

# Statistical Matching for Combining the European Survey on Income and Living Conditions and the Household Budget Surveys

An Evaluation of Energy Expenditures in Bulgaria

*Britta Rude*  
*Monica Robayo-April*



**WORLD BANK GROUP**

Poverty and Equity Global Practice  
June 2024



**Reproducible Research Repository**

A verified reproducibility package for this paper is available at <http://reproducibility.worldbank.org>, click [here](#) for direct access.

## Abstract

Energy poverty has gained attention in the context of increasing energy prices and the recent energy crisis in Europe. However, measuring energy poverty and characterizing the energy poor are challenging, given that expenditure surveys (household budget surveys) often need more information to characterize the energy poor. Additionally, there is no consensus on how to measure and monitor energy poverty. It is also unknown how and why it differs from income poverty. While income poverty relies on a well-defined poverty line, energy poverty does not have a clearly defined energy poverty line that indicates the minimum energy necessary for satisfying basic needs. In addition, monetary poverty and other welfare measures are measured with income in EU countries using the European Survey of Income and Living Conditions. Therefore, it is not straightforward to characterize energy affordability among the monetary income poor or to estimate the overlap between official income poverty and energy poverty. This paper explores statistical matching as a potential strategy to overcome these

data challenges in the context of Bulgaria. Via data fusion, a unique dataset is generated that contains information on energy spending shares, income-based indicators of poverty and inequality, and additional variables on households' living conditions and welfare. For this purpose, the paper first generates a harmonized dataset, which consists of the European Survey of Income and Living Conditions and household budget survey data. It then employs different imputation models and chooses the best-performing one to impute energy spending shares into data. Based on the resulting dataset, it overlays energy poverty with monetary poverty. The findings show that a large share of the energy poor is not income poor, calling for differentiated policy measures to tackle energy poverty. Importantly, these findings depend on the underlying definition of energy poverty. This paper contributes to a growing body of literature exploring the potential of statistical matching to improve the current data environment in the European Union.

---

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://www.worldbank.org/prwp>. The authors may be contacted at [mrobayo@worldbank.org](mailto:mrobayo@worldbank.org) and [brude@worldbank.org](mailto:brude@worldbank.org). A verified reproducibility package for this paper is available at <http://reproducibility.worldbank.org>, click [here](#) for direct access.



*The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.*

# **Statistical Matching for Combining the European Survey on Income and Living Conditions and the Household Budget Surveys: An Evaluation of Energy Expenditures in Bulgaria<sup>1</sup>**

Britta Rude

Monica Robayo-April

JEL classification: O13, P28, Q42, D12, C15, C52

Keywords: energy poverty, statistical matching, poverty, data fusion, imputation, EU-SILC, Bulgaria

---

<sup>1</sup> This paper was prepared as part of the Poverty Program for Bulgaria in the World Bank's Global Poverty and Equity Practice. The study was carried out by a team composed of Mónica Robayo-Abril (Senior Economist, World Bank) and Britta Rude (Young Professional, World Bank). We thank the peer reviewer Paul Corral (Senior Economist, World Bank), for helpful comments and feedback.

## I. Introduction

**Rising energy prices over recent years demonstrate that many households are vulnerable to energy poverty.** The recent energy crisis in Europe raised concerns about energy affordability, especially among low-income households. Households with lower disposable income might spend a higher share of their overall income on energy. Therefore, they might be more affected by the energy crisis or rising energy prices, more generally speaking, than high-income households. The fact that poor people spend a larger share of their income on energy and might even decrease their energy usage in response to rising prices is often described by the concept of energy poverty. Energy poverty has negative welfare implications for poorer households. Energy poverty might result in severe health deterioration, especially in cold climates and during winter months. According to data by Eurostat, nearly one-tenth of the population in Europe, a total of 35 million people, is affected by energy poverty (European Commission, 2023).<sup>2</sup>

**Tackling energy poverty is one of the commitments of the European Union (EU), but identifying the energy poor is challenging, partly because of data limitations.** The EU targets both the mitigation and reduction of energy poverty (European Commission, 2023). To achieve this goal, policymakers apply a variety of tools, such as increasing energy efficiency, decarbonization strategies, and clean energy transitions. In addition, to facilitate knowledge exchange and good practices on energy poverty, the EU has launched the Energy Poverty Advisory Hub, the Horizon 2020 Energy Efficiency Calls, the LIFE Clean Energy Transition Program, and other initiatives. Still, identifying those who are most vulnerable to energy poverty remains challenging, both due to a lack of common definitions and reliable data sources (Robayo-Abril and Rude, 2023a). One important limitation is that monetary measures of energy poverty rely mainly on data gathered via household budget surveys (HBS), which often consist of small sample sizes and are only representative at high geographic levels. In addition, these surveys gather limited information on individual and household characteristics. Consequently, it is difficult to

---

<sup>2</sup> In this case, energy poverty is the share of households unable to keep their homes warm, which is a non-monetary measure of energy poverty.

take advantage of these datasets for targeting strategies, such as proxy means testing or geographic targeting. Another important limitation is that HBS often only collects limited information on disposable income, which makes it difficult to deduce reliable poverty indicators from these types of surveys.

**The current data environment in Bulgaria imposes serious restrictions on studying energy poverty.**

Detailed information on income, living conditions, and access to social protection programs in Bulgaria is mainly gathered as part of the European Union Statistics on Income and Living Conditions (EU-SILC). Expenditure data, on the other hand, is primarily collected by HBS. Only relying on HBS to study energy poverty would result in very limited analyses due to the following reasons. First, the information on household income is less reliable and rigorous in the case of the HBS, as the survey is mostly designed to capture expenditure patterns, not income patterns. Figures A.1 and A.2 in the Appendix show that the distribution of equivalized household income, as reported in the HBS and EU-SILC, differs in the case of Bulgaria. Second, the HBS does not contain information on social protection programs. Consequently, it is not possible to study if energy poor households are sufficiently covered by social protection schemes. Third, the HBS only has very limited information to study potential drivers behind energy poverty. Detailed information on housing conditions, for example, is only available in the EU-SILC. Ideally, to achieve efficient and effective policy interventions, one would want to observe all this information together. In addition, the current data environment also imposes limitations on identifying individuals who might be vulnerable to energy poverty.<sup>3</sup>

**Given that the sample of households between different types of surveys does not overlap, one cannot combine both datasets based on common household identifiers.** Due to the data environment described previously, one can only observe energy spending shares and energy poverty indicators in the HBS, while we can only observe reliable monetary poverty indicators as well as access to social

---

<sup>3</sup> The concept of vulnerability to poverty more broadly speaking has gained more traction in recent years due to populations being increasingly exposed to both natural and human hazards (examples are by Klasen and Waibel (2015); Gao et al. (2020); Rude and Robayo (2023)).

protection programs and other welfare indicators in the EU-SILC. Traditionally, it is possible to combine information via record linking by relying on common household identifiers between two surveys. This is not possible in the case of HBS and EU-SILC, as the sample of households included in each survey does not overlap.

**In this paper, we explore statistical matching techniques to combine both datasets and impute energy spending shares into the EU-SILC.** While datasets are traditionally merged via record linking, an increasing number of researchers has explored the potential of statistical matching for combining datasets without a common identifier, especially in the European Union (see, for example, Donatiello et al. (2016b) for an application to Italy; Serafino and Tolkin (2017) for an example using data from six European countries; Lamarche et al. (2020) for the European context more broadly speaking; Schaller (2021) for an application in Germany, France, and the Netherlands; Emmenegger et al. (2022) for an application in Germany). We build upon the existing body of literature and apply statistical matching techniques to impute the energy spending share in Bulgaria reported in the HBS 2019 into EU-SILC data collected in 2020. We use EU-SILC data from 2020, as income data in the EU-SILC always refers to the previous year (so 2019 in this case).

**Imputing energy expenditure shares and energy poverty into the EU-SILC facilitates future research questions, such as exploring the impact of rising energy prices on the poorest households.** Our imputation exercise is helpful given that we can identify if those households belonging to the lowest income quintiles are those most affected by energy poverty. In addition, having information on energy expenditure shares in surveys that provide information on monetary poverty indicators allows us to evaluate the impact of rising energy prices on the poorest households. Moreover, given that the EU-SILC contains detailed information on social protection programs, we can explore the impact of potential mitigation strategies in response to rising energy prices by combining information from both surveys. Lastly, given that EU-SILC survey questionnaires also contain extensive information on

individual characteristics of household members, more effective targeting strategies might become possible by relying on the imputed dataset.<sup>4</sup>

**We follow the approach developed by Ruben (1986) and identify matching variables, based on which we concatenate both datasets to ultimately employ multiple imputation methods.** There are many different approaches to statistical matching. Lewaa et al. (2021) provide an overview and divide the different methodologies into parametric, nonparametric, and mixed approaches. We follow early work by Ruben (1986) and first identify potential matching variables. These are variables that are part of both surveys and help to identify households that resemble each other. Nevertheless, not all overlapping variables should form part of the matching variables, but only those relevant to the target variable and similar in distributions across surveys (Serafino and Tonkin, 2017). We harmonize these variables with each other and apply a lasso regression to identify those variables that are most relevant to explain energy spending shares. Next, we concatenate both datasets based on these variables. While the household budget survey contains information on energy spending shares and energy poverty, these variables have missing values in the EU-SILC. After concatenating both datasets, we apply multiple imputation methods.

**To address one of the shortcomings in the literature raised by Lewaa et al. (2021), namely the lack of quality assessments, we explore several different imputation models.** We employ linear regression imputation models, predictive mean matching (PMM), and truncated regression imputation models. For each model, we explore three different imputation specifications by varying the inclusion of survey weights and the survey sampling design as additional matching variables. To evaluate the different models, we apply several validity tests. We compare the joint distribution of the imputed and observed energy spending share (overall and by subgroups), analyze if results differ across imputations, and compare the consistency of mean imputed energy expenditure by variables used in the statistical matching and those not used.

---

<sup>4</sup> We address these type of research questions in Robayo-Abril and Rude (forthcoming).

**We find that a weighted predictive mean matching model is the best-performing model.** The weighted PMM approximates the underlying model well and allows us to weight regressions by survey weights. This result is in line with the previous literature, showing that PMM works well in the context of imputations (Kleinke, 2018). Given that we plan to use survey weights in future analyses performed on the generated dataset, we consider the weighted PMM most appropriate. We show that results across the multiple imputations are consistent. In addition, the distribution of the average imputed energy spending share closely resembles the observed distribution of energy spending shares when restricting the sample by households' gender composition and income categories. Results differ slightly by matching variables but are sufficiently similar. The same applies to results by non-matching variables. They also hold when using an alternative number of neighbors for the PMM.

**We use the generated dataset to demonstrate that the energy-poor and monetary-poor overlap.** Based on the synthetic EU-SILC, we show that all monetary poor, defined via the relative at-risk of poverty measure, are also energy poor. At the same time, not all the energy poor are at risk of poverty. Moreover, we show that energy expenditure shares are higher for lower-income quintiles. An important limitation of our analysis is that these insights highly depend on our chosen measure of energy poverty. Our main results rely on an income-based measure of energy poverty and a 10 percent threshold to identify the energy poor<sup>5</sup>. The overlap between energy and monetary poverty measures differs when using different thresholds or measuring energy poverty from a different angle.

**Our paper makes an important contribution to recent efforts exploring the potential of statistical matching techniques to improve the data environment in the EU.** Over the years, there has been an increasing effort in the EU to explore the potential of statistical matching to combine information from different datasets (Leulescu and Agafitei (2013); Serafino and Tonkin (2017); Moretti and Shlomo (2022)). Our paper contributes to this emerging literature by exploring the potential of statistical matching techniques in the case of energy expenditure shares and energy poverty in Bulgaria. To the best of our knowledge, we are the first ones to explore the potential of statistical matching techniques

---

<sup>5</sup> Under this definition, energy poor households are those that spend more than 10 percent of their income on energy.



for the imputation of energy poverty indicators and energy spending shares. The literature so far has mainly focused on income-based or education-based indicators.<sup>6</sup> Moreover, to the best of our knowledge, we are the first to explore statistical matching techniques in Bulgaria's case.

**Moreover, we make an important contribution by investigating three different multiple imputation approaches and comparing them to each other.** This assessment can shed light on the quality, validity, and sensitivity of data fusion methods. To facilitate statistical matching in the future, we recommend including auxiliary variables – variables that are highly correlated with energy expenditure – in both surveys. Elevating data on energy spending for a small subsample of the EU-SILC could also improve the quality of data fusion in the context of energy spending shares.

**The paper at hand is subject to some important limitations.** First, we show that significant differences in the distribution of the matching variables between surveys persist after the intent to harmonize them. Future research could address this shortcoming by analyzing how to better harmonize variables between both surveys. Second, similar to most papers in the area of statistical matching, our results rely on the assumption of conditional independence, which is untestable in the current setup. Third, future research could analyze if our results differ from an approach that combines EU-SILC 2020 (with income reference year 2019) with HBS 2020.

**The paper is organized as follows.** Section II describes the methodology and data. Section III summarizes the main results of the imputation and how energy and income poverty overlap. Section IV concludes.

---

<sup>6</sup> Most of the work to date focuses on imputing income data collected as part of EU-SILC surveys into HBS. Leulescu et al. (2013) apply a similar research question to the EU-SILC and EU-LSF while Kaplan and McCarty (2013) use statistical matching techniques in the case of educational data from PISA and TALIS surveys. Similarly, Wiest et al. (2019) explore the potential of statistical matching for the analysis of wider benefits of learning in later life.

## II. Methodology and Data

### Empirical Model and Methodology

**Statistical matching techniques, also known under the term data fusion, require at least one donor and recipient dataset and overlapping variables.** Traditionally, in statistical matching, there is a donor dataset that contains information one wants to add to a second dataset lacking this information (here denoted as variable C). The second dataset often goes by the name recipient dataset and has information on a second variable of interest not included in the donor dataset (here denoted as variable I). Moreover, for statistical matching techniques to work, there needs to be a set of overlapping variables X, also often called matching variables. Overlapping (or matching) variables are variables that are included in both the recipient and donor datasets. Consequently, they facilitate the matching of both datasets.<sup>7</sup> Ideally, one would have a third dataset, which has information on all variables under consideration. This dataset is often called auxiliary data. If there is no such data available, then the data fusion relies on the assumption of conditional independence (Donatiello et al., 2016b). The assumption of conditional independence means that the overlapping variables X fully explain the relationship between C and I, a strong assumption that often does not hold in practice (D’Orazio et al., 2006).

