# Comparative Panel File:

## Manual for CPF v.1.5

Konrad Turek

Isabel Voets

Matthijs Kalmijn

www.cpfdata.com

Comparative Panel File - Open Science Project

The Netherlands

May 2023

## Abstract

The Comparative Panel File (CPF) harmonises the world's largest and longest-running household panel surveys from seven countries: Australia (HILDA), Germany (SOEP), Great Britain (BHPS and UKHLS), South Korea (KLIPS), Russia (RLMS), Switzerland (SHP), and the United States (PSID). The project aims to support the social science community in the analysis of comparative life course data. The CPF is not a data product but an open-source code that integrates individual and household panel data from all seven surveys into a harmonised three-level data structure. In this manual, we present the design and content of the CPF, explain the logic of the project, workflow and technical details. We also describe the CPF's open-science platform. The first version of CPF was prepared by Konrad Turek, Thomas Leopold and Matthijs Kalmijn, and published in December 2020.

CPF version 1.5 was built on data versions availibe in early 2023. They were released in:

- 2020 (PSID ver. 2019)
- 2021 (HILDA ver. 2000)
- 2022 (SOEP ve. 37, RLMS ver. 2021, UKHLS ver 12, SHP ver. 22)
- and 2023 (KLIPS ver. 24)

## CPF v.1.5 team

*Core team:*

**Konrad Turek**, Tilburg University & Netherlands Interdisciplinary Demographic Institute

**Matthijs Kalmijn**, Netherlands Interdisciplinary Demographic Institute

**Thomas Leopold**, University of Cologne

*Research Assistant:*

**Isabel Voets**, Netherlands Interdisciplinary Demographic Institute

## Comparative Panel File web addresses:

- Website: cpfdata.com
- Forum: cpfdata.com/forum
- GitHub: github.com/cpfdata
- OSF: osf.io/h3yxq

## Citing this Manual

Turek Konrad, Isabel Voets, Matthijs Kalmijn (2023). *Comparative Panel File: Manual for CPF v.1.5* DOI: 10.31219/osf.io/9fhwg

## Citing the main article about the CPF

Turek, Konrad, Matthijs Kalmijn, and Thomas Leopold (2021), The Comparative Panel File: Harmonized Household Panel Surveys from Seven Countries, *European Sociological Review*, Vol. 37(3): 505–523, https://doi.org/10.1093/esr/jcab006

## Citing the CPF code and project

We kindly ask you to include the following information in publications that use the CPF code:

*This paper uses the code from the Comparative Panel File (CPF) version 1.5 available at www.cpfdata.com created by Konrad Turek, Matthijs Kalmijn, Thomas Leopold, and Isabel Voets. The project was supported by the NORFACE Joint Research Programme on the Dynamics of Inequality Across the Life-course. DOI:10.17605/OSF.IO/H3YXQ.*

# Contents

# List of symbols and abbreviations

**Symbols**

| | |
|---|---|
| Folder directory: | `D:\CPF\11_CPF_in_syntax\` |
| Syntax do-file: | `1_Folder_setup` |
| Data file: | **`CPF_v1.5.dta`** |
| Variable: | ***`education`*** |
| Syntax code: | `global your_dir "D:\CPF"  // <--inster your directory` |

**Abbreviations**

| | |
|---|---|
| CPF | Comparative Panel File |
| CNEF | Cross-National Equivalent File |
| | |
| BHPS | British Household Panel Survey |
| HILDA | Household, Income and Labor Dynamics in Australia Survey |
| KLIPS | Korean Labor and Income Panel Study |
| PSID | Panel Study of Income Dynamics |
| RLMS | Russian Longitudinal Monitoring Survey |
| SHP | Swiss Household Panel |
| SOEP | German Socio-Economic Panel |
| UKHLS | Understanding Society – The UK Household Longitudinal Study |
| | |
| AUS | Australia |
| GER | Germany |
| KOR | South Korea |
| RUS | Russia |
| SWT | Switzerland |
| UK | United Kingdom |
| US | United States |

# Information about the new version of CPF v.1.5

CPF version 1.5 was published in Aplril 2023. Compared to the previous version it includes:

1. Updated code to integrate new waves of source panels studies. Specifically, CPFv.1.5 was built on data versions availibe in early 2023. They were released in:
    - 2020 (PSID ver. 2019)
    - 2021 (HILDA ver. 2000)
    - 2022 (SOEP ve. 37, RLMS ver. 2021, UKHLS ver 12, SHP ver. 22)
    - and 2023 (KLIPS ver. 24)
2. A set of new variables covering the topics of:
    - Ethinicy
    - Migration
    - Religion
3. Many code modifications, e.g.:
    - Reorganisation of many parts of the lower-level code due to changes in the structure of the source data (e.g., changes in variables names, new variables)
    - Improvements of the code based on the team evaluation and users suggestions

CPF is an ongoing open-source project which aims to support the community of social researchers. Please note that users must take the responsibility for the final harmonization and analytical decisions, while CPF provide flexible tools and coding framework only. If you find an error, want to suggest and improvement or extension – please contact us contact@cpfdata.com or suggest the changes on GitHub https://github.com/cpfdata.

# The idea of CPF

Comparative Panel File (CPF) is as an ongoing, open science project to harmonise the world's largest and longest-running household panel surveys from seven countries (Turek, Kalmijn & Leopold 2021). The project aims to support the social science community in the analysis of comparative life course data. By harmonising individual repeated data covering long periods and several general population surveys, researchers can analyse both time trends and country differences. The CPF is not a data product, but an open-source Stata code that integrates individual and household panel data into a harmonised three-level data structure. he open-source character of the code allows for developing and extending areas of application. Currently, CPF includes seven studies:

- Australia (The Household, Income and Labor Dynamics in Australia Survey, HILDA),
- Germany (The German Socio-Economic Panel, SOEP),
- the United Kingdom (The British Household Panel Survey, BHPS, and Understanding Society – The UK Household Longitudinal Study, UKHLS),
- South Korea (The Korean Labor and Income Panel Study, KLIPS),
- Russia (The Russian Longitudinal Monitoring Survey, RLMS),
- Switzerland (The Swiss Household Panel, SHP), and
- the United States (The Panel Study of Income Dynamics, PSID).

The idea originated in 2019 in the context of the project "Critical Life Events and the Dynamics of Inequality: Risk, Vulnerability, and Cumulative Disadvantage" (CRITEVENTS). CRITEVENTS is funded by NORFACE through the transnational research programme "Dynamics of Inequality Across the Life-Course: Structures and Processes (DIAL)."[1] CPF is managed and developed by Konrad Turek (Tilburg University & Netherlands Interdisciplinary Demographic Institute, NIDI-KNAW), Matthijs Kalmijn (NIDI-KNAW) and Thomas Leopold at the University of Cologne. The CPF code and entire open-science platform were designed and prepared by Konrad Turek and will be continuously developed and improved by the CPF team and the community of users. Updates and developmenets for the current version 1.5 were prepared mainly by Isabel Voets (NIDI-KNAW) under supervision of Konrad Turek and Matthijs Kalmijn. We would

[1] This article forms part of the CRITEVENTS project. The CRITEVENTS project is financially supported by the NORFACE Joint Research Programme on the Dynamics of Inequality Across the Life-course, which is co-funded by the European Commission through Horizon 2020 under grant agreement No 724363.

also like to thank Daniel van Wijk (NIDI-KNAW) for precious comments and suggestions for improving and extending the code.

CPF originated as an attempt to move the data harmonisation process to open science and crowdsource cooperation and provide novel functionalities (Turek 2023). We recognised that the conventional, institutionalized approach to harmonisation can be limited in some aspects, and the open-source model can supplement it in certain areas. The idea of harmonising household panel studies was not new. The most well-known, long-running and successful harmonisation project is the Cross-National Equivalent File (CNEF) (Burkhauser et al. 2001; Frick et al. 2007). CNEF has been developed since 1990 under the lead of researchers from Cornell University. Over the years, the project was managed primarily by Dean R. Lillard and administered by Cornell University and Ohio State University.[2] Initially, in 1991, the dataset harmonised only a limited set of variables for two countries, the US and Germany.[3] Over the years, the project expanded by adding countries, such as the UK and Canada[4] in 1999, Australia and Switzerland in 2007, and Russia and Japan in later years. The set of topics and variables has been gradually extended, but the main focus remains on income and earnings. CNEF has been used primarily in income-related research in economics (Allanson 2011; Büchel & Frick 2004; Chen 2009), sociology (DiPrete & McManus 1996; Ehlert 2013; McCall & Percheski 2010; Musick, Bea & Gonalons-Pons 2020), or demography (Cooke et al. 2009), and less often in research on other topics, such as life satisfaction (Cho & Lee 2013), and self-employment (McManus 2003).

The initial motivation to harmonise data from national panel surveys was the fact that the CNEF release did not include measures for job loss and unemployment. Instead of harmonising only these variables, we decided to extend the approach pioneered by CNEF to a larger set of key variables of social science research and make the result available to the broader scientific community. Specifically, the novelty of CPF is the open-source model of harmonization that allows to makes several important steps forward (Turek 2023). First and foremost, CPF has a broad focus, including information about education, family

[2] The CNEF involved a cooperation with national source data administrators, an international group of researchers from US, Germany, UK, Switzerland, Australia, Korea, Russia and Canada. CNEF was funded by several institutions, including the US National Institute on Aging, the German Institute for Economic Research, and Cornell University.

[3] The CNEF was built on the model implemented in the Luxembourg Income Study (LIS), which harmonizes micro-level household surveys data from over 25 countries Burkhauser, R.V., Butrica, B.A., Daly, M.C., and Lillard, D.R. (2001). The Cross-National Equivalent File: A product of cross-national research. in *Social Insurance in a Dynamic Society*, edited by Becker, I., Ott, N., and Rolf, G. Frankfurt: Campus Fachbuch, Frick, J.R., Jenkins, S.P., Lillard, D.R., Lipps, O., and Wooden, M. (2007). The Cross-National Equivalent File (CNEF) and Its Member Country Household Panel Studies. *Schmollers Jahrbuch : Zeitschrift für Wirtschafts- und Sozialwissenschaften* 127: 627-54.. LIS was limited by the cross-sectional character of the data, and difficulties in accessing the data due to confidentiality issues. CNEF aimed at harmonizing more accessible panel data and including a broader range of research topics than LIS.

[4] The Canadian Survey of Labour and Income Dynamics (SLID) is discontinued.

and marital relationship, labour market status, subjective wellbeing and work satisfaction, social origin, and socio-economic status, in addition to the classic economic variables also present in CNEF. For several of these variables, CPF also offers more detail; for example, it allows distinguishing between unemployed, retired, self-employed or entrepreneurs.

Second, CPF is open and flexible, thereby facilitating a genuine bottom-up approach. CPF fully supports modifications in harmonised variables or adding new variables from the source database, depending on researchers' needs. Our code is available in full and for all selected countries. It also facilitates work with single surveys, as it instructs how to go from a large set of raw files to an integrated and ready for analysis panel data set (for some surveys, e.g., PSID, this is a complex process). In contrast, CNEF is a data product that offers a set of separate data files, but only parts of the code are available.

Third, CPF does not depend on direct government funding, which greatly facilitates the speed and direction of its further development. New waves can be added as soon as these are released by the national data centres. CNEF files are released differently by country, and most do not cover recent waves.

Fourth, procedures to obtain and use the data are streamlined and greatly simplified. The only administrative step needed is obtaining permission to use the national data from each of the seven national data centres. Once the permission is obtained, the openly available code can be run, and the CPF is readily available on the user's computer. CNEF requires an additional application for accessing some surveys, separate CNEF files are provided partly online and partly on a CD sent by mail, and they still have to be integrated.

In sum, we build on the approach pioneered by CNEF and other cross-national data harmonisation projects (Dubrow & Tomescu-Dubrow 2016), but overcome the main limitations for users who require a broader set of variables and more flexibility and control over the data management process.

The CPF provides free and full access to a code that generates a comparative dataset based on these household panel surveys. The code and complete documentation are available at www.cpfdata.com. After securing access to the national panel surveys, users can run our code which combines datasets and waves within a country, constructs harmonised variables and merges these into one data set for all countries and all waves. Users can either follow the default workflow and run the code unchanged or modify and improve it for their use (e.g., select countries, add new waves, add new or modify existing variables). The file is organised in a long format which contains one record for each person in each wave. The merged file

of CPF version 1.5 contains around 3 million observations from almost 400 thousand individuals observed for an average of 7.7 waves (up to 41 waves).

The CPF is organised as an open science platform that integrates several tools that support open collaboration, management, documenting and sharing all materials. The main elements of the platform are the website (central platform) with forum (general communication), GitHub code repository (code development), and Open Science Framework (general management of scientific research). User's improvements and suggestions will be recorded, incorporated and shared using open online tools to allow continuous development and regular updates to the official versions of the code.

# CPF data sources

Data for the CPF come from household panel studies – general population repeated surveys with household as the primary sampling unit. They regularly (mostly yearly) interview all or selected adult members of sampled households over a period of years and collect information about the entire household and its members (Rose 1995). Version 1.5 of the CPF combines seven most established and longest-running household panel studies globally. All studies are representative of the population of households. As ongoing panel studies, they continuously renew their samples by including new household members (e.g. grown-up children, newly married partners), following new independent household established by respondents (e.g. children leaving parents' homes), by refreshments (e.g. including a new set of households), or by extensions (e.g. including a new type of households, such as new migrant families). Many panels included systematic oversamples of subgroups; these are included in the CPF but identifiable with country-specific variables.

## *United Kingdom: BHPS and UKHLS*

The UK's CPF sample consists of two studies: The British Household Panel Survey (BHPS) covering years 1991-2008 (18 waves) and Understanding Society: The UK Household Longitudinal Study (UKHLS) from 2009 onwards (Buck & McFall 2012; Platt et al. 2020). BHPS began in 1991 as a multi-purpose panel survey for social and economic research. It was run on a yearly basis with the same individuals re-interviewed each successive year. The first wave of BHPS consisted of ca. 5,500 households and 10,300 individuals

drawn from 250 areas of Great Britain. In following years additional samples were added, including 1,500 households in Scotland and 1,500 households in Wales (1999), and 2,000 households in Northern Ireland (2001).

