WP. 25
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (v): Successful strategies for implementing new editing and imputation methods

DECENTRALISED MULTI-COUNTRY IMPUTATION: THE CASE OF THE EURO AREA HOUSEHOLD FINANCE AND CONSUMPTION SURVEY (HFCS)

Supporting Paper

Prepared by Claudia Biancotti, European Central Bank, Arthur Kennickell, Federal Reserve Board, Washington and Carlos Sánchez Muñoz, European Central Bank[1]

I. INTRODUCTION AND OUTLINE OF THE PAPER

1. Household surveys can provide information that is difficult or impossible to obtain in other ways. However, surveys also introduce the possibility of missing information: survey respondents may be unable or unwilling to answer a question, there may be errors in collecting the data, and there may be errors detected after data collection. Although it is possible to analyse data with missing values (Little and Rubin, 1987), it can be quite complicated to do so. In addition, there is often important specialised information or technical expertise available to the data compilers that is not available to later users of the data.

2. These and other such factors are typically used to justify imputation of missing data in surveys. Besides the sensitivity of the topics covered by the survey presented in this paper, there is an additional unusual complication: as discussed in more detail below, this survey is actually a decentralised multi-country effort with a set of common principles and some degree of central coordination. This paper describes the nature of the challenge and presents a strategy to tackle it.

3. The paper is in five sections. Section II provides a short description of the HFCS. Section III deals with the features of the imputation methodology selected for the HFCS and the motivation behind. Sections IV reviews the conceptual challenges imposed on imputation by the multi-country nature of the HFCS. Section V concludes.

II. THE HOUSEHOLD FINANCE AND CONSUMPTION SURVEY (HFCS)[2]

4. In September 2008, the ECB Governing Council[3] approved the conduct of the HFCS, a set of household surveys co-ordinated by the European Central Bank (ECB). The HFCS is conducted in all euro

[1] Corresponding author: claudia.biancotti@ecb.int. The views expressed in this paper are those of the authors, and do not necessarily represent the views of their respective institutions nor those of the Household Finance and Consumption Network

[2] General information about the HFCS, including the blueprint questionnaire, the list of output variables and other background documents, is available on the public HFCS website (http://www.ecb.europa.eu/home/html/researcher_hfcn.en.html)

[3] From www.ecb.int: "The Governing Council is the main decision-making body of the ECB. It consists of the six members of the Executive Board and the governors, respectively presidents of all national central banks (NCBs) of the euro area."

area countries[4] by national central banks (NCBs), in a few countries in collaboration with national statistical institutes (NSIs) or other research partners.

5. The HFCS comprises structural micro-level information on euro area households' financial and real assets and liabilities coupled with information on demographic characteristics of households, consumption expenditure, future pension entitlements, employment, income and savings, intergenerational transfers and gifts, expectations and attitudes. Besides person-specific demographic, employment and pension information, most parts of the survey target the household as a whole.

6. The European System of Central Banks (ESCB) has a twofold motivation for carrying out this type of survey. First, it is instrumental in formulating monetary policy decisions, as micro-level data can provide fundamental insights on the transmission channels and the ultimate effects of interest rate decisions. Secondly, it can help to safeguard financial stability by providing a better picture of how households allocate and manage their savings, how various types of assets are distributed in the population of non-professional investors, how groups of households differ in terms of financial literacy and attitudes toward risk, etc. The use of this information is expected to become an important element in assessing the vulnerability of the household sector to shocks and its role in the transmission of financial instability to the real economy.

7. The launch of the HFCS is also envisaged to open various areas of policy-relevant research beyond the confines of the ESCB, as appropriately anonymised micro-level data from the survey become available to the academic community. For example, new opportunities for analysis of the portfolio composition of individual households in the Euro area may provide hints as to how much sharp increases/declines in residential property prices may push up/down consumption expenditure, the sustainability of household indebtedness, riskiness and liquidity of household portfolios, signs of financial distress, etc. This may in turn provide powerful structural policy tools to assess/micro-simulate how income, interest-rate, stock-price, or exchange-rate shocks may propagate and affect macro aggregates. Another relatively uncharted field where the HFCS data may prove to be crucial encompasses the effects of the crisis on the investment and borrowing decisions of households (e.g. tendency to opt for safer assets, possible re-balance between real and financial assets, collateralised versus uncollateralised debt, financial constrains, etc.) as well as on their savings and consumption patterns.

