THE USE OF R
IN OFFICIAL STATISTICS

12th INTERNATIONAL CONFERENCE

Estimation of population parameters in EL-SILC with the R packages: the experience of ELSTAT

Irene Sarantou, Anastasia Mnimatidou

uRos2024, 27-29 November 2024

HELLENIC STATISTICAL AUTHORITY

Key Points Covered in the Presentation

- Overview of the survey scope and methodology
  (compliance with the new stringent requirements of the 2019/1700 IESS Regulation)
- Contribution to bridge a gap in estimation of population parameters using entirely R software
- Key benefits
- The significant achievement of Transition from the SAS macro “Calmar” to the R package “icarus”

Overview of the survey scope and methodology

1. **Survey Methodology**: SILC is an annual household survey targeting all private households in Greece, using a stratified two-stage area sampling design with rotating panels. The sample consists of four subsamples, each used for four years.

2. **Sampling Design:**

   Stratification: stratified by region + degree of urbanization

   Sampling Frame: Based on Census, with PSUs selected proportionally to their size.

   Rotating Sample Structure: Four panels rotate annually, ensuring a 75% overlap between consecutive years.

   Sample Size and Allocation: ~15,300 households, with 3,820 households per panel.

Overview of the survey scope and methodology
Cross and Long scheme

Stages of Sampling:

- **First Stage**: Random selection of PSUs within each stratum.
- **Second Stage**: Systematic sampling of households within selected PSUs.

Weighting:

- **Design Weights**: Inverse of the probability of selection.
- **Adjustments**: For nonresponse, attrition, and combining panels to improve estimate precision.
- **Calibration**: Adjusting weights to match known population totals for certain variables.

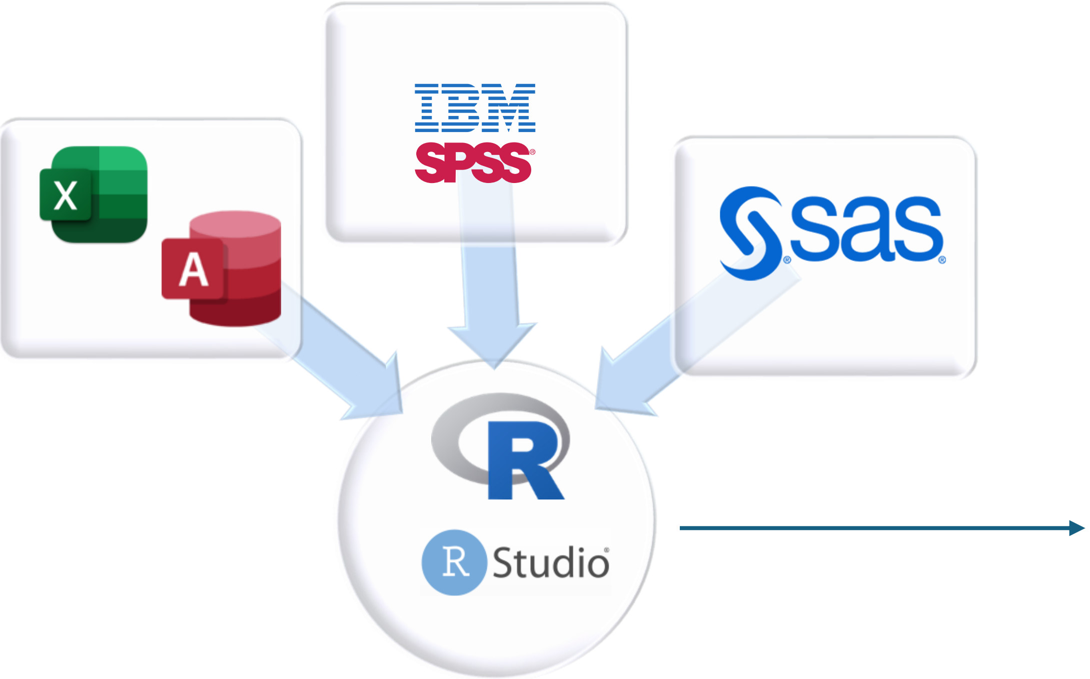
Survey year				
2020	1			
2021	1	2		
2022	1	2	3	
2023	1	2	3	4

