Gaston Schaber
Günther Schmaus
Gert Wagner

Oct, 9, 1991
Revised Jan. 15, 1993

BUILDING UP A CROSS-NATIONAL COMPARATIVE HOUSEHOLD PANEL DATABASE

1. Introduction

Our aim is to build a comparative database from existing household panel studies.

2. Elements of the Panel Database

2.1 Variables

The database should contain **as many comparable variables as possible**. Each panel carries a set of variables which is identical from wave to wave, **the core questions**. These core questions are the candidates for variables to be standardized. By comparing the questionnaires of the different panel studies we will identify the core questions which are available in each of the studies. This comparison will produce **a list of the core variables available in all the panel studies**.

List of core variables:

- Demographic variables
- Income variables
- Labour Force variables
- Unemployment variables
- Education
- Housing
- Calendar variables

A second set will contain the variables related to the **history of the individuals** before their entering the panel. Following topics are available in most files and can be harmonized:

- Family Background
- Education History
- Employment History
- Marriage History
- Fertility History

Particular topics in one or some of the waves in a given panel study (e.g. items on assets or specific questions for elderly widows and divorced) are normally very specific to the given study and therefore are very poor candidates for harmonization.

Ideally **our result file** should contain all the variables susceptible to be standardized. However the user of this result file will have the possibility to access also **all the original variables** in the panel studies which for different reasons have not been made comparable . Our procedure allows the researchers to access simultaneously original and harmonized variables.

2.2 File structure

We will create files on the **Household-** , on the **Individual-** and on the **Income group level** (whenever the latter is possible). Each file will contain variables for one year and one panel data set. Additional identifiers will guarantee that **matches** and **aggregations** between the single files are possible.

All files will be held in a relational **DATA BASE MANAGEMENT SYSTEM**. This allows to create **standard work files** for researchers and to create **special files** for very specific analyses.

2.3 Analysis programs

We will standardize the variables, the file structures and the access system in such a form that the analysis of different panel studies in a cross-national and longitudinal context will be possible with a **minimum need for modification of programs which have been written for one country**. This will hold true at least for standard tabulations and standard analyses. More complicated analyses could probably not be standardized in such a manner but would be supported efficiently by our form of data organization.

2.4 Documentation

In order to use such a panel data bank properly, documentation must be available. It is planned to integrate all the necessary information about original and standardized variables in a **META-DATA BANK**, amely into the documentation system which CEPS has developed on PC for its own Household Panel Study. Additional documentation about the new created comparable variables will be prepared both in machine readable and in written form.

2.5 Institutional Database

The interpretation of results from cross-national research with panel studies requires **additional information** about Social Security, Tax and School Systems. So while setting up the micro database we will also develop an **INSTITUTIONAL DATABASE**. Here we can use expertise and techniques developed by the LIS PROJECT.

2.6 Countries

The first version of the comparative panel database will include only household panel studies with the following common properties:

- first wave available around 1985
- at least 5 waves completed
- similar questions and file structures
- one year panel interval
- no or only marginal acquisition costs

Four Panels fulfill these requirements: Germany, Lorraine, Luxembourg and the USA (PSID). The panel dataset from the Netherlands can not be included at present because of its high data acquisition costs. Panel studies from Ireland and Sweden have achieved only a few waves or have been interrupted. However these two panels are candidates for the PANEL DATA ARCHIVE. At a later stage the present project will integrate additional panel studies from the Netherlands, Ireland and Sweden and all also panel studies which are starting now in Europe.

3. Technical issues of comparability

This chapter explains the methods used for creating comparable variables. **Comparability here means that the variables are to be prepared according to a common plan based on common definitions**.