8. The decentralised organisation of the HFCS takes advantage of some existing country wealth surveys, thus maximising cost-effectiveness of the effort. The existence of independent country samples is also aimed at permitting that results are representative not only for the whole euro area, but also at the individual country level. Inference at the country level is considered to be particularly important for investigating how institutional differences across euro area countries affect the financial and economic behaviour of households, the distribution of wealth, the propagation of shocks, etc.

9. Although countries are expected to develop their questionnaires around a commonly agreed reference blueprint, there will be inevitable asymmetries reflecting differences in institutional structures and in the importance of particular topics. The primary common product across the country surveys is defined in terms of a set of output variables based on common definitions.

10. All country surveys should also comply with other common technical characteristics. For instance, they will all be conducted every two or every three years[5] and will have a probabilistic sample design[6]. Countries should also strive to oversample wealthy households, where possible[7].

[4] The NCB of Slovakia, the latest country which has joined the euro area in January 2009, has recently decided also to conduct the HFCS.

[5] A two-year frequency is seen as more convenient for those countries which intend to maintain a panel component.

[6] All households in the target population should have a non-zero probability of being selected in the sample, and the selection probability of each household should be known beforehand.

[7] In a survey like the HFCS, this is deemed of special importance to ensure good quality results given that the distribution of wealth is highly skewed, meaning that a small percent of households (the wealthy) holds a disproportionate amount of assets (and virtually all the most sophisticated financial products), thus exerting a significant influence in the evolution of variables such as business wealth, asset prices, etc. Considering also the

III. IMPUTATION REQUIREMENTS FOR THE HFCS AND METHODOLOGY CHOSEN

11. Imputation consists in the process of assigning a value to a variable when it was not collected or not correctly collected.[8] Users typically have difficulties in processing incomplete datasets with most standard econometric packages. Therefore, imputing missing values is almost always a pre-requisite to be able to use the data.

12. Leaving imputation tasks to the users of survey data is an option, but probably not the most efficient one in a number of respects. Data producers typically have access to some auxiliary information which cannot be further disseminated for confidentiality reasons. Such additional information may encompass more detailed geographical and other demographic information, supplementary information provided by interviewers on inter alia the ability of respondents to answer numerical questions, factors behind survey and/or item non-response, etc. By using such auxiliary information, data producers are in a much better position to impute than general users. [9]

13. At the same time, as long as imputed variables are appropriately flagged, it is transparent to users which part of the information has been modified in which way. Thus, they may opt for either using the imputed data provided to them or working out their own imputation mechanisms. Indeed, there may be times when users with specific hypotheses might want to develop their own imputation models.

14. For the HFCS, a multiple stochastic imputation strategy has been chosen. Stochastic imputation for a given missing value can be defined as a random selection from the distribution of the variable imputed, conditional on a set of relevant observed variables.[10] Multiple imputation repeats the imputation process a number of times, as a means of expressing the latent variability of draws from the conditional distribution. In the HFCS, as in many other surveys, the multiple imputations are to be stored in a set of replicated complete data records, where the number of records corresponds to the number of multiple imputations.

15. Given the complexity of imputation tasks (and of the HFCS per se), it would be inefficient and potentially damaging for the comparability of the individual surveys if imputations were done by countries in an uncoordinated fashion. At the same time, efficient imputation models may have strong country-specific features as a consequence of cross-national heterogeneity, or rely intensely on variables that cannot be made available to the ECB for confidentiality reasons, especially where detailed information about the physical residence of respondents is concerned. For this reason, a harmonised yet decentralised approach is preferred over a centralised one: countries will strive to impute as much information as possible themselves, adhering to common methodological principles and guidelines indicated by the ECB.

