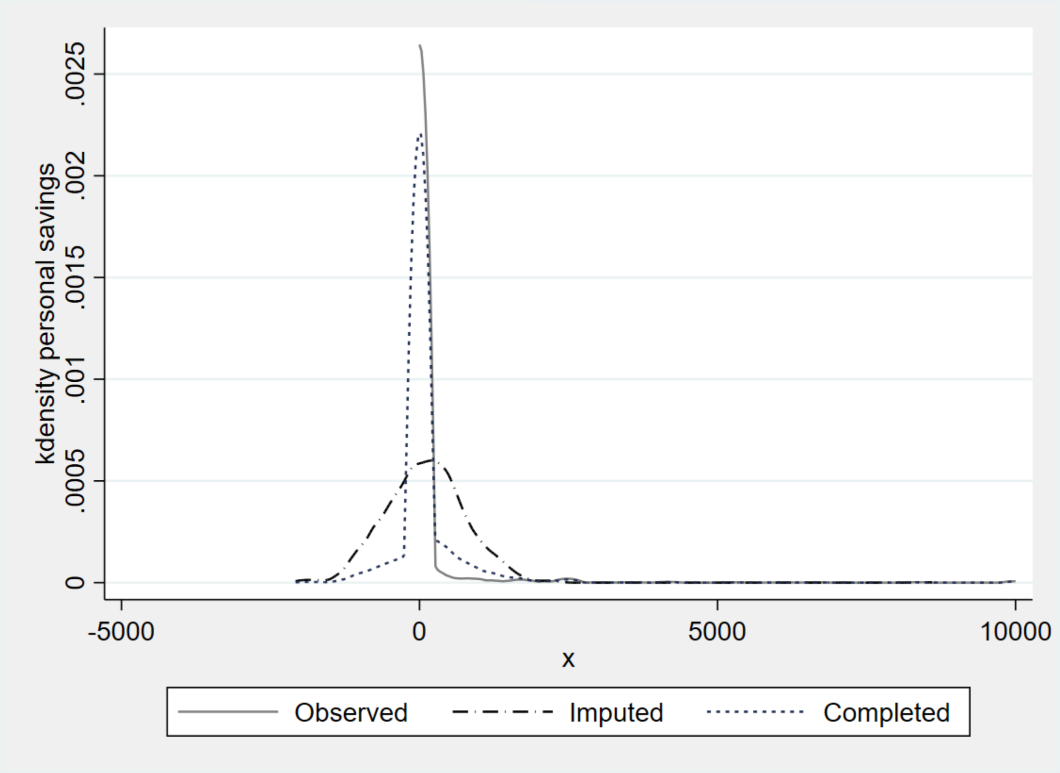


Figure 6 - Distribution of amount of social benefits received in the observed, imputed, and completed samples (using PMM as imputation method)