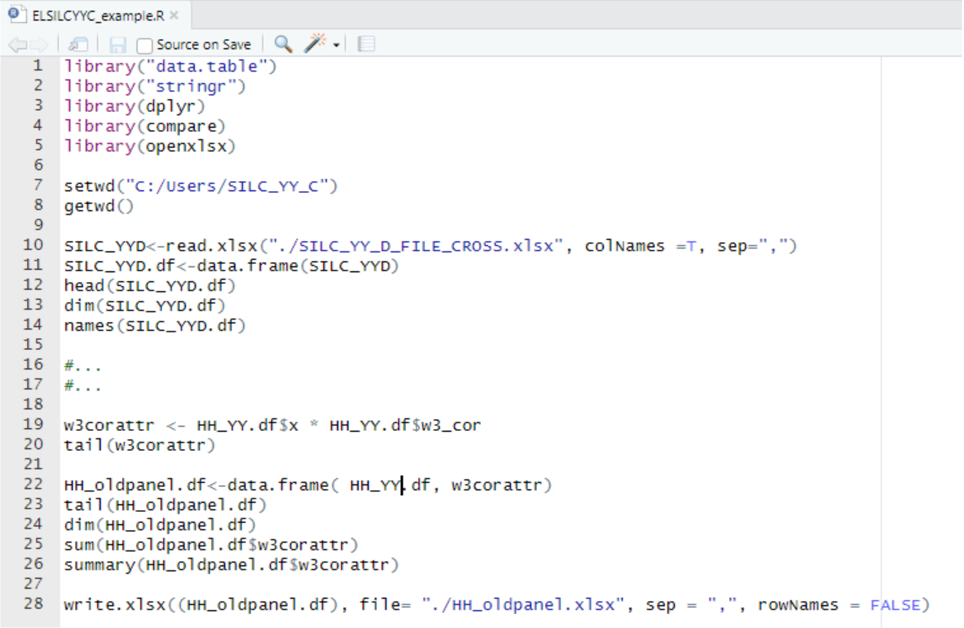
□ Cross –sectional weights (1,2,3,4 panels)

▨ Longitudinal weights (panels 3,2,1 of duration two, three and four years respectively)

⚠ Transmission of the cross data collection of year N, 11 months earlier .

Contribution to bridge a gap in estimation of population parameters using entirely R software





```r
library("data.table")
library("stringr")
library(dplyr)
library(compare)
library(openxlsx)

setwd("C:/Users/SILC_YY_C")
getwd()

SILC_YYD<-read.xlsx("./SILC_YY_D_FILE_CROSS.xlsx", colNames =T, sep=",")
SILC_YYD.df<-data.frame(SILC_YYD)
head(SILC_YYD.df)
dim(SILC_YYD.df)
names(SILC_YYD.df)

#...
#...

w3corattr <- HH_YY.df$x * HH_YY.df$w3_cor
tail(w3corattr)

HH_oldpanel.df<-data.frame( HH_YY.df, w3corattr)
tail(HH_oldpanel.df)
dim(HH_oldpanel.df)
sum(HH_oldpanel.df$w3corattr)
summary(HH_oldpanel.df$w3corattr)

write.xlsx((HH_oldpanel.df), file= "./HH_oldpanel.xlsx", sep = ",", rowNames = FALSE)
```

KEY BENEFITS:

- flexibility in reading, manipulating, processing datasets and writing data,
- availability of recent statistical methodology,
- a particularly economical solution (Matthias Templ & Valentin Todorov, 2016),
- a significant reduction in the time needed to estimate and disseminate the national results,
- in the case of a correction, when all the work is done, the corrected file is loaded, and the code is re-run,
- allows communication to all users.

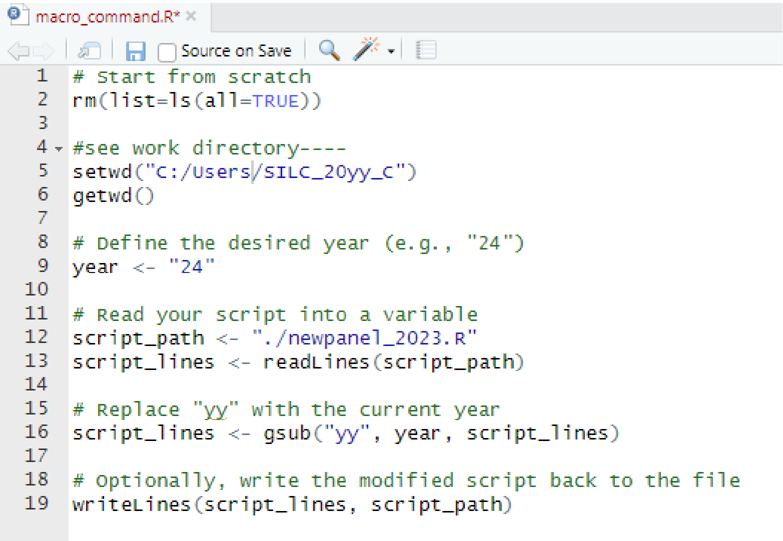
Contribution to bridge a gap in estimation of population parameters using entirely R software(2)

Main packages:

data.table; stringr; dplyr; laeken; tidyverse; foreign; sampling; haven; compare; openxlsx

Advantage of using macro scripts / functions in R.