Since 2009, BHPS has been integrated into UKHLS. With a target sample size of 40,000 households in wave 1, UKHLS became the largest nationally representative household panel study worldwide. Young people aged 10-15 complete a youth questionnaire. Respondents aged 16 and over complete the adult survey and continue to be interviewed when they leave their original households. Data continue to be collected every, yet the fieldwork is extended to around 2,5 year for each wave (e.g. the 1st wave of UKHLS covers years 2009-2011, the 2nd covers years 2010-2012). Both BHPS and UKHLS have been developed and carried out by the Institute for Social and Economic Research at the University of Essex.

For more information, see:

- BHPS: www.iser.essex.ac.uk/bhps,
- UKLHS: www.understandingsociety.ac.uk
- Data are available via the UK Data Service after granting access: www.ukdataservice.ac.uk.

For citing, please follow instructions of UKHLS, e.g.:

> University of Essex, Institute for Social and Economic Research. (2022). Understanding Society: Waves 1-12, 2009-2021 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 17th Edition. UK Data Service. SN: 6614, DOI: 10.5255/UKDA-SN-6614-18

## *Germany: SOEP*

The German data come from the German Socio-Economic Panel (SOEP). SOEP began in 1984 as a representative longitudinal study of private households in Germany for social, behavioural, and economic research. It is designed to measure disparities in resources across individuals over the life course by re-interviewing the same household members aged 17 and older annually. Initially, SOEP included only Western Germany and since 1990 after reunification it also covers eastern parts of the country, being the only database worldwide covering such a political unification (Giesselmann et al. 2019; Goebel et al. 2019; Siegers, Belcheva & Silbermann 2020).

The first wave of SOEP consisted of two samples and ca. 6,000 households from the western states of Germany: (1) with a German household head and (2) with a migrant Greek, Italian, Spanish, Turkish, or Yugoslavian household head. In 1990 the panel data was expanded to include a representative sample

from East Germany. The data was further advanced by adding immigrant (1994/95, 2013/2015 and 2020) and refugee (2016 and 2017) samples. Now ca. 15,000 households and 30,000 individuals participate in the SOEP.

SOEP has been developed by Research Center 'Sonderforschungsbereich' at the Universities of Mannheim and Frankfurt/Main together with the German Institute for Economic Research (DIW Berlin). Since 1990, SOEP has been fully delegated to DIW under the umbrella of Leibniz Association with financing from the state governments and Federal Ministry of Education and Research (BMBF).

For more information, see:

- Official website: www.diw.de/en/soep
- SOEPcompanion: companion.soep.de
- Additional resources (including variables-search system): paneldata.org
- Data are available via the Research Data Center SOEP after granting access: www.diw.de/en/diw_02.c.242211.en/criteria_fdz_soep.html

For citing, please follow instructions of SOEP, e.g.:

```
Socio-Economic Panel (SOEP), data for years 1984-2020, SOEP-Core v37, EU
Edition, 2022, doi:10.5684/soep.core.v37eu
```

## *United States: PSID*

The US data come from the Panel Study of Income Dynamics (PSID). It covers a period from 1968 and remains the oldest national panel survey worldwide. The study was initially created for evaluating poverty and economic wellbeing dynamics in the US. Currently, PSID aims to study the dynamics of income and poverty by interviewing only one person per family regularly.

From 1968 to 1997, the data were conducted every year. Since 1998, interviews are biennial. In 1968, ca. 5,000 families and 18,000 individuals participated in the survey. Over the decades, the PSID sample has grown though its genealogic design that allows gathering data from up to seven generations of the same family. It has collected survey information on more than 80,000 individuals in total (Johnson et al. 2018; McGonagle et al. 2012). Respondents had entered the study in three ways. Demographic inflows (birth, adoptions and marriages) brought up new members to the families. Formation of new independent households as a result of children splitting off their parents' homes provided new unit measures for PSID.

Additionally, post-1968 immigrant families extended the original sample in 1997/1999. Now ca. 10000 families participate in the PSID.

To enrich data, PSID collects supplement studies. Children development (CDS) for 18 years old and younger, transition into adulthood (TAS) of those over 18, disability and use of time (DUST) for 60 and older are monitored. PSID is managed by faculty at the University of Michigan. The project's major funders are the National Science Foundation, National Institute on Aging and National Institute of Child Health and Human Development.

For more information, see:

- Official website: simba.isr.umich.edu/
- Data are available via the official website for registered users: simba.isr.umich.edu/Zips/ZipMain.aspx

For citing, please follow instructions of PSID, e.g.:

> Panel Study of Income Dynamics, public use dataset [restricted use data, if appropriate]. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (year data were downloaded)

## *Australia: HILDA*

The Australian data come from the Household, Income and Labor Dynamics (HILDA). It began in 2001 as a nationally representative longitudinal survey of Australian households. HILDA is developed to study family and labour market dynamics, economic and subjective wellbeing over the life-course (Watson & Wooden 2020). All working-age members of the household (over 15 years old) are re-interviewed annually. In 2001, ca. 7,682 households and 13,969 individuals participated in the survey. In 2011, 2,153 households (5,477 individuals) were additionally selected and expanded the original sample size. Since then, HILDA has followed over 17,000 Australians each year.

On a less frequent basis, further information on wealth, health care utilisation, eating habits, cognitive functioning and retirement is collected. HILDA does not conduct interviews with foreign residents in Australia, people living in outlying areas, members of non-Australian defence forces and foreign diplomatic officers. The HILDA is directed by the Melbourne Institute of Applied Economic and Social

Research at the University of Melbourne and funded by the Australian Government Department of Social Services (DSS).

For more information, see:

- Official website: melbourneinstitute.unimelb.edu.au/hilda
- Data are available via the National Centre for Longitudinal Data Dataverse (Australian Government Department of Social Services): https://dataverse.ada.edu.au/dataverse/ncld

For citing, please follow instructions of HILDA, e.g. all publications must carry the following:

> "This paper uses unit record data from Household, Income and Labour Dynamics in Australia Survey [HILDA] conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this paper, however, are those of the author[s] and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. DOI: ####"

## *South Korea: KLIPS*

The South Korean data come from the Korean Labor and Income Panel Study (KLIPS). It began in 1998 as a national longitudinal survey of households and individuals living in urban areas in South Korea. Interviews with all household members aged 15 and older are conducted annually. KLIPS monitors individuals' economic and labour activities, income and expenditures, education and job training. An original sample was developed using stratified clustering method to select districts. Out of 7 metropolitan cities and urban areas in 8 provinces, households were derived based on equal probability technique. KLIPS set out to re-interview ca. 5,000 households and 13,000 individuals every year.

Over the waves, the sample expands by adding individuals who form family ties with original panel respondents. Such sample growth enables to track demographic dynamics (e.g. marriages, births, divorces) of initial sample members. In 2009, an additional consolidated sample of 1,415 households was added to overcome the household attrition and representability issues. Since 2009, KLIPS has contained two panels within one dataset. Following families over decades allows the KLIPS data contributing to vital improvements in Korean employment policies. The project is carried out by the Korea Labor Institute and Center for Labor Statistics Research. The main funder of the study is the Ministry of Employment and Labor.

For more information, see:

- Official website: www.kli.re.kr/klips_eng
- Data are available via the official website for registered users: www.kli.re.kr/klips_eng

## *Russia: RLMS*

The Russian data come from the Russian Longitudinal Monitoring Survey (RLMS) and cover a period from 1994. It is the longest-running panel survey of households in Eastern Europe and Asia (Gerry & Papadopoulos 2015; Kozyreva & Sabirianova Peter 2015). RLMS is a series of household-based nationally representative surveys which re-interview the same individuals almost every year. It aims to measure the effects of Russian reforms in economic and social sectors on individuals' welfare and health.

RLMS was initiated in 1992, and the first research phase (1992-1994) aimed to develop a sample of households that would meet statistical standards (it is not included in the CPF). The main phase of the research (RLMS–Phase II) begun in 1994 with a multi-stage probability sample covering eight regions of the Russian Federation (with 3975 households and 11290 individuals surveyed). Since then, the data has been collected annually (with two observation periods missing due to funding issues in 1997 and 1999).

The RLMS is conducted by the National Research University Higher School of Economics (HSE) in Moscow, the "Demoscope" team in Russia, and the Carolina Population Center at the University of North Carolina at Chapel Hill. Additionally, the project is co-financed by the US National Institutes of Health via a subcontract from Cornell University.

For more information, see:

- www.hse.ru/en/rlms
- www.cpc.unc.edu/projects/rlms-hse
- The data are available via the Higher School of Economics without application: www.hse.ru/en/rlms. Additionally, the data may be downloaded at the Carolina Population Center without application: www.cpc.unc.edu/projects/rlms-hse/data.

For citing, please follow instructions of RLMS, e.g.:

> "Russia Longitudinal Monitoring survey, RLMS-HSE", version 2023, conducted by National Research University "Higher School of Economics" and ZAO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology of the Federal Center of Theoretical and Applied Sociology of the Russian Academy of Sciences. (RLMS-HSE web sites: http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms)

## *Switzerland: SHP*

Swiss data come from the Swiss Household Panel (SHP) and cover a period from 1999. SHP was run as a longitudinal survey of private households based on a random representative sampling. The project aims to report the dynamics of living conditions change, income, quality of life and population representation in Switzerland. All household members aged 14 years and over are asked to complete an individual questionnaire every year by telephone. In 1999, ca. 5,074 households and 12,931 individuals participated in the survey.

Over a period of time, a high percentage of non-responses have accumulated in SHP. Thus, in 2004, new 2,538 households were added to the original sample to overcome the issue. In 2013, further 4,093 households were refreshed the sample size, although different measures were implemented to return those who refused to participate.

The SHP has been developed and funded by the Swiss National Science Foundation. The project was carried out by the Swiss Federal Statistical Office and the University of Neuchâtel. Since 2008, SHP has been integrated into the Swiss Centre of Expertise in the Social Sciences (FORS) and hosted by the University of Lausanne.

For more information, see:

- Official website: https://forscenter.ch/projects/swiss-household-panel/
- Data are available via FORSbase for registered users: https://forscenter.ch/projects/swiss-household-panel/data/

For citing, please follow instructions of SHP, e.g. all publications must carry the following:

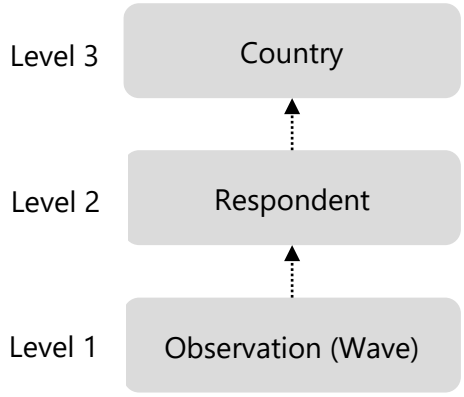
> "This study has been realised using data collected by the Swiss Household Panel (SHP), which is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is supported by the Swiss National Science Foundation".

# Basic information about CPF

## *Data structure and time frame*

CPF is a comparative panel dataset with a hierarchical structure of data. The structure has three levels (Figure 1): repeated individual observations from multiple waves (level-1) are clustered within individuals (level-2), and individuals are clustered within countries (level-3).

**Figure 1.** *CPF data structure*




The CPF version 1.5 covers up to 41 waves (between 1968 and 2021), combines seven countries, and includes around 3 million observations from almost 400 thousand respondents (Table 1). The oldest survey is PSID starting in 1968 and collecting 41 waves. The second oldest is SOEP which started in 1984 and so far, collected 37 waves. The youngest panel study in CPF is HILDA with 20 waves since 2001.

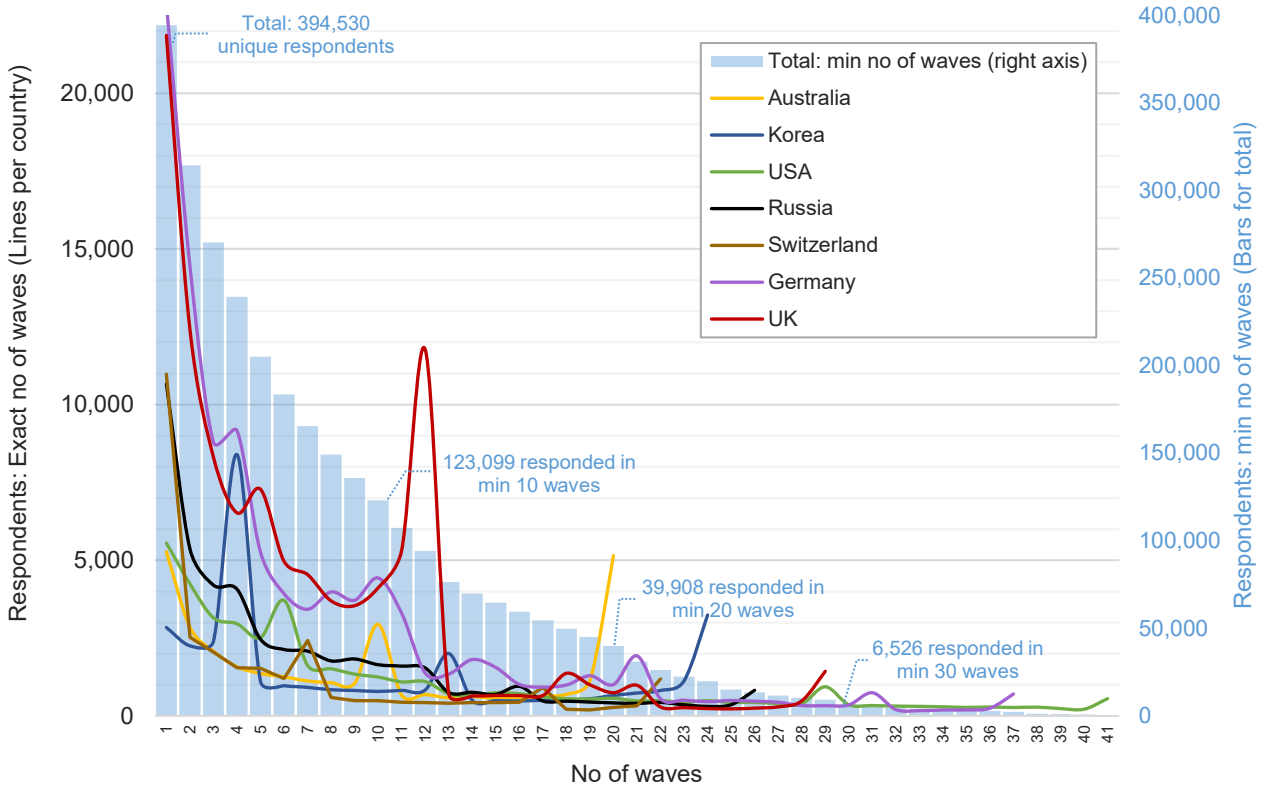
**Table 1.** *Number of waves, observations and respondents (CPF v1.5)*