**There are several possibilities to address the assumption of conditional dependence in the absence of auxiliary data (Donatiello et al., 2016b).** First, previous estimates on the relationship between C and I might be available (for example, correlation coefficients between C and I). Is it then possible to use these estimates to avoid the assumption of conditional independence. In the absence of such data, one can rely on at least one matching variable that is highly correlated with the target variables. If X and C are perfectly correlated, we can rely on the regression function in the donor dataset between both variables to predict C in the recipient dataset. While the set of matching variables that are

---

<sup>7</sup> Examples are information on the number of household members, the age and gender of household members, or their labor market status.



perfectly correlated with C might be limited in practical applications, a highly correlated matching variable might be sufficient to ensure that the relationship between C and I, given X, is close to its true underlying relationship (Donatiello et al., 2016b).

**The concept of statistical matching draws from the concept of imputations often applied to missing observations in survey data (Bacher and Prander, 2018).** There are three overarching approaches to imputing missing values of the added variable from the recipient data (C) in the donor data: parametric approaches, nonparametric approaches, and mixed methods (European Commission, 2014; Lewaa et al., 2021). Parametric approaches are regression imputations. They make use of the functional form between X and C in the donor dataset to estimate the missing values of C in the recipient dataset (Bacher and Prander, 2018). While they are more parsimonious than nonparametric methods, they rely on pure predictions and model specifications, so the accuracy of the underlying model. Nonparametric approaches rely on distance functions and identify similar observations to impute actual observed values from the donor to the recipient dataset (European Commission, 2014). They often involve the segmentation of observations by socio-economic characteristics or geography as well as choosing a suitable distance function (Bacher and Prander, 2018). Examples of nonparametric approaches are hot deck approaches. Mixed methods combine parametric with nonparametric approaches, such as the estimation of a stochastic regression imputation followed by a nearest-neighbor hot deck.

**Table 1 describes the basic concept behind the application in this paper: imputing energy expenditure shares to the EU-SILC in Bulgaria.** The goal of this paper is to create a synthetic dataset that contains information on income-based indicators, households' access to social protection programs, other indicators of households' welfare and living conditions, and energy expenditure shares, as well as energy poverty. To this end, we follow the approach in Table 1, which is highlighted in blue. More concretely speaking, we first create our variable of interest, the energy expenditure

share per household (denoted as C), in the HBS.<sup>8</sup> We then follow Rubin (1986) and concatenate the HBS and the EU-SILC.<sup>9</sup> To this end, we take advantage of overlapping variables X, which we harmonize previously and describe later in this paper. We decided to impute energy spending shares into the EU-SILC, as the EU-SILC is the richer dataset and contains many variables we are ultimately interested in to describe the energy-poor population in Bulgaria.

Table 1: Conceptualization of data fusion application in this paper (EU-SILC and HBS in Bulgaria)

EU-SILC (Imputation for expenditure C)	<b>Variables, which form part of EU-SILC and HBS</b>	<b>EU-SILC with information on I</b>
 <b>HBS with information on C</b>	Socio-economic information (X)	HBS  (Imputation for income I)

Source: Own elaboration based on Bacher and Prander (2018)

**Before applying multiple imputation techniques to the missing information on energy spending shares in the concatenated dataset, we address the conditional independence assumption.** As we do not dispose of an auxiliary dataset or auxiliary information on the relationship between I and C, we follow Donatiello et al. (2016b) and identify matching variables that are highly correlated with our variables of interest, the energy expenditure share, later in this paper. We identify one highly

---

<sup>8</sup> While we could also impute income data into the HBS (the approach in Column 3), we decide to do the reverse as the EU-SILC gathers more extensive information on individual characteristics, household wellbeing, and social protection programs. EU-SILC data contains a richer set of information and is therefore more attractive as a recipient dataset. Nevertheless, the European Commission (2014) recommends using the larger dataset as the donor. We do not follow this approach here, as we are interested in conducting microsimulations based on EU-SILC related variables in future projects.

<sup>9</sup> Both Rubin (1986) and Renssen (1998) recommend concatenations in settings of complex survey designs.

correlated variable: the income category<sup>10</sup> a household belongs to. By including this variable as a matching variable, we are confident that the conditional independence assumption holds, although it is not possible for us to test it.

**We next apply multiple imputation methods to the concatenated dataset.** Multiple imputation (MI) methods originated in the 1970s as a response to missing survey observations (Rubin, 1972). Since then, researchers have increasingly relied on them to respond to missing data entries (Carpenter, 2013). An imputation  $M_i$  is a set of plausible values  $m_i$  for missing observations. In multiple imputations,  $M$  is a set of imputations that consist of  $i$  individual imputations, each of which consists of a set of imputed values  $m_i$ . While a single imputation might overstate the precision of the underlying imputation, given that it treats the generated values as known, MI accounts for sampling variability by imputing the missing values several times. This approach allows us to consider the between-imputation variability that emerges from the sampling variability of the imputed values (Little and Rubin, 2020). After one imputes  $i$  sets of missing observations, one typically pools the resulting sets to generate one final data set.<sup>11</sup> In this paper, we follow the approach taken by Abayomi et al. (2008) and generate the final imputed energy expenditure shares in the synthetic dataset by taking the average of the different imputations.

**Based on the recommendations of the literature, we apply 20 different imputations.** The theory behind multiple imputations relies on an infinite number of imputations. In practice, the precision of the final imputed variable depends directly on the number of imputations performed. The number of minimum imputations  $i$  needed to generate reliable imputed values also relates to the number of missing data entries, the data itself, and the model applied for the imputation (StataCorp, 2021). While there is no consensus on the minimum number of imputations needed, recommendations range from two over five to a maximum of 20 (Rubin (1987); StataCorp (2021)). Importantly, given that all

---

<sup>10</sup> While household income is part of the HBS in Bulgaria, this information is gathered in a less detailed and rigorous manner than information on household income gathered as part of the EU-SILC.

<sup>11</sup> For the methodological details and the justification behind the MI approach see Rubin (1987).

observations of energy spending shares and energy poverty are missing in the EU-SILC by nature, we assume that data is missing completely at random (MCAR) and that missing information is not related to observable or unobservable characteristics of households. This data property allows us to ignore the underlying process that generates missing data (Rubin, 1976). Based on the literature, we decided to run a set of 20 imputations.

**The reliability of imputed values depends on choosing a proper imputation model.** A proper imputation model is a model that generates proper multiple imputations (Rubin, 1987; Binder and Sun, 1996). A proper imputation model considers all variables that are relevant to the missing data generation. In addition, the model should include all variables that researchers plan to use in future analyses conducted with the imputed dataset. In our case, given that we intend to analyze the overlap between monetary and energy poverty, it is crucial to include income in the imputation model. Lastly, survey-related variables, such as weights, strata, and cluster identifiers, should also be part of the imputation model.

**We compare three multiple imputation models to each other.** We address concerns raised by Lewaa et al. (2021) on the lack of quality assessments about data fusion methods and employ several multiple imputation techniques to identify the model that works best. This approach also helps us to gain a better understanding of the quality of our matching exercise. We start with a simple linear regression imputation method. This method is fully parametric and relies on the assumption of normality (Schenker and Taylor, 1996). Under scenarios in which the assumption of normality might be violated, other imputation approaches might be more suitable. One approach recommended in the literature is predictive mean matching (PMM) (Little, 1988; Rubin, 1986). PMM is a semiparametric approach that combines nearest-neighbor imputations and linear regressions (Laqueur et al., 2022). PMM first makes linear predictions and then uses these predictions as a distance measure to choose possible donors from the observed values. Based on that, PMM draws imputed values from the observed ones. The range of observed values persists in the imputed set of values. The last imputation model we consider is the truncated regression imputation model. Truncated regressions can be used for the imputation

of variables that are restricted by a certain range. In our application, we impute a share, so it is worth exploring the potential of truncated regression imputations.<sup>12</sup>

**We follow the literature and assess if the underlying imputation models are proper.** Abayomi et al. (2008) recommend three different approaches to assess the reliability of multiple imputations: plotting the completed data to detect unusual patterns, comparing the distribution of the imputed variable in the observed and unobserved data<sup>13</sup>, and analyzing the fit of observed data to the imputation model. The first two approaches are simple reasoning tests. Tests for model fits consist of standard toolkits, such as the analysis of residual plots. Most of the literature seems to rely on a simple comparison of the distribution of observed and imputed values. We follow this approach and plot the distribution of the observed energy spending shares against one of the imputed energy spending shares. For the final imputation, we choose the model that resembles the original distribution most closely.

**After choosing the best-performing model, we apply several additional quality checks to confirm the quality of the data fusion method.** Serafino and Tolkin (2017) propose several other validity checks. Based on their recommendations, we also compare the consistency of mean energy expenditure shares by variables used in the statistical matching versus those not used. We also duplicate the HBS and act as if energy expenditure shares were missing in this duplicated test. We conclude that these tests verify that our matching procedure works well.

---

<sup>12</sup> There are several other imputation models researchers can explore. For an overview of the models available in Stata see: <https://www.stata.com/features/multiple-imputation/>

<sup>13</sup> While one would ideally want to compare the distribution of the imputed variable to its true underlying distribution, the true underlying distribution cannot be observed. Therefore, the approach proposed by Abayomi et al. (2008) is an approximation and a reasoning test. While deviations in the distribution of the imputed and observed variable values do not necessarily imply that the imputation model is improper or that the MCAR assumption is violated, extreme deviations might serve as a reasonability test. Abayomi et al. (2008) propose a visual comparison of the underlying density and a Kolmogorov-Smirnov (KS) test.

**The result of this paper is a synthetic dataset that contains all variables in the EU-SILC plus energy spending shares and an energy poverty indicator, which we can ultimately use to overlap energy with monetary poverty, or to describe the energy poor population in Bulgaria.** The resulting dataset from our data fusion exercise is a synthetic version of the EU-SILC, which has information on income-related variables I, households' access to social protection programs, imputed expenditure-related information C (the energy spending share and energy poverty indicator), a set of overlapping variables, and all other variables included in the EU-SILC. In the last step, we use the generated dataset to analyze if the energy and monetary-poor overlap. We also explore energy spending shares by income quintiles. Future research can use the generated dataset to conduct detailed studies about the energy poor population in Bulgaria.<sup>14</sup>

## Data Description and Preparation

### Data and Target Variable Description

**In this paper, we use two different data sources: the 2019 HBS and the 2020 EU-SILC with reference income from 2019<sup>15</sup>.** We focus on the pre-COVID year (2019) because we want to make sure that our data is representative of spending patterns in Bulgaria and not affected by idiosyncratic crisis-related behavior induced by the COVID-19 pandemic.<sup>16</sup> For this reason, we rely on the HBS from 2019. The HBS is a survey conducted every year and collects data on expenditure patterns by households in Bulgaria. The survey consists of 2,952 households and is a quarterly survey, meaning that the information is representative of each quarter of the year.

**The HBS relies on a two-stage random sample design of households and uses expenditure diaries.** In the first stage of the sampling design, census enumeration areas are selected randomly, while in the

---

<sup>14</sup> For a first study of this type, see Robayo-Abril and Rude (forthcoming b).

<sup>15</sup> In EU-SILC, the reference period of income refers to the calendar year before the year in which the survey took place.

<sup>16</sup> Examples could be increased expenditure on medical expenses, for example, or decreased expenditure on restaurant visits given the strict stay-at-home policies.



second stage, households from chosen census enumerations are sampled. Every quarter, 1,020 households are part of the survey each month. These 1,020 households are interviewed every quarter. Each household completes a diary that collects information on its respective expenditures twice a month. Interviewers visit each household twice a month to collect this information. They also gather additional data on a limited set of household characteristics and a broad measure of household income. The sample is generally representative at the national level and at the residence level (urban-rural) but not at lower levels of geographic disaggregation.

**Expenditure data in the HBS follows the classification of COICOP, developed by Eurostat, and allows us to identify energy-related spending components clearly.** We refer to the variables detailed in Table 2 to measure energy spending. To distinguish between different types of energy expenditures, we rely on the 4-digit categories of spending components, which is the most disaggregated category detailed in the HBS codebooks. Table 2 indicates that we can identify spending on electricity, natural gas and town gas, liquefied hydrocarbons, butane, propane and similar, liquid fuels, coal, other solid fuels, and heat energy.

**We define energy spending shares, our target variable, as the ratio of energy spending over household income, as observed in the HBS.** To calculate the energy expenditure share, we aggregate the different energy spending components at the household level and divide the resulting overall energy expenditure by household income. We do not scale energy spending or household income by the number of household members or the adult equivalent of household members, as they would cancel out in any case. The resulting indicator is the energy spending share in overall income at the household level. Table 2 reveals that the average energy spending share observed in the HBS 2019 in Bulgaria is 13.1 percent. The largest energy spending component is on electricity (59.3 percent of all energy expenditure), followed by other solid fuels (27.5 percent) and heat energy (9.3 percent).

**There are several definitions of energy poverty from a monetary perspective.** One stream of literature defines the energy poor as all households that spend more than 10 percent of their total expenditure on energy (Boardman, 1991). Thema and Vondung (2020) explore two additional energy poverty indicators. The first one is based on the national median of the energy spending share (often

denoted as 2M). Based on this definition, a household is energy poor when its share of energy expenditure is above a specified national threshold, defined as twice the national median value. In this case, energy expenditure shares are measured as the ratio of equalized energy expenditure over equalized disposable household income. The second indicator explored by Thema and Vondung (2020) refers to the national median of the absolute energy expenditure (often denoted as M/2). Following this definition, a household is affected by energy poverty as soon as its absolute spending on energy is below a certain threshold (half the national median of absolute energy spending). The objective of these indicators is to measure the relative insufficiency of basic energy services caused by a lack of financial resources or to recognize situations where the consumption of these essential energy services disproportionately burdens households in relation to their available income (see Robayo-Abril and Rude (2023a) for more details).

**Table 2 demonstrates that the share of households affected by energy poverty varies significantly by definition.** Applying the 10 percent threshold results in an energy poverty incidence rate of 53.0 percent, while only over every 10<sup>th</sup> household is considered energy poor when following the methodology detailed by Thema and Vondung (2020). These differences are significant and demonstrate how important it is to rely on a unified definition of energy poverty when making comparisons across geographic units or over time. Moreover, the underlying population might differ significantly in observable characteristics depending on the definition used, which might severely affect the design of targeting strategies. For the rest of this paper, we rely on the definition of energy poverty that uses the 10 percent threshold.

