New Techniques and Technologies for Statistics Conference (NTTS 2021)

9-11 March 2021, Brussels

Overcoming the Conditional Independence Assumption in matching income, consumption and wealth

Irene Rioboo and Friderike Oehler

F1 – Social Indicators, methodology and development, relations with users

**Keywords:** Income, Consumption, Wealth, Economic wellbeing, Statistical matching, CIA, Auxiliary information.

**Abstract**

# Summary

In view of the lack of harmonised EU statistics covering the distributional aspects of households' income, consumption and wealth in a single data set, an alternative commonly used to estimate that joint distribution is statistical matching. A main problem in statistical matching is that the relationship between income, consumption and wealth can be estimated only indirectly, relying on a strong assumption: the Conditional Independence Assumption (CIA), which might not hold and it is difficult to test. This work presents the tools to overcome the CIA, making it a justifiable hypothesis and testing it with the help of auxiliary information.

# Introduction

Income, consumption and wealth (ICW) are key dimensions to determine households' economic behaviours. The study of their joint distribution provides a better picture of the economic wellbeing and material inequality of households, showing characteristics hidden in a unidimensional analysis.

In view of the lack of harmonised EU statistics covering the distributional aspects of households' income, consumption and wealth in a single data set, an alternative commonly used to estimate that joint distribution is statistical matching [1]. This technique aims to integrate in a synthetic dataset specific variables from different independent sources referring to the same target population. The matching is done by using common information shared among the data sources as a link. The data sources used in the present study are the EU Statistics on Income and Living Conditions (EU-SILC), the Household Budget Survey (HBS) and the Household Finance and Consumption Survey (HFCS).

A main problem in statistical matching is that the relationship between the target variables (income, consumption and wealth) can be estimated only indirectly, relying on assumptions that are difficult to test. The most common one is the Conditional Independence Assumption (CIA), which states that the relationship between the target variables can be fully explained by the values of the variables that those sources have in common. This is a very strong assumption, which might not hold and which is difficult to test trough the synthetic dataset created.

In this work, we present the tools used to overcome the CIA, making it a justifiable hypothesis and testing it with the help of an additional data source.

# Methods

Statistical matching is a technique to join variables of interest, known as target variables, from different data sets without common identifiers based on a set of common variables that explain the target variables, known as matching variables. This technique has been well described in literature, for example in Renssen [2], Rässler [3], D’Orazio et al. [4] and Eurostat [5].

We match data of the three original ICW data sources in two steps: First, we merge the consumption expenditure of households from HBS 2015 with EU-SILC household income of the same year. For data available countries, we then experiment with matching the SILC-HBS fused data with HFCS data, in order to join wealth data to the income-consumption distribution previously created. We thus obtain estimates of the full distribution of income, consumption and wealth.

In both steps, we have to deal with the CIA, which is a hypothesis that, on one hand, does not seem plausible working with income, consumption and wealth, and on the other hand, cannot be tested from the fused data set SILC-HBS-HFCS. However, we were able to address and overcome both limitations in our matching exercise as follows:

Zhang [6] describes a possibility to make the CIA a more justifiable and plausible assumption by including in the set of matching variables a proxy of one of the target variables. Adopting this approach, we used in the first step the total income divided in 20 quantiles, available both for EU-SILC and HBS, as proxy for the EU-SILC target variable total disposable income. In the second step, we used gross income, which is available both in EU-SILC and HFCS, as a proxy of total disposal income.

As mentioned before, the synthetic dataset obtained does not allow to test whether the CIA is fulfilled. The only possibility to do so is using additional samples in which the target variables are jointly observed. We have used the Over-indebtedness, Consumption and Wealth (OCW) testing module for EU-SILC 2017, which offers joint information about both income, consumption and wealth. This auxiliary information allows to check if independence between the target variables can be assumed while controlled by the matching variables.

# Results

The use of the chosen proxy variables for the targets as matching variables (i) has made the independence between income, consumption and wealth plausible given the matching variables, i.e. it has justified maintaining the CIA.

Moreover, the 2017 OCW module has allowed to test the CIA (ii) and to affirm that this hypothesis holds, meaning that independence between the target variables can be assumed while controlled by the matching variables. The results indicate that the partial correlation between the target variables given the proxy variables are very low, being in most of the cases not statistically significant anymore, which shows that the existing relationship among those target variables is mainly explained by the proxy variables.

# Conclusions

Considering that the CIA is a very strong assumption, which might not hold, an option to improve the matching process is the use of both proxy variables as matching variables and auxiliary information to test the relationship between the target variables. For our study, such auxiliary information is available for some countries in the 2017 OCW module, which allows to say not only that the CIA is a justifiable assumption, but also that it holds in our matching exercise, which is an essential aspect to assess the global quality of the statistical matching of ICW statistics.

# References

1. Eurostat (2017). Measuring Income, Consumption and Wealth jointly at the micro-level. <https://ec.europa.eu/eurostat/documents/7894008/8074103/income_methodological_note.pdf>
2. Renssen, R. H. (1998). Use of statistical matching techniques calibration estimation. Survey Methodology, vol 24, n. 2, pp. 171-183.
3. Rässler, S. (2002). Statistical Matching, a Frequentist Theory, Practical Applications and Alternative Bayesian Approach. Springer.
4. D'Orazio M, Di Zio M, Scanu M. (2006). Statistical matching: Theory and practice. John Wiley & Sons.
5. Eurostat (2013). Statistical matching: a model base d approach for data integration. Luxembourg: Publications Office of the European Union. <https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF>
6. Zhang, Li-Chun (2015). On proxy variables and categorical data fusion. Journal of Official Statistics, 31 (4), pp. 783-807.