Extending Simulation Populations Using Additional Survey Datasets

Florian Ertz and Ralf Thomas Münnich

Economic and Social Statistics Department
Trier University

CELSI Data Forum
Data Pooling: Opportunities and Challenges
11 October 2019

The HFCS

- National central banks of the Euro area use Eurosystem Household Finance and Consumption Survey (HFCS) to collect ex-ante harmonised micro data on private households' wealth, debt, income, and consumption
- First wave (2010 as main reference year):
  - 15 member states
  - 62,521 households
  - 154,247 individuals
- Oversampling of (likely) wealthy households in some surveys
  $\rightarrow$ Influence on econometric estimates?
- Problem:
  No *real* design variables or regional identifiers in scientific use file (SUF) provided by European Central Bank (ECB)

AMELIA

- AMELIA is close-to-reality synthetic simulation population, constructed within the project Advanced Methodology for European Laeken Indicators (AMELI)
- www.amelia.uni-trier.de
- Characteristics:
  - 30 states as basis
  - 3,781,289 households
  - 10,012,600 individuals
  - 4 large *multi-state* regions
  - Detailed spatial structure
- Since 2017 extension within research infrastructure InGRID-2
- Problem:
  No wealth variables in data set

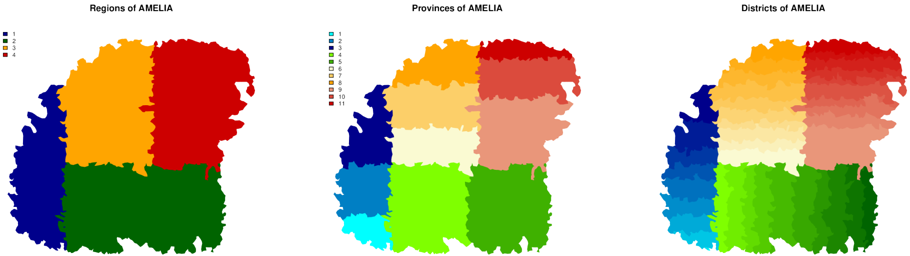
Spatial structure of AMELIA

- Original structure:

REG	Regions	4	NUTS 1
PROV	Provinces	11	NUTS 2
DIS	Districts	40	NUTS 3
CIT	Cities/communities/municipalities	1,592	LAU 1

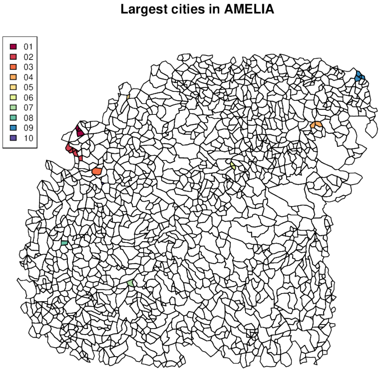
- Large cities and metropolitan areas:
  - Use of variable on degree of urbanisation
  - Definition of 10 large cities, including 2 metropolitan areas
  - Useful for implementation of HFCS survey designs
  - Scenario variables to mimic urban inequality patterns

Spatial structure of AMELIA (ctd.)



Source: Ertz (2020).

Large cities and metropolitan areas in AMELIA

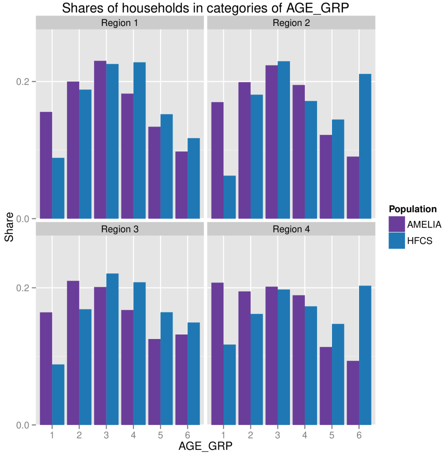


Source: Ertz (2020).

Synthesis of AMELIA and HFCS

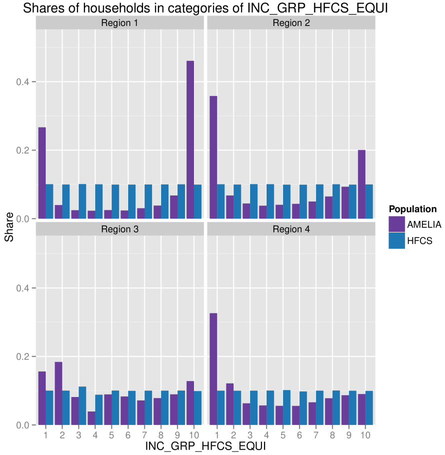
- Advantages of both datasets should be exploitable in design-based Monte Carlo simulation studies
- Aim is generation of synthetic wealth variables on household level in AMELIA using the HFCS SUF (6 broad asset classes like in *Arrondel et al., 2016*)
- EU Statistics on Income and Living Conditions (EU-SILC) is data source (AMELIA) and *template* (HFCS), respectively
- Modifications/recodings in both datasets
  $\Rightarrow$ 13 variables in *intersection* of datasets
- Synthesis of (survey) data sources for the generation/extension of simulation populations as an interesting methodological problem
- Highly relevant for microsimulations, where very detailed simulation populations are needed

Age groups - AMELIA a. HFCS



Data sources: HFCS (2017) and Trier University (2017).