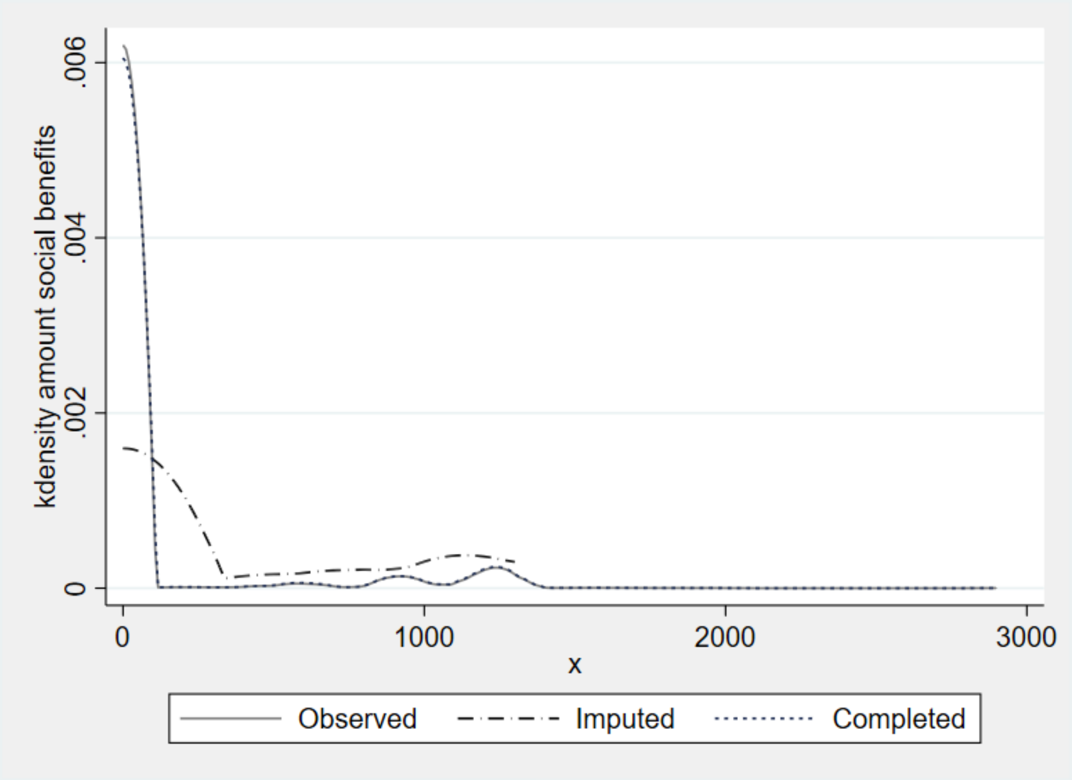


Figure 7 - Distribution of personal savings in the observed, imputed, and completed samples (using linear regression as imputation method)

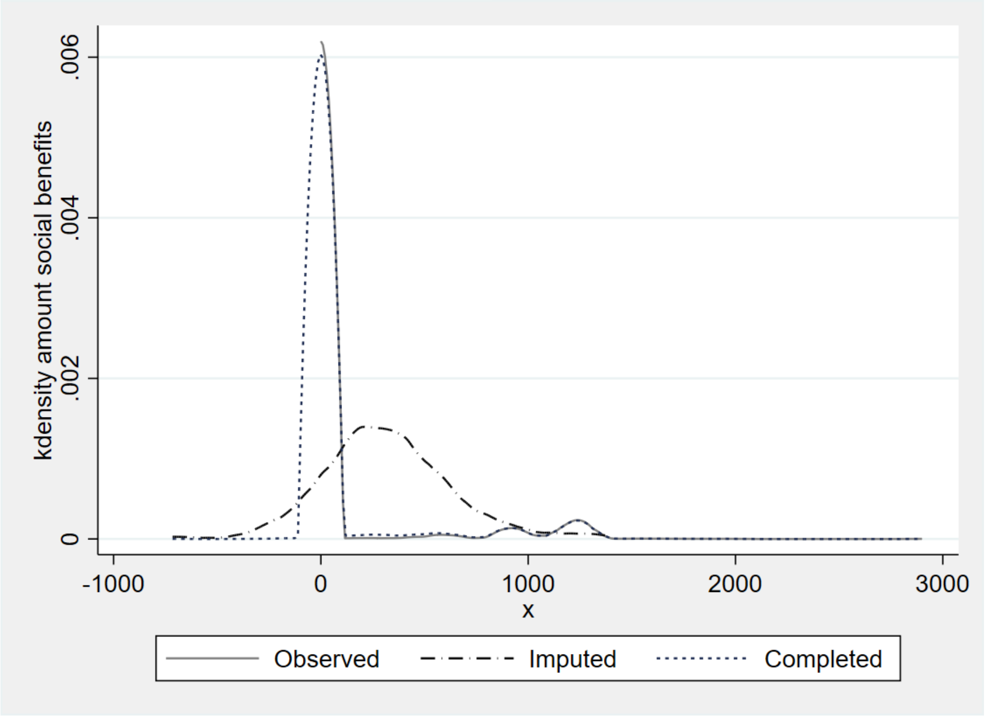


Figure 8 - Distribution of amount of social benefits received in the observed, imputed, and completed samples (using PMM as imputation method, with condition that imputed values must be > 0)

