Title: Data-driven estimation of building owners' budget restrictions on investing in deep renovation

Authors: Iná E.N. Maia, Rosimary Almeida, Lukas Kranzl, Ricardo Moraes, Andreas Müller and Fabian Schipfer

EU's deep renovation rates have not been sufficiently fast enough to significantly reduce building stocks' emissions. As a reaction against these facts, the newly published policy packages "Fit to 55" aim to boost renovation activities and to decarbonise the building stock until 2050. Nevertheless, building owners' affordability to pay for renovation has been recognised as a significant barrier. Therefore, a high amount of investments and adequate financing schemes are important instruments to enable the achievement of these goals.

The present paper contributes to this context by classifying households natural gas expenditure and their budget restrictions, relevant information to develop more user-targeted financing instruments. Budget restrictions are expressed through the household's savings (income minus expenditures). In this paper, the authors develop an approach to statistically match and test it using HBS and SILC data for 2015 Spain. Following research questions will be answered: what is the replicability of a method to merge HBS and SILC datasets? What can we learn about household annual natural gas expenditures, savings and incomes of four different household types? To carry out this analysis, mainly two databases of EU-households were used: EU-SILC (European Union Statistics on Income and Living Conditions) and HBS (Household Budget Surveys). The method consists of two steps. First, the application of a logistic regression model to perform a statistical match of both datasets. Then, statistical describing the income, savings and natural gas expenditure for four household types: single-family house owner-occupied, single-family house rented, multi-family house rented and multi-family house owner-occupied. The whole approach was carried out, tested and validated for data from Spain. 16% of the total households spend annually more than 600 euro for natural gas. Rented single family houses were identified as the most vulnerable household type, due to their low income and saving. Next steps are replicating the workflow to other countries. And, using the estimated budget restrictions as input data in building stock models.

Title: Statistical matching applied to EU-HBS/SILC data of households

Authors: Iná E.N. Maia, Rosimary Almeida, Lukas Kranzl, Ricardo Moraes, Andreas Müller and Fabian Schipfer

Focus: Statistical matching part

Abstract

There are many studies that classifies building's according to their technical characteristics. Other studies assess household's according to socio-economic characteristics. However, the literature review showed a gap on studies that combine both buildings' and households' characteristics. This is the initial motivation for the present paper: to derive techno-socio-economic household's archetype based on a data-driven analysis using Household Budget Surveys (HBS) data. However, preliminary analysis of existing European Union datasets concluded that the information of interest exists in two independent surveys: Household Budget Surveys (HBS) and European Union Statistics on Income and Living Conditions (SILC). Both surveys collects information across the European Union at household and person's level. Then, before working on the data-driven household's archetypes, the first methodological step is to carry out a statistical matching of both surveys - main scope of the present paper.

Statistical matching is a method to be applied, when piecewise information should be drawn from different independent surveys. This is a well-established statistical method, and can be applied in varies areas. Then, the research question is: can a statistical matching between EU-HBS and SILC surveys be carried out and provides accurate results? In the present paper, both HBS and SILC surveys are matched using a logistic regression model. Firstly, the available datasets were analysed, to compare and understand the existing variables. Secondly, the matching concept was drawn setting the receipt and donor datasets, defining the variables of interest and the matching variables. The datasets of different countries were pre-analysed, in terms of completeness of data and the HBS/SBS data was compared for each interest variable. Finally, the logistic regression model was drawn and delivered the results. The logistic model to predict the variable dwelling type from the SILC uses seven variables available in both datasets. The model presents an accuracy of 77%.

Next steps: apply the model to HBS data and provide the clustering to derive techno-socio-economic household's archetype.

1. State of the Art

1.1. Statistical matching and logistic regression

Statistical matching is a method that gains importance when dealing with available data, however from different data sources, being the desired information though the joining of both sources. One of the first applications of this method were performed by matching the Tax File and the Survey of Economic Opportunities (Okner 1972) – and creating a new data set with information on socio-demographic variables, income and tax returns by family. Nowadays, computational tools (for example: R and Python) facilitate the data manipulation and performance of the matching. However, these tools do not exclude the need of in-depth analysis of the data and understanding of the datasets through statistical analysis. Although, the statistical matching is not a new method, the literature review (as described below) could not find many papers that explicitly deal with the problem of merging HBS and SILC data sets, what is considered main contribution of the present paper. For the present analysis, HBS and SILC longitudinal and cross-sectional datasets for different EU-countries were available.