Income groups - AMELIA a. HFCS



Data sources: HFCS (2017) and Trier University (2017).

Approaches taken in AMELI

- *Alfons et al. (2011)* suggest three approaches to generate (semi-)continuous synthetic micro data, where regression models are first estimated using survey data and then used for predictions within the simulation population:
  1. Multinomial logistic regression model and draws from value intervals or certain distributions
  2. Sequence of logistic and linear regression model and draws from residuals
  3. Generation of aggregate variable and draws from shares of subcategories of aggregated variable
- Approach 1 and combination of approaches 1 and 3 clearly dominated in our case

$\Rightarrow$ Approach 2 with certain modifications and extensions

Two-step regression model

Each variable is separately generated in every region. The models are estimated on HFCS data using final household weights.

1. Estimation of logit models for participation in asset class (j^{th} variable) using up to $j-1$ *common* exogenous variables
2. Prediction of conditional probabilities in AMELIA:

$$\widehat{p}_{ij}^{P} = \frac{\exp\left(\widehat{\beta}_0^{\text{Logit}} + \widehat{\beta}_1^{\text{Logit}} x_{i1}^{P} + \ldots + \widehat{\beta}_{j-1}^{\text{Logit}} x_{i,j-1}^{P}\right)}{1 + \exp\left(\widehat{\beta}_0^{\text{Logit}} + \widehat{\beta}_1^{\text{Logit}} x_{i1}^{P} + \ldots + \widehat{\beta}_{j-1}^{\text{Logit}} x_{i,j-1}^{P}\right)}.$$

Two-step regression model (ctd.)

3. Drawing of participation dummys in AMELIA
4. Estimation of linear models for amounts held in asset class within subgroup of participating households using up to $j-1$ common exogenous variables
5. Prediction of amounts in AMELIA:

$$\widehat{x}_{ij}^{P} = \widehat{\beta}_0^{\text{OLS}} + \widehat{\beta}_1^{\text{OLS}} x_{i1}^{P} + \ldots + \widehat{\beta}_{j-1}^{\text{OLS}} x_{i,j-1}^{P} + e_i^{P} \,.$$

- Error term e_i^P prevents identical amounts for all *identical* households
- Error term e_i^P drawn from HFCS residuals

Two problems in this application

1. Multiple imputation (MI)
   - Potentially large non-response in wealth surveys
   - HFCS uses MI with 5 implicates
   - Finding *good* models not trivial (cf. *Wood et al., 2008*)

   $\Rightarrow$ Model selection has to be modified
2. Differences in datasets
   - AMELIA and HFCS *only* similar
   - AMELIA uses more states (30 vs. 15)
   - AMELIA is *older* (2005 vs. 2010)

   $\Rightarrow$ Drawing of error terms has to be modified

Model averaging after multiple imputation (MAMI)

Schomaker and Heumann (2014) estimate $\boldsymbol{\theta}$ as follows:

1. Step 1 - MA

   Computation of point estimates *averaged* over K models

   $$\widehat{\overline{\boldsymbol{\theta}}} = \sum_{k=1}^{K} w_k \cdot \widehat{\boldsymbol{\theta}}_k \,,$$

   using, e.g., the following weights (see *Buckland et al., 1997*):

   $$w_k = \frac{\exp(-0.5 \cdot \mathrm{AIC}_k)}{\sum_{k=1}^{K} \exp(-0.5 \cdot \mathrm{AIC}_k)} \,.$$

2. Step 2 - MI

   Computation of MI point estimates over D implicates using

   $$\widehat{\boldsymbol{\theta}}_{\mathrm{MI}} = \frac{1}{D} \sum_{d=1}^{D} \widehat{\boldsymbol{\theta}}^{(d)} \,.$$

Model averaging after multiple imputation (MAMI) (ctd.)

3. Step 3 - Combination
   - Determination of model averaged point estimates for each implicate
   - Combination using Rubin's combination rule:

$$\widehat{\widehat{\boldsymbol{\theta}}}_{\mathrm{MI}} = \frac{1}{D}\sum_{d=1}^{D}\widehat{\widehat{\boldsymbol{\theta}}}^{(d)} \quad \text{with} \quad \widehat{\widehat{\boldsymbol{\theta}}}^{(d)} = \sum_{k=1}^{K} w_k^{(d)} \cdot \widehat{\boldsymbol{\theta}}_k^{(d)} .$$

Implementation in our application:

- Use of `dredge` from R package `MuMIn` (see *Barton, 2016*)
- Selection of K candidate models (including transformations): Top percentile (AIC) → Cross-validation → Top quartile
- Generation in descending order of empirical participation rates
- Use of previously generated variables in candidate models