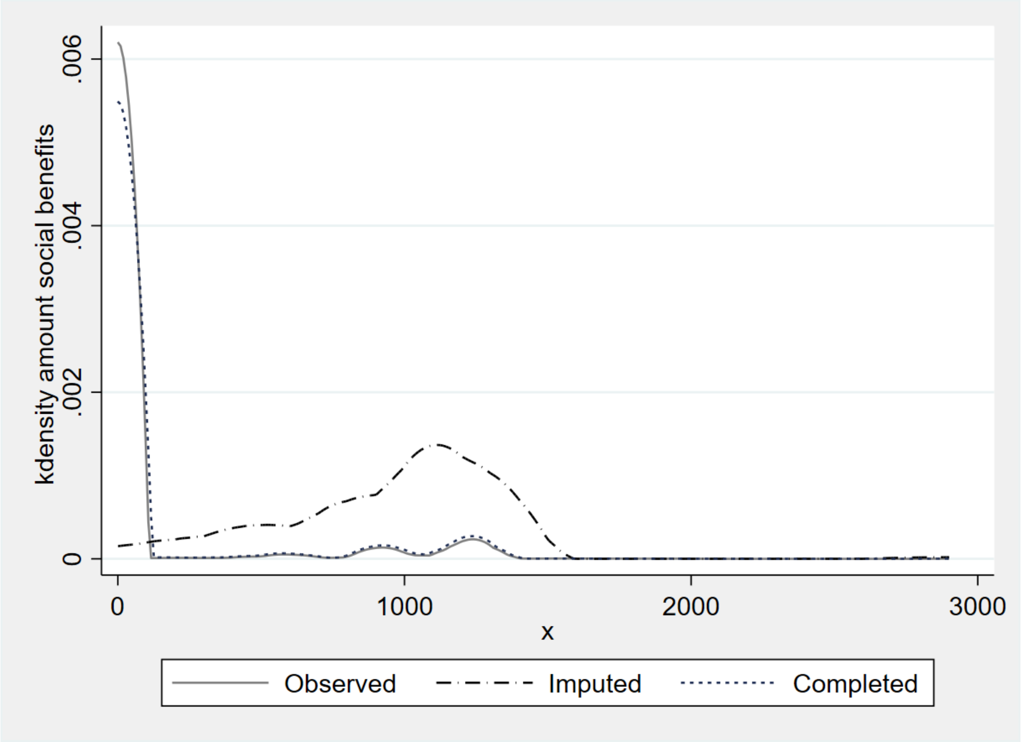


Figure 8 clearly shows that the shape of the distribution of the imputed values of social benefits differs from that of the observed values (but the distribution of observed values and completed values are similar). However, Marchenko and Eddings (2011) explain that this is not necessarily a problem. Indeed, the distributions should be similar only if the data are missing completely at random (MCAR), i.e., the missingness of the data is independent of both observed and unobserved data. This means that they may be different if the data are missing at random (MAR) or missing not at random (MNAR). Since the imputed values from figure 8 seem more reasonable, given the information reported by the main respondent in the survey, than the imputed values from figure 6, we kept the results from figure 8 for the final input dataset.

4 Conclusion

The input dataset used for the simulation of our non-take-up estimates of various means-tested benefits includes 4986 observations: 1909 TAKE survey's main respondents, 860 other TAKE survey's respondents (main respondent's household members who participated to the survey), and 2217 household members who did not participate to the survey. The dataset contains both administrative and survey data. Furthermore, we had to impute the missing values of three survey variables for the simulations: 'personal monthly disposable income', 'personal savings', and 'amount of social benefits received'. For this purpose, we used the technique of multiple imputation by chained equation (MICE).

5 References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), 40-49.

Bailey, B. E., Andridge, R., & Shoben, A. B. (2020). Multiple imputation by predictive mean matching in cluster-randomized trials. *BMC medical research methodology*, *20*(1), 1-16.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001), 'A comparison of inclusive and restrictive strategies in modern missing data procedures', Psychological Methods 6(4), 330–351.

Eddings, W., & Marchenko, Y. (2012). Diagnostics for multiple imputation in Stata. *The Stata Journal*, *12*(3), 353-367.

Goedemé, T. (2022). The TAKE Sample design and the TAKE sample: basic features of a new sample to study take-up and non-take-up of social benefits in Belgium. Report, Antwerpen: Centrum voor Sociaal Beleid – Herman Deleeck, Universiteit Antwerpen.

Goedemé, T., Janssens, J., Derboven, J., et al. (2022) TAKE: Reducing poverty through improving the take up of social policies. Final Report. Report, Brussels: Brussels: Belgian Science Policy Office.

Hardt, J., Herke, M. & Leonhart, R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med Res Methodol* **12**, 184 (2012). https://doi.org/10.1186/1471-2288-12-184.

