STATISTICAL INFORMATION DATABASES DEVELOPMENT SUPPORT GUIDE

Maria Helena C. Guerra[1] *and Ana Cristina M. Costa*[2]

Abstract — *The main objective of the Development Support Guide is to establish procedures to help statistic technicians to construct the Production Database (PDB) and the User Database (UDB) for Portugal's data on the European Community Household Panel (ECHP). These databases are constructed through several steps. The Development Support Guide concerns the imputation and weighting procedures steps as well as the final production of the PDB and the UDB. The PDB and UDB development process is based on specific programs developed by the Statistical Office of the European Communities (Eurostat) with the Statistical Analysis System software (SAS) and C++ programming language applications developed by the University of Michigan. The SAS programs require adjustments every time a new wave (annual survey) is carried out. As a result, with the Development Support Guide it is intended to describe and characterize the SAS programs modifications and the structure of the data files engaged on this process.*

Index Terms — *European Community Household Panel, SAS programming, Statistical Information Databases, Statistics Technicians Guide.*

INTRODUCTION

Information on income and social indicators is getting more and more relevance to support economic and social policies within the European Union. Consequently, the Statistical Office of the European Communities (Eurostat) decided to implement a longitudinal survey called European Community Household Panel (ECHP).

A panel design allows following up and interviewing the same private households and persons over several consecutive years. The questionnaire of the ECHP gathers information on income, labour, health and other social indicators concerning living conditions of private households and persons.

This information is processed through several steps such as data checking, imputation and weighting. Afterwards, the data is stored in the Production Database (PDB). Based on the PDB a process of anonymisation is carried through and a user-friendly longitudinal User Database (UDB) is constructed.

The purpose of the Development Support Guide is to establish procedures to help statistic technicians to construct the Production and the User Databases for Portugal's data on the ECHP [1]. This document concerns all the mentioned steps to create the PDB and the UDB except data checking.

The next section describes the ECHP and the several stages that lead to the statistical databases construction. After referring the most relevant databases technical issues, the structure and content of the Development Support Guide is presented in detail. The last section points out the main conclusions.

THE EUROPEAN COMMUNITY HOUSEHOLD PANEL

The ECHP is a longitudinal annual survey conducted by Eurostat in cooperation with the National Data Collection Units, witch are National Statistics Institutes or Research Centres from each country. The annual surveys will be referred to as waves. The first wave was implemented in 1994.

The questionnaire development was carried out by Eurostat and a standard version was established. Although a few modifications to the questionnaire were made for each country, the information gathered allows the development of comparable social statistics across Member States on income, labour, poverty and social exclusion, housing, health and other social indicators.

Portugal's National Statistic Institute collects the data by means of the national questionnaire. The raw data is compiled and processed through several stages (Figure 1). In the first stage, data checking, data cleaning and cross wave consistency checks are made. By providing multiple observations on the same individuals, panel data makes possible to reveal some inconsistencies that may damage the quality of the results.

After the data checking, there is the imputation stage where several techniques are employed for a number of crucial variables to minimize the amount of missing data. The imputation is more difficult in panels than in cross-sectional surveys since plausible cross-sectional imputations of a given variable in successive waves may produce a highly implausible measure of change [2]. Then, in the weighting stage, several sample weights are constructed

[1] Maria Helena C. Guerra, Assistant Professor at Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Campus de Campolide 1070-312 Lisboa, Portugal, helena@isegi.unl.pt

[2] Ana Cristina M. Costa, Assistant Professor at Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Campus de Campolide 1070-312 Lisboa, Portugal, ccosta@isegi.unl.pt

taking into account the cross-sectional and the longitudinal weights and adjusting the sample distribution to agree with the known distribution of population totals.