The book "Statistical matching: Theory and Practice" (Orazio, Di, and Scanu 2006) provides a very solid description of the method, containing theoretical and methodological chapters together with practical examples. In the literature, statistical matching is also called data fusion or synthetical matching.
Basically, the statistical matching consists of integrating different data-sources. In the reality, it may be less-costly to match datasets then funding new surveys or spending large amount of time to plan and execute new surveys. Therefore, this method is a practical solution to support data-driven analysis. Nevertheless, for applying statistical matching one of the conditions is that the different datasets contain a set of common variables (also calls matching variables). In the present paper, section 2.2 presents the matching variables.
In the current workflow, it is assumed a nonparametric framework – which means that the sample data was randomly drawn and therefore any particular probability distribution can be described. According to (Orazio, Di, and Scanu 2006) and the literature review, this is a commonly used assumption which was also made in the present paper, because parametric assumptions lead to misleading results.
The reviewed literature showed that "*hot deck imputation* procedures" are the commonly used nonparametric method. It consists of filling missing values in a dataset based on observations from the other dataset. Therefore, we defined the framework to apply the "*distance hot deck imputation*", starting from the recipient and donor datasets. The recipient dataset is the one with absence of variable of interest. The donor dataset has the variable of interest. Finally, the matched dataset consists imputing the missing data in the recipient dataset. In the present paper, the SILC data sets is the donor. The variable of interest is the dwelling type and the HBS is the recipient dataset. Before the matching was performed, we performed a pre-processing analysis of the completeness of the data sets for different EU-countries – presented in section 2.1. Based on that pre-processing, it was decided to focus the analysis on the case study for Spain, due to the completeness and comparability as presented in the section 2.2.
There are three types of hot deck methods: *random*, *rank* and *distance* hot deck.

Then, a statistical model has to be established to estimate the probability distribution function of the variable of interest. The main challenge of this step is that the model addresses the distribution function for the recipient dataset based on the donor dataset, which means that the distribution function of the variable of interest in the recipient dataset is actually unknown. Therefore, the accuracy of the model estimator is a useful indicator to measure model's performance – meaning, the ability to be very close to the true but unknown distribution. In the macro approach (Orazio et al. 2006) followed the data from the donor dataset (SILC) will be used to estimate the joint distribution function of the variable of interest – the categorical variable "dwelling type". Therefore, a logistic regression model described in the section 3 is used. This model generates the key characteristics and correlation between variables, from each the coefficients will estimate the probability distribution function, to be applied to the recipient dataset.

QUESTION: parametric or nonparametric framework -> what is the one chosen here?

MISSING LINK: parametric versus non-parametric versus regression model

Parametric statistics is a branch of statistics which assumes that sample data comes from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters.[1] Conversely a **non-parametric model** does not assume an explicit (finite-parametric) mathematical form for the distribution when modeling the data. However, it may make some assumptions about that distribution, such as continuity or symmetry (Geisser and Johnson 2006).

1.1. HBS/SILC analysis applied to the building energy systems

Although, there common information between these surveys, they have different objectives. The HBS focuses on information about household's final consumption expenditure on goods and services (detailly divided in sub-categories); while the SILC collects data on income, poverty, social exclusion and living conditions. Besides that, also information about labour, education and health is obtained. HBS and SILC are both information-rich surveys. The HBS datasets consist of basically two files, that summed present data for about 630 variables; the SILC consists of four files, and together, mostly 600 variables are provided.

https://en.wikipedia.org/wiki/Nonparametric_statistics

2. Data description

2.1. Datasets pre-processing

2.2. Matching variables: the case of Spain

The authors (D'Orazio, Di Zio, and Scanu 2017) wrote that usually the selection of matching variables is performed based on expert knowledge. The same authors described a method to select the variables and reduce their uncertainties in the estimation. This is especially relevant in cases where the number of variables is very high. In the present paper, the matching variables were selected firstly through observing and comparing the availability of them in both datasets. Hence, also the common definition and the categorisation (especially for categorical variables) were taken into account. Second, the comparison of each variable in both datasets (HBS and SILC) samples were performed for the selected country. Below, you find the list of the matching variables, and the comparison between both datasets for the selected country Spain.