Grouping and drawing of HFCS residuals

Drawing of error terms within grid cells:

1. Prediction in **HFCS** using MAMI coefficients
2. Grouping of residuals based on binary/categorical variables with largest average relative importance
   $\rightarrow$ Grids built with combinations of up to four variables
3. Addition of sampled residuals to predictions within cells
4. Sorting of grids in descending order of fit
5. Prediction in **AMELIA** using MAMI coefficients
6. Drawing from HFCS residuals pooled across implicates in cells
7. Addition of HFCS residuals to predictions in *smallest* cell
8. Editing

Example: Grid built using 3 variables

X = 1

	Z = 1	Z = 2
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	Z = 2
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	Z = 2
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	Z = 2
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

Example: Grid built using 3 variables

X = 1

	Z = 1	Z = 2
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	Z = 2
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	Z = 2
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	Z = 2
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

→ *Smallest* cell reached (combination of 3 variables)

Example: Grid built using 3 variables

X = 1

	Z = 1	Z = 2
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	Z = 2
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	Z = 2
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	Z = 2
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

Example: Grid built using 3 variables

X = 1

	Z = 1	**Z = 2**
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	**Z = 2**
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	**Z = 2**
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	**Z = 2**
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

→ *Medium* cell reached (combination of 2 variables)

Example: Grid built using 3 variables

X = 1

	Z = 1	Z = 2
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	Z = 2
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	Z = 2
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	Z = 2
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

Example: Grid built using 3 variables

X = 1

	Z = 1	Z = 2
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	Z = 2
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	Z = 2
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	Z = 2
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

Example: Grid built using 3 variables

X = 1

	Z = 1	Z = 2
Y = 1	286	4
Y = 2	221	121
Y = 3	270	83
Y = 4	102	237

X = 2

	Z = 1	Z = 2
Y = 1	219	65
Y = 2	71	238
Y = 3	249	234
Y = 4	163	242

X = 3

	Z = 1	Z = 2
Y = 1	20	23
Y = 2	72	103
Y = 3	38	123
Y = 4	67	46

X = 4

	Z = 1	Z = 2
Y = 1	16	3
Y = 2	3	106
Y = 3	19	256
Y = 4	8	147

→ *Largest* cell reached (1 variable)

Alternative: Cell-based generation

- Problems of regression-based approach:
  - Results not satisfying in our application
  - Model selection process computationally intensive
  - Considerable user discretion
- Proposed alternative method: Cell-based generation (CBG)
  - Grids (see slide on HFCS residual grouping) as starting point
  - Use of monotone cubic splines (see *Hyman, 1983*) to replicate distribution functions within grid cells

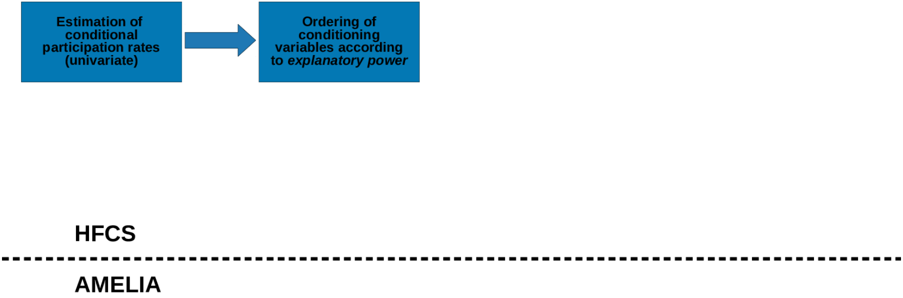
Generation of participation dummys

Estimation of conditional participation rates (univariate)

HFCS

AMELIA

Generation of participation dummys



Generation of participation dummys

Estimation of conditional participation rates (univariate) → **Ordering of conditioning variables according to *explanatory power*** → **Selection of 7 *most important* conditioning variables**

HFCS

AMELIA

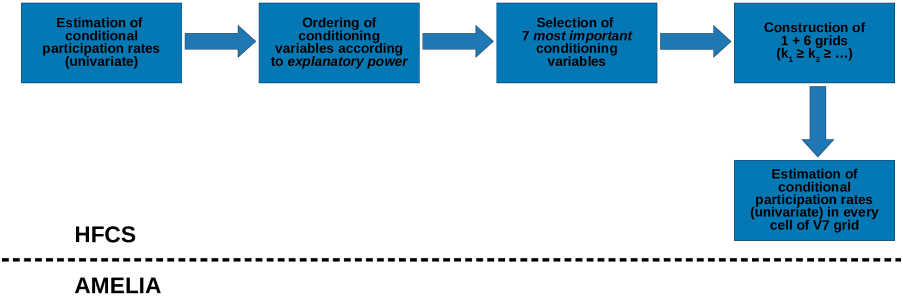
Generation of participation dummys

Estimation of conditional participation rates (univariate) → **Ordering of conditioning variables according to *explanatory power*** → **Selection of 7 *most important* conditioning variables** → **Construction of 1 + 6 grids ($k_1 \geq k_2 \geq \ldots$)**