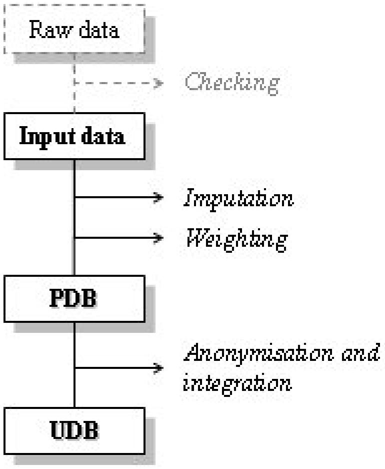


FIGURE. 1
PDB AND UDB DEVELOPMENT STAGES.

After these various stages, the data is stored in the Production Database (PDB). For each wave, this cross-sectional database is composed of four files containing *household register records*, *membership roster records*, *household questionnaire records* and *personal questionnaire records*.

Besides of the complex structure of the PDB files, the information stored is considered confidential. Therefore, a process of anonymisation takes place and a user-friendly longitudinal database is created (UDB).

DATABASES TECHNICAL ISSUES

File Management

The PDB and UDB development process is based on specific programs developed by Eurostat with the Statistical Analysis System software (SAS) and C++ programming language applications developed by the University of Michigan.

Physically, the data files and the SAS programs files are organized in six directories for each wave and three directories common to all waves. There are also SAS programs, C++ applications and other files stored in SAS software directories (for example: *c:\sas*, *c:\sas\sasuser* and *c:\sas\srclib*). The physical organization of all files was setup by Eurostat and any change to the directories structure or file names implies extra modifications to all programs involved.

The approximate number of files involved in the databases development and their functional organization is illustrated in Figures 2 and 3. It is important to point out that as the number of waves increases the file management becomes more complex.

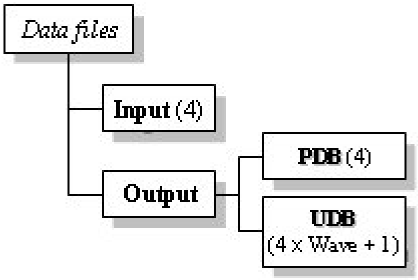


FIGURE. 2
APPROXIMATE NUMBER OF DATA FILES INVOLVED IN THE DATABASES DEVELOPMENT AND THEIR FUNCTIONAL ORGANIZATION.

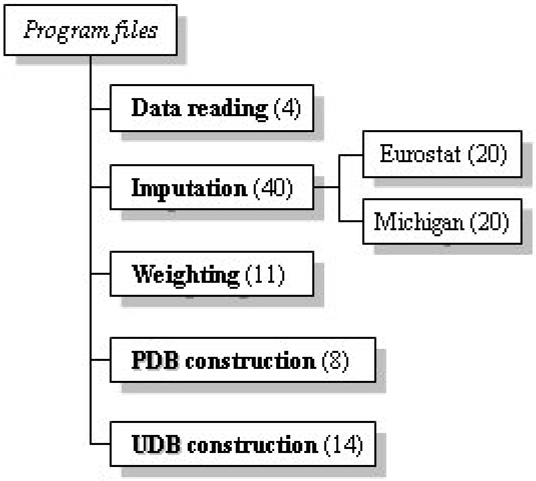


FIGURE. 3
APPROXIMATE NUMBER OF PROGRAM FILES INVOLVED IN THE DATABASES DEVELOPMENT AND THEIR FUNCTIONAL ORGANIZATION.

Program Execution

The program execution sequence follows, in general, the steps described by Figure 1. The way that SAS code was produced does not allow any deviation to this sequence and therefore the running order is extremely important. For example, some programs are run more than once over the process and, during the imputation stage, Eurostat programs are executed alternately with Michigan programs.

THE DEVELOPMENT SUPPORT GUIDE

The main objective of the Development Support Guide is to establish procedures to help statistic technicians on the process of construction of the statistical databases, PDB and UDB, for each new wave. This document describes and characterizes the SAS programs modifications and the structure of the files engaged on that process.

