Public Use Files of EU-SILC and EU-LFS data

Peter-Paul de Wolf

Statistics Netherlands

Introduction

Eurostat provides access to EU microdata:

- Secure Use Files
- Scientific Use Files

Getting access takes time (up to 10 weeks...)

It it worth the effort?

Perhaps a PUF could help?

Introduction

Eurostat provides access to EU microdata:

- Secure Use Files
- Scientific Use Files

Getting access takes time (up to 10 weeks...)

It it worth the effort?

Perhaps a PUF could help?

Specific Grant Agreement launched to produce PUFs

Introduction

Why Public Use Files?

- Aid in decision on effort
- Start with research
- Training file?

Introduction

Why Public Use Files?

- Aid in decision on effort
- Start with research
- Training file?

Approach EU-SILC

EU-SILC = EU Statistics on Income and Living Conditions

Cross-sectional as well as longitudinal sample survey

Sensitive variable income in PUF?
Able to recontruct households in PUF?
Many member states: NO!

Synthetic data? ‘Fake’ data?

Approach EU-SILC

EU-SILC = EU Statistics on Income and Living Conditions

Cross-sectional as well as longitudinal sample survey

Sensitive variable income in PUF?
Able to recontruct households in PUF?
Many member states: NO!

Synthetic data? 'Fake' data?

Fully synthetic data

Approach EU-SILC

EU-SILC = EU Statistics on Income and Living Conditions

Cross-sectional as well as longitudinal sample survey

Sensitive variable income in PUF?
Able to recontruct households in PUF?
Many member states: NO!

Synthetic data? 'Fake' data?

Fully synthetic data

Only cross-sectional data

Approach EU-SILC

General idea:

- Estimate models from original data
- Create synthetic population using these models
- Draw a sample of the size of the original data

Approach EU-SILC

General idea:

- Estimate models from original data
- Create synthetic population using these models
- Draw a sample of the size of the original data

Per regional stratum:

- Setup household structure
- Simulate categorical variables
- Simulate (semi) continuous variables
- Split (semi) continuous variables into components

Approach EU-SILC

Setup hh-structure:

- Estimate number of hh by hh-size (HT-estimate)
- Generate that number of hh to construct the population
- For each hh of size *h*, using resampling, draw hh-structure from hh of size *h* in original data

To prevent illogical hh-structures (age/sex distribution)

Approach EU-SILC

Simulation categorical variables

- Sequentially; conditionally on previously simulated variables
- Multinomial logistic regression fitted on original data with previously generated variables as predictors
- Variables: economic status, citizenship, marital status, education, occupation (1 digit, second drawn randomly), NACE (1 digit)

Approach EU-SILC

Simulation (semi) continuous variables

- Mapped to discretized version (e.g. income classes)
- Apply method like with categorical variables
- Draw randomly within category/class to obtain continuous value

Approach EU-SILC

Split into components

- Use *proportions* of donor record
- Independently for hh income and person income

Construct sample

- Stratified simple random sampling with replacement
- Stratum is region
- Sampling unit is hh

Approach EU-SILC

Practical issues:

- Sparseness of variables $\Longrightarrow$ no stratification
- Population size $\Longrightarrow$ generate smaller population
- Too many variables $\Longrightarrow$ generate some variables unconditionally from (weighted) distribution in original data
- R-package simPop and some additional R-scripts

Approach EU-LFS

EU-LFS = EU Labour Force Survey

Cross-sectional and longitudinal (4Q + Y + rotating panel)

Start with 4Q files and construct Y file from these

Approach EU-LFS

General approach (starting point: SUF)

- Remove some variables (globally set to ‘missing’)
- Global recoding
- Local suppression
  - based on *k*-anonymity on specific subset of all identifying variables, PRAM on remaining variables
  - based on all-*m* approach

Approach EU-LFS

Removing variables

- Variables that could reconstruct households
- Region
- Some complexly related variables
- To keep format/structure of corresponding SUF, all scores set to Missing

Resulted in 13 identifying variables remaining
(12 in Q-files, one additional in Y-file)