*Table 2: Summary statistics of energy expenditure shares (income-based) and energy poverty (2019)*

VARIABLES	(1)	(2)	(3)	(4)	(5)
	N	mean	sd	min	max
Share of energy expenditure	2,952	0.131	0.128	0	3.662
Energy poverty incidence rate (share)	2,952	0.134	0.341	0	1

Energy poverty incidence rate (absolute)	2,952	0.106	0.308	0	1
Energy poverty (> 10 percent)	2,952	0.530	0.499	0	1
Electricity (share)	2,952	0.0762	0.0792	0	3.063
Natural gas and town gas (share)	2,952	0.00132	0.00942	0	0.170
Liquefied hydrocarbons/butane/propane (share)	2,952	0.00111	0.00428	0	0.0630
Liquid fuels (share)	2,952	3.00e-05	0.00104	0	0.0519
Coal (share)	2,952	0.00329	0.0352	0	1.625
Other solid fuels (share)	2,952	0.0374	0.0790	0	1.075
Heat energy (share)	2,952	0.0118	0.0292	0	0.243

---

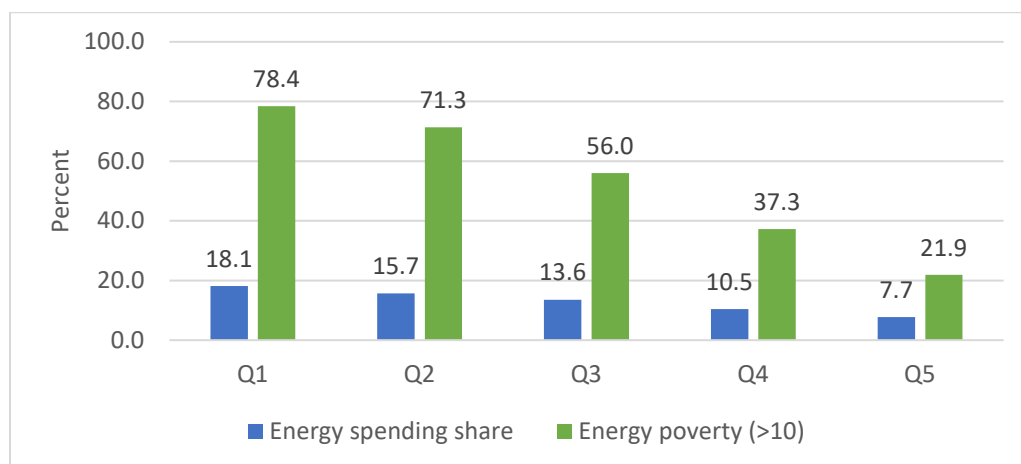
*Notes: The indicators in this table are at the household level (N=2,952). We weigh each observation by household survey weights. The energy poverty incidence rate (share) is the 2M energy poverty indicator. The energy poverty incidence rate (absolute) is the M/2 energy poverty indicator. All variables are reported as shares.*

*Source: Own estimates based on HBS (2019).*

**Lower welfare quintiles report higher energy expenditure shares and higher energy poverty rates.**

Figure 1 demonstrates that those households that report the lowest average income spend a higher share of their overall household expenditure on energy (18.1 percent). The energy spending share falls by income quintile, with the highest income quintile reporting the lowest share (7.7 percent). Figure 1 reveals similar patterns for the case of energy poverty rates. Households in lower income quintiles report higher incidence rates than those in higher income quintiles.

Figure 1: Energy spending shares (income-based) and energy poverty rates (10 percent threshold) by income quintiles (in percent, 2019):



Notes: The graph shows the energy spending share in overall household income at the household level (in percentages) by income quintiles in Blue and the related energy poverty rates in Green. Energy poverty rates use a 0.1 threshold by income quintiles. Energy spending shares are income-based. Income quintile 1 is the 25 percent of households with the lowest average per capita income, while income quintile 5 is the 25 percent of households with the highest average per capita income. We weigh each observation by household survey weights. N=2,952.

Source: Own estimates based on HBS (2019).

**We impute energy spending shares from the HBS 2019 into the EU-SILC 2020 (with the income year 2019).** Ultimately, we are interested in generating information on energy poverty for households sampled in the EU-SILC. While we could also impute energy poverty indicators directly, imputing energy spending shares gives us the flexibility to explore several expenditure-based measures at once because these measures directly result from energy spending shares.

**One important caveat is that energy poverty indicators that take a purely monetary perspective suffer from conceptual limitations.** Energy poverty indicators mostly reflect that poorer households potentially spend a larger share of their overall resources on energy and are, therefore, more vulnerable; they suffer from conceptual deficiencies. For example, households that do not have access to energy in the first place might be the most vulnerable ones of all but will likely report lower energy spending shares. Similarly, certain households might access energy generated from illegal logging or private forest fields. These costs are generally not reflected in HBS and might distort official indicators.

Moreover, households might simply report high energy spending shares due to energy-inefficient behavioral patterns. One would want to correct these behaviors instead of incentivizing them with social protection strategies, which mainly target households that face serious constraints in energy consumption. Relative indicators, such as those summarized by Thema and Vondung (2020), do not correct for the limitations mentioned above and additionally suffer from the fact that they are based on relative measures and not a fixed threshold based on minimum energy consumption needs, which is difficult to monitor over time. Consequently, if energy prices increase for all households, the energy poverty rate might not increase when relying on the latter two indicators, which might underestimate increasing vulnerabilities to energy poverty.

**There are several alternatives to monetary-based measures of energy poverty, but we abstract from them in this paper.** While there are many alternative ways to measure energy poverty (for a recent overview of indicators used in the EU, see Gouveia et al. (2022)), we restrict our analysis to expenditure-based measures.<sup>17</sup> Hence, we narrow down our analysis of energy poverty to a subset of its multidimensional nature.<sup>18</sup> The results derived later in this paper are dependent on our chosen definition of energy poverty, which is an important limitation when thinking about the external validity of our findings and the potential to generalize them more broadly speaking. While we might find a significant overlap of energy and monetary poverty, the results might not hold under alternative measures of energy poverty.

#### Potential Matching Variables

**Statistical matching requires a series of data processing steps.** Serafino and Tonkin (2017) outline eight steps that should be applied when preparing the statistical matching of two datasets. These steps include adjusting the time frame and categories of variables. In statistical matching, harmonization can

---

<sup>17</sup> Gouveia et al. (2022) divide available indicators into three categories: expenditure-based measures, consensual approaches, and direct measures.

<sup>18</sup> Many stress that energy poverty is a multidimensional problem that is linked to a multitude of underlying drivers (see for example Energy Poverty Advisory Hub (2022)).

take various forms, such as aligning the definition of units, reference period, population, variables, classification, measurement error, missing data, and derivation of variables. For example, in the case of statistical matching between household budget surveys, we harmonize the reference period by annualizing variables whenever possible and appropriate. Additionally, we harmonize the categorization of variables to ensure they are classified in a consistent manner across different surveys. The harmonization process is facilitated by the fact that both the Household Budget Survey (HBS) and the European Union Statistics on Income and Living Conditions (EU-SILC) share a similar definition of a household. This definition, as described by Serafino and Tonkin (2017), states that a household comprises individuals who live together in the same dwelling and share meals or jointly provide living conditions. Essentially, a household refers to a group of people residing in the same place, with a degree of interdependence in their daily living arrangements, such as cooking and sharing meals together.

**We next identify potential matching variables and harmonize them with each other.** Table A.2 in Appendix 1 gives an overview of the variables that are included both in the HBS 2019 and the EU-SILC 2020. We mark those variables that we include in the set of potential matching variables in green. We also depict the respective categorization of these variables. As there are some important differences in the categorization of most of the variables, we harmonize them with each other. Next, we aggregate all indicators at the household level. Based on our assessment of potential matching variables, we harmonize the variables as follows:

- Harmonize HBS regions to EU-SILC regions (BG3 and BG4)
- Create a dummy variable for urbanization, both in HBS and EU-SILC
- Create a categorical income variable of 20 categories, the lowest category representing those households with the lowest income
- Generate household characteristics based on individual-level characteristics, which are as follows:
  - Household with at least one child (<15); with at least one pensioner; with at least one elderly (+64); female-headed household; a household with at least one unemployed

- Number of household members, the adult equivalent, number of children, number of pensioners, number of female members, number of unemployed members, number of self-employed members, number of members working in the primary sector, number of members in part-time employment, number of members with primary, secondary, and tertiary education, number of members attending an educational facility, number of foreign citizens in the household

**When combining the harmonized variables of both datasets, some significant differences between both surveys become evident.** Table 3 compares the harmonized variables of both datasets to each other. The table reveals that there are some slight deviations in most of the variables and some more important ones in the case of several of the household characteristics, which is expected given that the underlying sampling designs differ from each other.<sup>19</sup> While both datasets should be representative of the underlying population in Bulgaria by design, the table demonstrates how difficult it is to achieve this in practical applications. The differences in the harmonized variables could also be related to the fact that a full harmonization between both datasets is challenging, given that some variables rely on different underlying classifications, such as in the case of the educational variable, for example. We also compare the similarity in the distribution of potential matching variables across surveys. To this end, we plot histograms and compare them to each other.<sup>20</sup> Appendix 3 presents the resulting graphs, which confirm that there are some important differences. The systematic differences in household characteristics between both datasets are an important caveat in the analysis to come.

---

<sup>19</sup> First, the EU-SILC consists of a significantly larger sample than the HBS. Moreover, the time dimension of both surveys differs. In addition, the survey design of the HBS consists of strata and primary sampling units while the one of the EU-SILC consists of primary and secondary sampling units.

<sup>20</sup> While some researchers propose additional methods, such as calculating the Hellinger Distance (HD), there is no fixed rule regarding what degree of similarity would be suitable in the context of statistical matching, which is why we abstract from these methods (Serafino and Tonkin, 2017).

Table 3: Household characteristics of households in HBS (2019) and EU-SILC (2020) – potential matching variables

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
	HBS N	Mean	sd	EU-SILC N	mean	sd
Household with children (<15)	2,952	0.166	0.372	7,313	0.230	0.421
Household with pensioner	2,952	0.557	0.497	7,313	0.482	0.500
Households with elderly (>64)	2,952	0.532	0.499	7,313	0.414	0.493
Female-headed household	2,952	0.441	0.496	7,313	0.495	0.500
Household with unemployed	2,952	0.131	0.338	7,313	0.124	0.329
Urban	2,952	0.733	0.442	7,313	0.687	0.464
Income cat. (1-20)	2,952	10.50	5.767	7,313	8.669	5.833
No of female	2,952	1.179	0.745	7,313	1.043	0.656
No of self-employed	2,952	0.0717	0.285	7,313	0.101	0.352
No in the primary sector	2,952	0.0513	0.276	7,313	0.0706	0.284
No of part-time employed	2,952	1.587	1.808	7,313	0.0521	0.247
No with primary educ.	2,952	0.339	0.818	7,313	0.101	0.424
No with secondary educ.	2,952	1.372	1.031	7,313	1.398	1.111
No with tertiary educ.	2,952	0.454	0.713	7,313	0.876	1.115
No attending educ.	2,952	0.281	0.634	7,313	0.159	0.410
No of foreign citizens	2,952	0.00365	0.0745	7,313	0.0113	0.149
No of children	2,952	0.240	0.595	7,313	0.345	0.734
No of pensioners	2,952	0.733	0.744	7,313	0.612	0.707
No of unemployed	2,952	0.167	0.482	7,313	0.151	0.447
No. of household members	2,952	2.165	1.216	7,313	2.375	1.463
Adult equivalent	2,952	1.832	0.961	7,313	1.624	0.639
Northern and Eastern Bulgaria	2,952	0.499	0.500	7,313	0.502	0.500
South-West and South-Central Bulgaria	2,952	0.501	0.500	7,313	0.498	0.500

Notes: The indicators in this table are at the household level. We weigh each observation by household survey weights.

Source: Own estimates based on HBS (2019) and EU-SILC (2020).



**In addition to these systematic differences in overlapping variables between both surveys, the sample design slightly differs.** According to information published by the National Statistical Office (2023a), in the case of the EU-SILC in Bulgaria, primary sampling units are census enumeration units. From these units, five secondary sampling units (households) are selected. The HBS follows a similar design but chooses six secondary sampling units (households) from each census enumeration unit (National Statistical Office, 2023b).

**We identify potential auxiliary variables to approximate the underlying conditional independence assumption of statistical matching.** For this purpose, we correlate each of the potential matching variables with energy expenditure shares in the HBS. Table A.3 in the Appendix details the results. The highest correlation coefficient is above -0.37 for the income category variable. There is also a strong correlation between the number of part-time employed household members (-0.2561) and those having tertiary education (-0.192). There is no consensus in the literature on when a correlation coefficient is high enough to approximate the conditional independence assumption. The rule of thumb identified previously in the literature is that auxiliary variables are those variables with a correlation coefficient higher than 0.4 (Johnson and Young, 2011). The income category variable is close to this threshold, although the correlation is negative.<sup>21</sup>

#### Final Matching Variables via Feature Selection

**To achieve the most parsimonious model possible, we apply lasso regressions and identify those variables most relevant to explaining the energy expenditure share.** In addition to similarity, relevance is another important criterion when choosing matching variables (Serafino and Tonkin, 2017). As the inclusion of too many matching variables could result in unnecessary noise (Donatiello et al.,

---

<sup>21</sup> This approach is in line with Donatiello et al. (2016b) who argue that, based on income being measured in both the HBS and EU-SILC, income is a strong auxiliary variable as households who belong to lower income categories in the HBS likely coincide with those in the EU-SILC.

2016b), we apply a lasso regression for feature selection.<sup>22</sup> The lasso regression allows us to follow a data-driven approach when selecting the most relevant variables from the potential set of overlapping variables. We apply a 10-fold cross-validation (CV) procedure. The lasso estimation results in 7 out of 22 possible matching variables, which are households with pensioners, households with elderly, households with unemployed, a female-headed household, urban, its income category, and the number of household members with secondary education. The resulting out-of-sample R-squared is 0.151, and the CV mean prediction error is 0.014.

**In addition to the variables identified via feature selection, we include the variables that describe the underlying sample design in the set of matching variables.** This is the variable that identifies the primary sampling unit (census enumeration units in the case of the HBS). We denote this variable as `psu1` in the rest of the paper. In an alternative model specification, we include the sample weights by running weighted regressions. Additionally, we explore empirical specifications, in which we abstract from weighting regressions by the given survey weights but include survey weights as an additional matching variable.