Following items need clarification:

- Income variables:
    - Definition of Income
    - Income accounting period
    - Standardization of income streams
    - Income Units
    - Imputation
    - Missing values

- Standardization of time variables
    - variables with time periods
    - variables with date variables

- Qualitative variables:
    - Definition of variables
    - Recodes for Qualitative variables

- Missing value codes for all variables

In a panel study there are many different types of variables to be found:

- Cross-sectional variables
- Variables for change of status between two waves
- Length of spells lasting longer than one wave
- Detailed history of transitions (dates, calendar variables) in all waves

3.1 Income

The definition of income variables will be very **similar to those used by LIS**, but will be changed for improvement in some aspects. There will be

- more different income sources (e.g social transfers)
- no apriori aggregation of income variables
- more variables available on the individual level

In most panel studies the monthly amount of income for some sources is queried, but no information is available for each single month. For other income sources only annual amounts are available. The **income accounting period** must therefore be the **year**. But we will divide the yearly income amounts by 12 in order to get **monthly incomes**. It is more practical to use monthly values because many social transfers regulations use monthly income lines in order to determine if an individual is eligible for participation in a specific program.

In all panel studies the income of a special reference year can only be derived by combining results from different waves (two years for Lorraine, Luxembourg and the PSID; three years for SOEP). This **standardization of incomes** will be done to set up the annual panel files.

Primarily we will proceed to create income variables for the household level and for the individual level. For two of the panel studies (Lorraine and Luxembourg) it will be possible to get income information at the group level. For comparability reasons it will be possible to aggregate this group income information onto the household level.

The **missing income values problem** can not be solved in the present context. The donor of the data set has to proceed to the necessary imputations for missing values before the data can be integrated. What is necessary here is to develop a set of variables for all countries **documented** in a way which indicates if imputation and what type of imputation has taken place.

3.2 Time variables

Variables which measure the length of a special status (e.g. working hours and length of unemployment spells) need to be standardized. For some variables the smallest unit is hours, for others we have only time periods of full years which we can identify. We plan to standardize these variable to **month-units** with the possibility to have floating point variables to accurately report the value (e.g. 1 day = 0.2 weeks, 1 month = 4.3 weeks). For working hours we want to adjust the amount of hours worked into **working hours per week**.

Furthermore it is necessary to convert in some cases the birth years into age values.

There are some more problems with date variables and corresponding length of a spell. The length of a spell, e. g. unemployment period can be alternatively given in
a) number of months,
b) date of begin of first unemployment spell and date of end of first unemployment spell
c) date of begin and end of the first three unemployment spells d) and by calendar variables.
What we need here is a **consistent procedure to measure the length of spells** in a year and total length of this spell for all waves.

3.3 Qualitative variables:

As with the income variables common definitions must be made here. **Recodes of values** must be made if the original qualitative variables in the panel studies are not compatible. The easiest case here is when the categories are identical and only the numerical value is to be recoded (e.g. family status). Another case: one panel study uses a very detailed code, while the remaining studies use a less detailed code for the same variable (e.g. relationship to head). Harmonization can be achieved in those cases by **aggregating the more detailed categories into the less detailed categories**.

Another possibility to achieve comparability for qualitative variables can be done by creating **new variables** (generated variables) and not by recoding existing variables. By combining several original variables into one new variable comparability can be reached for variables where it is otherwise impossible (e.g. Family type).

Experience with LIS has shown that the attempt to standardize all existing qualitative variables is not successful for all variables (e.g. brackets for size of firm). In such cases it makes more sense to use original variables.

3.4 Missing values

The missing value codes as "Not applicable", "Don't know", "Record missing", etc. must be made identical for all data sets.

4. Data Processing Approach

4.1 Panel Data Archive

Target of this chapter is to explain the **technical procedures** to create a data bank with Panel data from different countries. The panel data from each single study are not available in a compatible data format.

- Germany: SIR-Files
- Lorraine: SPSSX/SAS Files
- Luxembourg: SPSSX files/SQL-DS
- USA: OSIRIS-Files

Cross-national research with such kinds of data sets is difficult to effectuate because each of the data sets is organized in a different manner and uses a different data format:

- no common variables names
- no common format
- no common software
- not managed by identical database management systems
- not stored as SPSSX/SAS system files

The first step in achieving comparability would be to establish a data archive of available panel data **without harmonizing the existing variables**, and to store the data in a consistent manner and document the data sets. The variables of each file would be stored under their original variables names. This data archive represents **the first -intermediate-step** on the way to a fully harmonized panel data bank.

