Statistical matching of EU-SILC and HBS at the European level: a flexible strategy based on the optimisation of the nearest neighbour distance hot deck method

Tomás, Manuel [a]*; Mariel, Petr [b]; Arto, Iñaki [a]; Kratena, Kurt [c]

[a]*Basque Centre for Climate Change (BC3)*
[b] *University of the Basque Country (UPV/EHU)*
[c] *Centre of Economic Scenario Analysis and Research (CESAR)*

Corresponding author information*

E-mail: manuel.tomas@bc3research.org; Tel: +34 944 014 690 Ext. 183

ABSTRACT

In recent years, there is an emerging consensus on the need to observe micro-data on households' income, consumption, saving, and wealth along with other relevant socio-economic variables in a single dataset. Statistics collecting the joint distribution of all those dimensions could provide a full picture of households' economic situation either in terms of inequality or poverty. At the same time, it could be used as an instrument to build Distributional National Accounts by describing the economic behaviour of many types of households at the macro level. Unfortunately, in most countries of the European Union, data on income and consumption are compiled by Member States through two separated surveys: the European Union Statistics on Income and Living Conditions (EU-SILC) and Household Budget Survey (HBS). A critical question is, therefore, how to integrate the information of these two datasets in order to provide a complete picture of household income and consumption dynamics.

This paper presents a new strategy to merge the microdata of the EU-SILC and the HBS databases. To that end, we use statistical matching techniques which encompass a range of well-known statistical methods aimed at integrating different surveys referred to the same target population in a new synthetic dataset. We use the Nearest Neighbour Distance Hot Deck non-parametric method which employs a set of common variables observed in both surveys and distance functions to match the records from the donor survey (HBS) to the recipient survey (EU-SILC) minimising the aggregate absolute distance of all pairings.

Our strategy shows a set of desirable properties. First, it can be applied to all European countries in a harmonised way, regardless of each country's survey sample size. Second, both samples are stratified using the income distribution which enables to hold the Conditional Independence Assumption, something that is essential for guaranteeing the reliability of the results (it should be noted that the HBS includes auxiliary information on household income). Third, our algorithm optimises the matching procedure by testing all the possible matching alternatives (i.e., we take into account all possible combinations of matching variables, different stratification sizes and several distance functions). The algorithm looks for the lowest distance between the original cumulative distribution of total consumption in the HBS and that of the total consumption imputed in the matched data.

This exercise presents some limitations; therefore, the results obtained should be used with caution. However, the resulting joint EU-SILC-HBS dataset shows an improvement in terms of quality over other

approaches described in the literature. We also observe that the different choices related to the stratification size, matching variables and distance functions have a critical impact on the goodness of the matching. Finally, we argue that, when integrated surveys are not available or record-linkage techniques are not feasible, statistical matching techniques can be useful and necessary instruments to create joint-micro-datasets aimed at improving households representation in macroeconomic models or creating different types of inequality and poverty indicators. This information is critical to analyses the distributional impacts of policies and to monitor the evolution of poverty or (in)equality.