| | | First | No of | Observations | | Unique respondents | |
|---|---|---|---|---|---|---|---|
| **Country** | **Survey** | **wave** | **waves** | n | % | n | % |
| [1] Australia | HILDA | 2001 | 20 | 290,616 | 9.6 | 31,897 | 8.1 |
| [2] Korea | KLIPS | 1998 | 24 | 333,699 | 11.0 | 34,609 | 8.8 |
| [3] USA | PSID | 1968 | 41 | 472,054 | 15.5 | 43,348 | 11.0 |
| [4] Russia | RLMS | 1994 | 26 | 316,497 | 10.4 | 47,013 | 11.9 |
| [5] Switzerland | SHP | 1999 | 22 | 170,268 | 5.6 | 29,491 | 7.5 |
| [6] Germany | SOEP | 1984 | 37 | 735,913 | 24.2 | 102,947 | 26.1 |
| [7] UK | BHPS/UKHLS* | 1991 | 30 | 720,001 | 23.7 | 105,225 | 26.7 |
| Total | | | | 3,039,048 | 100 | 394,530 | 100 |

** BHPS: 1991-2008, 18 waves, UKHLS: from 2009, 12 waves*

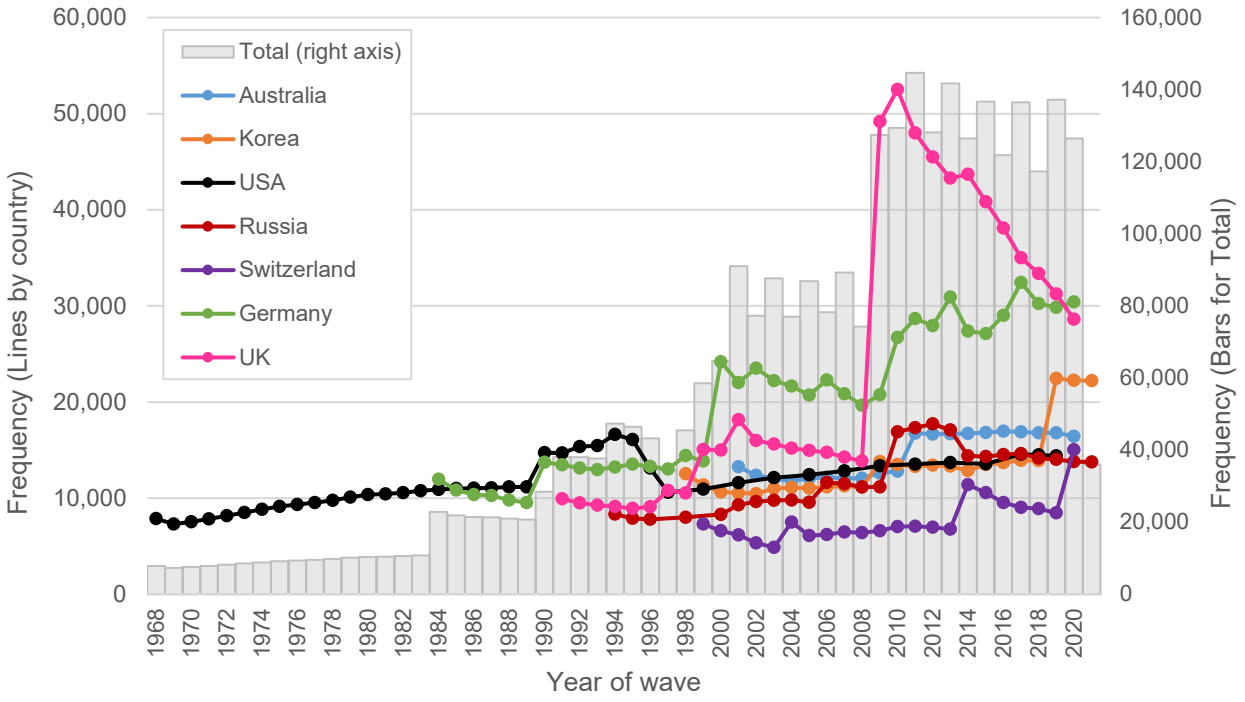
An average respondent participated in 7.7 waves (between 5.8 in Switzerland and 10.9 in the US). As shown in Figure 2, out of all 394,530 respondents, 123,009 participated in a minimum of 10 waves and 39,908 in a minimum of 20 waves. Country-specific lines indicate the exact number of waves for which the dataset provides information on sample members. For example, there are almost 22,000 respondents in the UK who participated in only one wave and around 3,500 who participated in exactly nine waves. In Australia, 16.2% of the sample (ca. 5000 respondents) participated in all 20 waves of HILDA.

**Figure 2.** *Number of waves in which individuals participated: exact number by survey (left axis) and minimum number for the total sample (right axis) (CPF v1.5).*




The oldest survey in CPF is PSID covering a period from 1968 and collecting 41 waves until now (Figure 3). The youngest panel study in CPF is HILDA with 20 waves since 2001. Since the wave of 2000, the number of participants in SOEP has grown significantly. CPF includes four countries since 1994, five countries since 1999, and all seven countries since 2001. A particularly large increase observed for the UK sample in 2009 is related to the transition from BHPS to UKHLS. For most of the surveys, data have been collected yearly (after 1997, PSID has switched to 2-year intervals in data collection).

**Figure 3.** *Timeline of the data and number of observations by wave (CPF v1.5)*




## *Variables*

The goal of CPF is to harmonise variables across surveys. We based our approach on the CNEF but aimed at extending the range of variables included. For example, instead of a simple indicator for being employed or not employed, CPF provides a full range of labour market statuses, including being unemployed, retired, in education, or inactive. Instead of years of schooling, CPF focuses on education level according to the ISCED classification, a measure that has been designed for cross-national comparison. We provide more detailed information on marital status. CPF provides a set of additional variables, such as training participation, satisfaction with different domains, social origin, labour market experience, self-employment and entrepreneurship, work-education skill fit, and perception of job security. An overview of all variables is presented in Table 2.

For harmonisation, we explored the items available in the source data for their comparative potential. Some questions had a very similar form across all surveys (e.g., self-rated health), but many differed in the wording or number of answer categories. In the latter case, we assessed the comparative value and

compared distributions of responses. It is important to keep in mind that descriptive statistics of harmonised variables may differ across countries if the original variables had different numbers of answering categories in different countries (Revilla, Saris & Krosnick 2014). Similarly, differences in the wording of questions may produce (probably small) differences in the frequency distributions. Correlations with other variables are not necessarily affected by such differences (Kaminska & Lynn 2017; Slomczynski & Tomescu-Dubrow 2019; Wolf et al. 2017). We advise users to read the *Codebook* closely to be aware of such differences.

Full harmonisation was not always possible so that some variables are available for a subset of countries. Many of the CPF variables are composed of multiple source variables. For example, retirement is based on information about working status, self-reported retirement status, receiving retirement pension, and age. In many cases, the CPF's code includes data cleaning, such as updating contradictory entries with the most reliable information, filling missing values based on information from other waves or other variables (e.g., for education, age, year of birth, marital status). Users can modify existing or add further variables from the source data by developing the open CPF code (the procedure is described below in *Workflow D*). Detailed information on all variables is provided in the CPF Codebook (www.cpfdata.com).

**Table 2.** *Variables available in the CPF version 1.5*

| Group of variables | Description | Main variables |
|---|---|---|
| Technical | Respondent identifiers, information about wave and interview and other technical information | - Country<br>- Personal and household identification numbers<br>- Wave's number and year<br>- Interview status<br>- Year and month of interview<br>- Sample identifiers |
| Demographic | Basic demographic characteristics. | - Gender<br>- Age<br>- Year of birth |
| Education | Education level is harmonised using the ISCED classification in four different versions with three, four, and five levels. For example, three levels are: [0-2] Low, [3-4] Medium, [5-8] High. Variables also include years of education, participation in training, self-assessment of qualifications. | - Education: 3/4/5 levels<br>- Participation in training in the past 12 months<br>- Work-education skill fit<br>- Qualifications for job |
| Marital and relationship status | CPF distinguishes between formal marital status and partnership living-status, which also accounts for living with the partner. Additionally, it includes less precise primary partnership status equivalent to the one used in CNEF. Also, it provides indicators for specific statuses (e.g. divorced) and being never married. | - Formal marital status<br>- Partnership living-status<br>- Primary partnership status<br>- Living with the partner<br>- Never married<br>- Widowed<br>- Divorced<br>- Separated |
| Number of children and | There are several children-related variables to account for differences in questionnaires in: | - Number of children in household (aged 0-15, 0-17)<br>- Number of children ever had |

| Group of variables | Description | Main variables |
|---|---|---|
| household members | - the definition of children, e.g. own-born, adopted, of other family members, any children<br>- the situation of children, e.g., living currently in the household, living elsewhere, children ever had<br>- age of children, e.g., any age, below 18, and below 15 years old | - Has own children (yes/no)<br>- Number of people in household |
| Labour market situation and employment | An important goal of the CPF is to provide a comprehensive view of individuals' labour market situation. These include the following areas: | |
| | Labour market situation: employed, unemployed, retired or disabled, in education, not active, employed but on leave. CPF also identifies maternity leave. | - Labour market situation (5/6 categories)<br>- Currently working (self-reported)<br>- Working in the previous year (based on reported working hours)<br>- Being on maternity leave<br>- Never worked |
| | Level of employment: full- or part-time, number of working hours (several versions, including actual and contracted hours) | - Full- or part-time work (based on working hour / self-reported)<br>- Number of working hours (per year, month, week, day)<br>- Work hours per week: contracted |
| | Occupation - classified according to the International Standard Classification of Occupations (ISCO). KLIPS and PSID use different classifications than ISCO. In these cases, crosswalk algorithms were developed. ISCO level 1 and 2 are harmonised for all countries, but if available, CPF provides a more detailed classification in versions ISCO-88 or ISCO-08 at 3- or 4-digit levels. | - Occupation: ISCO level 1: 1 digit, 10 categories<br>- Occupation: ISCO level 2: 2 digits, 50+ categories<br>- Additionally, ISCO-08/ ISCO-88 with 3 or 4 digits<br>- Supervisory position |
| | Characteristics of the employee's organisation. | - Industry: 3 major, 10 sub-major and 17 minor groups<br>- Sector (public)<br>- Size of organisation |
| | More precise and specific identification of actively unemployed, self-employed, entrepreneurs (with employees), and retirees. These indicators are built on information from several variables. For example, individuals are classified as retired when they are not working and meet any of the following criteria:<br>- Self-categorisation as retired & age 50+<br>- Receives old-age pension & age 50+<br>- Age 65+ | - Unemployed: actively looking for work<br>- Self-employed<br>- Entrepreneur (including or not including farmers)<br>- Retired fully<br>- Receiving old-age pension |
| | Labour market experience measured as years of employment/work | - Total Labour market experience (total/ full time / part time)<br>- Tenure with current employer |
| | Perception of job security - Whether the respondent is worried about job security (in two versions) | - Secure /Insecure<br>- Secure /Insecure / Hard to say |
| Incomes | Incomes of individuals and households. Depending on the original data, information on individual income is included in several variables based on:<br>- source of income (total income from jobs and benefits, from all jobs, from the main job)<br>- type of income (gross, net)<br>- reference period for income (year, month, per hour)<br><br>This approach results in multiple variables but provides clear definitions. For analytical purposes, users can combine particular variables using the nominal values or relative values (e.g., percentiles). CPF provides values as they are included in the source data, without any additional cleaning, imputation, conversion or inflation-adjustments. Values are in local currency.<br>Depending on the type of monthly household income in the original data, information is provided in two versions: before taxes and deduction (gross, pre), after taxes and transfers (net, post). Some datasets provide a negative household income indicating a loss or debit (e.g. PSID since 1994). Values are in local currencies. | - Individual Income (All types)<br>• year, net<br>• month, net<br>- Individual Labor Earnings (All jobs)<br>• year, gross<br>• year, net<br>• month, net<br>• month, gross<br>- Salary from the main job<br>• year, net<br>• year, gross<br>• month, gross<br>• month, net<br>• per hour, gross<br>- Household income (month)<br>• gross<br>• net |

| Group of variables | Description | Main variables |
|---|---|---|
| Health and wellbeing | Self-rated health status is based on the standard 5-point scale.<br>There are three versions of disability-related questions.<br>Variable for chronic diseases is in a working version: it is not fully harmonised and should be modified by the users according to specific conceptual framework (e.g. defining chronic conditions).<br>CPF provides several dimensions of subjective wellbeing, which can be harmonised for at least several countries. We include two versions of each variable due to differences in original answer scales: with a 5-point scale (1-5 range) and 11-point (0-10 range). If required, the original values were rescaled. | - Self-rated health<br>- Receiving disability pension<br>- Disability: any type (physical, mental or nervous condition)<br>- Disability: min. category 2 or >30%<br>- Chronic diseases (yes / no)<br>- Satisfaction with<br>• Life<br>• Work<br>• Financial situation of household<br>• Individual income<br>• Family<br>• Health |
| Parental background | Parents' education level is coded in 3- and 4-categorical variables similarly to respondent's education level. | - Mother's / Father's education: 3 /4 levels |
| Socio-economic position | Socio-economic position scales are based on respondents' work status and occupation's ISCO code. | - International Socio-Economic Index of occupational status (ISEI)<br>- Treiman's international prestige scale (SIOPS)<br>- German Magnitude Prestige Scale (MPS) |
| Ethnicity | Self-reported ethnicity based on broad categories. For the US, a separate variable is available for hispanicity | - Ethnicity<br>- Hispanicity (US only) |
| Migration | Country of birth of respondents and (if available) their parents. For foreign-born individuals, country of birth is categorised in global regions. Derived from the country of birth variables is a set of variables relating to the respondents migration status. | - Country of birth<br>- Migration status<br>- Migrant generation (derived) |
| Religion | A binary variable is available indicating whether the respondent self-identifies as belonging to any religious group. Further information is available on attendance of religious services which is placed on a 4-point scale. If required, the original values were rescaled. Additionally, a separate variable is available for Korea only reflecting religious participation, which may be relevant as a proxy for attendance (which is not available for Korea) | - Religiosity<br>- Attendance religious services<br>- Religious participation |