16. Given the lack of familiarity of most data users with multiple imputation, it is also important that the ECB as coordinator of the HFCS encourages countries to provide researchers with sufficient guidance

non-random distribution of survey non-response, a purely random selection of units would yield a statistically very inefficient estimate of the distribution of wealth.

[8] According to the glossary of the United Nations Statistical and Economic Commission for Europe (UNECE), data imputation is the “substitution of estimated values for missing or inconsistent data items (fields). The substituted values are intended to create a data record that does not fail edits”.

[9] According to Rubin (1996) “correctly” modelling the missing data must be, in general, the data constructor’s responsibility: *“in general, ultimate users have neither the knowledge nor the tools to address missing data problems satisfactorily. […] Data-base constructors typically know more about reasons for nonresponse and have better access to confidential and detailed information not released for public use […] Ultimate users should be focused on their substantive scientific analyses and for these, missing data are generally simply a nuisance*”.

[10] The survey literature stresses that imputation models should consider not only the obvious variables that may explain the missing values but also variables that may be correlated or be used later in analysis. Otherwise the consistency of such correlations in the post-imputation dataset could be impaired and the resulting data would be less useful for analysis. The broad conditioning approach outlined in the following ensures that a large majority of the variables that might be of interest to analysts is included in the imputation models.

on how to use the multiple imputed datasets alongside the actual survey data, and monitors the results of this effort.

17. Bearing in mind the lack of experience in some countries and the associated costs of setting up imputation models, a minimum set of variables to be imputed at country level has been established. Such a minimum comprises those variables which are central to the HFCS, namely balance sheet, consumption and income variables.

18. In sum, the imputation strategy envisaged for the HFCS and underwritten so far by participating countries comprises the following five elements: (1) imputation mostly conducted on a decentralised basis; (2) imputation of all missing information taken as a longer-term goal; (3) initial imputation of at least a set of critical variables central to the HFCS (balance sheet, consumption and income variables) by all countries; (4) provision of information to users on how to use multiply imputed datasets; (5) creation of flag variables to denote whether variables were directly collected, edited or imputed.

19. Where software tools are concerned, the ECB is currently considering the possibility of developing a version of the Federal Reserve Imputation Technique Zeta routines (FRITZ; Kennickell, 1998) for the HFCS. FRITZ is a toolkit originally prepared by the Federal Reserve for the Survey of Consumer Finances, and subsequently also used for the U. S. Survey of Small Business Finances, It was recently adapted by the Bank of Spain for the Encuesta Financiera de las Familias (EFF; Bover, 2008). Although FRITZ has some technical limitations in terms of the models available, it has the advantage of the insights gained in its use in wealth surveys. The particular advantage of the EFF experience is that this survey is a component of the HFCS and the relevant computer code is available as a starting point for developing a generic HFCS approach. Some of the effort saved could be devoted to improving some of the modelling techniques available.

20. One possible initial approach might be to take the imputation framework corresponding to one of the existing wealth surveys and adapt it to include the core HFCS variables. From that point, the software could be made progressively more generic and extensive documentation could be added within the program. Countries that were interested in using this software could then make their own modifications to account for institutional differences and differences in their surveys. Because the sample sizes will differ considerably across countries and because underlying data structures may differ in important ways, a degree of model re-specification would be expected; for that reason, it would be important to identify a set of variables that are recommended to be included in any case. Sample designs will also differ across countries, and at least some controls should be introduced to account for those differences.

21. Other characteristics of the imputation process have not been agreed yet by the network of experts in charge of developing the HFCS, namely the Household Finance and Consumption Network (HFCN). They have been the object of reflection on the part of ECB experts and their consultants at the Federal Reserve Board of Governors though. These characteristics especially relate to the optimal set of conditioning variables and details concerning how the software platform will be made available to interested countries.

22. The proposals that are going to be put forward and discussed by the HFCN in the upcoming months are outlined in the next Section.