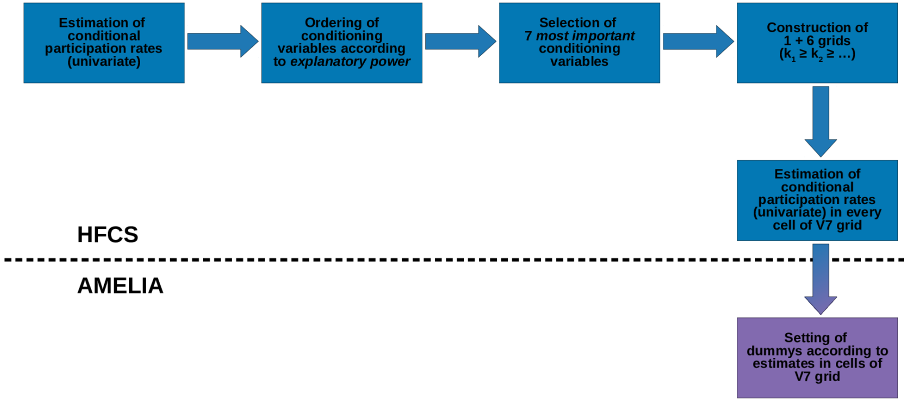
HFCS

AMELIA

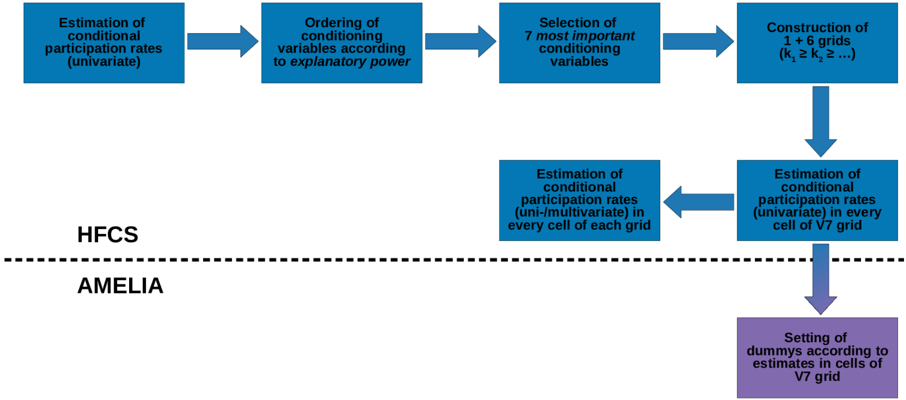
Generation of participation dummys



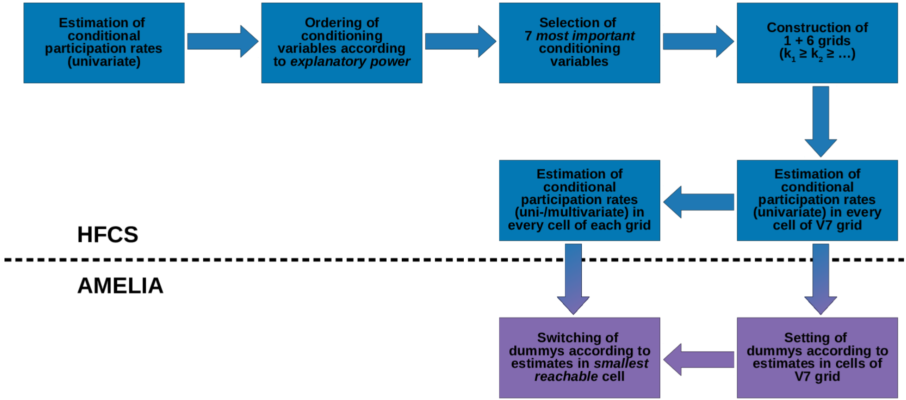
Generation of participation dummys



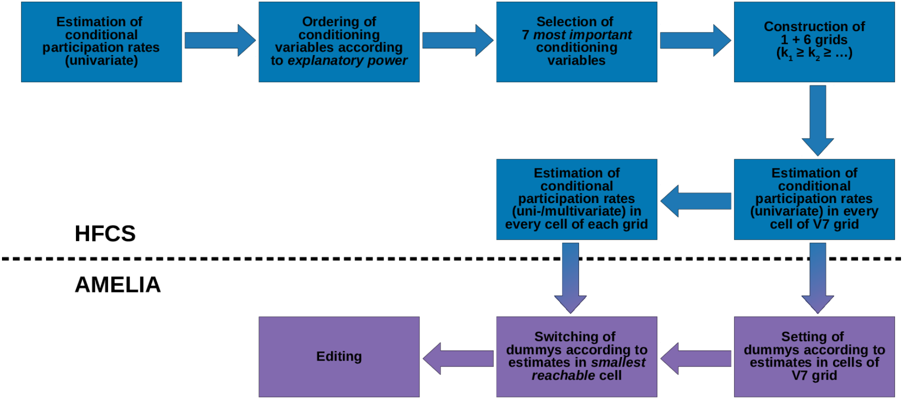
Generation of participation dummys



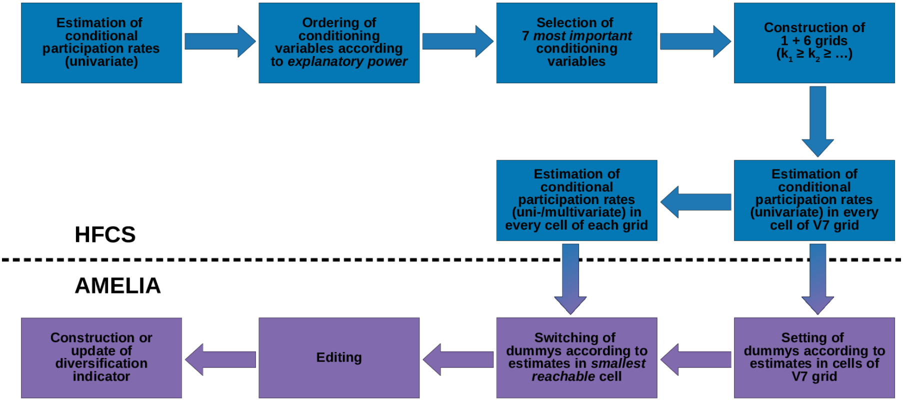
Generation of participation dummys



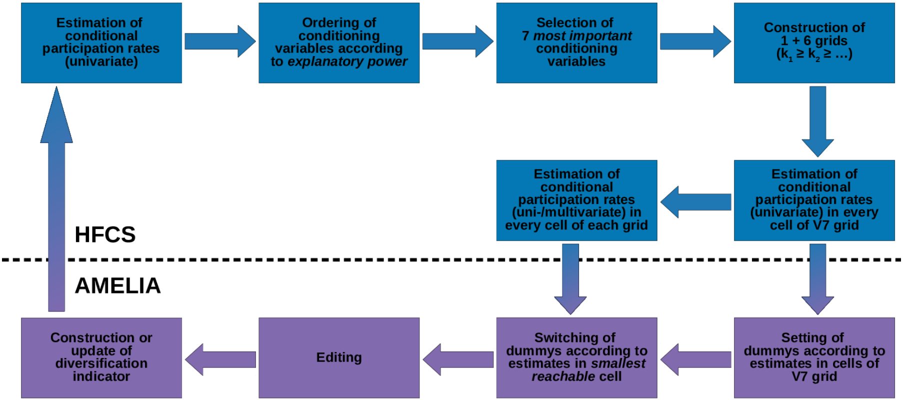
Generation of participation dummys



Generation of participation dummys



Generation of participation dummys



Generation of amounts

Estimation of conditional holding means (univariate)