## *Samples*

In the default settings, CPF includes observations from individuals aged 18 and older and meet the following criteria:

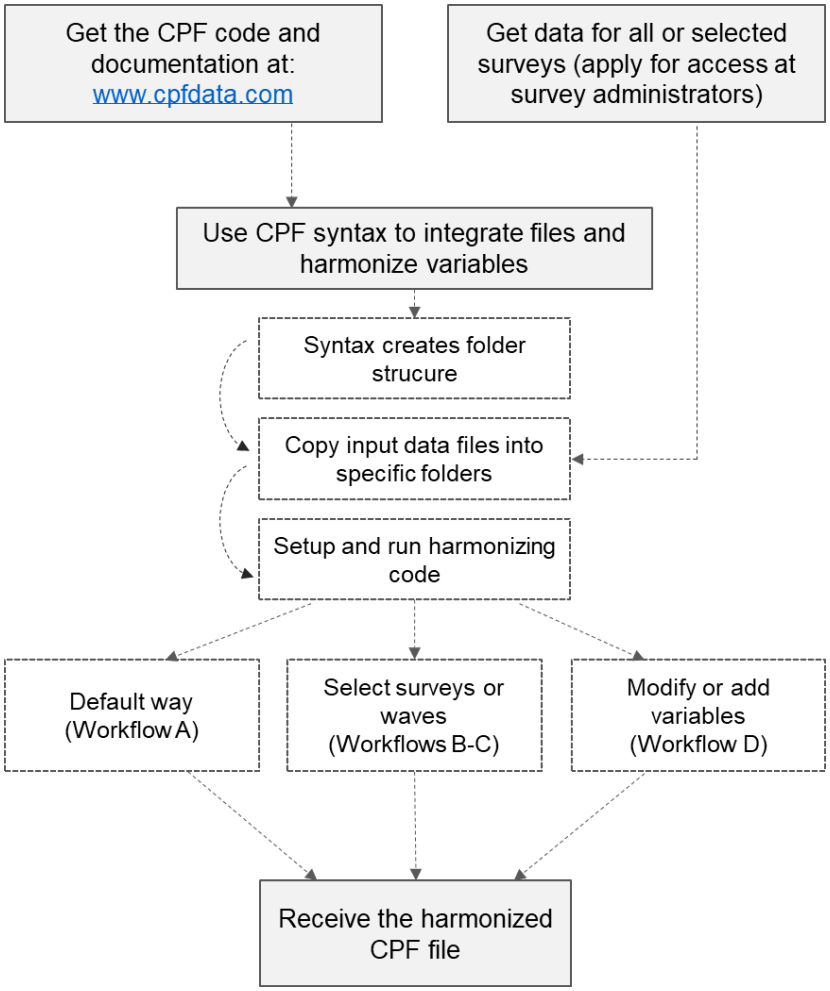
1. **Interview status**:
    a. South Korea, Russia, Switzerland, the UK: keep all observations (including proxy responses)
    b. Australia, Germany: keep direct respondents only
    c. The US: only reference persons (heads) and partners (spouses) (see details on PSID for explanation)
1. **Age**: 18 and older
2. **Delete observations** with missing values for gender and age (a minor correction)

Users can easily modify these selection criteria (see: *Workflow D - Adjustments to sampling criteria*).

# How to work with the CPF syntax

The CPF provides the syntax (programming code in Stata do-file format), the *Manual* explains how to work with the syntax, and the *Codebook* describes all variables. The code allows combining the separate raw survey data into a single harmonised data file. A step-by-step guideline of how to work with the CPF is presented in Figure 4. Users must first apply for access to each of the original datasets (see: *Obtaining the original data*). When access is granted, the first syntax can be run to set up a folder structure where original survey files can be extracted. Then, users can easily follow the instructions to build the comparative file in the default way or modify the procedures according to their needs. In the latter case, the hierarchical design of the code allows locating all the steps in the algorithms easily. Country-specific syntaxes are commented and organised in a similar way to facilitate the work.

**Figure 4.** *A step-by-step guide through using the CPF code*



There are four general ways of working with the CPF syntaxes (workflows). *Workflow A* describes the basic approach which constructs the data without any modifications. *Workflows B*, *C* and *D* refer to different modifications of the existing syntaxes, with *Workflow D* being the most flexible and advanced. All approaches are described in the following paragraphs. More details are in the syntax's comments which additionally include references to the *workflows A-D* to highlight places which might require adjustments.

## *Getting started*

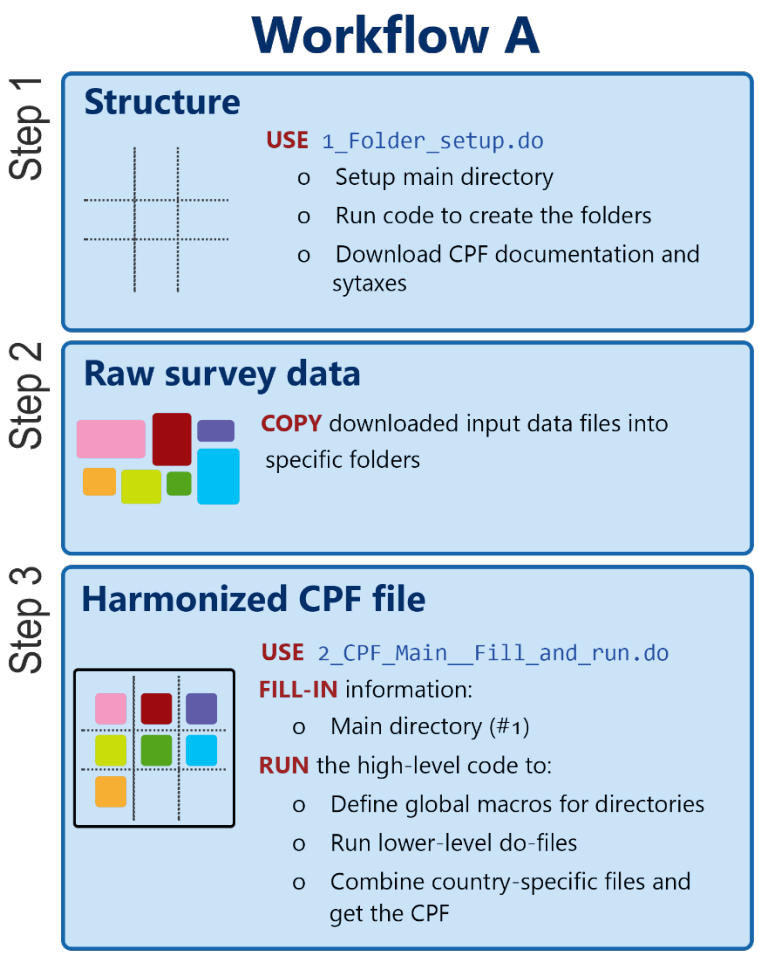
How to start using the CPF:

1. Download the latest CPF-code from the Github, OSF or the CPF webpage (all sources are synchronized and provide the latest version of the code). For example, to code can be downloaded as **CPF-Code-main.zip**. Unpack the entire folder structure with CPF-do-files to `11_CPF_in_syntax`. Or simply rename the un-packed `CPF-Code-main` as `11_CPF_in_syntax`.
2. Run `1_Folder_setup.do` according to instruction in the do-file.
3. Copy the original input datafiles (e.g. from SOEP) to specific folders (e.g. `C:\CPF\02_Country_Data_Origin\06_SOEP\data`).
4. Go to `2_CPF_Main__Fill_and_run.do` and follow instructions regarding the worfklows.

## *Basic workflow A: Three steps to get the CPF data*

The basic way of working with the CPF syntax leads from the raw data to a CPF harmonised dataset without any modifications (such as modifying variables, adding new variables, adding new waves, or selecting countries – for these see the next part). The approach requires only to use two higher-level syntaxes (1 and 2). Workflow A consists of three basic steps, as presented in Figure 5:

**Figure 5.** *The basic way of working with the CPF syntax (workflow A)*




Users first have to fill in the necessary information, such as the directory in the first syntax (`1_Folder_setup.do`) and run it to create an appropriate folder structure (see the part on *Folder Structure*). Then, they can place the downloaded data in specific folders of the `02_Country_Data_Origin` main folder. There is a separate subfolder for each survey (e.g. `01_HILDA`) with a subfolder `Data`, where the files should be copied, e.g. `02_Country_Data_Origin\01_HILDA\Data` (see details in *Obtaining original data*). The next step uses the second syntax (*2_CPF_Main__Fill_and_run.do*)

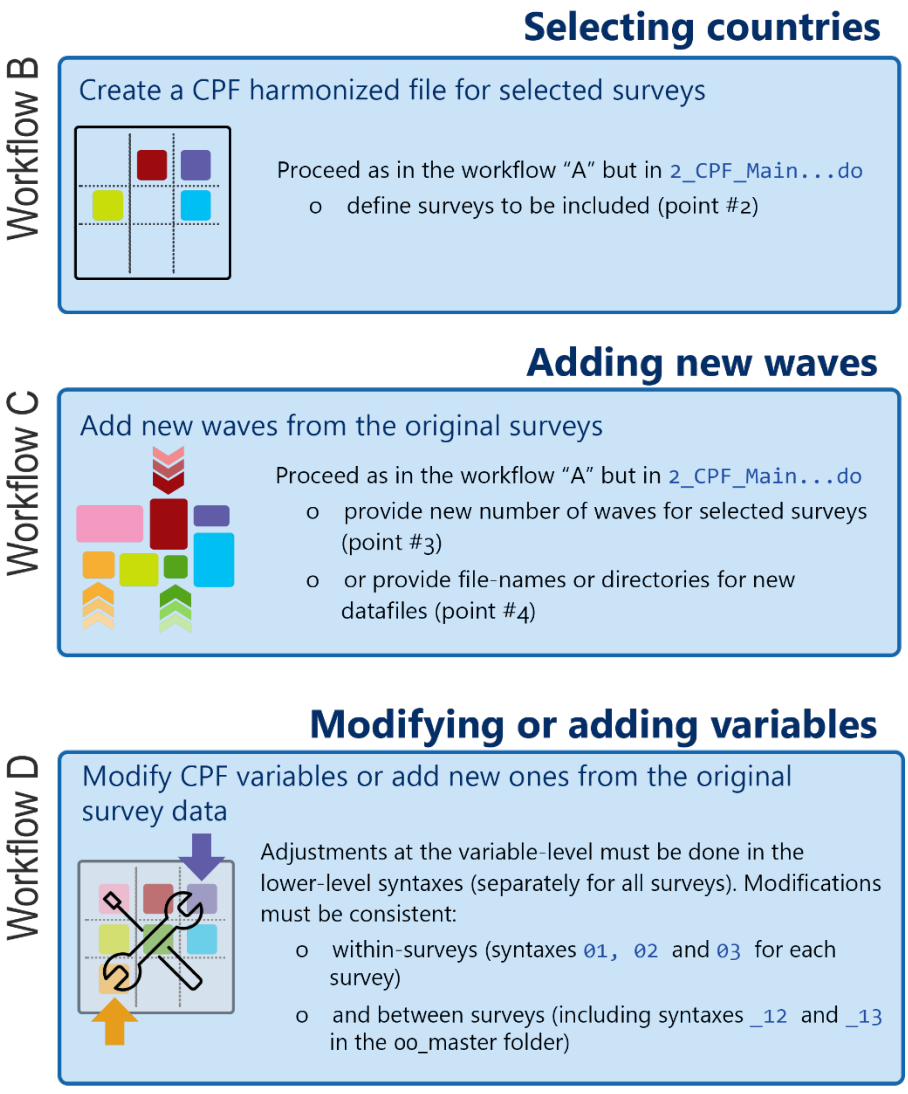
The third step uses syntax `2_CPF_Main__Fill_and_run.do` to call all lower-level syntaxes (*_10*, *_11*, *_12*, and *_13*). Users only have to fill the address of the main directory in #1. (Also check the information on the number of waves in #3 and file names in #4 – see Workflow C). With these information, the code in parts #5-#7 can be simply run to activate lower-level syntaxes (`_10`, `_11`, `_12`, and `_13`). Within this structure, syntax *_11* calls all country-specific syntaxes (multiple do-files numbered *_01, _02, _03*). Note

that operations – especially in #6 – are complex and running the code can easily take an hour or two on an average computer. More details are in the syntax's comments. For any modifications or problems, refer to workflows B, C or D.

## *Advanced workflows B, C and D: Modifying and adding data*

The other workflows can be used if users wish to modify or add data (Figure 6).