### III. Results

This section presents the results of the different imputation models applied in the paper and analyzes the quality of the resulting imputation.<sup>23</sup> For all regressions, we apply 20 imputations and set the seed to 12345 to ensure the reproducibility of our results.

---

<sup>22</sup> The least absolute shrinkage and selection operator (lasso) forms part of penalized least square regressions. It was originally developed by Robert (1996) and performs both automatic variable selection and continuous shrinkage. For details on the lasso regression see Meinshausen and Bühlmann (2006).

<sup>23</sup> Before applying the different imputation models we prepare our data for the usage of the `mi` commands in Stata. We use the data in marginal long style (`mlong`) as it is more memory-efficient. We then register the energy spending share and the energy poverty indicator as imputation variables. We also make sure that there are no missing observations in any of the matching variables.

## Linear Regression Imputation Method

### **We apply linear regression imputation methods after log-transforming the energy spending share.**

Given that the distribution of energy spending shares is slightly skewed to the right (Figure A1 in Appendix 1), we first log-transform the variable and then impute the missing values. After the imputation, we transform energy spending shares back to their original scale by exponentiation. We apply three different versions of the linear regression imputation model. The first specification includes variables that describe the sample design of both surveys (psu1) as matching variables and weighs the regression by survey weights. Appendix 2 details the graphical results.<sup>24</sup> The figures show that the imputation performs poorly in the case of linear regressions. Next, we repeat the linear regression imputation method but do not consider psu1 as a matching variable. We still weigh observations by their survey weights. While the imputations generated by this empirical specification are closer to the true underlying values, they suffer from outliers. The imputed energy spending share assumes values below 0 and above 1 in some cases. Lastly, in the third empirical specification, we do not weigh observations by survey weights but include them as an additional matching variable. The results indicate that the latter specification outperforms the other two specifications when looking at the joint distribution of observed and imputed values (see Appendix 2). In general, restricting the imputed energy spending shares to values between zero and one shows that the distributions resemble each other better in this case.

## Predictive Mean Matching

### **We next explore three empirical specifications of predictive mean matching imputation methods.**

Predictive mean matching imputation methods might be more appropriate in settings where the normality assumption might be violated. In addition, due to the nature of PMMs, many researchers consider them more robust. PMMs draw information from observations that are as similar as possible.

---

<sup>24</sup> We also report the model estimates for the linear regression and truncated regression in Appendix 6. However, as they purely rely on HBS they are less informative. In addition, model estimates are not presented for PMM model specifications because this methodology is semi-parametric.

The similarity is the smallest possible absolute difference between the linear prediction for the missing value and that for the observed value (Morris et al., 2014). We employ the same three model specifications outlined for linear regression imputation methods. A critical choice in PMMs is the number of nearest neighbors used for the donor set. The optimal number of nearest neighbors is subject to a tradeoff between the bias and the variance. There is no consensus in the literature on the recommended number of donors. We follow recommendations by Morris et al. (2014) and choose the ten nearest neighbors. The graphs presented in Appendix 2 reveal that all the specifications resemble the distribution of the observed energy spending share closely.

### Truncated Regression Imputation Method

**Lastly, we employ three empirical specifications of the truncated regression imputation model.** We again employ the same three model specifications outlined for linear regression imputation methods. We restrict the range of the imputed variable to 0 and 1, given that we impute a share. The graphs presented in Appendix 2 show that the model's performance is lower than the PMM. The average energy spending share is above the true observed one. In addition, the distributions do not look alike.

### Choosing the Final Model

**We choose the weighted PMM as our best-performing model based on a comparison of the generated distributions.** After implementing different imputation models, we compare the best-performing model specification of each of the three models to each other. We choose the model that best replicates the distribution of the observed energy spending share as our final model. Figures 2 to 4 present the distribution of the true and the imputed energy spending share for 1) a linear regression model using survey weights as a matching variable (Specification 3), 2) a weighted PMM (Specification 2), and 3) a truncated regression model using survey weights as a matching variable (Specification 3). We choose the weighted PMM as our best-performing model, given that we plan to weigh observations by survey weights in future analyses. Our decision is also informed by the previous literature, showing that PMM works well in the context of imputations (Kleinke, 2018).

Figure 2: True and imputed distribution of energy spending share (Linear regression model using survey weights as a matching variable)

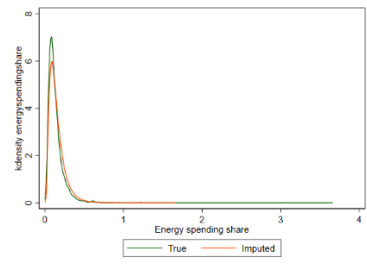


Figure 3: True and imputed distribution of energy spending share (weighted PMM)

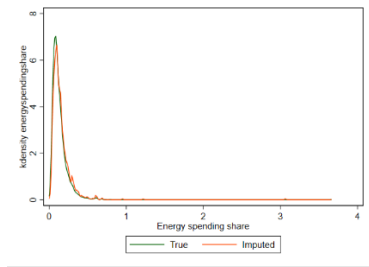
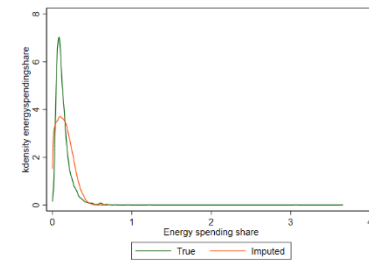


Figure 4: True and imputed distribution of energy spending share (Truncated regression model using survey weights as a matching variable)



Notes: We do not weigh observations by survey weights in these graphs. Source: Own estimation based on a harmonized synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

We also compare the generated distribution to the true observed distribution by subgroups to validate that the imputation performs well. Figures 5 and 6 present the distributions for households with at least one female member compared to households without female members. Figures 7 and 8 plot the results for households in the lowest five income categories compared to the highest five income categories. The distribution of the imputed energy spending share is close enough to the distribution of the observed energy spending share. We also validate that the results from the 20 different imputations performed do not differ significantly from each other. Appendix 2 reports the mean and standard deviation of the imputed energy spending share for each of the 20 imputations. The table demonstrates that values are close to each other. The final energy spending share is the average of each of these 20 imputations at the household level. The simulated mean (weighted by survey weights) is close to the observed mean of energy spending shares in the household budget survey (14.7 percent versus 13.1 percent).

Figure 5: True and imputed distribution of energy spending share (weighted PMM) - Female

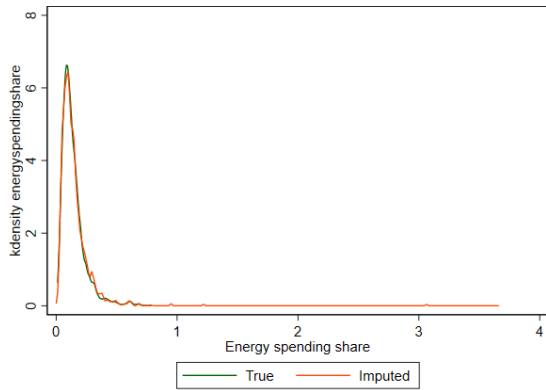


Figure 6: True and imputed distribution of energy spending share (weighted PMM) – No female

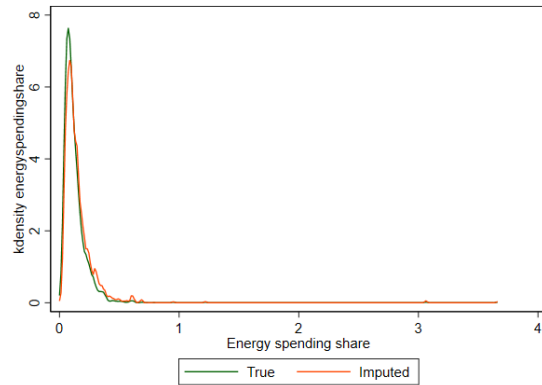


Figure 7: True and imputed distribution of energy spending share (weighted PMM) - High-income

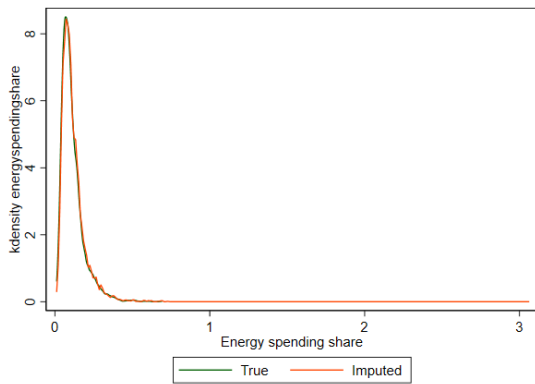
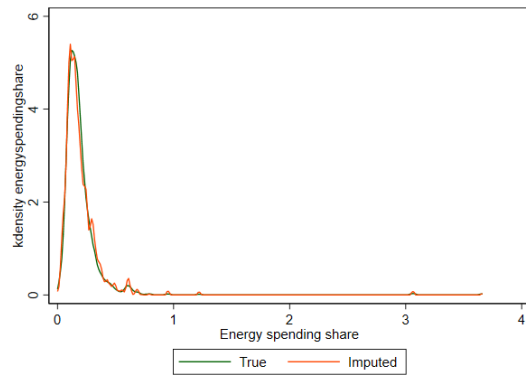


Figure 8: True and imputed distribution of energy spending share (weighted PMM) - Low-income



Notes: We do not weigh observations by survey weights in these graphs. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

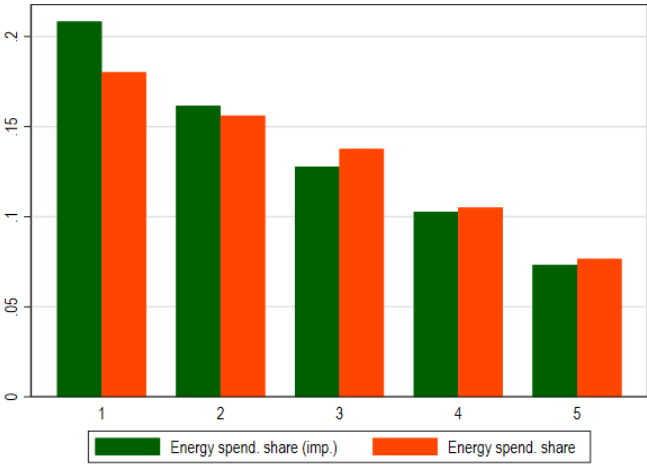
**We apply additional validity checks to the final model.** First, we duplicate the HBS and act as if the energy spending share is missing in this duplicated dataset. This check confirms that the weighted PMM results in an appropriate replication of the observed energy spending share. Next, we analyze the average imputed and observed energy spending share by matching variables. The results in Appendix 4 show that averages of imputed values are consistently slightly above the observed value, with a few expectations but sufficiently close. Lastly, we compare the average energy spending share for a variable not used in the matching procedure, namely the number of household members.

Appendix 5 shows that deviations are more significant in this case. Lastly, we verify that the distribution remains similar when choosing a lower number of neighbors, namely five neighbors (Appendix 2).

Overlaying Monetary and Energy Poverty

The dataset generated in this paper allows us to estimate the incidence of energy poverty and overlay official monetary poverty with energy poverty. Figure 9 depicts the energy spending share by income quintiles. The graph demonstrates that those in the lowest income quintile spend a larger share of household expenditure on energy than those in the highest income quintile. To validate our results further, we compare them to data from the EU-SILC 2015. In 2015, Eurostat gathered energy expenditure data as part of the EU-SILC and published the respective energy spending as a share of income by income quintiles and found that shares vary between 15 and 8 percent (Gouveia et al., 2022). This range is close to the one presented in Figure 9. Shares by quintiles are also similar when using alternative imputation methods.

Figure 9: Imputed average energy spending shares versus observed average energy spending shares by income quintiles



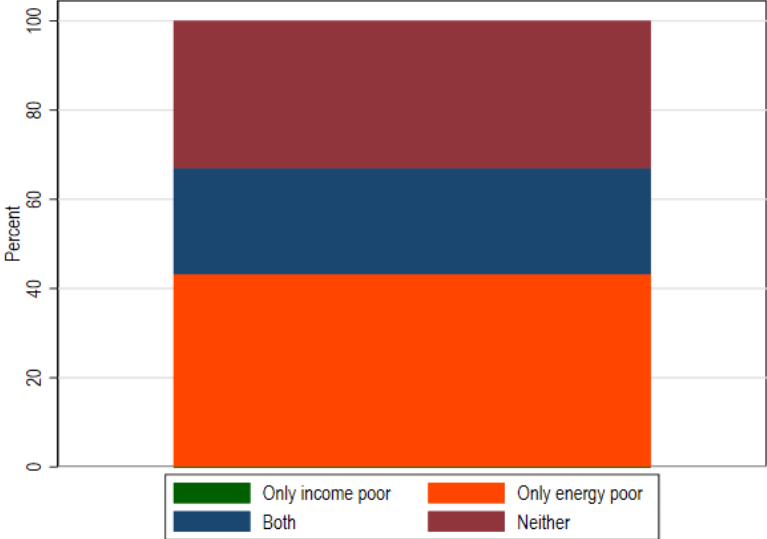
Note: The figure depicts energy spending shares by income quintiles for observed (in Green) and imputed (in Blue) values. Q1 is the lowest, and Q5 is the highest income quintile based on equivalized household income. Imputed values apply a weighted multiple imputation PMM to a synthetic dataset consisting of HBS (2019) and EU-SILC (2020), while observed values are reported energy spending shares in the HBS (2019). Estimates shown in these figures are at the individual level. We weigh each observation by the respective survey weights.

Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

**The incidence and overlap of energy and monetary poverty differ by definition.** Both the definition of monetary poverty and energy poverty impact the extent to which monetary and energy poverty affects the Bulgarian population. The identification of distinct segments within the population—those at risk of poverty but not experiencing energy poverty—holds significant policy implications. Understanding the nuanced relationship between income poverty and energy poverty enables policymakers to design more targeted and effective interventions.

**For example, in terms of energy poverty measured by the 10 percent estimator, we estimate that, with the synthetic database, a significant proportion of the population falls into this category (Figure 10).** Nearly all individuals who are income-poor are also energy-poor. Moreover, more than four out of 10 people are energy poor although they are not at risk of poverty. This subset likely includes individuals with substantial energy expenditures, who, despite not being classified as income-poor, fall below the poverty line when accounting for their energy costs. Consequently, initiatives aimed at providing income support to those in poverty may not effectively reach this specific group (Figure 10 panel a). Other interventions should address the unique challenges faced by these individuals.