HFCS

AMELIA

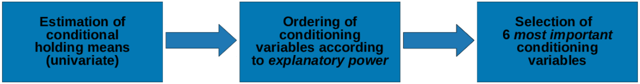
Generation of amounts

Estimation of conditional holding means (univariate) → **Ordering of conditioning variables according to *explanatory power***

HFCS

AMELIA

Generation of amounts



HFCS

AMELIA

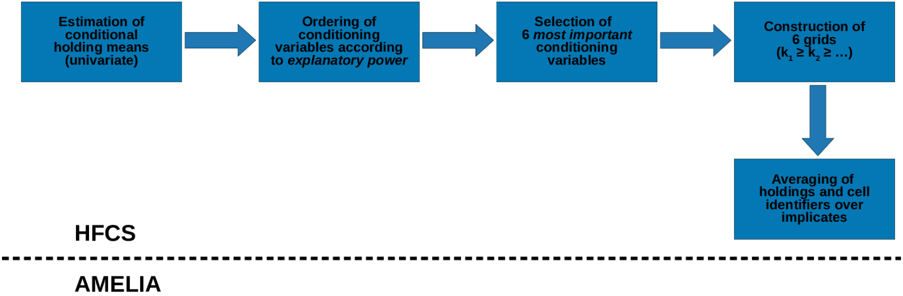
Generation of amounts

Estimation of conditional holding means (univariate) → **Ordering of conditioning variables according to *explanatory power*** → **Selection of 6 *most important* conditioning variables** → **Construction of 6 grids ($k_1 \geq k_2 \geq \ldots$)**