As a complement to Eurostat documentation, this guide constitutes a very useful support for any experienced SAS programmer to make the necessary adjustments to the programs every time a new wave is carried out.

Document Structure

The Development Support Guide is organized in two major sections: *Programs, data files and requirements* and *Program modifications*.

The first one includes:

- Hardware and software requirements
- Software configuration
- Files organization
- Data files description
- Program files description
- External information
- Programs execution sequence

On the second section, all the modifications to the SAS programs are described by program execution sequence. In this part of the document, every time a more complex program adjustment takes place, the text is illustrated with SAS code examples for the fifth wave. Finally, at the end of the document, the most important bibliographic references are listed.

The main sections of the Development Support Guide referred before will be analysed in detail.

Document Content

On the first section, the *Hardware and software requirements* subsection describes the software specifications and the minimum requirements of free disk space, RAM memory and computer processor. The Development Support Guide users must take into account these restrictions in order to construct the statistical databases successfully.

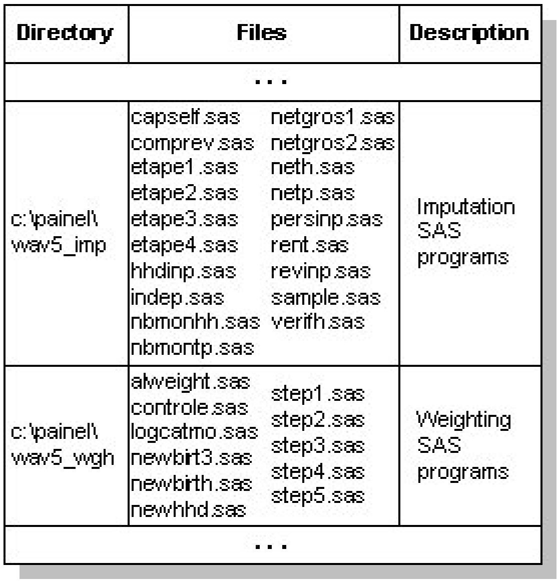
All the software configurations are clearly specified and, for example, when it concerns to the *config.sas* file the necessary changes are explicitly presented on the guide.

In the *Files organization* subsection the information concerning the path and the description of each file is illustrated through a table format and follows the fifth wave situation. The decision of making this kind presentation was due to the fact that there are a great number of files engaged on the databases construction process and because this number increases every time a new wave is implemented. Figure 4 shows a fraction of that tabular form.

The fourth subsection, *Data files description*, summarises the contents of the PDB and UDB files whereas a full description is available on the *ECHP UDB Manual – Waves 1, 2 and 3* [3].

The *Program files description* subsection is organized by program execution sequence:

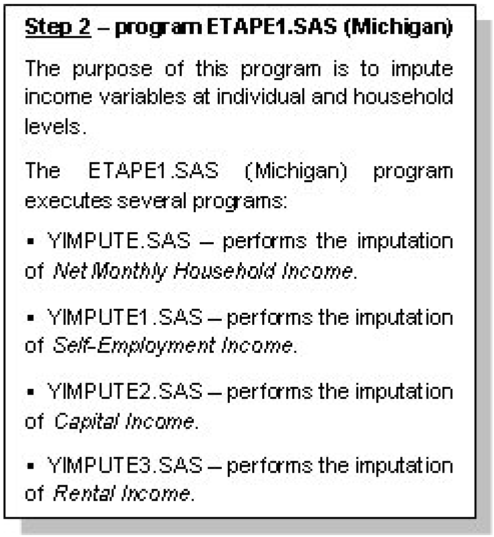
- SAS programs that convert the input data files in ASCII format to SAS Data Set format
- SAS programs of the imputation stage
- SAS programs of the weighting stage
- SAS programs for PDB construction
- SAS programs for UDB construction
- SAS programs that convert the UDB data files to ASCII format