The main target is increasing efficiency, thus saving time and effort by automating repetitive tasks.



```
# Start from scratch
rm(list=ls(all=TRUE))

#see work directory----
setwd("C:/Users/SILC_20yy_C")
getwd()

# Define the desired year (e.g., "24")
year <- "24"

# Read your script into a variable
script_path <- "./newpanel_2023.R"
script_lines <- readLines(script_path)

# Replace "yy" with the current year
script_lines <- gsub("yy", year, script_lines)

# Optionally, write the modified script back to the file
writeLines(script_lines, script_path)
```

Future Challenges:

Develop an overarching process that includes all the stages of the survey using R eg. Design, Sample, Collect, Analyse, Disseminate and Evaluate.

THE USE OF R
IN OFFICIAL STATISTICS

12th INTERNATIONAL CONFERENCE

uRos 2024 ATHENS urosconf.org

CALIBRATION

Part I

HELLENIC STATISTICAL AUTHORITY

Hellenic Statistical Authority

@StatisticsGR

ELSTAT

statistics.gr

“Integrative” Calibration

- EL-SILC follows Eurostat’s recommendation using an ‘**integrative’** calibration.
- The calibration variables are defined at both household and individual level and the process ensures that the survey estimates are more accurate and reflective of the actual population characteristics.
- The calibration is done at the household level using the household variables (hh size, tenure status, Region - NUTS II) and the individual variables (distribution of population by five-year age group and sex in their aggregate form).

Advantages

- This technique ensures «consistency» between household and individual estimates, by making the household and the individual weights equal.
- Both household and individual information is taken into account in a single calibration.

Successful switch from SAS CALMAR to R icarus

It permits the calibration of a sample by re-weighting the units, using auxiliary information from external sources (eg Estimated Population on the 1st of January for each reference year, Population-Housing Census).

Aim: Increase the precision of the estimates.

EL- SILC uses the « LOGIT » method (M=3)

Advantages:

The ratio between the new weights and the former weights are bounded (by L and U)

The calibrated weights always take positive values

References

- [1] EUROSTAT. (2023). Methodological Guidelines and Description of EU-SILC target variables, 2023 operation (Version 6: Draft), Eurostat.

- [2] Merkouris, P. (2018). Study of the current sampling design of the Survey of Income and Living Conditions with the objective to increase/adjust the sample at regional (NUTSII) level (Part I & II), Athens: AUEB Research Center.

- [3] Merkouris, T. (2001). Cross-Sectional Estimation in Multiple Panel Household Surveys, Canada: Statistics Canada, Vol. 27, No. 2, pp. 171-181.

- [4] Templ, M., & Todorov, V. (2016). The Software Environment R for Official Statistics, Austrian Journal of Statistics(45), 97–124.

- [5] Verma, V., Betti, G., & Ghellini, G. (2017). Cross-sectional and longitudinal weighting in a rotational household panel: Applications to EU-SILC, Statistics in transition-new series, 8(1), 5-50.

Acknowledgments

- Sincere thanks are due to Dr. Aik. K. Gkolia, University of Nicosia, for her valuable contribution to the project throughout the migration to R.

uRos 2024 ATHENS urosconf.org

THE USE OF R
IN OFFICIAL STATISTICS

12th INTERNATIONAL CONFERENCE

CALIBRATION

Part II

The significant achievement of transition from the SAS macro “Calmar” to the R package “icarus”

Version 0.3.2 May 27, 2023

HELLENIC STATISTICAL AUTHORITY

Hellenic Statistical Authority

@StatisticsGR

ELSTAT

statistics.gr

Implementation of ‘icarus’ in EL-SILC

The second part of the presentation, which deals with the implementation of the "icarus" R package and the comparison of the results with the corresponding results of the SAS Calmar software, is illustrated in the following pdf file.

uRos-2024_Mnimatidou.pdf

CONCLUSIONS

The successful and exclusive use of R software:

✓ **marks** a major step forward in bridging gaps in the estimation of population parameters for the EL SILC survey,

✓ **offers** significant advantages in terms of both functionality and time savings,

✓ **provides** an up-to-date and effective framework for handling complex sampling surveys,

✓ **represents** a noticeable reduction in the time needed to estimate and disseminate national results,

✓ becomes more efficient,

✓ **facilitates** better communication with users,

✓ **making** the system more economical and user-friendly.