**Figure 6.** *Advanced ways of working with the CPF syntax (workflows B, C and D)*



### *Workflow B*

*Workflow B* allows selecting surveys to be included in the harmonised CPF dataset by changing the list of surveys in syntax 2. This option can be useful for users who do not have access to all surveys or have no need to harmonise all of them. The only difference in the procedure compared to Workflow A, is to define surveys to be included in `2_CPF_Main__Fill_and_run.do` (point #2). For example, to keep all surveys, simply leave all their names in the global macro (use lowercase):

```
global surveys “hilda klips psid rlms shp soep ukhls”
```

To select only PSID, RLMS and UKHLS, keep only respective names in the code:

```
global surveys “psid rlms ukhls”
```

The rest of the procedure is limited to the default running of the entire code in syntax 2.

### *Workflow C*

*Workflow C* serves to add new waves when they become available for the surveys. The CPF code will be regularly adjusted to incorporate new waves, however, some users might want to modify the syntaxes on their own. In most of cases, the procedure should be easy and limited to filling-in information in `2_CPF_Main__Fill_and_run.do` on the number of waves in #3 (for HILDA, KLIPS, SHP, SOEP and UKHLS) and file names in #4 (for RLMS, UKHLS and PSID).

The new number of waves for selected surveys should be filled in in point #3, e.g.:

```
global hilda_w "18" // for 18 waves
global ukhls_w "9"  // for 9 waves (it refers only to the UKHLS, not BHPS waves)
```

This approach does not apply to RLMS and PSID, for which the procedure is different. Data for RLSM is downloaded as a multi-wave dataset, so updating it with new waves only requires only to update the filename (point #4), e.g.:

```
global rlms_dataIND "USER_RLMS-HSE_IND_1994_2018_v2_eng_STATA.dta"
global rlms_dataHH "USER_RLMS-HSE_HH_1994_2018_eng.dta"
```

For PSID, the adding new waves in more complex and must be done manually for each variable (see: *Survey-specific details on PSID*). The latest name of the individual dataset has to be added in:

```
global psid_ind_er "${psid_in}\pack\IND2017ER.txt", clear
```

Additionally, new waves of UKHLS may require to update a directory for the data files (point #4), e.g.:

```
global ukhls_data "UKDA-6614-stata\stata\stata11_se"
```

Note that releases of the new waves can also bring changes to the original data structure which do not fit the current CPF algorithms, such as changes in names of variables, files or directories. In such cases,

additional adjustments have to be made in higher-level syntax 2 and/or lower-level syntaxes _01 (see *Workflow D*). For PSID, adding new waves is more complicated because the names of variables in new waves change and must be updated in the code manually (see: *Survey-specific details on PSID*).

## *Workflow D*

Workflow D refers to all other modification of the existing structure of the CPF data. Users can modify variables, add new ones, or modify the criteria for sample selection. Any adjustments of this type must be made in the lower-level syntaxes, separately and consistently for all surveys and for the master-syntaxes. Depending on the character of modifications, the procedure can be easy or complicated.

**Adjustments to variables**

1. When adding or modifying variables, users should:
    - Carefully explore questionnaires, codebooks and data
    - Assess the consistency of original variables across waves
    - Assess the consistency of new variables between countries
    - Perform check-up of the new variables within a country (logical rules, distributions, cross-tabulations)

2. Adjustments to variables must be first introduced in the survey-specific syntaxes 01 and 02 stored in the survey-folders (e.g. `11_CPF_in_syntax\01_HILDA`). Note, that for some surveys there are multiple syntaxes at levels 01 and 02. The main steps include:

- New variables can be added to the main CPF dataset using do-files 01. The procedure depends on the structure of the original data. For most of the surveys (HILDA, KLIPS, RLMS, UKHLS), all original variables are already available in the **`xx_01.dta`** file created with the 01 syntaxes (note that code for UKHLS drops some variables at the end). Note that for UKHLS, variables that need to be kept in the harmonized file need to be listed under the isvar command for both BHPS and UKHLS. SOEP and SHP require to add variables from multiple source datafiles using 01 code. The procedure is more complex for PSID, where a whole set of variable names

has to be added (so-called item-blocks) using 01_3. More details can be found in survey-descriptions and instruction in the do-files.

- Modifications of the new or existing variables (such as recoding, combining multiple variables, renaming etc.) can be done in do-files 02. This is a place to harmonise the variables across surveys.
- Always include new or modified variable names in the *keep* commands at the end of the file.
- Syntaxes 03 do not have to be adjusted in case of variable-level modification.

3. Further on, users have to adjust the master (between-surveys) syntaxes _12 – _13 stored in the 11_CPF_in_syntax\00_master folder as follows:

- New or modified variables must be added to the *keep* code after appending data in _12.
- Then, appropriate labels must be included in _13.

**Adjustments to sampling criteria**

Adjustments to the sampling criteria can be done in syntaxes 03 separately for each survey. They should not require additional modifications in other syntaxes.

## *Troubleshooting*

**Computer requirements**

- The CPF code is available in Stata, and it has been prepared in Stata 17. Running the entire code can easily take an hour or two on an average computer. Faster processor and more working memory will speed up the process.
- It is recommended to have at least 80 GB storage hard drive space if all countries are included (the original data files require minimum 50 GB and the CPF working and output files need additional 25 GB).

**Large size of files and computer-power limitations**

- Some surveys, particularily UKHLS and SOEP, have some very large files and operations may be challenging whith limited disk space or computer-power.
- In such case, users can add option to keep only the necessary variables at the end of 01_Prepare_data.do for particular countries. As default, the CPF code does not drop variables

(with some exception, e.g. UKHLS) to provide an easier way to add new variables to the harmonized CPF file.

- Users can also run the code step-by-step, in smaller pieces, or delete unnecessary files (e.g. files **01** and **02** which are used to create the final **03** file).

**Searching for errors**

- Although, the structure of the CPF code is complex, there are ways to locate errors if the expected outcome does not appear.
- It's recommended to run the code step-by-step, in smaller pieces, also to learn the procedure better. When the code is run from the higher-level syntax, Stata does not print errors by default. Running the code from the lower-level syntax (following the order of the files, i.e. **01**, **02** and **03**), allows to control it better.
- For each country, the code produces three files (apart from other temporary or working files) numbered **01**, 02 and **03** what correspond to respective syntax names. If the code does not produce file **02.dta**, it suggests that an it's best to search for errors in syntax 02.do.
- Many errors may be related to problem with defining global macros in sytaxes 1, 2 and 10. Especially when adding new waves or working on older releases of the source data.

**Older versions of the source files**

- CPF version 1.5 was built on data versions released in 2020 (PSID ver. 2019), 2021 (HILDA ver. 2000), 2022 (SOEP ve. 37, RLMS ver. 2021, UKHLS ver 12, SHP ver. 22) and 2023 (KLIPS ver. 24).
- New waves will be continuously integrated into the CPF code; users can also do this independently (see Workflow C in Using the CPF syntax).
- Backward compatibility with older releases may not be available for some variables or surveys due to changes in variables names and file structure in the original data sources
- To work with older versions of the source data, first change information in 2_CPF_Main__Fill_and_run.do in point A.1. (e.g. change *global soep_w "35"* to "34" or older). However, names of some variables could have changed between waves. In such case, the syntax (mostly syntax 01 for particular survey) would have to be modified. In such a case, GitHub provides access to older releases of the CPF code (https://github.com/cpfdata/CPF-Code/releases) and documentation (https://github.com/cpfdata/CPF-Documentation/releases).

**Questions and answers**

- The best place to questions and answers is the Web Forum (www.cpfdata.com/forum).

# CPF syntax: design, details and advanced options

## *Folder structure*

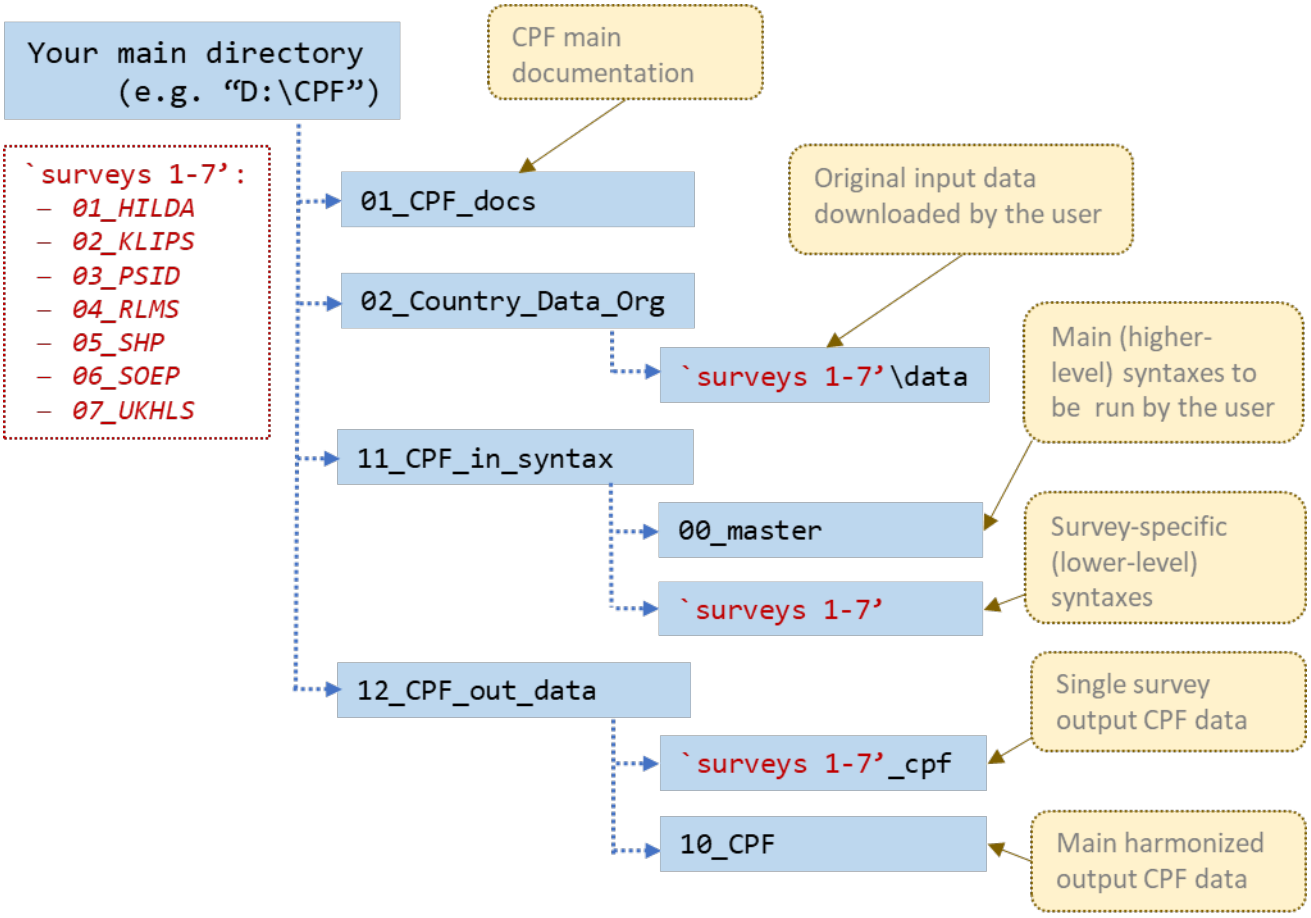
In order to run the CPF code, users must first create a folder structure as presented in Figure 7. It is done automatically by running `1_Folder_setup.do`:

1. First, insert the directory to your CPF folder in #1 ("Your local directory"), e.g.:

```
global your_dir " D:\CPF"  // <--insert your directory
```

2. Running the rest of the code (#2-#3) will create appropriate folders.
3. Then you can copy the original input datafiles to specific folders

**Figure 7.** *Structure of the folders required for the CPF algorithms*



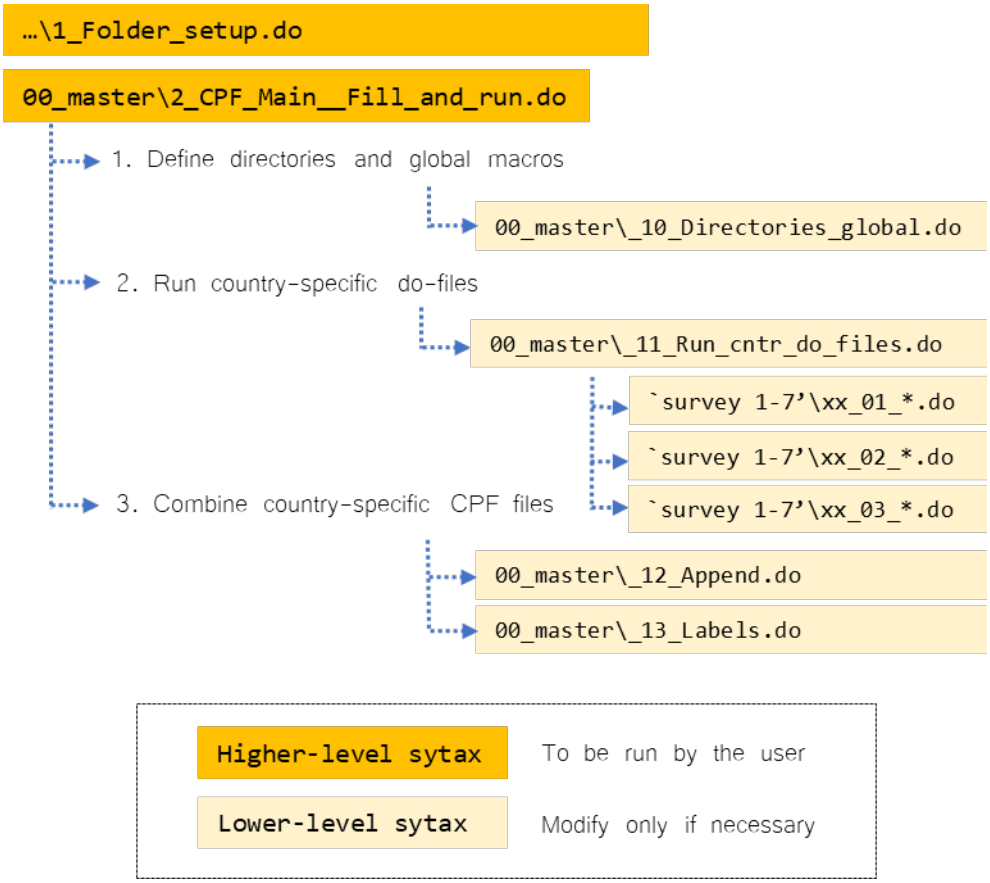
## *Syntax: higher and lower-level code*

CPF syntaxes (codes) are designed at two levels: higher and lower (Figure 8). Two *higher-level syntaxes* are: `1_Folder_setup.do` to set up the directory and folder structure, and `2_CPF_Main__Fill_and_run.do` to harmonise the dataset. These are short meta-code and do not refer directly to variables or data files. Instead, they work as an interface and allow to fill in the necessary information (e.g. file directory) and setup options for harmonisation (e.g. which surveys to include). Higher-level syntaxes call all the required code of a more complex structure of lower-level syntaxes.