IV. CONCEPTUAL CHALLENGES OF IMPUTATION IN A MULTI-COUNTRY SETTING

A. The goals to be achieved

23. When the set of surveys corresponding to every single wave of the HFCS is completed for all participating countries, the commonly defined output variables are expected to be analysed together, with necessary econometric adjustments for differences in timing, sample size and design, and other objective technical heterogeneity. It is essential that, to the degree possible, this analysis is not substantially affected meaningfully by artefacts of measurement and data processing. In light of the experience in surveys in many countries, it is reasonable to expect that there may be nontrivial amounts of missing or

partially missing financial information in the HFCS. For that reason, as previously noted different approaches to imputation could have serious implications for the analysis of the imputed data.

24. For example, suppose missing values of a given variable were imputed using the mean of the non-missing values within cells defined by another variable. The two most direct consequences of this approach would be that analysis using a different set of cells could be biased and there would be insufficient variability in the distribution of the data. In regressions using such data, coefficients on variables other than that defining the conditioning cells would tend to be biased toward zero and the variability of the residual would be artificially reduced. If missing data are not missing completely at random within the cells, then even the overall mean would be biased. Furthermore, if different countries used different conditioning cells, the biases would be different for different countries. Thus, the utility of data imputed in this way would be highly limited.

25. For a variety of reasons, it would be difficult or impossible to eliminate all such artefacts of imputation entirely. The information available in each country may vary and the sample sizes available to support the modelling for imputation may also differ. But adherence to a common set of standards, including some common approaches to imputation modelling, can do much to minimise the level of artificial differences.

26. In the ideal, imputation would be based on draws from the estimated joint distribution of all observed information. Anything short of this imposes a structure on the data that could influence analytical results. Both to improve the efficiency of estimation and to reflect the effects of missing data on the precision of estimates, a multiple imputation technique is proposed.

B. Challenges and strategies

27. The major challenges currently faced by the ECB and the countries participating in the HFCS are as follows:

(a) The maximum possible degree of methodological commonality must be achieved, while preserving enough flexibility to accommodate country-specific phenomena, different response patterns and other idiosyncratic components;

(b) The solutions proposed must ensure maximum cost-effectiveness both on the side of the ECB as co-ordinator of the project and on the national side, especially in the case of newcomer countries where relevant set-up costs exist alongside ordinary operational costs.

28. The joint consideration of issues arising under (a) and (b) leads to a fundamental choice. Consider possible imputation strategies as a continuum in the dimension of specification parsimony, ranging from structural econometric models based on economic theory, to black-box, heavily nonparametric machine-learning exercises based on tools such as neural networks or support vector machines. For the reasons argued below, an appropriate response to the HFCS problem should probably fall more toward the latter. That is, it should be based less on a particular causally driven structure and more on an a-theoretic structure that encompasses as much as possible what might be included in broad families of possible structural models. Thus, it would be more likely to avoid a type of omitted variable bias that might occur if the possibility of a particular model had not been anticipated in a narrow specification of an imputation model.

29. While the former (structural econometric) approach might be thought by some to be preferable from the conceptual point of view, as it would embody a definite hypothesis on the data generation process for each variable, the practicality of that approach is also heavily conditioned by the pattern of missing data and by the behavioural specificities of households in each individual survey. If a structural econometric approach were chosen, the sequence of imputations would have to be dataset-specific as blanks must be filled based on the quota of dependent-variable variance explained by each equation. Conditioning variables might vary widely depending on the country, which in turn would create a broad difference in the reliability of estimates. For example, data sets where variables with few missing values (e.g. geographical location, other demographic characteristics etc.) are strongly correlated with the variables that have to be imputed (e.g. in countries where significant differences exist in wealth and income levels across regions or specific subpopulations) imputation will be easier to handle compared to

surveys where no such correlations are present. While the advantages of finding the optimal models and sequence thereof for each country are obvious—assuming universal agreement on the structure of the model—the setup, coordination and monitoring of costs and resources would be far too large in the context of the HFCS, especially given the decentralised nature of the process and the fact that some countries with little experience and large sample sizes would be involved.