It is important that panel data sets should be integrated only once they are 'clean.' Here a clean data set means that the owner of the data set has run consistency checks on all of his variables and his identifiers. Furthermore necessary imputation and creation of imputation flags have to be done by the donor of the data set before the original panel data can be integrated.

Sufficient technical documentation about file structures, variable and value labels is also necessary. It would be very helpful if a machine readable description of each record type of the raw data files (variable names, columns, variable type: Integer, Real, Character) were available.

Because the computer system which CEPS/INSTEAD uses, has no database software as SIR or OSIRIS the data must be transmitted as **raw data files** (or as SPSSX-Files).

The form of sequential raw data files (e.g. PSID) is a possible form to store and analyze data, but it is not very efficient for handling panel data. Therefore it is necessary to transform the data which come as sequential files into a more appropriate format.

One approach to handle panel data more efficiently is to store each wave into separate files. Furthermore it is wise to store each basic unit (Households/Groups/Individuals) of each wave into separate files. A further subdivision of the individuals into children and adults can be done, because the number and type of available variables are different for both groups.

It would also practical to put special information (e.g. marriage history) - which is asked only once for a individual and could come for different individuals from different years - into separate files.

This would be the case for following topics:

- Family background
- Employment history
- Education history
- Marriage history
- Fertility history

The within-family relationships should be well defined. The original panel data files should prepare for each individual **identifiers of his/her head, father, mother and spouse**, if such relatives live in the same family as the individual. If these identifiers are not yet available, they could easily be created with the help of the variable " Relationship to Head". Those identifiers allow to link the individual records within a family.

Experience has shown that work with panel data is considerably easier when for each household and each individual who have not participated in all waves, **dummy records** have been created for those years where no information is available. These dummy records need not to be stored permanently, they could be temporarily created while matching waves (e.g. Match File command SPSS).

Characteristics of Panel Data Archive files:

- contain all original variables
- use original variable names
- use common format
- use common software
- are stored in a relational database structure
- are accessible as SPSSX system files
- offer possibility of raw data output

This data archive could be used to analyze panel data separately in a cross-national aspect. These files are not 'comparable' in a strict sense, but they are represent a real improvement for international research because they are stored on the same software, use similar file organization and the same computer hardware. Some panel studies (e.g. from Ireland and Sweden) which for different reasons are not good candidates to be standardized in the early phases of this project could be stored and accessed via such a procedure.

*** The special case of PSID:**

A special problem - which has to be solved for the present data set - arises from the structure of the PSID file. The family-file of PSID is a mixed file with information for the family and the individuals (head, wife, first, second and third and further income earner). These complex file structure should be transformed into separate records for head, wife , first ... further income earner and to be matched with the individual file of PSID.

4.2 Comparable Panel Database

Here we show the methods to use for creating comparable data for the base. As we said under point 3, comparability means that the data sets are prepared according to a common plan based on common definitions.

This has to be done through the **transformation programs which have to be written**. Ideally a strict **modularisation approach** should be used. For each panel study a set of programs must be developed. Each programs will handle one specific topic:

- program a: Demographic variables at the household level
- program b: Demographic variables at the individual level
- program c: Labour force variable at the individual level
- program d: Income variables at the individual level
  " .....
- program x: .....

One subset of these programs would handle the cross-section variables, another set would handle the transition variables between two waves. An additional set of programs would create the set of LIS-variables. Each transformation program would create harmonized variables, which are named in a consistent way in each panel and between all panels.

These transformation programs should be written in such a manner that the modifications necessary for each additional wave could be minimized. There are **software tools as parameters, macros and include commands** available to reach these goals. Also the job control language files to run the programs should be kept permanently. These organizational measure will help to update the files when necessary and to implement new waves more easily.