3. Method

1) Logistic regression versus nonparametric macro/micro methods

4. Results

5. Conclusions

6. References

D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu. 2017. "The Use of Uncertainty to Choose Matching Variables in Statistical Matching." *International Journal of Approximate Reasoning* 90:433–40. doi: 10.1016/J.IJAR.2017.08.015.

Geisser, Seymour., and Wesley O. Johnson. 2006. "Modes of Parametric Statistical Inference." 192.

Okner, Bejamin. 1972. "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File." Retrieved April 1, 2022 (https://econpapers.repec.org/bookchap/nbrnberch/9435.htm).

Orazio, M., Zio Di, and M. Scanu. 2006. *Statistical Matching Statistical Matching: Theory and Practice*.

Description of the work: SILC and HBS merge

Summary

this work is about understanding how to merge the SILC and HBS datasets to cluster them and deliver techno-socio-economic archetypes and calculate the ranges of budget restriction (BR)
S: I – Sum(Exp)
I: income
Exp: expenditures

Description of the datasets:

SILC database are Social Income and Living Conditions Survey for EU-MS in the 2004-2018 (long and cross)
HBS database is the Household Budget Survey for EU-MS in 2010 and 2015

File with the overview per country and year: C:\Users\maia\ownCloud3\phd\silc_hbs_analysis\available_data_countries.xlsx

Data used

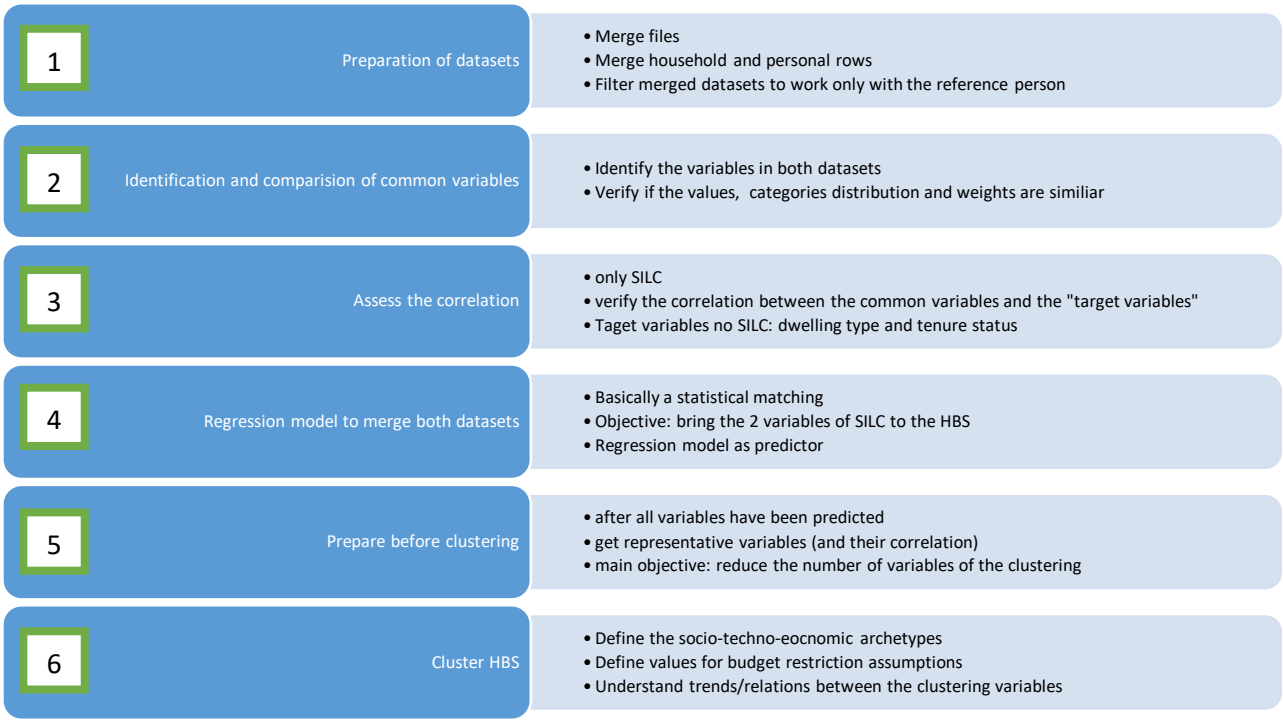
DE 2015 HBS and SILC

- **HBS**

Number of households (file H)	52412
Number of variables (file H)	605
Number of individual (file M)	110236
Number of variables (file M)	26