Janssens, J. & Derboven, J. (2022) Measuring non-take-up with TAKEMOD. Technical report. Report, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.

Marchenko, Y. V., & Eddings, W. (2011). A note on how to perform multiple-imputation diagnostics in Stata. *College Station: StataCorp*.

Morris, T.P., White, I.R. & Royston, P., 2014. Tuning multiple imputation by predictive mean matching and local residual draws. BMC medical research methodology, 14(1), p.75

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: an overview and case study. *Emerging themes in epidemiology*, *14*(1), 1-12.

Rhoads, C. H. (2012). Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy*, *3*(1).

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American journal of epidemiology*, *169*(9), 1133-1139.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Vergauwen, J., Linchet, S. & Thiry, B. (2022) The TAKE Survey. Fieldwork Report. Report, Antwerp: University of Antwerp.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377-399. DOI: 10.1002/sim.4067

Wulff, J. N., & Jeppesen, L. E. (2017). Multiple imputation by chained equations in praxis: guidelines and review. *Electronic Journal of Business Research Methods*, *15*(1), 41-56.

6 Appendix

Table A1 – Coefficient estimates and standard errors from the univariate models obtained in the final imputation cycle.

	Personal monthly disposable income		Personal savings		Amount of social benefits received	
Predictors	**Coefficient**	**Standard error**	**Coefficient**	**Standard error**	**Coefficient**	**Standard error**
Age	12.169	1.091	1.478	.872	2.234	1.537
Gender	-80.341	26.893	-16.812	21.299	17.513	34.965
Household type	-100.092	15.072	-5.940	12.216	-35.360	20.501
Number of children in hh						
1	432.968	83.127	-59.004	68.364	429.012	120.640
2	397.838	91.972	-35.714	74.533	414.174	128.747
3	378.137	100.924	-27.108	81.173	405.564	139.870
4	420.185	105.147	-36.704	84.523	550.254	150.681
5	328.261	118.092	-64.842	94.296	781.197	160.644
6	252.950	155.983	-13.044	123.936	431.059	202.449
7	133.028	214.306	-85.759	165.107	-7.235	250.040
8	166.967	216.927	-103.179	171.169	625.530	351.143
9	380.139	250.458	21.683	197.050		
Number of adults in hh	234.625	55.539	-72.301	42.194	140.936	64.855
2	378.472	80.170	-96.612	62.596	163.035	112.501
3	336.973	85.443	-112.513	66.367	401.019	123.970
4	146.593	98.113	-151.644	76.503	411.017	148.713
5	663.815	128.424	-96.473	101.616	423.358	205.260
6	3.020	194.248	-129.517	151.430		
7						
Marital status						
2	99.693	46.916	-99.124	35.954	8.403	45.759
4	68.909	53.092	59.177	40.278	-13.853	41.709
5	-40.676	89.238	344.170	70.424	-115.907	91.115
Global net taxable income	.402	.022	.008	.018	-.071	.046

Personal net taxable income	-.002	.077	.158	.057	-.444	.615
Income from property	3.775	.685	7.840	.533	20.520	7.110
Amount paid on private transfers	.843	.353	5.241	.261	1.438	2.888
Amount of social benefits received (admin source)	.467	.049	-.077	.036	.097	.040
Household disposable income	.225	.013	.031	.011	-.020	.034
Personal monthly disposable income	-	-	.035	.013	.021	.026
Personal savings	.060	.020	-	-	.022	.109
Amount of social benefits received	.117	.052	-.028	.038	-	-

The TAKE project

Reducing poverty through improving the take-up of social policies (TAKE) is a Belgian research project financed by Federal Science Policy (Belspo). It aims to significantly improve the measurement and understanding of non-take-up of social policies in Belgium and to contribute to practical solutions. It is carried out by a research consortium consisting of the University of Antwerp (Coordinator), the University of Liège, the Federal Planning Bureau and the Federal Public Service for Social Security. The project makes use of a mixture of research approaches, including in-depth interviews with administrations, large-scale field experiments, microsimulations as well as a survey which brings together a unique blend of information collected through register data and face-to-face interviews. More information can be found on http://takeproject.wordpress.com.

For more information, please contact the Coordinator:

Tim Goedemé, PhD

University of Antwerp

tim.goedeme@uantwerpen.be