The result of these transformation programs would be kept in **new harmonized result files**. The format of these files would enable the researchers **to do comparable analysis runs with different panels without the need to adjust their program for each individual panel** or at least to minimize program modifications.

Characteristics of a Comparable Panel Database:

- access to harmonized Panel variables
- access to LIS variables
- possibility to access the original variables
- standardized variables names
- common format
- common software
- stored in a relational database management system
- stored as SPSSX system files
- possibility of raw data output

In the first stages of the project not all required harmonized variables will be available. Each researcher may have the need to create some harmonized variables from the not standardized data archive files and to match these variables with those from the harmonized panel database. This will be not difficult because unique identifiers will allow to match both type of files.

Another problem which has to be solved relates to the decision which harmonized variables should be stored permanently and which variables should be created by ad hoc software only when needed.

5. Work plan

5.1 First Phase: Creating a 'two waves' set

We want to start with just two complete years of panel information - taking the panels from Germany (SOEP), Luxembourg (PSELL) and the USA (PSID) and, possibly, the Lorraine one. We need for this purpose four waves from Germany, three waves from the other panel studies.

The starting point in time will be year 1985, because it is the year when the Luxembourg and the Lorraine panels started.

Necessary Waves for the first two years period:	85	86
	---	---
Lorraine:		
- Standard variables	85	86
- Income	86	87
- Taxes	-	-
PSELL:		
- Standard variables	85	86
- Income	86	87
- Taxes	-	-
PSID:		
- Standard variables	85	86
- Income	86	87
- Taxes	86	87
SOEP:		
- Standard variables	85	86
- Income	86	87
- Taxes	87	88

5.1.1 First Step: Building the Panel Data Archive

The original panel data sets will be transformed to be fitted into the Panel Data Archive format. For this transformation process we have to prepare the commands to read the original files and to check if data input has been done correctly. Then the original panel data sets will be split into the proposed data structure of the Panel Data Archive format.

5.1.2 Second Step: Building the Comparable Panel Database

At this point the variables will be made comparable. Because each panel has many topics and questions in each wave it appears to be wiser **to begin with a subset of variables**. The most frequently used variables in panel analyses are the Demographic, the Labour Force and the Income variables. In the first phase we want to include all income variables, but only the most important Demographic and Labour Force variables.

The reason for this approach is to get as quickly as possible comparable data sets which can be analyzed in order to test the validity of the harmonized panel data set. This initial work will help us to **improve the panel comparability process** in relation to **conceptual problems** and **software techniques**. With the first version of the Comparable Panel Database it will be possible to set up files for end use (e.g. flatform files, SPSSX files) for research.

5.2 Second Phase: Integration of all available waves

In the second phase all available waves from all four panel studies will be prepared and integrated into the comparative Panel Database. In this phase there will be no preparation of additional comparable variables.

Experience with panel data has shown that detailed analysis of social and economic problems need a sufficient number of cases to get significant statistical results. The requirement for greater numbers of cases can be satisfied by cumulating waves or events, if enough waves are accessible. Therefore it is more important to get more waves than to get more additional harmonized variables.

5.3 Third Phase: Integration of all necessary variables

After the creation of the minimum variables set for all available waves we plan to enlarge the set of variables. First there will be an integration of further Labour Force and Demographic variables. We plan to introduce then LIS Compatibility. Some selected variables from other topics could also be harmonized at that stage.

5.4 Fourth Phase: Analysis Phase

A comparable panel database with all available waves and a fairly complete set of variables on Income, Labour Force and Demography will be ready for analysis. Practical work with the data sets will show if our approach is a genuine tool for researchers who want to do panel analysis in a cross-national context. The experience will help us to improve our tools.

5.5 Next Phases

In the later phases it is planned to include panels which are starting now in Europe. Also topics not yet standardized will be prepared.

Additionally more complex types of variables (such as longitudinal variables and event types of variables) will be created.

And obviously the database will have to be updated with the newly incoming waves from the respective national panel studies.