*Figure 10: Imputed Overlap between Income and Energy Poverty.*



*Notes: The figures depict the overlap of energy poverty and official monetary poverty in Bulgaria. Energy poverty relies on the 10 percent measure. A household is considered energy-poor if energy spending shares are above 10 percent. Monetary*



*poverty is the at-risk poverty rate using the income measure reported in SILC, and therefore, corresponds to the official poverty measure. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).*

**The overlap between both concepts varies significantly by the metric of energy poverty used.** This reflects that energy poverty is a complex concept that lacks a common definition and is multidimensional by nature.<sup>25</sup>

## **IV. Conclusion**

**In this paper, we explore the potential of statistical matching techniques to combine information gathered on households' living conditions and welfare with information on households' energy expenditure from separate surveys.** More concretely speaking, we apply data fusion methods to the EU-SILC 2020, with income reference year 2019, and the HBS 2019 in Bulgaria to ultimately add information on energy spending shares to the EU-SILC. Statistical matching is necessary in this case, as traditional data linkage is not possible given that households sampled into the HBS and EU-SILC do not coincide. We first determine several overlapping variables that can serve as matching variables. We next choose the set of relevant matching variables via lasso regressions. Based on these variables, we generate a harmonized and concatenated dataset containing information from both HBS and EU-SILC data. We then apply three different imputation models to impute energy spending shares for households from the EU-SILC.

**We choose the best-performing model by analyzing the distribution of imputed and observed energy spending shares.** Our results show that weighted PMM imputation models generate the closest distribution of energy spending shares within the set of weighted imputation models. We verify this result by a number of robustness checks. We choose the imputed values from the weighted PMM to generate imputed energy spending shares for households in the EU-SILC.

---

<sup>25</sup> For the detailed discussion on energy poverty measurement, see Robayo-Abril and Rude (forthcoming b).

**Based on the final synthetic dataset, we overlay monetary with energy poverty indicators and investigate energy poverty rates by income quintiles.** These types of analyses are not possible without data fusion methods, given that reliable and detailed expenditure and income data are only generated separately in Bulgaria, like in most countries in the EU. By employing statistical matching techniques, we can generate valuable information on who the energy poor are. We find that more than one-third of the energy-poor individuals (using the low-income high-cost measure) are not considered at risk of poverty. As a result, efforts directed at offering financial assistance to individuals in poverty might not adequately reach this particular subgroup. Importantly, these findings vary significantly by the underlying definition of energy poverty.

**This paper makes an important contribution to the increasing body of literature exploring statistical matching techniques.** Researchers and Statistical Offices have become increasingly interested in alternatives to record linkage to combine different data sources, and the number of papers in this area is growing. To the best of our knowledge, we are the first ones exploring the potential of statistical matching techniques in Bulgaria. We are also the first to apply data fusion methods to the area of energy poverty. Lastly, we make a methodological contribution by exploring several different imputation models and comparing their performance to each other.

**Future research can take advantage of the generated dataset to study energy poverty in Bulgaria in more detail.** Our generated dataset allows us to describe the energy poor in more detail. Based on the new dataset, it is possible to analyze if the energy poor have access to social protection programs, for example, or what could drive their energy poverty status. In addition, having energy expenditure information in the EU-SILC allows for microsimulations that study energy price increases or potential mitigation measures.

**To improve the possibilities for statistical matching between EU-SILC and HBS, policymakers could push for greater harmonization between both surveys and include expenditure-based questions in a subsample of the EU-SILC.** In line with previous recommendations at the EU level (see Serafino and Tokin, 2017), we recommend better harmonization of surveys. In addition, including a roster of expenditure patterns for a subsample of the EU-SILC would increase the reliability of imputation and

matching methods. These initiatives could be achieved by greater statistical and methodological cooperation at the EU level.

**Finally, the results on energy and income poverty have several policy implications for the targeting of social programs, energy subsidies, and energy efficiency programs.** First, as the monetary poor and energy poor overlap, policymakers may consider integrating energy assistance programs within existing social safety nets. This could help ensure that those struggling with monetary poverty also have access to affordable and reliable energy services, which are essential for their well-being and socio-economic development. Second, recognizing that not all energy-poor individuals are monetary poor, targeted energy subsidies could be implemented to support vulnerable populations who may not qualify for traditional monetary poverty assistance but still face challenges in accessing adequate energy services, especially in times of crisis. Finally, measures to encourage energy efficiency in the medium term can be beneficial for both the monetary poor and energy-poor individuals. These initiatives can help reduce energy costs for low-income households and contribute to overall energy sustainability (see more details in Robayo-Abril and Rude (2023b)). Therefore, policymakers should invest in comprehensive data collection and research to better understand the dynamics of energy poverty and its relationship with monetary poverty, as this information can lead to more effective and targeted policy interventions.

## References

- Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3), 273-291.
- Bacher, J., & Prandner, D. (2018). Datenfusion in der sozialwissenschaftlichen Wahlforschung–Begründeter Verzicht oder ungenutzte Chance? Theoretische Vorüberlegungen, Verfahrensüberblick und ein erster Erfahrungsbericht. *Österreichische Zeitschrift für Politikwissenschaft*, 47(2), 61-76.
- Binder, D. A., and W. Sun. 1996. Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Survey Research Methods Section, American Statistical Association* 281–286.
- Boardman, B. (1991). *Fuel poverty: from cold homes to affordable warmth*. Pinter Pub Limited.
- Carpenter, J. R., and M. G. Kenward. 2013. *Multiple Imputation and its Application*. Chichester, UK: Wiley.
- Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., & Spaziani, M. (2016b). The role of the conditional independence assumption in statistically matching income and consumption. *Statistical Journal of the IAOS*, 32(4), 667-675.
- D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.
- Emmenegger, J., Männich, R., & Schaller, J. (2022). Evaluating Data Fusion Methods to Improve Income Modelling (No. 2022-03). University of Trier, Department of Economics.
- Energy Poverty Advisory Hub (2022). *Introduction to the Energy Poverty Advisory Hub (EPAH) Handbooks: A Guide to Understanding and Addressing Energy Poverty* Published by the Energy Poverty Advisory Hub
- European Commission (2014). *Micro-Fusion-09-M-Statistical Matching Methods v1.0.pdf*. Link: [Micro-Fusion - Statistical Matching Methods \(pdf file\) | CROS \(europa.eu\)](#)
- European Commission (2015). *Household Budget Survey 2015 Wave EU Quality Report*. Link: [72d7e310-c415-7806-93cc-e3bc7a49b596 \(europa.eu\)](#)

European Commission (2023). Energy Poverty in the EU. Link: [https://energy.ec.europa.eu/topics/markets-and-consumers/energy-consumer-rights/energy-poverty-eu\\_en](https://energy.ec.europa.eu/topics/markets-and-consumers/energy-consumer-rights/energy-poverty-eu_en)

Gao, J., Vinha, K., & Skoufias, E. (2020). World Bank Equity Policy Lab Vulnerability Tool to Measure Poverty Risk.

GESIS (2021). Study: EU-SILC 2021. Link: [GESIS: Missy - Study: EU-SILC 2021](#)

Gouveia et al. (2022). Energy Poverty. National Indicators. Insights for a more effective measuring. Link: [https://energy-poverty.ec.europa.eu/system/files/2023-01/EPAH\\_Energy%20Poverty%20National%20Indicators%20Report\\_0.pdf](https://energy-poverty.ec.europa.eu/system/files/2023-01/EPAH_Energy%20Poverty%20National%20Indicators%20Report_0.pdf)

Johnson and Young (2011). Towards Best Practices in analyzing Datasets with Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family*, 73(5): 926-45

Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys. *Large-scale assessments in education*, 1(1), 6.

Klasen, S., & Waibel, H. (2015). Vulnerability to poverty in South-East Asia: drivers, measurement, responses, and policy issues. *World Development*, 71, 1.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology*.

Lamarche, P., Oehler, F., & Rioboo, I. (2020). European household's income, consumption and wealth. *Statistical Journal of the IAOS*, 36(4), 1175-1188.

Hannah S Laqueur and others, SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations, *American Journal of Epidemiology*, Volume 191, Issue 3, March 2022, Pages 516–525, <https://doi.org/10.1093/aje/kwab271>

Leulescu, A., & Agafitei, M. (2013). Statistical matching: a model-based approach for data integration. *Eurostat-Methodologies and Working papers*, 10-2.

Lewaa, I., Hafez, M. S., & Ismail, M. A. (2021). Data integration using statistical matching techniques: A review. *Statistical Journal of the IAOS*, (Preprint), 1-20.

Little, R. J. A. 1988. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* 6: 287–296. <https://doi.org/10.2307/1391878>.

Little, R. J. A., and D. B. Rubin. 2020. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken, NJ: Wiley.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3), 1436-1462.

Moretti, A., & Shlomo, N. (2022). Improving Statistical Matching when Auxiliary Information is Available. *Journal of Survey Statistics and Methodology*.

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, 1-13.

R.H. Renssen, Use of Statistical Matching Techniques in Calibration Estimation, *Survey Methodology* 24 (1998), 171–183.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996. 5, 9, 10

Rubin, D. B. 1972. A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Journal of the Royal Statistical Society, Series C* 21: 136–141. <https://doi.org/10.2307/2346485>.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1), 87-94.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rude, Britta Laurin; Robayo, Monica. Quantifying Vulnerability to Poverty in El Salvador (English). Policy Research working paper; no. WPS 10289 Washington, D.C.: World Bank Group. <http://documents.worldbank.org/curated/en/099642102012330604/IDU0a3d39af50f12704d0d0889c0f48b6edbbdd>

Robayo-Abril, and Rude, Britta, 2023a "Conceptualizing and Measuring Energy Poverty in Bulgaria," forthcoming.

Robayo-Abril, and Rude, Britta, 2023b "Energy Affordability in Bulgaria: Effects of a Crisis and Potentials for a Pro-poor Clean Energy Transition," forthcoming.

Schaller, J. (2021). Datenfusion von EU-SILC und Household Budget Survey—ein Vergleich zweier Fusionsmethoden. *WISTA—Wirtschaft und Statistik*, 73(4), 76-86.

Schenker, N., and J. M. G. Taylor. 1996. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* 22: 425–446. [https://doi.org/10.1016/0167-9473\(95\)00057-7](https://doi.org/10.1016/0167-9473(95)00057-7).

Serafino, P., & Tonkin, R. (2017). Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey. Eurostat: Statistical Working Papers. Luxembourg: Publications Office of the European Union. Doi, 10, 933.

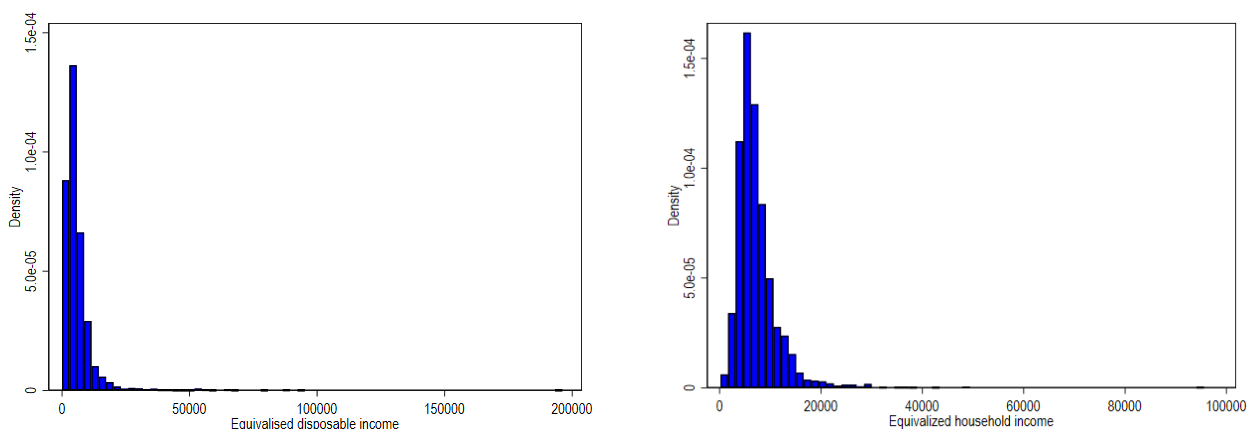
StataCorp LLC (2021). Stata Multiple-Imputation Reference Manual. Release 17. Stata Press. Link: [\[MI Multiple Imputation \(stata.com\)\]](#)

Thema, J., and Vondung, F. (2020) *EPOV Indicator Dashboard: Methodology Guidebook*. Wuppertal Institut für Klima, Umwelt, Energie GmbH.

Wiest, Maja; Kutscher, Tanja; Willeke, Janek; Merkel, Julie; Hoffmann, Madlain; Kaufmann-Kuchta, Katrin; Widany, Sarah: The potential of statistical matching for the analysis of wider benefits of learning in later life - In *European journal for Research on the Education and Learning of Adults* 10 (2019) 3, S. 291-306 - URN: urn:nbn:de:0111-pedocs-180965 - DOI: 10.25656/01:18096

## Appendix 1 – Additional Information on Data and Variables

Figure A 1: Distribution of equivalized household income (EU-SILC)    Figure A 2: Distribution of equivalized household income (HBS)



Source: Own estimates based on EU-SILC (2020) and HBS (2019).