Methodology

Objective: bring SILC variables to the HBS and cluster variables of the HBS, taking into according dwelling type and tenure status



1- Preparation of the datasets

Studied countries

Country	Dataset2	Samples	Variables	Dataset1	Samples3	Variables4
BG	SILC	4947	602	HBS	2966	631
DE	SILC	12902	602	HBS	52412	631
ES	SILC	12142	607	HBS	22130	631
FR	SILC	10882	607	HBS	16978	631
IT	SILC	17983	602	HBS	15013	631
NL	SILC	9806	607	HBS	14408	631
PL	SILC	12183	602	HBS	37148	631
RO	SILC	7411	607	HBS	30625	631
SE	SILC	5859	602	HBS	not filtered	631

2- Identification and comparison of common variables

Sweden: not included due to unavailability of the variable MA05 (reference person)

Continuous variables: Total housing costs and Disposable income

Conclusion:

- Total housing costs : not possible to be used in the statistical matching
- Disposable income: possible to be used in 5 countries

Country	Variable	SILC	HBS	SILC comparision results	Other comments	Variable	SILC	HBS	SILC comparision results	Other comments
BG	Total housing costs	Available	Available	Not similar	SILC values are lower	Disposable income	Available	Available	Not similar	HBS values very low
DE	Total housing costs	Not available	Available	Comparison not possible		Disposable income	Available	Available	Similar values	
ES	Total housing costs	Not available	Available	Comparison not possible		Disposable income	Available	Available	Similar values	
FR	Total housing costs	Not available	Available	Comparison not possible		Disposable income	Available	Available	Similar values	Disposable income boxplots very close to zero
IT	Total housing costs	Available	Available	Not similar	SILC values very low	Disposable income	Available	Not available	Comparison not possible	
NL	Total housing costs	Not available	Available	Comparison not possible		Disposable income	Available	Available	Similar values	Disposable income boxplots very close to zero
PL	Total housing costs	Not available	Available	Comparison not possible		Disposable income	Available	Available	Similar values	Disposable income boxplots very close to zero
RO	Total housing costs	Not available	Available	Comparison not possible		Disposable income	Not available	Available	Comparison not possible	

Categorical variables

- Marital status
- Employment
- Economic status

File: ownCloud3\phd\silc_hbs_analysis\statistical match datasets\variaveis HBS_SILC_v8.xlsx

3- Assess the correlation between common variables and variable of interest

Variable of interest

SILC: HH010: Dwelling type

	New coding
1 Detached house	*0*
2 Semi-detached or terraced house	0
3 Apartment or flat in a building with less than 10 dwellings	1
4 Apartment or flat in a building with 10 or more dwellings	*1*
5 Some other kind of accommodation	*2*

SILC HH021: Tenure status

	New coding
1 Outright owner	*0*
2 Owner paying mortgage	*1*
3 Tenant or subtenant paying rent at prevailing or market rate	1
4 Accommodation is rented at a reduced rate (lower price than the market price)	*1*
5 Accommodation is provided free	*1*

4- Regression model to merge both datasets

5- Data preparation before clustering

6- Techno-socio-economic clusters

Cluster parameters:

1st possibility

techno (household type)
economic (tenure status)
socio (age)
economic status: working and retired
urbanisation degree (HBS,HA09)

2nd possibility

Set parameters according to correlation (data-driven) – check with Guillermo

Conclusions

HBS much higher number of samples than SILC (5 times more)
SILC basically working and retired peopled, while HBS many other economic groups (disabled etc)
Degree of urbanisation: empty in SILC but would be very interesting