Approach EU-LFS

Global recoding

- Age into 6 classes
- Nationality into 3 classes
- Country of birth into 3 classes
- Occupation into 1 digit
- Years of residence into 3 classes
- Level of education into 3 classes
- Professional status one less category
- Country of work into 4 classes
- Degree of urbanisation one less category
- NACE into 7 classes

Approach EU-LFS

Local suppression

- using k-anonymity on key of 7 variables (Degree of Urbanisation, Sex, Age, Nationality, ILO working status, Years of residence, Highest level education) with $k = 5$
- using all-m approach with $m = 4$ and threshold 10

Disclosure risk

- Synthetic data
  - Fully synthetic data $\Longrightarrow$ 'Fake' data $\Longrightarrow$ safe data?
  - AMELI project considered several disclosure scenario's (linkage)
  - Unique, large households may be found …

Disclosure risk

- Synthetic data
  - Fully synthetic data $\Longrightarrow$ 'Fake' data $\Longrightarrow$ safe data?
  - AMELI project considered several disclosure scenario's (linkage)
  - Unique, large households may be found …
    but with synthetic income, etc.

Disclosure risk

- Synthetic data
  - Fully synthetic data $\Longrightarrow$ 'Fake' data $\Longrightarrow$ safe data?
  - AMELI project considered several disclosure scenario's (linkage)
  - Unique, large households may be found …
    but with synthetic income, etc.

- Traditional approach
  - Limited *k*-anonymity (7 out of 12 variables)
  - All-*m* approach
  - Suppress Age, Sex and ILOSTAT with low priority
  - Count uniques on full *k*-anonymity

Disclosure risk

First preliminary results:

- all-*m* approach may lead to many more suppressions compared to *k*-anonymity
- many more uniques under 13 identifying variables with *k*-anonymity compared to all-*m* approach

NB:

- under all-*m* approach usually multiple suppressions per record
- application of PRAM influences number of uniques with *k*-anonymity

Utility

Relative error:

$$\frac{\text{Value}(\text{Indicator} \in \text{PUF}) - \text{Value}(\text{Indicator} \in \text{SUF})}{\text{Value}(\text{Indicator} \in \text{SUF})} \times 100\%$$

Confidence interval overlap:

$$\log(\text{equivalenced disposable income}) \sim$$

$$\text{age} + \text{gender} + \text{education} + \text{citizenship} + \text{hhsize}$$

Utility

First results (all-*m* approach, Finland):
Unemployment rate (ILOSTAT=2, 15-74 years old)
Relative difference in precentages

		Q1	Q2	Q3	Q4
Total	Total	-0.28	-0.19	-0.56	-0.67
Sex	Male	-0.11	-0.23	-0.16	-0.49
	Female	-0.43	-0.08	-1.02	-0.81
Age	15-24	0.42	-0.03	0.58	0.47
	25-54	-0.40	-0.12	-0.60	-1.03
	55-74	0.17	0.22	-1.08	0.06
HATLEV1D	H	-7.88	-9.58	-9.84	-7.71
	M	0.37	0.25	-1.43	-1.76
	L	2.62	6.72	5.96	2.72

Utility

First results (*k*-anonymity approach, Slovenia):
Unemployment rate (ILOSTAT=2, 15-74 years old)
Relative difference in precentages

		Q1	Q2	Q3	Q4
Total	Total	-0.13	-1.02	-0.93	-0.23
Sex	Male	-0.27	-0.76	-0.17	-0.28
	Female	0.01	-1.29	-1.60	-0.18
Age	15-24	0.12	0.17	0.83	-3.58
	25-54	0.07	-1.21	0.93	-0.01
	55-74	-1.10	-9.09	-2.46	-6.80
HATLEV1D	H	-12.94	-8.21	-13.01	-8.40
	M	-0.14	-1.12	-1.36	-1.44
	L	-3.14	-2.11	-5.88	-6.81

Conclusions

- First results look promising
- Need more detailed look at
  - Utility (different measures)
  - Risk (two approaches to same dataset)