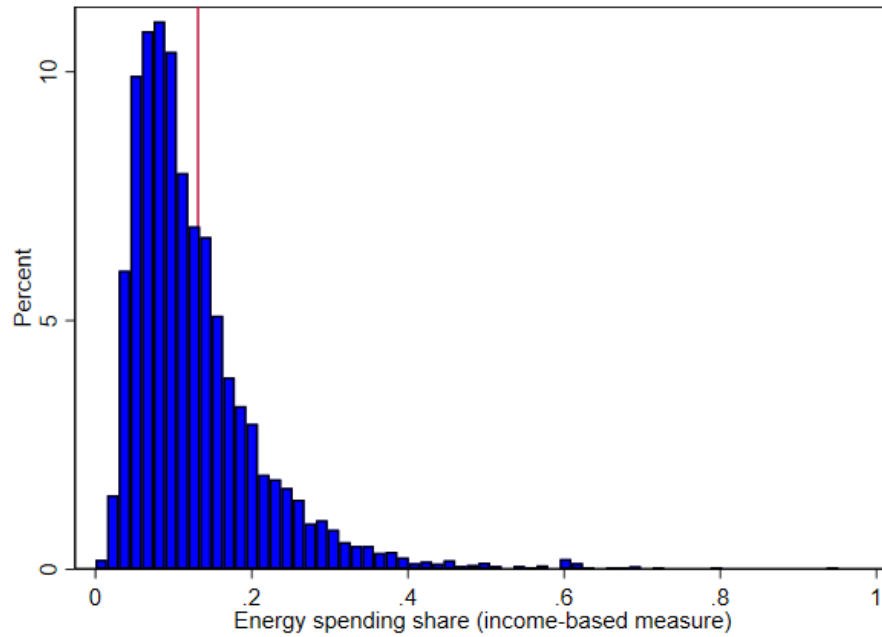
Table A 1: Variables identifying energy expenditure

HE_04_5	Electricity_gas, and other fuels	HE_A_04_5
HE_04_5_1	Electricity	HE_A_04_5_1
<b>HE_04_5_1_0</b>	<b>Electricity</b>	<b>HE_A_04_5_1_0</b>
HE_04_5_2	Gas	HE_A_04_5_2
<b>HE_04_5_2_1</b>	<b>Natural gas and town gas</b>	<b>HE_A_04_5_2_1</b>
<b>HE_04_5_2_2</b>	<b>Liquefied hydrocarbons, butane, propane, etc.</b>	<b>HE_A_04_5_2_2</b>
HE_04_5_3	Liquid fuels	HE_A_04_5_3
<b>HE_04_5_3_0</b>	<b>Liquid fuels</b>	<b>HE_A_04_5_3_0</b>
HE_04_5_4	Solid fuels	HE_A_04_5_4
<b>HE_04_5_4_1</b>	<b>Coal</b>	<b>HE_A_04_5_4_1</b>
<b>HE_04_5_4_9</b>	<b>Other solid fuels</b>	<b>HE_A_04_5_4_9</b>
HE_04_5_5	Heat energy	HE_A_04_5_5
<b>HE_04_5_5_0</b>	<b>Heat energy</b>	<b>HE_A_04_5_5_0</b>

Source: HBS Questionnaire provided by the NSI (2023).



Figure A 3: Histogram of energy spending share (2019)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimates based on HBS (2019)

Table A 2: Potential matching variables based on HBS 2019 and EU-SILC 2020

HBS 2019	Coding	EU-SILC 2020	Coding
Region of residence (HA08)	NUTS 2016 level 2 (6 categories)	Region (db040)	Only has two categories (BG3 and BG4)
Degree of urbanization (HA09)	3 categories	Degree of urbanization (DB100)	Three categories
Net income (total income from all sources minus income tax) (HH099) → without imputed rent	Refers to the same year	Total disposable household income (HY020)	Refers to the previous year

Sex (MB02)	2 values (1=male; 2=female)	Sex (PB150)	2 values (1=male; 2=female)
Age (MB03)	Continuous, but for some individuals only 5-year-bans	Age (PX020)	Continuous
Marital status (MB04)	4 categories (1=never married; 2=married; 3=widowed; 4=divorced; 5=not specified)	Marital status (PB190)	5 categories
Household members	Can be calculated from MA04 (Member ID)	Household members	Can be calculated
Nationality (MB01)	4 values (1=National; 2=Non-national and EU; 3=Non-national and non-EU; 9=Not stated)	Country of birth (PB210)	3 values (EU, LOC, OTH) // Not same as nationality
Country of main citizenship (MB011)	4 values (1=National; 2=Non-national and EU; 3=Non-national and non-EU; 9=Not stated)	Citizenship 1 alphanumeric (PB220A)	3 values (EU, LOC, OTH)
Highest education attained (ISCED 2011) (MC01)	9 categories	Highest ISCED level attained (PE040)	19 categories

Participation in formal education/training (MC02_A)	3 categories (1=Yes; 2=No; 9= not stated)	Highest ISCED level attained (PE010)	2 categories (1=Yes; 2=No)
Main activity (self-defined) (ME01_A)	8 categories	Self-defined current economic status (PL031)	11 categories
Main job: Full- or part-time (ME02)	2 categories (1=full-time; 2=part-time; 9=not stated)	Self-defined current economic status (PL031)	11 categories
Permanency of main job (ME03_A)	3 categories (1=permanent; 2=fixed-term contract; 9=not stated)	Type of contract (PL140)	2 categories (1=permanent; 2=temporary)
Economic activity (ME04)	NACE Rev.2	NACE (Rev 2) (PL111)	NACE Rev.2
Occupation in main job (ME0908)	ISCO08	Occupation (PL051)	ISCO08
Status in employment in main job (ME12)	5 categories (1=self-employed with employees; 2=self-employed without employees; 3=employee; 4=unpaid worker; 9=not stated)	Self-defined current economic status (PL031)	11 categories

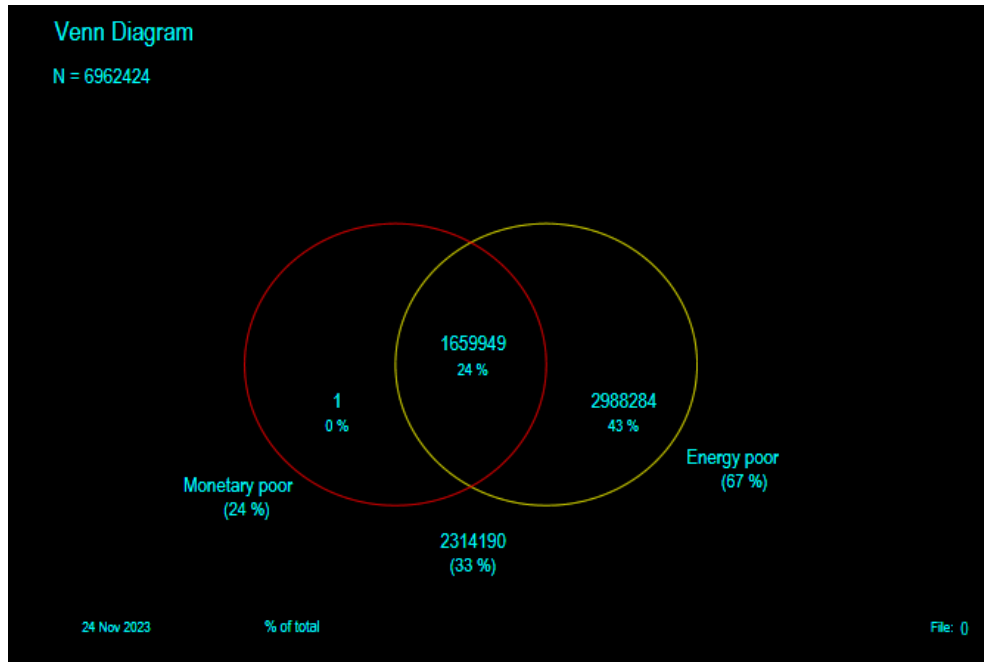
Source: Own elaboration based on HBS codebook and EU-SILC codebook.

Table A 3: Correlation coefficient table

	(1)
	Energy spending share
Household with children (<15)	-0.104***
Household with pensioner	0.166***
Household with elderly (>64)	0.158***
Female-headed household	0.063***
Household with unemployed	0.094***
Urban	-0.193***
Region	-0.050***
Income cat. (1-20)	-0.373***
No of female	-0.090***
No of self-employed	-0.059***
No in primary sector	-0.017***
No of part-time employed	-0.265***
No with primary educ.	-0.034***
No with secondary educ.	-0.012***
No with tertiary educ.	-0.202***
No attending educ.	-0.098***
No of foreign citizens	-0.010***
No of children	-0.092***
No of pensioners	0.114***
No of unemployed	0.089***
No. of household members	-0.152***
Adult equivalent	-0.154***
Observations	2952

Notes: The table depicts correlation coefficients of correlating each variable shown with the energy expenditure share. We weigh each observation by their respective survey weights. Source: Own elaboration based on HBS (2019)

Figure A 4: Venn Diagram of overlap between energy and monetary poverty



Note: Energy poverty is measured with the 10 percent measure.

## Appendix 2 – Detailed results

### Linear Regression Imputation – Specification 1

Figure A 5: Distribution of true and imputed energy spending share

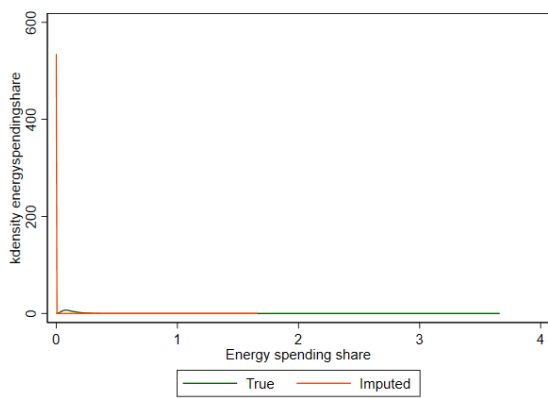
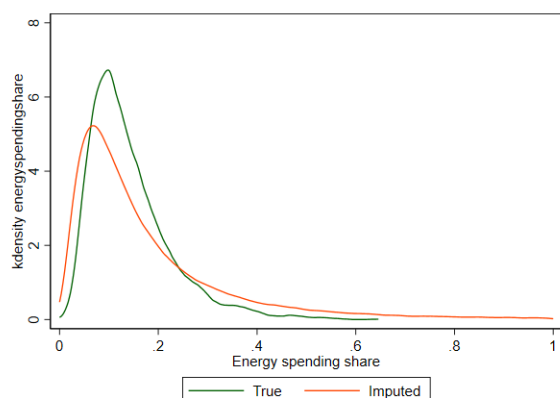


Figure A 6: Distribution of true and imputed energy spending share (shares restricted to values between 0-1)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Linear Regression Imputation – Specification 2

Figure A 7: Distribution of true and imputed energy spending share.

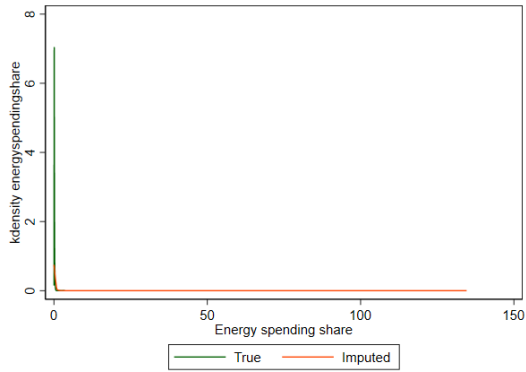
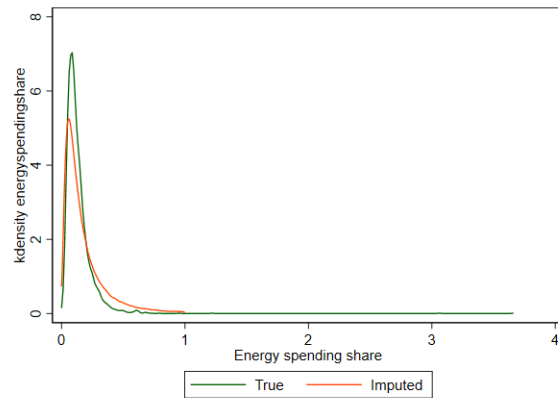


Figure A 8: Distribution of true and imputed energy spending share (imputed share restricted to values between 0-1)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Linear Regression Imputation – Specification 3

Figure A 9: Distribution of true and imputed energy spending share.

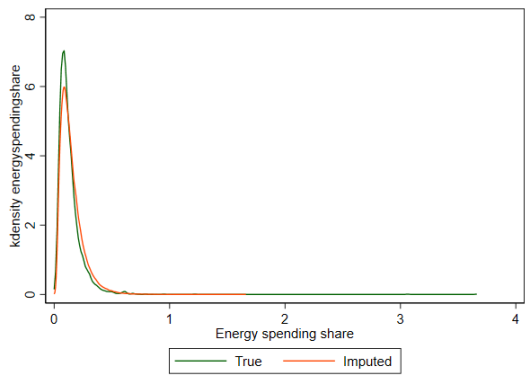
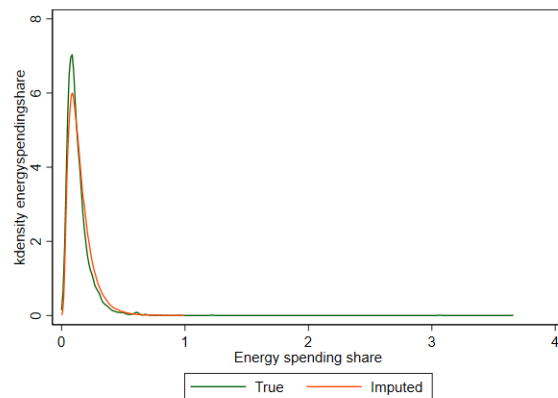


Figure A 10: Distribution of true and imputed energy spending share (imputed share restricted to 0-1)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Predictive Mean Matching – Specification 1

Figure A 11: Distribution of true and imputed energy spending share.

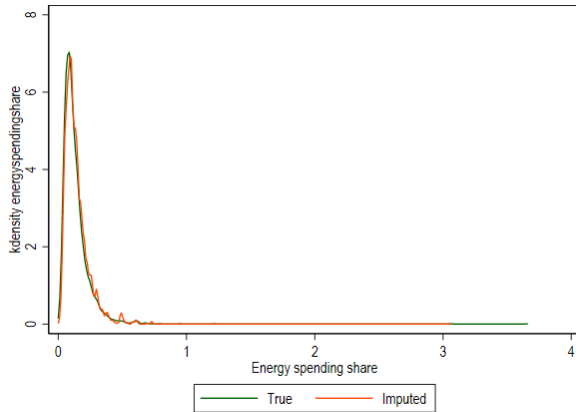
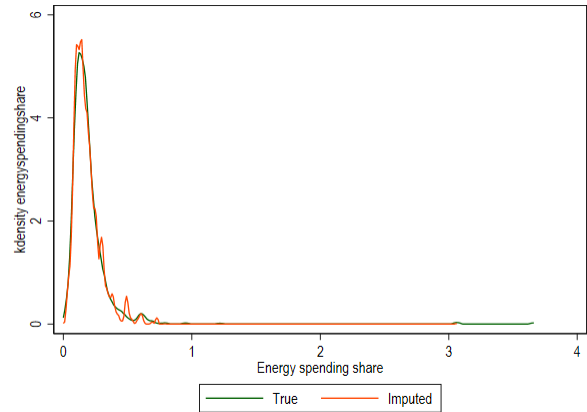


Figure A 12: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Predictive Mean Matching – Specification 2

Figure A 13: Distribution of true and imputed energy spending share.