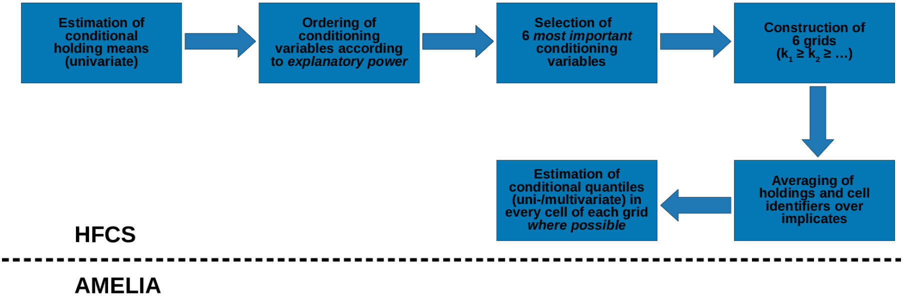
HFCS

AMELIA

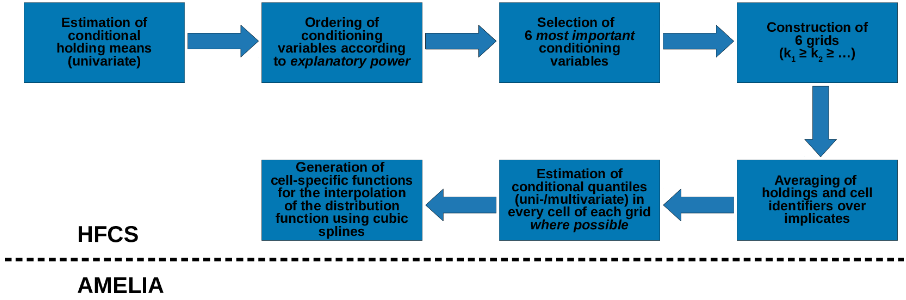
Generation of amounts



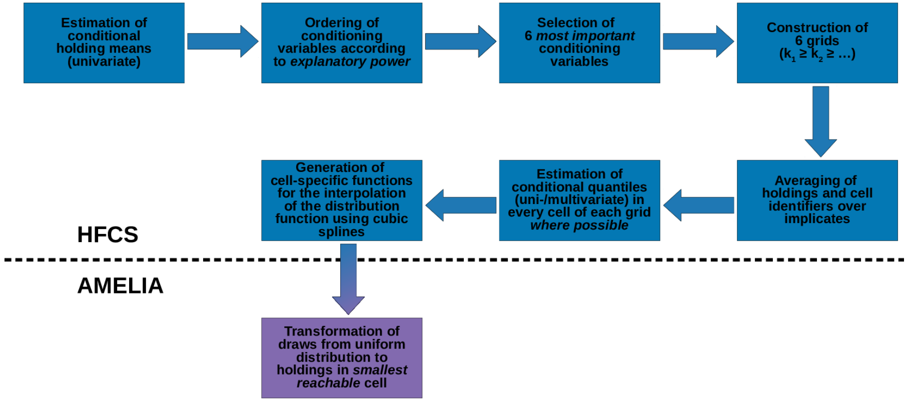
Generation of amounts



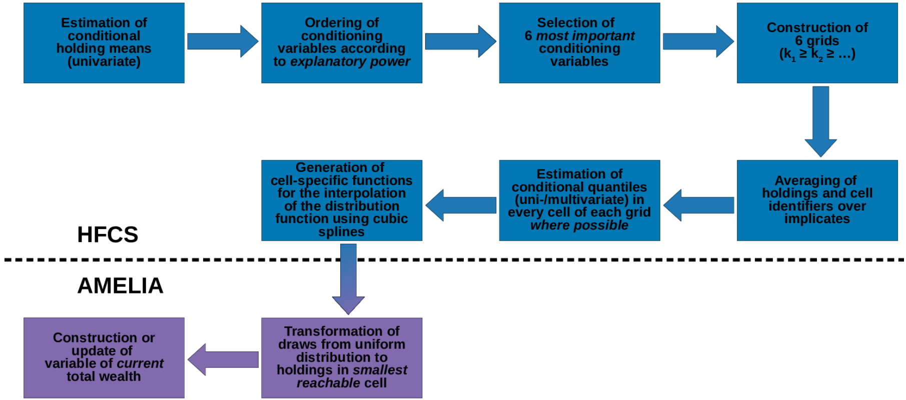
Generation of amounts



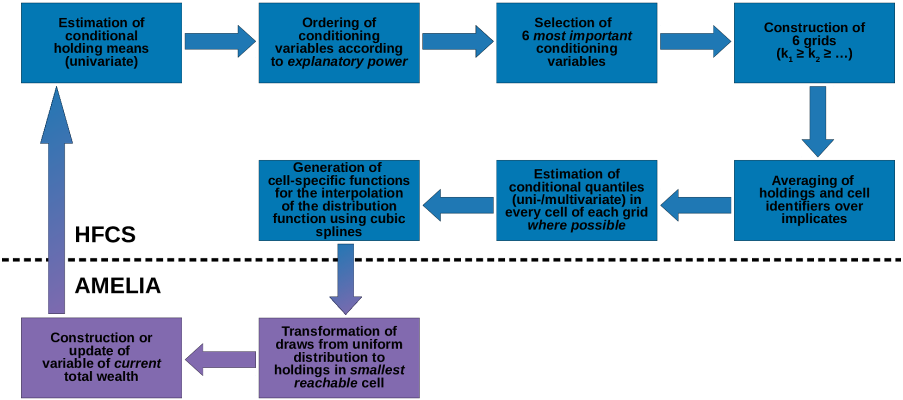
Generation of amounts



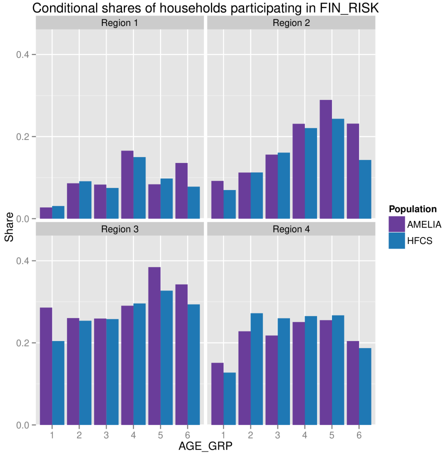
Generation of amounts



Generation of amounts

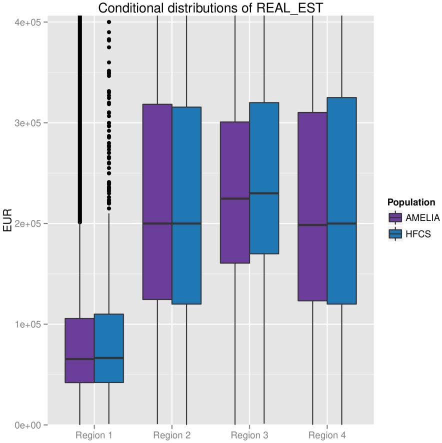


Participation - FIN_RISK a. age - AMELIA a. HFCS



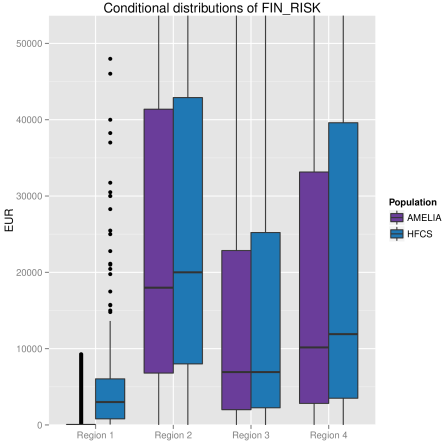
Data sources: HFCS (2017) and Trier University (2017).

Cond. distribution - REAL_EST - AMELIA a. HFCS



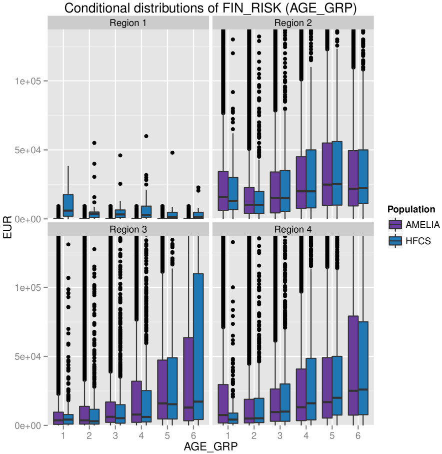
Data sources: HFCS (2017) and Trier University (2017).

Cond. distribution - FIN_RISK - AMELIA a. HFCS



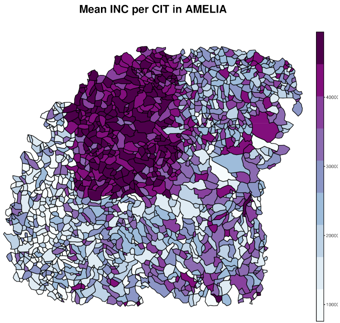
Data sources: HFCS (2017) and Trier University (2017).

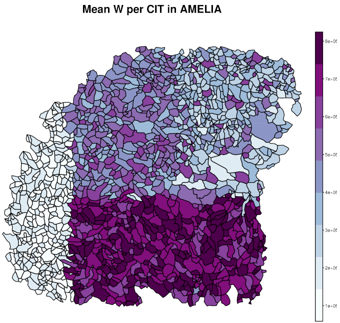
Cond. distribution - FIN_RISK a. age - AMELIA a. HFCS



Data sources: HFCS (2017) and Trier University (2017).

Personal income and wealth in AMELIA





Source: Ertz (2020).

Regression- (R) vs. cell-based (C) approach

	R	C
Participation rates (Unconditional)	0.18	0.11
Holding means (*Unconditional*)	0.14	0.28
Fraction means (Unconditional)	0.22	0.27
Holding quantiles (Unconditional)	20.66	0.32
Diversification indicator shares (Unconditional)	0.37	0.27
Participation rates (Univariate)	0.48	0.30
Holding means (*Univariate*)	0.32	0.36
Fraction means (Univariate)	0.73	0.45
Holding quantiles (Univariate)	14.33	6.84
Diversification indicator shares (Univariate)	1.21	0.80
Participation rates (Multivariate)	19.87	13.58
Holding means (Multivariate)	110.16	20.46
Fraction means (Multivariate)	574.19	387.09
Directional error shares - Participation rates	0.23	0.21
Directional error shares - Holding means	0.25	0.27
Directional error shares - Fraction means	0.34	0.23
Directional error shares - Diversification indicator shares	0.27	0.22