30. The alternative, i.e. using a broadly conditioned a-theoretic model, has a foundation in statistics (Little and Raghunantan, 1997). This approach attempts to reproduce the correlation structure of variables in a dataset to the maximum extent possible, as opposed to attempting to extrapolate a particular data-generating process. Missing data are imputed conditional on a very large set of observed variables, i.e. all information that is expected to be somewhat correlated with at least one of the variables that have to be imputed. In this sense, every variable is represented as a linear combination of all remaining variables, making the sequence of imputations immaterial, in principle, to the final results, provided the incidence of missing values is not extremely high on single items.

31. Where this approach may become difficult to apply is when the number of variables that could be included in a given model exceeds the available degrees of freedom or where variables may be collinear or nearly collinear. In such instances, judgment—fortified by statistical criteria—is needed to specify a workable model[11].

32. Broad conditioning in a multi-country setting, however, poses some additional problems that to the best of our knowledge have not been addressed in the literature so far. Not least, because for the HFCS different variables will be available in different countries. While there is a common set of core variables in all country surveys, participating countries are likely to also include in their surveys items that are only relevant for that country (or for a small number of countries). If the variables available in only one country, say, are correlated with the relevant common data only in that country, then including the special variables is clearly preferred. If the country-specific variables are also correlated with the common/core variables in other countries where these country-specific variables are not available, then there may be differences in the bias of a sort in the imputations. Thus, differences in imputed results between countries where the country-specific variables are available or not will depend on the selectivity in the patterns of item non-response that is correlated with the unobserved variables.

33. There is no general guarantee that broad conditioning will eliminate all biases, but if observed variables are correlated with the country-specific variables, broad conditioning should at least help even in situations where the patterns of missing data are not random with respect to the set of conditioning variables. Certainly, one should always inspect the data from a variety of angles for selection processes in reporting that are driven by unmeasured variables that are correlated with the measured variables.

V. CONCLUSIONS

34. It should never be forgotten that imputation may reflect a failure either for not making a question clear enough to be answered or for not making sufficient effort to persuade respondents to provide an answer to the question. Nothing should diminish the pursuit of better questions and better techniques for gaining cooperation.

35. Given the experience of most existing surveys on wealth or similar financial concepts, it is reasonable to expect that some degree of imputation is highly likely to remain necessary for the HFCS. Although the survey is centrally coordinated and has commonly agreed principles, the great majority of the implementation will be decentralised.

[11] For example, exact multicollinearity can be handled in linear regression models by using a sweep operator, rather than a normal inverse function. The sweep operator scans the moment matrix before computing an inverse to determine whether there is exact collinearity; where there is such, one of the row-column combinations is excluded. One might specify another sort of sweep operator that had as a parameter the nearness to collinearity. There are also numerical approximations that can be used to decompose a moment matrix to extract a more robust subset of variables to use as inputs in such models.

36. This paper, after reviewing decisions already taken with respect to the imputation process, proposes a largely a-theoretic technique to be implemented in the closest local approximation possible in each country. Because of the large number of countries involved, it is possible that coordination may fail in some aspects or that particular countries may choose to pursue different approaches. Thus, it is important that a program of evaluation be established to identify differences of approach and to investigate the analytical implications of those differences.

References

Bover, O. (2008), *The Spanish survey of household finances (EFF): description and methods of the 2005 wave*, Documentos Ocasionales 0803, Banco de España.
Kennickell, A. B. (1998), *Multiple imputation in the Survey of Consumer Finances*, http://www.federalreserve.gov/pubs/oss/oss2/papers/impute98.pdf.
Little, R. and T. Raghunathan (1997), *Should imputation of missing data condition on all observed variables?*, Proceedings of the Section on Survey Research Methods, American Statistical Association: 617–622.
Little, R. and D. Rubin (1987), Statistical Analysis with Missing Data, Wiley, New York.
Rubin, D. (1987), Multiple Imputation for Nonresponse in Surveys, Wiley, Hoboken.
Rubin, D. (1996), *Multiple Imputation after 18+ Years*, Journal of the American Statistical Association, Vol. 91, No. 434, pp. 473-489.