Directory	Files		Description
. . .			
c:\painel\ wav5_imp	capself.sas comprev.sas etape1.sas etape2.sas etape3.sas etape4.sas hhdinp.sas indep.sas nbmonhh.sas nbmontp.sas	netgros1.sas netgros2.sas neth.sas netp.sas persinp.sas rent.sas revinp.sas sample.sas verifh.sas	Imputation SAS programs
c:\painel\ wav5_wgh	alweight.sas controle.sas logcatmo.sas newbirt3.sas newbirth.sas newhhd.sas	step1.sas step2.sas step3.sas step4.sas step5.sas	Weighting SAS programs
. . .			

FIGURE. 4
FILES ORGANIZATION.

The description of each file includes the program purpose and its subprograms (program executed by other program). Figure 5 illustrates the document text regarding the second program to be executed during the imputation stage (Michigan program *etape1.sas*).



<u>Step 2</u> – program ETAPE1.SAS (Michigan)

The purpose of this program is to impute income variables at individual and household levels.

The ETAPE1.SAS (Michigan) program executes several programs:

▪ YIMPUTE.SAS – performs the imputation of *Net Monthly Household Income*.

▪ YIMPUTE1.SAS – performs the imputation of *Self-Employment Income*.

▪ YIMPUTE2.SAS – performs the imputation of *Capital Income*.

▪ YIMPUTE3.SAS – performs the imputation of *Rental Income*.

FIGURE. 5
DESCRIPTION OF THE PROGRAM FILE ETAPE1.SAS.

The *External information* subsection specifies which auxiliary information must be provided by the National Data

Collection Units for each wave and also details the structure of the external information control file.

The *Programs execution sequence* subsection presents thoroughly and step by step, the running order procedure of the databases construction. At this point, and for each step, the programs names, their location and a short technical description are indicated. Since the execution sequence is quite complex, it is important to point out that all the information presented in this stage is brief and very precise.

On the second major section, *Program modifications*, the changes to the SAS programs are examined in detail.

All the programs and subprograms are analysed by execution sequence and, for each one, it is mentioned the exact position in the program code where the modification takes place and the exact description of that adaptation. It is important to refer that, at this point, it is also explained the logic process of the modifications. Along the text, SAS code examples for the fifth wave are used to illustrate more complex program adjustments.

The Development Support Guide also draws attention to some efficiency issues. To create the SAS programs for the current wave it is advisable to use as reference the last implemented wave programs to perform the necessary changes. Thus, this process becomes more efficient and the occurrence of programming errors is minimized. Moreover, it is important to write descriptive comments along the SAS code every time an adaptation occurs. This makes easier the program modifications for the next wave.

CONCLUSION

The main objective of the Development Support Guide is to establish procedures to help statistic technicians to construct the Production and the User Databases for Portugal's data on the ECHP. The feedback to this document was positive since it turned out to be very useful for the statistical databases development for the sixth and following waves.

However, the users would like the SAS code to be presented more thoroughly in order to simplify the modification process. Unfortunately, the programs complexity does not allow an exhaustive description of the SAS code. Nevertheless, this guide is intended to be used by experienced SAS programmers and a previous analysis of all programs is advisable.

REFERENCES

[1] Teekens, R., Costa, A. C. and Guerra, M. H., "Manual de Apoio à Produção das Bases de Dados (PDB e UDB) do Painel das Famílias da Comunidade Europeia", Projecto - Preparação dos Micro-Dados Portugueses para a Constituição do Painel das Famílias da Comunidade Europeia, Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, June 2000.

[2] Duncan, G., "ECHP: Quality control measures, weighting and imputation in the EC panel project". Eds. Eurostat Development Group, European Community Household Panel, Doc. PAN 6/92, August 1992.

[3] Eurostat, "ECHP UDB Manual – European Community Household Panel Longitudinal Users' Database", European Commission, Eurostat, November 1999.