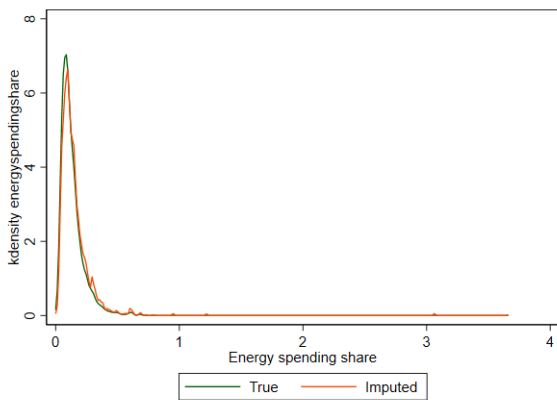
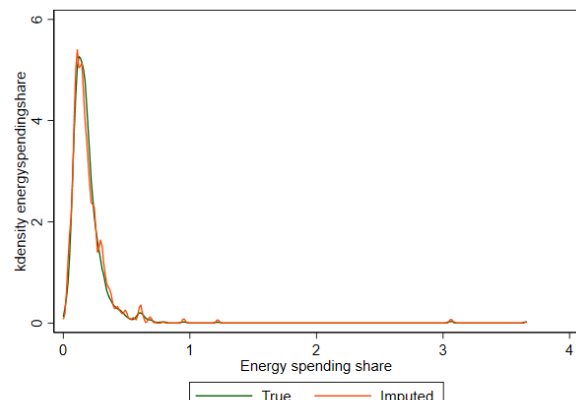


Figure A 14: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

### Predictive Mean Matching – Specification 3

Figure A 15: Distribution of true and imputed energy spending share.

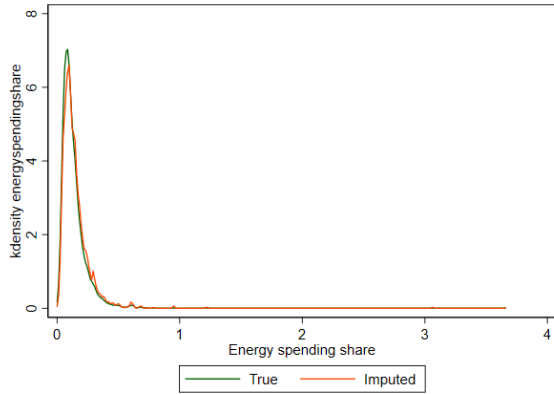
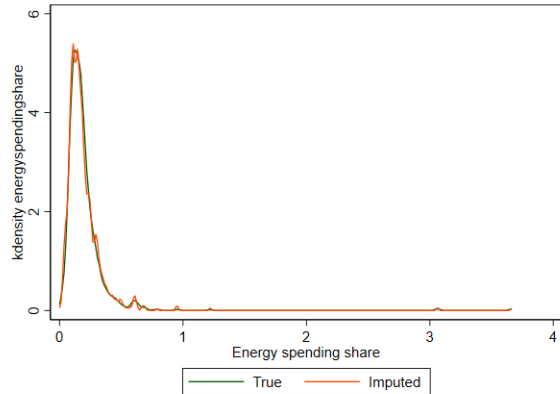


Figure A 16: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

### Predictive Mean Matching (5 nearest neighbors) – Specification 3

Figure A 17: Distribution of true and imputed energy spending share

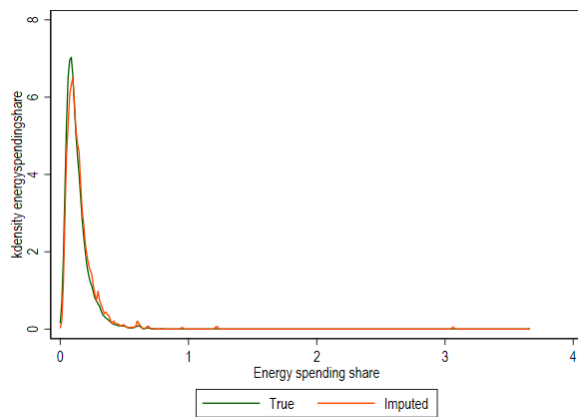
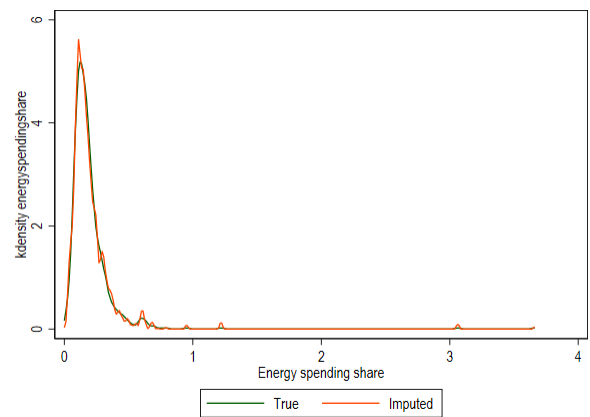


Figure A 18: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).



## Truncated Regression Imputation – Specification 1

Figure A 19: Distribution of true and imputed energy spending share.

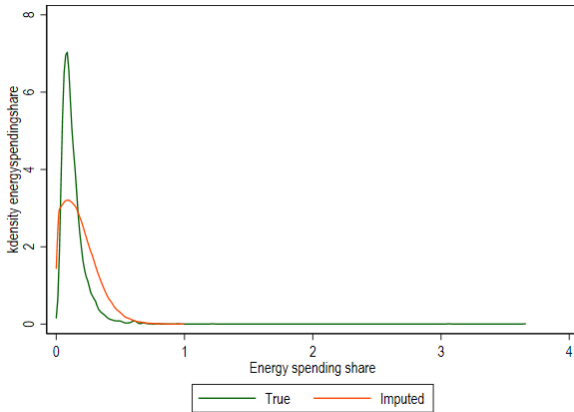
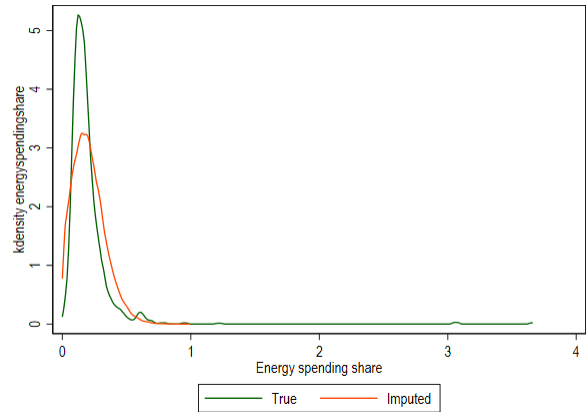


Figure A 20: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Truncated Regression Imputation – Specification 2

Figure A 21: Distribution of true and imputed energy spending share.

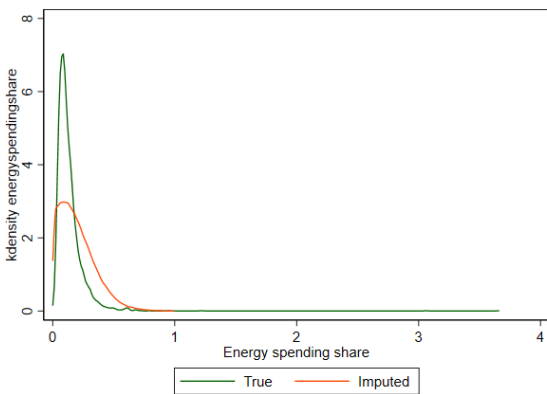
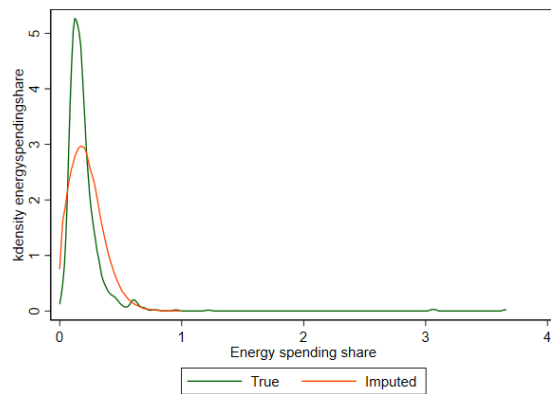


Figure A 22: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Truncated Regression Imputation – Specification 3

Figure A 23: Distribution of true and imputed energy spending share.

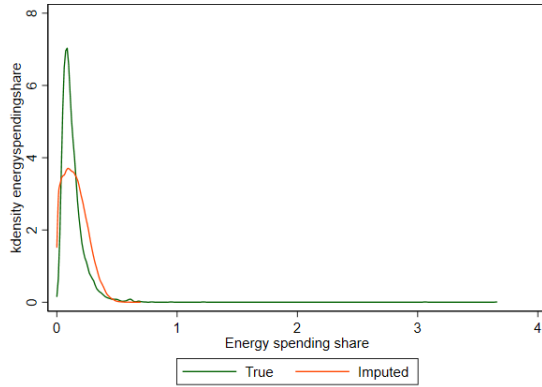
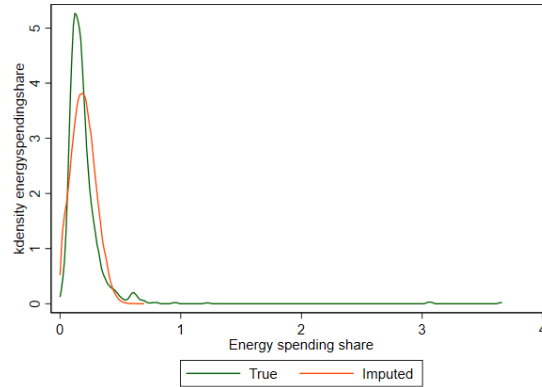


Figure A 24: Distribution of true and imputed energy spending share (5 lowest income categories)



Notes: We do not weigh observations by survey weights in this graph. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Mean and Standard Deviation of 20 simulations of weighted PMM

Table A 4: Mean and standard deviations of 20 simulations of weighted PMM

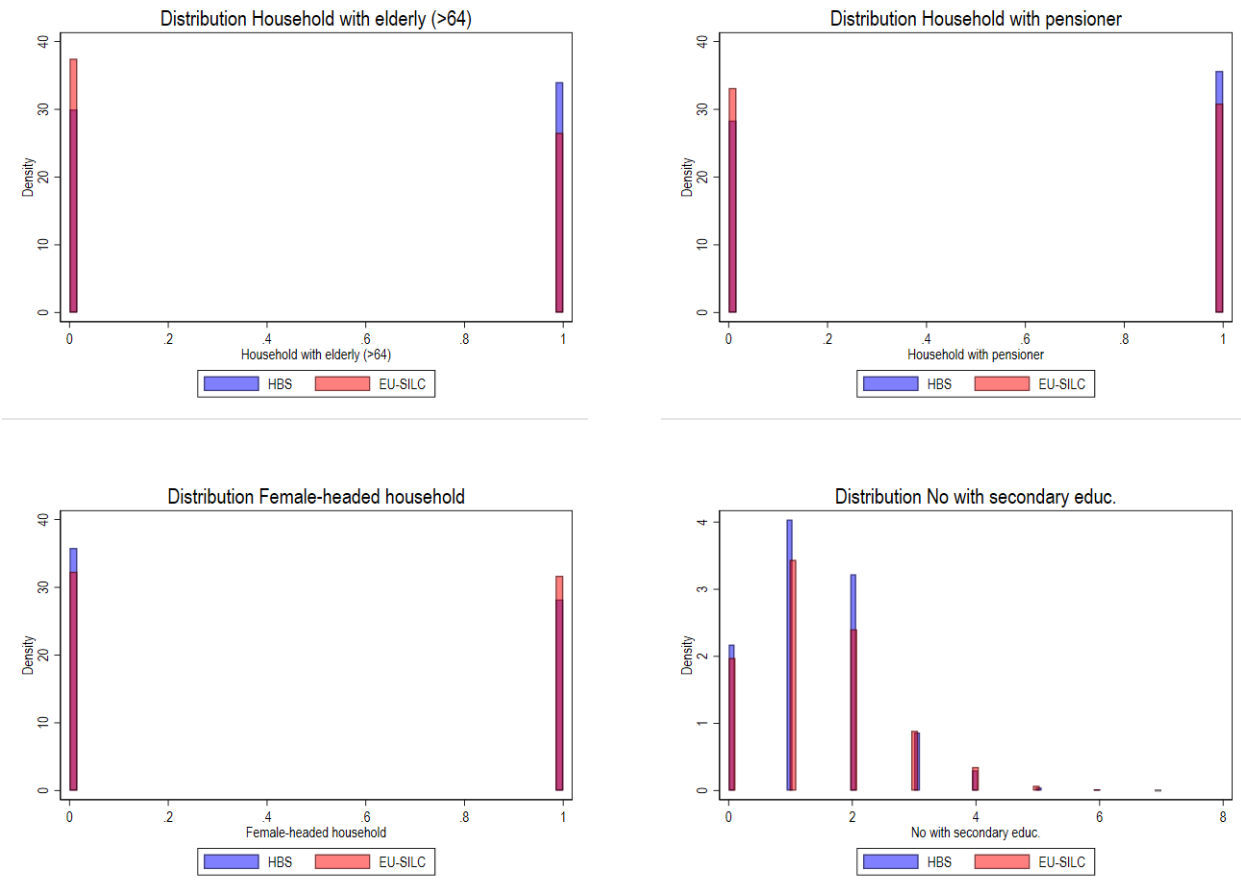
Number of imputation	Mean	Std. Dev.
0	0.146	0.139
1	0.149	0.181
2	0.146	0.154
3	0.143	0.114
4	0.148	0.163
5	0.147	0.145
6	0.147	0.147