For each survey, there are separate *lower-level syntaxes*, and the algorithms are designed differently. However, they all lead through the same three steps: the first constructs initial separate country data in a long format by merging original files, the second harmonises variables within countries according to the common template, and the third selects comparative samples. The process results in separate datasets with the same data structure for each country. Then, all country files can be combined into the single CPF harmonised dataset using a higher-level syntax.

Users can easily follow the instructions to build the comparative file in the default way (Workflow A) or modify the procedures according to their needs. In the latter case, the hierarchical design of the code allows locating all the steps in the algorithms quickly. Country-specific syntaxes are commented and organised in a similar way to facilitate the work. Many users can be interested in syntaxes *_01* which contain code that integrates all raw files into single and ready for analysis (yet un-harmonised) country data sets.

**Figure 8**. *Structure of the higher and lower level syntaxes*




## *Design of the lower-level code*

Harmonisation of the data is done in four steps using a lower-level syntaxes specific for each survey (stored in `11_CPF_in_syntax\`survey'\`). The first step is to construct the base separate-country data in a long format, the second – to harmonise variables within a country-files, the third – to select the sample, and the fourth – to combine all country files into a single CPF harmonised dataset. All of these steps can be run from the higher-level syntax `2_CPF_Main….do`.

**Step 1.** Preparation of the long format base-file for each country

- **Input:** original surveys' datafiles (stored in `02_Country_Data_Origin\`survey'\Data`)
- **Syntaxes**: `11_CPF_in_syntax\`survey'\xx_01_*.do`
- **Result**: data file(s) in a long format: `12_CPF_out_data\`survey'_cpf\xx_01_*.dta`

First, for each country, we must construct the base-file in a long format, which contains all (or selected) source variables, as provided by the data supplier. A long-format of panel data means that the repeated observations are clustered by individuals so that each row of data refers to respondent's information from a specific wave (contrary to wide data, where wave-specific information is provided in separate variables, e.g. health_wave1, health _wave2).

The procedure is different for each country, and its complexity depends on the data structure. Most of the datasets require to combine (append and merge) separate files for specific waves and/or types of surveys (e.g. individual, household or topic-specific questionnaires). Codes for this stage often include a group renaming of variables to a format which is required in a long-format file (e.g. removing wave-reference in variable names). For the PSID, the procedure is much more complicated, since it needs first to retrieve and combine sets of variables which refer to the same question or concept. The RLMS, on the other hand, is already provided in a long format for specific types of questionnaires. For some countries, a pre-selection of variables and initial cleaning is done at this stage. In addition to the raw data files, in some cases, CPF also uses selected CNEF variables (separate CNEF data files are provided for HILDA, SHP and SOEP).

Step 1 uses lower-level syntaxes `xx_01_*.do`. The result of this step is one or more datafiles **`xx_01_*.dta`** with a large number of non-harmonised variables in a long data format.

**Step 2.** Harmonisation of variables within a country

- **Input**: `12_CPF_out_data\`survey'_cpf\`**`xx_01_*.dta`**
- **Syntaxes**: `11_CPF_in_syntax\`survey'\xx_02_*.do`
- **Result**: a single data file for each country with a harmonized variable structure (`12_CPF_out_data\`survey'_cpf\`**`xx_02_CPF.dta`**)

In the second step, we use variables from the **`xx_01_*.dta`** base-file(s) to construct new harmonized variables. The harmonised variables are the same for all countries in terms of names, format and response categories. However, some of them are available only for selected surveys.

The harmonisation process involves:

- Recoding and combining original variables into new variables
- When necessary, creating additional versions of variables (not fully harmonised)

- Selecting additional variables to keep (e.g. weights, sample characteristics, other country-specific variables)
- Basic data cleaning (which is not a full preparation for analysis)

Step 2 uses lowe-level syntaxes xx_02_*.do. A result of this step is a single datafile **xx_02_CPF.dta** with a set of harmonised variables in a long data format.

**Step 3.** Sample selection

- **Input**: 12_CPF_out_data\`survey'_cpf**xx_02_CPF.dta**
- **Syntaxes**: 11_CPF_in_syntax\`survey'\xx_03_ Sample_selection.do
- **Result**: a single data file for each country with a harmonized variable structure (12_CPF_out_data\`survey'_cpf**xx_03_CPF.dta**)

During the third step, we select the final sample to be included in the main CPF dataset. The selection is based on the interview status (keeping different types of interviewed respondents or proxy-interviews), age criteria (age 18+) and missing values (keeping individuals with information on age and gender). For some surveys, e.g. PSID or SOEP, selection of the sample is not straightforward, and users might want to adjust the criteria based on their research needs. Step 2 uses lower-level syntaxes xx_03_ Sample_selection.do. A result of this step is a single datafile **xx_03_CPF.dta** with harmonised sample criteria.

**Step 4.** Combining country data into a one harmonised CPF dataset

- **Input**: harmonized separate survey datafiles 12_CPF_out_data\`survey'_cpf**xx_03_CPF.dta**
- **Syntaxes**: 11_CPF_in_syntax\00_master_12_Append.do and _13_Labels.do
- **Result**: a single CPF data file for all countries (12_CPF_out_data\10_CPF**CPFvX.X.dta**)

Finally, all separate country-files with a harmonised structure of the data are merged into a single harmonised CPF dataset. It is done by running _12 syntax. Additionally, labels for variables and categories are added in syntax _13. The result is the final CPF dataset, e.g. **CPFv1.0.dta.**

## *Obtaining the original data*

Users must first apply for access to each of the original datasets independently at national administrator institutions. Access is free of charge, but in most cases, users must describe their research goals and sign a contract. When access is granted, data can be extracted to specific CPF subfolders in the `02_Cntry_Data_Orgin\`survey'\Data` folders, as explained below. With new waves, users have to modify global macros in #3 of syntax 2 (see Workflow C), e.g.:

```
global klips_w        "21"          // number of waves
```

However, if the approach to naming variables, folders or datafiles in the original data changes in the future, additional adjustments have to be made in higher-level syntax 2 and/or lower-level syntaxes `xx_01`.

CPF version 1.5 was built on data versions released in 2020 (PSID ver. 2019), 2021 (HILDA ver. 2000), 2022 (SOEP ve. 37, RLMS ver. 2021, UKHLS ver 12, SHP ver. 22) and 2023 (KLIPS ver. 24). Backward compatibility with older releases may not be available for some variables or surveys due to changes in variables names and file structure (but the syntax can be modified). New waves will be continuously integrated into the CPF code; users can also do this independently (see *Workflow C* in *Using the CPF syntax*).

Note: the specific file and folder names for the datasets may change between waves. The names given below serve as an illustration of the data structure and may not correspond to the names given in the latest data release.

### 01_HILDA – Australia

Apply for the data via the National Centre for Longitudinal Data Dataverse (Australian Government Department of Social Services): https://dataverse.ada.edu.au/dataverse/ncld. Unpack downloaded files, such as `STATA 190c (1-Combined Data Files)` and `STATA 190c (2-Other Data Files)`, to subfolders indicated as "Combined" and "Other" in the "Data" folder. The final structure should look as follows:

02_Cntry_Data_Orgin\01_HILDA\Data
- ↳ STATA 190c (Combined)
  - ↳ **Combined_r190c.dta**
  - ↳ **…**
- ↳ STATA 190c (Other)
  - ↳ **Household_r190c.dta**
  - ↳ **…**

Note that names of downloaded folders change between waves. The CPF-names for Hilda are given in the syntax in 10. The number 190 in folders' names refers to the current version (number of waves) of HILDA which is filled in #3 of syntax 2 as, e.g. 19 (see Workflow C):

```
global hilda_w  "19"          // version of HILDA, number of waves
```

## 02_KLIPS – South Korea

Data are available via the official website for registered users: www.kli.re.kr/klips_eng. Unpack all downloaded files directly in the Data folder, e.g.:

02_Cntry_Data_Orgin\02_KLIPS \Data
↳ **eklips01h.dta**

With new waves, users have to modify global macros in #3 of syntax 2 (see Workflow C), e.g.:

```
global klips_w        "21"          // number of waves
```

## 03_PSID – US

The logic behind PSID differs from other datasets and is much more complex (see *Survey-specific details* for PSID). To organise the data, we use **psidtools** ado (Kohler, 2015)[5], which can be downloaded using:

```
ssc install psidtools
```

Data are available via the official website for registered users:
https://simba.isr.umich.edu/Zips/ZipMain.aspx:

1. Download all Family Files (one per wave, e.g. fam2019er.zip) and place them in into Family and Ind Files (zip). Do not unpack.
2. Download **Cross-year Individual: 1968-XXXX** zipped file and place it in Family and Ind Files (zip). Do not unpack.

[5] PSIDTOOLS is Stata' module to facilitate access to PSID, developed by Ulrich Kohler from the University of Potsdam. See: https://ideas.repec.org/c/boc/bocode/s457951.html.

3. Leave all files in the Family and Ind Files (zip) folder unpacked but additionally unpack the **`Cross-year Individual: 1968-XXXX`** zipped file (e.g. **ind2019er.zip**) to Data/Cross-year Individual 1968-XXXX/pack. It should contain a txt file with vales named, e.g. **IND2019ER.txt** (which is defined in 10 as `global psid_ind_er "${psid_in}\pack\IND${psid_w}ER.txt"` based on the latest PSID year indicated in 2 as `global psid_w`).

CPF syntax manages further reorganisation of the files. E.g. after running the lower-level syntaxes 01 for PSID, the code will unpack and combine required files into PSIDtools_files folder, and syntax 03 will create a number of item-specific files in the temporary directories.

After downloading, the PSID data-folder should look as following:

02_Cntry_Data_Orgin\03_PSID\Data
- Cross-year Individual 1968-2019
  - pack
    - **IND2019ER.txt**
  - **...** *(place for automatically created file **psid_crossy_ind.dta**)*
- Family and Ind Files (zip)
  - Cross-year Individual 1968-2019.zip
  - **fam1968.zip**
  - **ind2019er.zip**
  - **...**
- PSIDtools_files
  - **...** *(place for automatically unpacked files, e.g **fam1968.dta**)*

## 04_RLMS – Russia

The data are available via the Higher School of Economics (HSE) without application: www.hse.ru/en/rlms. Additionally, the data may be downloaded at the Carolina Population Center (CPC) without application: www.cpc.unc.edu/projects/rlms-hse/data. There should be two multi-wave files in long-data formats: for the individuals and households. Unpack them into the main Data folder, e.g.:

02_Cntry_Data_Orgin\04_RLMS\Data
- **USER_RLMS-HSE_IND_1994_2020_v3_eng.dta**
- **USER_RLMS-HSE_HH_1994_2020_3_eng.dta**

Names of the files can differ depending on the source (HSE or CPC). Both for the individual and household files, the names have to be properly included in syntax `2_CPF_Main__Fill_and_run.do` in #4, e.g.:

```
* RLMS
 global rlms_dataIND "USER_RLMS-HSE_IND_1994_2020_v3_eng.dta"
 global rlms_dataHH " USER_RLMS-HSE_HH_1994_2020_3_eng.dta"
```

## 05_SHP – Switzerland

Data are available via FORSbase for registered users: https://forscenter.ch/projects/swiss-household-panel/data. Unpack all folders from **Data_STATA.zip** into the main `Data` folder. It should then contain several folders with different types of datasets. The main source of the individual- and household-level data are files in `SHP-Data-W1-W21-STATA` folder (e.g. **shp99_p_user.dta**). Additionally, CPF refers to other folders, including `SHP-Data-CNEF-STATA` and `SHP-Data-WA-STATA`. After unpacking, the structure should look as follows:

```
02_Cntry_Data_Orgin\05_SHP\Data
    ↳ SHP-Data-Biography-STATA
        ↳ SHP0_bh_user.dta
        ↳ ...
    ↳ SHP-Data-CNEF-STATA
        ↳ shpequiv_1999.dta
        ↳ ...
    ↳ SHP-Data-Imputed-Income-Wealth-STATA
        ↳ imputed_income_hh_long_shp.dta
        ↳ ...
    ↳ SHP-Data-Interviewers-STATA
        ↳ shp00_v_user.dta
        ↳ ...
    ↳ SHP-Data-SHP-3-W1-STATA
        ↳ shpiii_cs_user.dta
        ↳ ...
    ↳ SHP-Data-W1-W21-STATA
        ↳ W1_1999
            ↳ shp99_p_user.dta
            ↳ shp99_h_user.dta
        ↳ ...
```

↳ SHP-Data-WA-STATA
↳ **shp_ca.dta**
↳ **...**

The number of waves in the folder name SHP-Data-W1-W21-STATA is accounted for automatically after filling it in in the syntax 2_CPF_Main__Fill_and_run.do in #3. Be aware, however, of any future changes in folder names.

### 06_SOEP – Germany

Data are available via the Research Data Center SOEP after granting access:
www.diw.de/en/diw_02.c.242211.en/criteria_fdz_soep.html

Data should be unpacked into Data keeping additionally the wave-specific subfolder (e.g. soep.v35), which contains then all the SOEP files.