Summary and outlook

Summary:

- Synthesis of data sources for the generation/extension of simulation populations is an interesting methodological problem
- Regression-based approaches reach their limits here
- Cell-based approach yields *better* results
- Relative re-identification risk measures of *Templ and Alfons (2010)* always, mostly considerably so, below 1%

Outlook:

- Use of newer waves of HFCS (potentially pooling of waves)
- Comparison of approaches using other datasets and target variables exhibiting smaller skew (e.g. consumption)
- Possibly contribution of R package

Focussing on Germany - Another look ahead

- New DFG Research Unit FOR 2559: Multi-sectoral regional microsimulation model (MikroSim) started last year
- http://gepris.dfg.de/gepris/projekt/316511172?language=en
- Consortium of Trier University, University of Duisburg-Essen, and the Federal Statistical Office
- Building of large-scale synthetic simulation population using 2011 German Census and many other German micro datasets
- Use of street maps
- Open dynamic microsimulation infrastructure (demographic change and changes in household composition)
- Initial focus on health care and migration
- Wealth information should be integrated as well

This talk is supported by:

- InGRID-2 Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy (European Commission G.A. no. 730998)
- Research Institute for Official and Survey Statistics (RIFOSS)

This talk uses data from the Eurosystem Household Finance and Consumption Survey. The results published and the related observations and analysis may not correspond to results or analysis of the data producers.

InGRID
Supporting expertise in inclusive growth

RIFOSS
Research Institute for Official and Survey Statistics

Thank you for your attention!

References

Alfons, A.; Kraft, S.; Templ, M.; Filzmoser, P. (2011): Simulation of Close-to-Reality Population Data for Household Surveys with Application to EU-SILC. Statistical Methods & Applications 20 (3), S. 383–407.

Arrondel, L.; Bartiloro, L.; Fessler, P.; Lindner, P.; Mathä, T.Y.; Rampazzi, C.; Savignac, F.; Schmidt, T.; Schürz, M.; Vermeulen, P. (2016): How Do Households Allocate Their Assets? Stylized Facts from the Eurosystem Household Finance and Consumption Survey. International Journal of Central Banking 12 (2), S. 129–220.

Bartoń, K. (2016): MuMIn: Multi-Model Inference. R package version 1.15.6.

Buckland, S.; Burnham, K.; Augustin, N. (1997): Model Selection: An Integral Part of Inference. Biometrics 53 (2), S. 603–618.

Ertz, F. (2020): Regression Modelling with Complex Survey Data: An Investigation Using an Extended Close-to-Reality Simulated Household Population. Ph.D. dissertation. Trier University. To be published.

Hyman, J.M. (1983): Accurate Monotonicity Preserving Cubic Interpolation. SIAM Journal on Scientific and Statistical Computing, 4 (4), S. 645–654.

Schomaker, M.; Heumann, C. (2014): Model Selection and Model Averaging after Multiple Imputation. Computational Statistics & Data Analysis 71, S. 758–770.

Templ, M.; Alfons, A. (2010): Disclosure Risk of Synthetic Population Data with Applications in the Case of EU-SILC, in: Domingo-Ferrer, J.; Magkos, E. (Editors): Privacy in Statistical Databases. Lecture Notes in Computer Science Number 6344. S. 174-186. Springer-Verlag Berlin Heidelberg.

Wood, A.M.; White, I.R.; Royston, P. (2008): How Should Variable Selection be Performed with Multiply Imputed Data?. Statistics in Medicine 27, S. 3227-3246.

Examples for measures used to gauge *proximity*

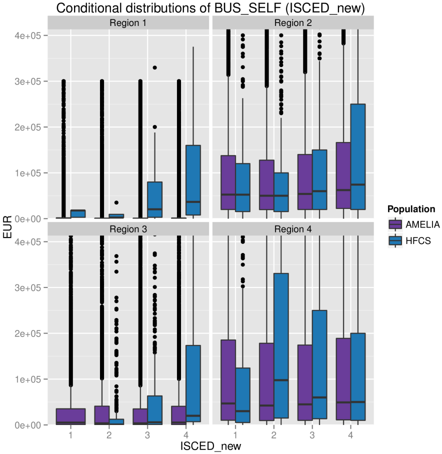
Unconditional participation rates:

$$\bar{\delta}_p = \frac{1}{6}\sum_{j=1}^{6}\frac{1}{4}\sum_{l=1}^{4}\left|\frac{p_{jl}^{\text{AMELIA}}}{\widehat{p}_{jl}^{\text{HFCS}}} - 1\right|$$

Unconditional holding quantiles:

$$\bar{\delta}_q = \frac{1}{6}\sum_{j=1}^{6}\frac{1}{4}\sum_{l=1}^{4}\frac{1}{17}\sum_{m=1}^{17}\left|\frac{q_{jlm}^{\text{AMELIA}}}{\widehat{q}_{jlm}^{\text{HFCS}}} - 1\right|$$

Sparsely-held asset classes and small sample sizes



Data sources: HFCS (2017) and Trier University (2017).