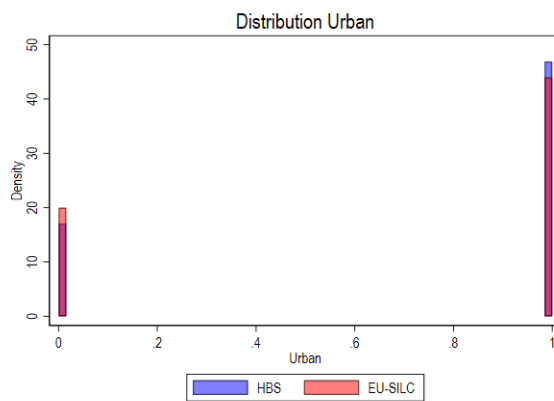
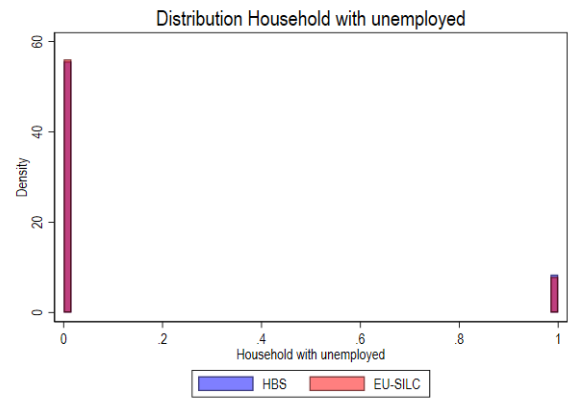
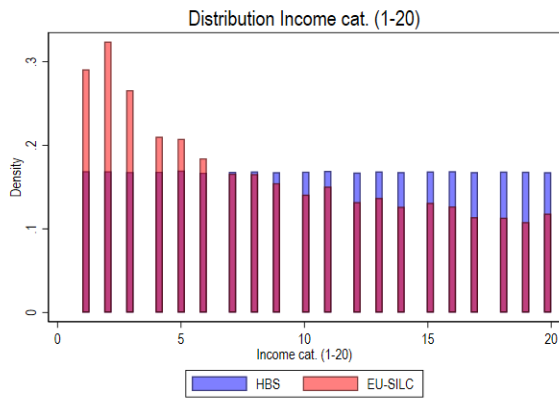
<b>7</b>	0.144	0.139
<b>8</b>	0.145	0.145
<b>9</b>	0.146	0.155
<b>10</b>	0.148	0.154
<b>11</b>	0.148	0.165
<b>12</b>	0.145	0.146
<b>13</b>	0.148	0.156
<b>14</b>	0.143	0.126
<b>15</b>	0.144	0.141
<b>16</b>	0.147	0.151
<b>17</b>	0.148	0.166
<b>18</b>	0.144	0.133
<b>19</b>	0.146	0.148
<b>20</b>	0.146	0.139

*Notes: Means and standard deviations reported in this table are not weighted by survey weights. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).*

# Appendix 3 – Distribution of matching variables across surveys

Figure A 25: Panel of covariates - HBS versus EU-SILC

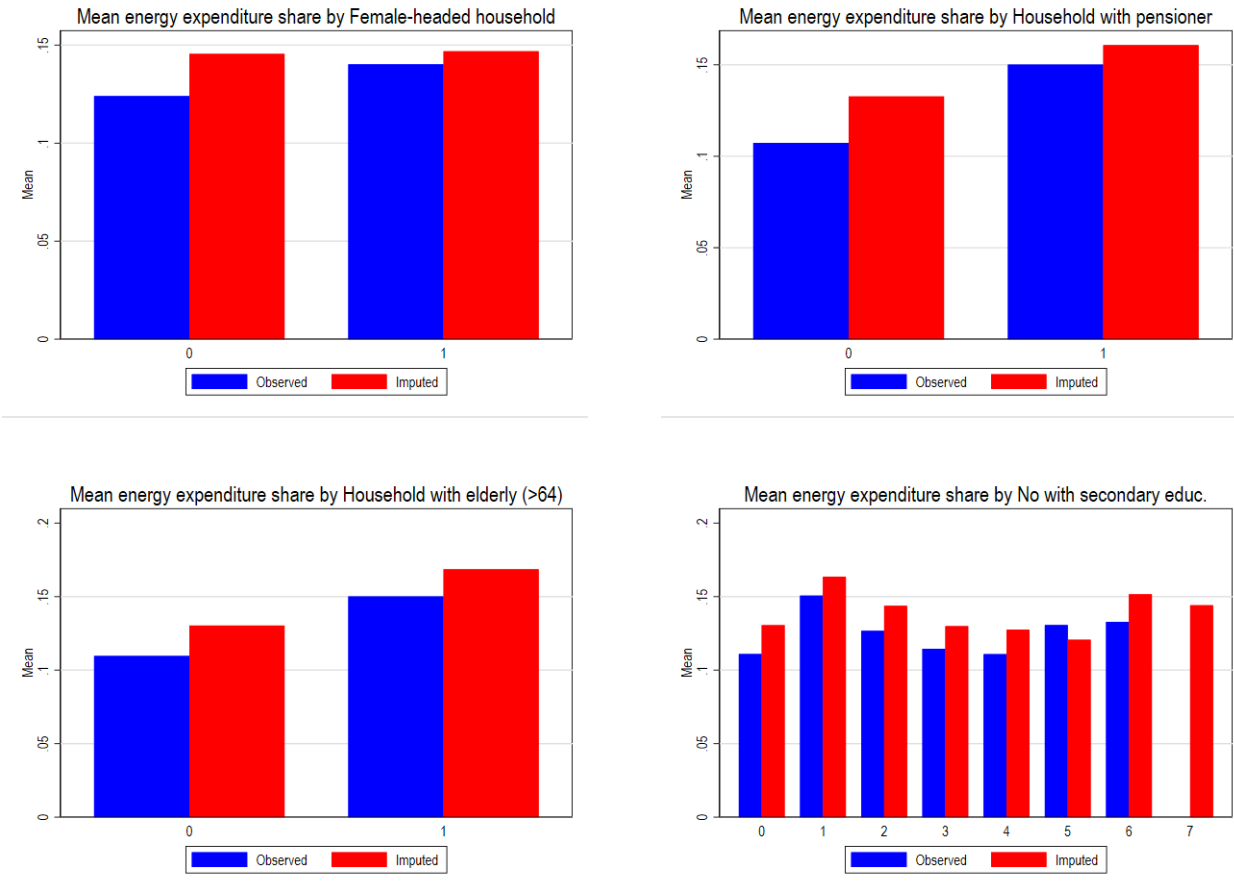


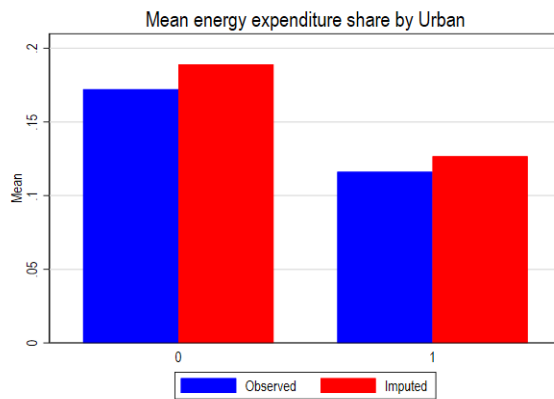
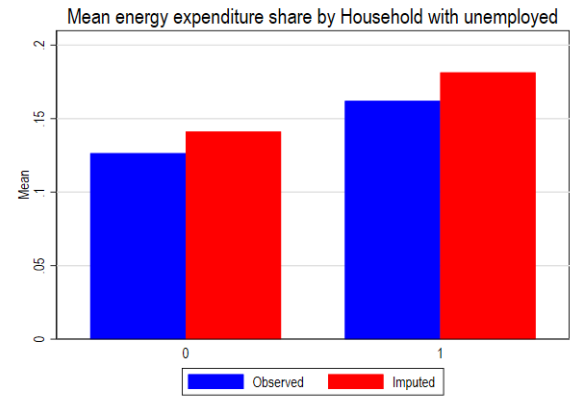
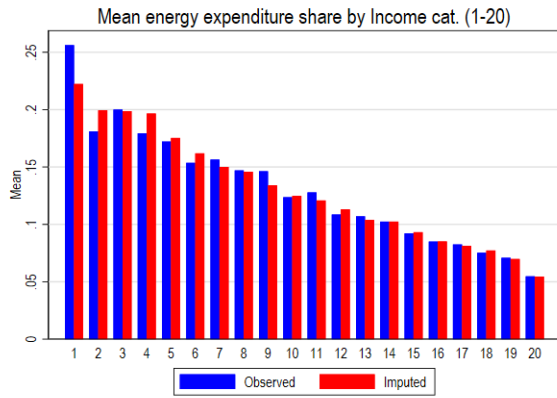


Notes: Observations are weighted by survey weights. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Appendix 4 – Mean household expenditure by matching variables and income quintiles

Figure A 26: Panel of average energy spending share (observed and imputed) by matching variables





Notes: Observations are weighted by survey weights. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

Figure A 27: Imputed energy spending share by income quintiles (Linear regression imputation method - Specification 3)

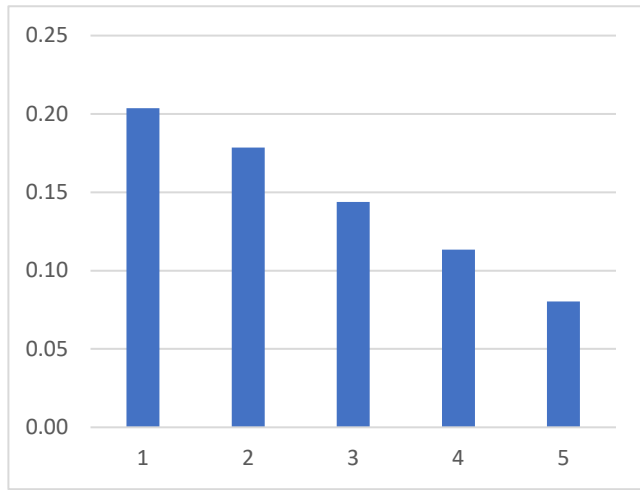
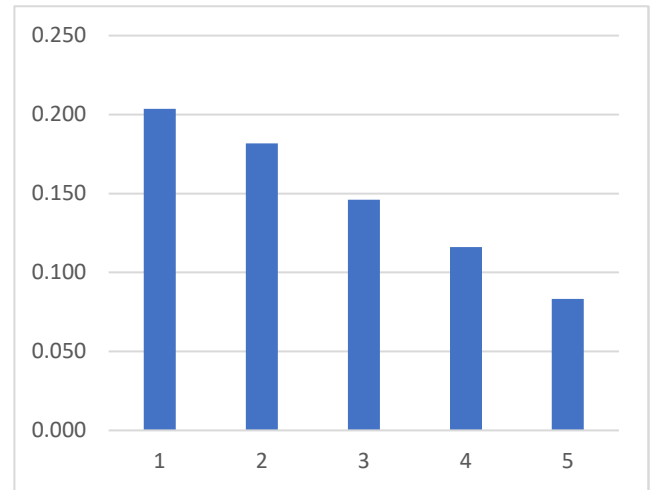


Figure A 28: Imputed energy spending share by income quintiles (Truncated regression imputation method - Specification 3)

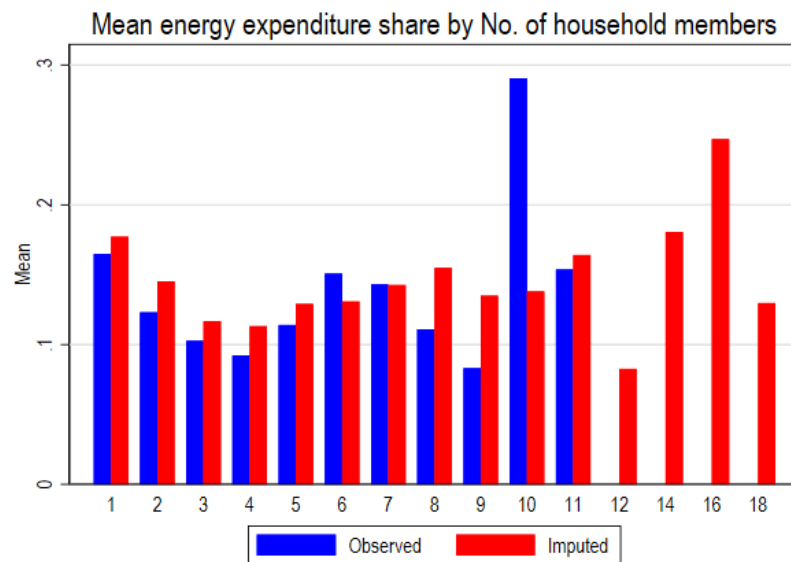


Notes: Observations are weighted by survey weights. 1 is the lowest income quintile and 5 the highest income quintiles.

Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Appendix 5 – Mean household expenditure by variable not used in matching

Figure A 29: Energy spending shares (observed and imputed) by household members





Notes: Observations are weighted by survey weights. Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

## Appendix 6

Table A 5: Regression Model (using HBS)

VARIABLES		(1)	(2)	(3)
		Energy spending share (log.)	Energy spending share (log.)	Energy spending share (log.)
Household pensioner	with	0.13*** (0.03)	0.13*** (0.03)	0.13*** (0.03)
Household with elderly (>64)		-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)
Female-headed household		-0.04** (0.02)	-0.04* (0.02)	-0.04* (0.02)
Household unemployed	with	0.04 (0.03)	0.04 (0.03)	0.03 (0.03)
Urban		-0.02 (0.14)	-0.17*** (0.02)	-0.16*** (0.02)
Income cat. (1-20)		-0.06*** (0.00)	-0.06*** (0.00)	-0.06*** (0.00)
No with secondary educ.		0.06*** (0.01)	0.07*** (0.01)	0.07*** (0.01)
HA10				0.00 (0.00)

Constant	-1.71*** (0.12)	-1.61*** (0.04)	-1.65*** (0.05)
Observations	2,951	2,951	2,951
R-squared	0.45	0.37	0.33
Sample FE	Yes	No	No
Weighted	Yes	Yes	No
MSE	0.471	0.498	0.519

Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).

Table A 6: Truncated Regression Model (using HBS)

VARIABLES	Expenditure share		
Household with pensioner	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Household with elderly (>64)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Female-headed household	-0.01* (0.00)	-0.01* (0.00)	-0.01* (0.00)
Household with unemployed	0.02*** (0.01)	0.02*** (0.01)	0.02** (0.01)
Urban	-0.01 (0.03)	-0.03*** (0.01)	-0.03*** (0.00)
Income cat. (1-20)	-0.01*** (0.00)	-0.01*** (0.00)	-0.01*** (0.00)
No with secondary educ.	0.00 (0.00)	0.00* (0.00)	0.01** (0.00)
sigma	0.11*** (0.00)	0.12*** (0.00)	0.12*** (0.00)

HA10			0.00 (0.00)
Constant	0.21*** (0.03)	0.23*** (0.01)	0.23*** (0.01)
Observations	2,952	2,952	2,952
R-squared	.2215	.1561	.1563
Sample FE	Yes	No	No
Weighted	Yes	Yes	No

*Source: Own estimation based on a harmonized, synthetic dataset consisting of HBS (2019) and EU-SILC (2020).*