02_Cntry_Data_Orgin\05_SHP\Data
↳ soep.v35
↳ **abroad.dta**
↳ **pl.dta**
↳ **...**

The wave-specific subfolder name is accounted for automatically after filling in the number of waves in the syntax 2_CPF_Main__Fill_and_run.do in #3.

### 07_UKHLS - UK

Data are available via the UK Data Service after granting access: www.ukdataservice.ac.uk.

Data should be unpacked into Data with keeping additionally the specific subfolders' path (e.g. UKDA-6614-stata\stata\stata11_se), which contains then all the wave-specific folders. These folders (e.g. bhps_w1, ukhls_w1) contain the data files for each wave.

The specific path may change from wave to wave and has to be properly included in syntax 2_CPF_Main__Fill_and_run.do in #4, e.g.:

```
* UKHLS
global ukhls_data "UKDA-6614-stata\stata\stata13_se"
```

The structure should look like following:

```
02_Cntry_Data_Orgin\07_UKHLS\Data
    ↳ UKDA-6614-stata
        ↳ stata
            ↳ stata13_se
                ↳ bhps
                    ↳ ba_indresp.dta
                    ↳ …
                ↳ ukhls
                    ↳ a_indresp.dta
                    ↳ …
```

## *Survey-specific details*

This part presents additional details and instruction for the lower-level (survey-specific) codes.

### 01_HILDA – Australia

1. au_01_Prepare_data– to prepare data
    a. Step 1: Rename datasets
    b. Step 2: Append waves of the original dataset
2. au_02_1_Harmonize (p1- cnef)– code for the CNEF data file
3. au_02_2_Harmonize (p2- combined)– code for the original variables from all waves
4. au_02_3_Combine_p1p2 – combines p1 and p2 into a **xx_02_CPF.dta**
5. au_03_Sample_selection – sample selection

### 02_KLIPS – South Korea

1. ko_01_Prepare_data – to prepare data
    a. Install ado **renvars** (for renaming)

```
net install http://www.stata-journal.com/software/sj5-4/dm88_1
```

    b. Step 1: Prepare p-files
    c. Step 2: Prepare h-files
    d. Step 3: Combine p & h files
2. ko_02_Harmonize– to prepare harmonised variables
3. ko_03_Sample_selection – sample selection

## 03_PSID – US

The philosophy behind PSID data differs from other surveys. PSID is the oldest ongoing research in the CPF set and the most challenging to incorporate. Unfortunately, unlike in other surveys, adding new waves of PSID to CPF cannot be achieved at the file-level (e.g. by updating a file name). The reason is that names of variables which refer to the same construct (e.g. age, employment status) change from wave to wave (e.g. the variable for employment status is named ER30509 in 1986, ER33111 in 1994, ER34317 in 2015, and ER34516 in 2017). Therefore, all variables must be retrieved separately from all waves by searching in PSID's online system (https://simba.isr.umich.edu). This is a challenge for users who would like to add new items or new waves to the CPF data. To add to the complexity, items are stored in separate variables for Reference Persons (called Heads in older waves) and Partners (called Spouses in older waves). Also, similar questions were sometimes framed differently for Reference Persons and Partners and included in different waves. Therefore, the syntax for PSID is more complex and requires additional clarification.

1. us_01_1_Create_psid_crossy_ind

This lower-level code is run from 2_CPF_Main… syntax to create **`psid_crossy_ind.dta`**. Note, however, that code us_01_1 uses the input code provided by PSID in the zipped file (e.g. IND2017ER) which might have to be updated with new waves. Also, the code us_01_1 refers to the individual data file at the end of the **infix** part, which has to be updated in the 2_CPF_Main.

2. us_01_2_Create_waves_psidtools

The **psidtools** implemented in this code will automatically unpack the files, prepare and copy them into PSIDtools_files (one family file per wave and one individual file). These are the base input files for the CPF dataset. After this, the zipped files in Family and Ind Files (zip) folder can be deleted.

3. us_01_3_Get_vars

When adding new variables or new waves to PSID, users must modify the us_01_3_Get_vars. It contains a **combvars** program which is a wrap up for **psidtools** command. **Combvars** is used to

- combine variables across waves - strings of variables related to the same item are different for each wave (e.g. age: ER30004, ER30023, etc.)
- reshape them into a long format (using psidtools)
- save them as separate files
- add names to global macro for further use (merging)

The **combvars** program uses files created by **psidtools** in PSIDtools_files folder. However, names of the variables have to inserted by hand (names can be found and copied from the PSID's online search tools at https://simba.isr.umich.edu/Zips/ZipMain.aspx) into specific item-lists. For example, the code to add and combine original variables which refer to the age of respondent is:

```
combvars age, list("[68]ER30004 [69]ER30023 [70]ER30046 [71]ER30070 [72]ER30094
[73]ER30120   [74]ER30141   [75]ER30163   [76]ER30191   [77]ER30220   [78]ER30249
[79]ER30286   [80]ER30316   [81]ER30346   [82]ER30376   [83]ER30402   [84]ER30432
[85]ER30466   [86]ER30501   [87]ER30538   [88]ER30573   [89]ER30609   [90]ER30645
[91]ER30692   [92]ER30736   [93]ER30809   [94]ER33104   [95]ER33204   [96]ER33304
[97]ER33404   [99]ER33504   [01]ER33604   [03]ER33704   [05]ER33804   [07]ER33904
[09]ER34004 [11]ER34104 [13]ER34204 [15]ER34305 [17]ER34504")
```

Unlike in other survey, names of variables in PSID changes from wave to wave. Thus, names must be first combined across waves. Each of the items on the list refers to a specific wave (e.g. [68] refers to wave from 1968). The last item on the list refers to the last wave (here: [17]ER34504). If new waves are to be added, users have to add an item to the list with a name of a variable in the latest wave (e.g. [19]ER…). It has to be done for all variables separately.

Thus, the syntax us_01_3 contains following steps:

- Step1: Define **combvars** program to be used in this syntax and run `global vars""`
- Step 2: Run **combvars** to combine vars across waves. This will create separate long files for each item
- Step 3: Combine single-item long files from step 2 into **us_01.dta**
- Step 4: Add variables which constant across all waves to **us_01.dta** (get them from long **psid_crossy_ind.dta)**

Command "Add new time-constant vars - only if necessary" can be used to add new time-constant variables once the **us_01.dta** is already created.

Command "Add new files - only if necessary" can be used to add a new block of items using **combvars** (after creating **us_01.dta**).

4. us_02_Harmonize
   - Selecting observations – the default option is "Keep 2: heads & partners". The current version of CPF is not adjusted to include other family members. However, users may choose different sets of observations if necessary.

- For ISCO variables, the code refers to external do-files (due to their length they are stored separately). Note that ISCO recoding can take a lot of time.

5. us_03_Sample_selection

In the default option, it keeps spouses and partners only (Keep 2), repeating the code in 02.

## 04_RLMS – Russia

1. ru_01_Prepare_data– to prepare data
    - Combine individual and household files (which are already combining waves in a long format)
2. ru_02_Harmonize– to prepare harmonised variables
3. ru_03_Sample_selection – sample selection

## 05_SHP – Switzerland

1. ch_01_1_Prepare_data_Equiv_to_long– to prepare supplementary CNEF variables
2. ch_01_2_Prepare_data_Waves_to_long– to prepare individual data

    This is also a place for adding new variables

    - there are 3-4 places you have to put the name of a new variable from the wave-specific files you want to add
    - these places are indicated as:

```
*>>>
*>>> NEW VARS [x*x; y*] 1/3:
*>>>
```

    - you must adjust the formatting of the name in each case
    - ***x*x*** - variables with year inside of the name, e.g. ***p17e50*** (3 places to add)
    - ***y**** - variables with the year at the end of the name, e.g. ***educat17*** (4 places to add)
    - please, verify if the results are correct, there are a few rules which help to check it
3. ch_02_1_Harmonize_ (p1- equiv)– to prepare supplementary variables from CNEF
4. ch_02_2_Harmonize_ (p2-waves)– to prepare the main variables

5. `ch_02_3_Combine_p1p2`– combine the main file with CNEF variables. Additionally, some missing values are cross-filed.
6. `ru_03_Sample_selection` – sample selection

## 06_SOEP – Germany

1. `ge_01_Prep_data`– to prepare data

   This is a place for adding new raw variables from the original SOEP datafiles. Users must identify the specific original data file and variables and add the name(s) in an appropriate place in the syntax. For example, if variable ***newvar*** comes from SOEP's **`health.dta`**, the original name of a variable must be added to the `KEEP` command under the headline `*# health.dta #`., e.g.:

```
*#############################
*#                           #
*#    health.dta             #
*#                           #
*#############################
*
use "${soep_in}\health.dta", clear
*
keep                ///
...                 /// already added variables
newvar              // the newvar added
*
```

   In case when the original data file is not listed in the syntax, it can be added in a similar way as the other files. First, open the new datafile, keep variables, and save the file under new name:

```
*#############################
*#                           #
*#    newdatafile.dta        #
*#                           #
*#############################
use "${soep_in}\newdatafile.dta", clear    // open the file

keep  ///
...   // add variables

rename persnr pid    // rename if neccessary
rename hhnr hid      // rename if neccessary
*
sort  pid syear
*
save "${soep_out_work}\gnewdatafile_1.dta", replace
```

Then, add the new file to the final `merge` command, e.g.:

```
*############################
*#                          #
*#    MERGING               #
*#                          #
*############################

***** HH + P
use "${soep_out_work}\gppathl_1.dta", clear
merge m:1 hid syear using "${soep_out_work}\ghpathl_1.dta" , keep(1 3) nogen

****
...
merge 1:1 pid syear using "${soep_out_work}\gnewdatafile_1.dta", keep(1 3) nogen
// adjust the code depending on the structure of the datafile
...
```

2. `ge_02_Harmonize` – to prepare harmonised variables
3. `ge_03_Sample_selection` – sample selection

## 07_UKHLS - UK

1. `uk_01_Prepare_data`– to prepare data
    - The code combines BHPS and UKHLS datasets
    - Note that operations require much disk space. Therefore, temporary files are deleted
    - Also, the combined file is large. For this reason, one of the last procedures in the syntax is `DROP` to delete variables which will not be used in further harmonisation.
      Users might have to adjust the command when adding new variables. Additionally, if the order of the variables changes with new editions, the `DROP` command must be modified (or deleted).
2. `uk_02_Harmonize`– to prepare harmonised variables
    - waves from BHPS and UKHLS have mostly separate sets of variables
3. `uk_03_Sample_selection` – sample selection

## *Doing analysis with the CPF*

To account for the hierarchical data structure, users can refer to the following variables:

- ***country*** – to identify countries (surveys)
- ***pid*** – to uniquely identify respondents (based the original id number from source surveys)
- ***wave***, ***wavey,*** or ***intyear*** – to include the time dimension:
    - ***wave*** – country-specific wave number (counting from 1)
    - ***wavey*** – the main (initial) year of data collection for a given wave
    - ***intyear*** – year of interview

There are different approaches to account for the entire 3-level hierarchical structure. For example, users can include countries as dummies, perform contextual analysis, run the separate analysis by country, or use robust (clustered) standard errors. Performing a multilevel model with all 3-levels is problematic due to the low number of countries (however, Bayesian approach can be considered in this case). For more information on multilevel and panel analysis, we recommend popular statistical handbooks:

- Gelman A., J. Hill (2007) "Data Analysis Using Regression And Multilevel/Hierarchical Models", New York: Cambridge University Press
- Snijders, T., R. Bosker (1999) "Multilevel Analysis: An introduction to basic and advanced multilevel modeling**",** London: Sage.
- Raudenbush, S.W., A.S. Bryk (2002) "Hierarchical Linear Models: Applications and Data Analysis Methods", Thousand Oaks, CA: Sage Publications
- Joop Hox (2002, 2010) "Multilevel Analysis: Techniques and Applications", Routledge
- Hoffman L., (2015) "Longitudinal Analysis: Modeling Within-Person Fluctuation and Change", Routledge
- Singer J., J. Willett (2003) "Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence", Oxford University Press
- Rabe-Hesketh S., A. Skrondal (2012) "Multilevel and Longitudinal Modeling Using Stata (3rd Edition)", Stata
- McElreath (2020). "Statistical Rethinking: A Bayesian Course (2nd Edition)", CRC Press
- Kruschke (2014) "Doing Bayesian Data Analysis: A Tutorial Introduction with R", Academic Press

For example, in Stata a simple regression model which accounts only for the country clustering by including a country-dummy can be written as:

```
reg satlife5 i.edu3 i.country
```

A panel model which accounts additionally for repeated observations (2-level model) can be written with the `mixed` command as:

```
mixed satlife5 i.edu3 i.country || pid:,
```

After defining a panel structure with `xtset`, a similar model can be written with the xt-command:

```
* Definie panel structure
xtset pid wave // counting waves from 1
* OR:
xtset pid wavey // counting waves by the calendar year
* Panel model
xtreg satlife5 i.edu3 i.country
```

General recommendations:

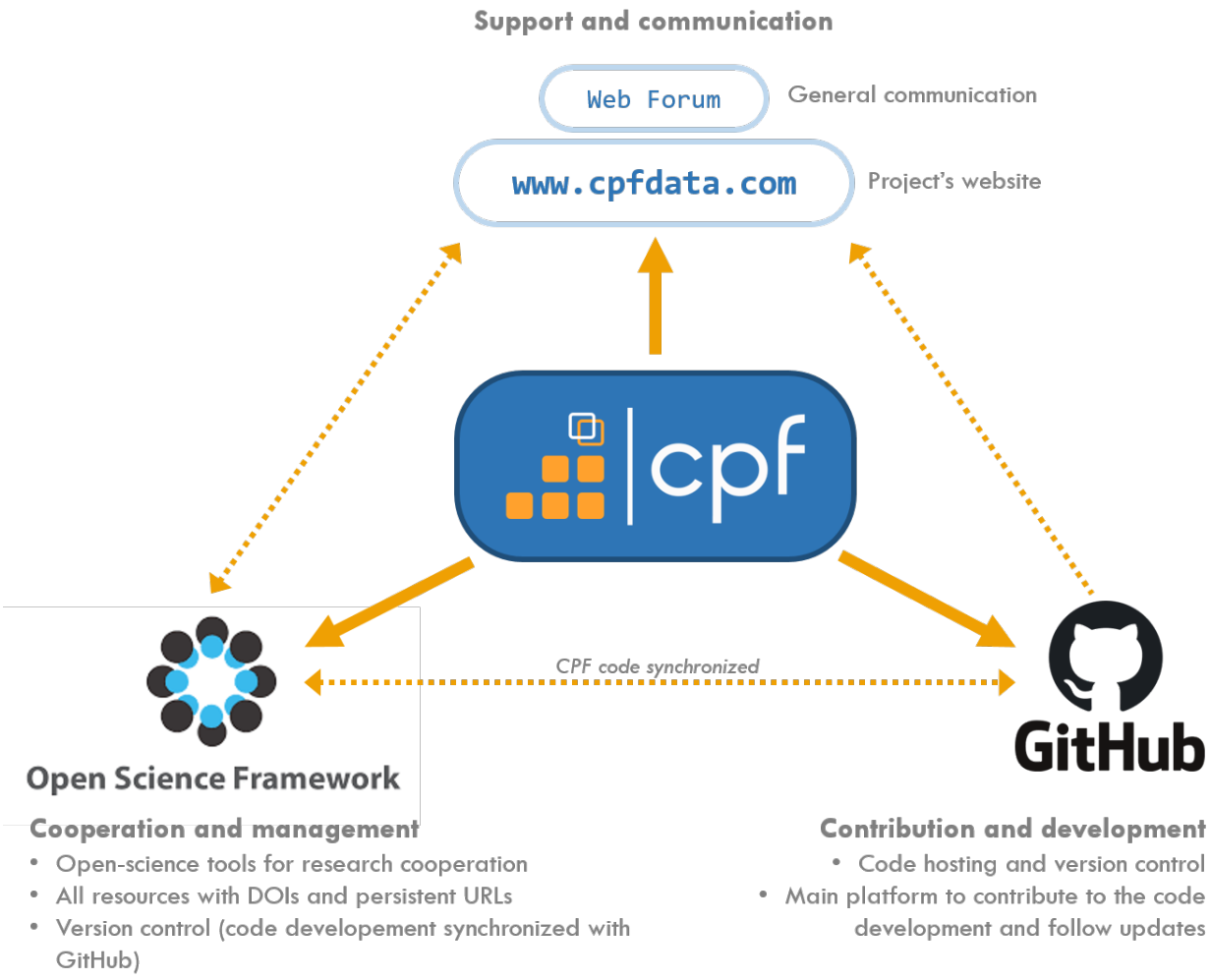
- Before any substantial analysis, consider the missing values (MV) in the used variables. In most cases, MV should be removed from the analysis, e.g. by recoding them into system-missing values (`.`), of applying `mvdecode` command.
- Given that the CPF data (in version 1.0) contains more than 2.5 million observations, complex statistical analysis can be running slowly. In such a case, run the initial analysis on subsamples. Alternatively, consider other statistical software, such as Mplus or R (especially for Bayesian analysis).
- Be aware of different time frames and differences in gaps between years of data collection (e.g. in PSID) when performing longitudinal analysis. Depending on the research question and method, consider using ***wave***, ***wavey*** or ***intyear***.

# Open-science platform for CPF

## *Tools and services*

CPF is an open-science project, which means that it provides access to all resources, including the programming code. Furthermore, the code can be improved and developed by anyone who wishes to contribute to the project. To allow the open access and community-based development, we have built an open-science platform that connects several tools: website, online Forum, GitHub and OSF (Figure 9).

**Figure 9.** The structure and tools of the CPF's open-science framework




The central element is the project's **website** (www.cpfdata.com) that contains all important information, documentation and the latest major version of the code. The website also includes an online forum. The

**Forum** serves general communication, discussions and suggestions related to the code. It may also be used for asking questions and providing answers.

**GitHub** (www.github.com) is precisely oriented at the development of the CPF code. GitHub is a code hosting platform for collaborations in code development, especially useful for managing open-source projects. It allows users to access the main and alternative versions of the code, share their modifications, track changes and continuously integrate them into consecutive versions. Extensions, improvements or alternative versions of the code can be offered by all researchers and programmers who register free of charge at the GitHub platform. Importantly, all changes are recorded, providing version control functionality.

**Open Science Framework** (OSF; www.osf.io) is one of the most popular open-science platforms, which facilitates open collaboration in research. OSF integrates many tools and services which support managing, organising, documenting and sharing all aspects of a project. Among others, OSF allows pre-registering studies, storing code and data; it is linked to preprint services and many scientific platforms. It facilitates collaborative workflow on projects, allows to document the work and progress. Similarly to GitHub, OSF uses a version control system, so all changes to the project are recorded. OSF allows additionally to register the project at each stage and creates an archival version of the project with a unique hyperlink. All materials can be registered this way, receiving permanent links and DOIs. Importantly, OSF includes a GitHub add-on which directly links files stored at GitHub repository into the OSF project. This way, changes to the code can be introduced either through GitHub or OSF, and they are synchronised so that the code at the OSF is always up to date.

Links to the resources:

- Website: www.cpfdata.com
- Forum: www.cpfdata.com/forum
- GitHub: www.github.com/cpfdata
- OSF: www.osf.io/h3yxq

## *Help and support*

The up-to-date documentation of CPF can always be found at the projects' website: http://www.cpfdata.com/download. Questions regarding the CPF code can be asked on the Forum or by email contact@cpfdata.com. CPF is an independent project developed on a voluntary basis. As such, it does not engage employees responsible for support and help. The CPF team will try to answer all the questions, but extensive support cannot always be provided.

## *Contribution and cooperation*

User's improvements and suggestions will be recorded, incorporated and shared using open online platforms (i.e. web forum and GitHub code repository) to allow continuous development and regular updates to the official versions of the code.

CPF is an open and ongoing project. We invite interested users to provide feedback (e.g. on the Forum) or contribute to the development of the code (through GitHub or OSF). We are also happy to cooperate in research or support the development of the research network by linking people and institutions. Do not hesitate to contact us!

# Acknowledgments

This study uses the following datasets, for which we are grateful to the data providers:

The British Household Panel Survey, BHPS, and Understanding Society – The UK Household Longitudinal Study, UKHLS. University of Essex, Institute for Social and Economic Research, NatCen Social Research, Kantar Public. (2019). Understanding Society: Waves 1-9, 2009-2018 and Harmonised BHPS: Waves 1-18, 1991-2009. 12th Edition. UK Data Service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-14.

Socio-Economic Panel (SOEP), data for years 1984-2020, SOEP-Core v37, EU Edition, 2022, doi:10.5684/soep.core.v37eu

Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (2020). https://psidonline.isr.umich.edu

Department of Social Services; Melbourne Institute of Applied Economic and Social Research, 2022, "The Household, Income and Labour Dynamics in Australia (HILDA) Survey, GENERAL RELEASE 21 (Waves 1-21)", doi:10.26193/KXNEBO, ADA Dataverse, V3

Korean Labor & Income Panel Study (KLIPS) version 24. Copyright Korea Labor Institute, 2020. www.kli.re.kr/klips_eng

Russia Longitudinal Monitoring Survey, RLMS-HSE, version 2023, conducted by National Research University "Higher School of Economics" and ZAO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS. (RLMS-HSE sites: http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms)

Swiss Household Panel (SHP), version 21, SHP is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is supported by the Swiss National Science Foundation. https://forscenter.ch/projects/swiss-household-panel.

The Cross-National Equivalent File project is sponsored by the National Institute on Aging (Grant: 5-R01AG040213-10) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Grants: 1-R03HD091871-01, 1-R03HD100924-01) and was conducted by The Ohio State University. www.cnef.ehe.osu.edu

The CPF team would also like to thank to the colleagues who supported development of the first version of CPF, in particular (in alphabetical order) Eldad Davidov, Dina Maskileyson, Aleja Rodríguez, Katya Sytkina, and Gordey Yastrebov.

# References


Allanson, P.F. (2011). On the characterization and economic evaluation of income mobility as a process of distributional change. *The Journal of Economic Inequality* 10(4): 505-28.

Büchel, F., and Frick, J.R. (2004). Immigrants in the UK and in West Germany ?Relative income position, income portfolio, and redistribution effects. *Population Economics* 17(3).

Buck, N., and McFall, S. (2012). Understanding Society: design overview. *Longitudinal and Life Course Studies* 3: 5-17.

Burkhauser, R.V., Butrica, B.A., Daly, M.C., and Lillard, D.R. (2001). The Cross-National Equivalent File: A product of cross-national research. in *Social Insurance in a Dynamic Society*, edited by Becker, I., Ott, N., and Rolf, G. Frankfurt: Campus Fachbuch.

Chen, W.-H. (2009). Cross-National Differences in Income Mobility: Evidence from Canada, the United States, Great Britain and Germany. *Review of Income and Wealth* 55(1): 75-100.

Cho, J., and Lee, A. (2013). Life Satisfaction of the Aged in the Retirement Process: A Comparative Study of South Korea with Germany and Switzerland. *Applied Research in Quality of Life* 9(2): 179-95.

Cooke, T.J., Boyle, P., Couch, K., and Feijten, P. (2009). A longitudinal analysis of family migration and the gender gap in earnings in the united states and great britain. *Demography* 46(1): 147–67.

DiPrete, T.A., and McManus, P. (1996). Institutions, Technical Change, and Diverging Life Chances: Earnings Mobility in the United States and Germany. *American Journal of Sociology* 102(1): 34-79.

Dubrow, J.K., and Tomescu-Dubrow, I. (2016). The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field. *Quality & Quantity* 50(4): 1449-67.

Ehlert, M. (2013). Job loss among rich and poor in the United States and Germany: Who loses more income? *Research in Social Stratification and Mobility* 32: 85-103.

Frick, J.R., Jenkins, S.P., Lillard, D.R., Lipps, O., and Wooden, M. (2007). The Cross-National Equivalent File (CNEF) and Its Member Country Household Panel Studies. *Schmollers Jahrbuch : Zeitschrift für Wirtschafts- und Sozialwissenschaften* 127: 627-54.

Gerry, C.J., and Papadopoulos, G. (2015). Sample attrition in the RLMS, 2001-10. *Economics of Transition* 23(2): 425-68.

Giesselmann, M., Bohmann, S., Goebel, J., Krause, P., Liebau, E., Richter, D., . . . Liebig, S. (2019). The Individual in Context(s): Research Potentials of the Socio-Economic Panel Study (SOEP) in Sociology. *European Sociological Review* 35(5): 738-55.

Goebel, J., Grabka, M.M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics* 239(2): 345-60.

Johnson, D., McGonagle, K., Freedman, V., and Sastry, N. (2018). Fifty Years of the Panel Study of Income Dynamics: Past, Present, and Future. *Ann Am Acad Pol Soc Sci* 680(1): 9-28.

Kaminska, O., and Lynn, P. (2017). Survey-Based Cross-Country Comparisons Where Countries Vary in Sample Design: Issues and Solutions. *Journal of Official Statistics* 33(1): 123-36.

Kozyreva, P., and Sabirianova Peter, K. (2015). Economic change in Russia: Twenty years of the Russian Longitudinal Monitoring Survey. *Economics of Transition* 23(2): 293-98.

McCall, L., and Percheski, C. (2010). Income Inequality: New Trends and Research Directions. *Annual Review of Sociology* 36(1): 329-47.

McGonagle, K.A., Schoeni, R.F., Sastry, N., and Freedman, V.A. (2012). The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research. *Longitudinal and Life Course Studies* 3(2): 268 - 84.

McManus, P.A. (2003). Parents, Partners, and Credentials: Self-Employment Mobility in the United States and Germany. Pp. 171-200 in *Inequality Across Societies: Familes, Schools and Persisting Stratification*.

Musick, K., Bea, M.D., and Gonalons-Pons, P. (2020). His and Her Earnings Following Parenthood in the United States, Germany, and the United Kingdom. *American Sociological Review* 85(4): 639-74.

Platt, L., Knies, G., Luthra, R., Nandi, A., and Benzeval, M. (2020). Understanding Society at 10 Years. *European Sociological Review*.

Revilla, M.A., Saris, W.E., and Krosnick, J.A. (2014). Choosing the Number of Categories in Agree-Disagree Scales. *Sociological Methods & Research* 43(1): 73-97.

Rose, D. (1995). Household panel studies: An overview. *Innovation: The European Journal of Social Science Research* 8(1): 7-24.

Siegers, R., Belcheva, V., and Silbermann, T. (2020). *SOEPcore v35 - documentation of sample sizes and panel attrition in the German Socio-Economic Panel (SOEP) (1984 until 2018)*: DIW/SOEP: SOEP Survey Papers, 826.

Slomczynski, K.M., and Tomescu-Dubrow, I. (2019). Basic Principles of Survey Data Recycling. in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC),*, edited by Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., and Dorer, B.: John Wiley & Sons.

Turek, K. (2023). Accelerating Social Science Knowledge Production with the Coordinated Open-Source Model. *OSF Preprint. doi:10.31219/osf.io/3brjv*.

Turek, K., Kalmijn, M., and Leopold, T. (2021). The Comparative Panel File: Harmonized Household Panel Surveys from Seven Countries. *European Sociological Review* 37(3): 505-23.

Watson, N., and Wooden, M. (2020). The Household, Income and Labour Dynamics in Australia (HILDA) Survey. *Journal of Economics and Statistics* Published online.

Wolf, C., Joye, D., Smith, T., and Fu, Y.-c. (2017). Harmonizing Survey Questions Between Cultures and Over Time. in *The SAGE Handbook of Survey Methodology*, edited by Wolf, C., Fu, Y.-c., Joye, D., and Smith, T. London: SAGE Publications.