Machine Learning Missing European Household Wealth

work in progress

Johannes Fleck

European University Institute, Florence

SEM Conference, Frankfurt

August 16, 2019

The aim of this project

- In most HFCS countries, all variables are collected in surveys
  - Variables are affected by item-non response Types of Missings
  - Some of the data are not observed but imputed
- This project explores an imputation approach which
  - uses tools from Machine Learning (ML)
  - shares benefits of *quasi-admin data* with *survey-only countries*
  - avoids collecting country-specific (non-harmonized) admin data

Outline

Missing Item Imputation

ML based Imputation for the HFCS

- Methodology
- Example: Value of Household Main Residence

Conclusion

Missing Item Imputation

Methods to impute missing items in surveys

1. Model-based
   - no well-specified model for household wealth decisions
   - imputed data cannot be used to estimate model parameters
2. Algorithmic
   - driven by data and 'theory-free'
   - sensitive to choice of algorithm

Multiple Imputation: gold standard of algorithmic method

- uses several stochastic simulations to impute specific item
- item distributions show imputation uncertainty

→ SCF and HFCS imputation follow this approach

Item Imputation in the SCF and HFCS

- SCF: FRITZ Model
  - contains a highly structured set of constraints:
    Sequential: follow predetermined path through survey variables imputing missing items
    Iterative: imputed values from each previous iteration treated as observed for consecutive iteration
- HFCS:
  - most countries use FRITZ derivatives ('€mir', ...)
  - but differ with respect to data collection
    - 15/20: surveys (true values of missing items unknown)
    - 5/20: 'quasi-admin' data for some variables
      EE, FI, FR, IE (registers); IT (contract)

Item Imputation with Machine Learning

- In some fields, imputing missing items with ML is common
  - medical science: Jerez et al [2010], Masconi et al [2015], ...
  - industrial research: Lakshminarayan et al [1996], ...
- Why ML for imputation?
  - easy comparison of many distinct algorithms ML algorithms
  - allows modeling relationships without priors (theories)
- For survey imputation:
  - Nordbottom [1998], Amer [2006], ...
  - Census Bureau (CPS, ASEC, ACS)
  - **Main challenge: 'True' data not available**
    $\Rightarrow$ cannot train, validate, test

ML based Imputation for the HFCS

HFCS imputation using ML: three step procedure

1. Create training data with true but *most likely missing values*
   1.1 Survey: identify determinants of non-response to specific item
   1.2 Quasi-admin dataset: identify hhs most likely to miss item
   $\rightarrow$ use this group as survey's artificial counterfactual

2. Select and train algorithm using training data
   - Experiment with those used in papers listed earlier

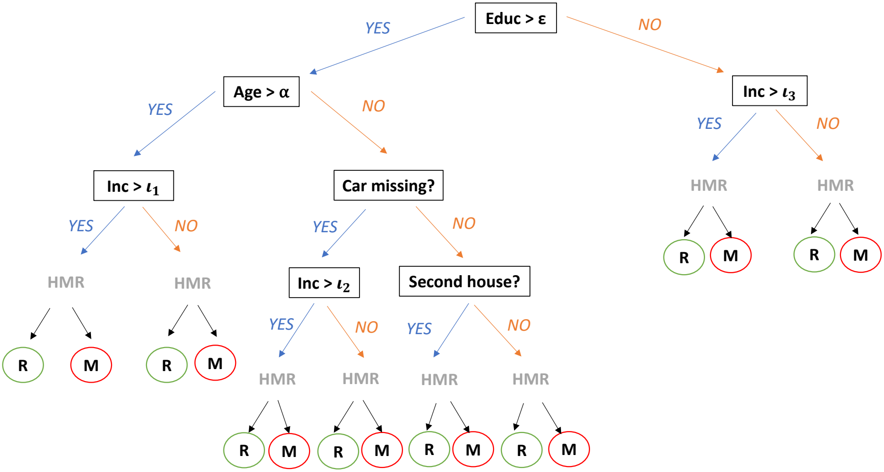
3. Apply trained algorithm to survey country
   - Current imputation is benchmark to assess results

Example: value of hh main residence (HMR; HB0900)

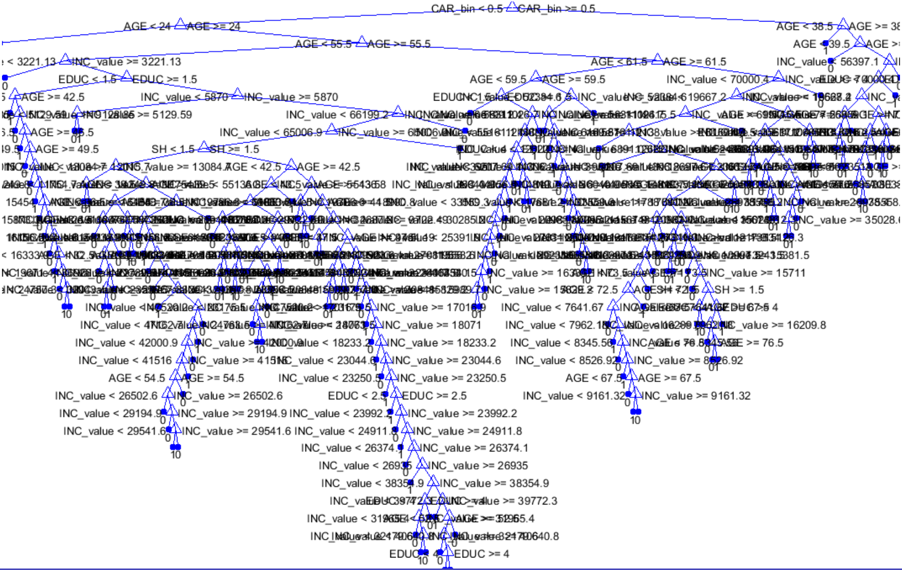
- Step 1: I am working on three options:
    1. Decision Trees (DT; supervised ML classification method)
    2. Statistical Matching
    3. Item Response Theory Modelling

  DT minimizes classification error using hyperparameters
    - number of branches
    - branching variables
    - branching thresholds

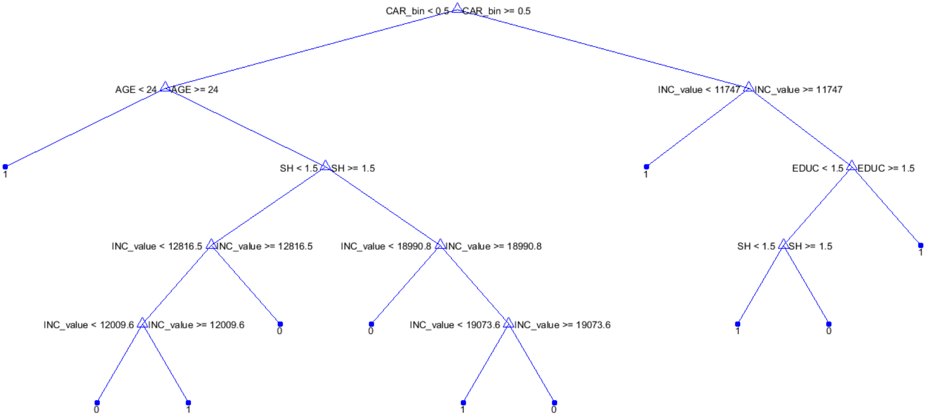
Decision Tree: Illustration for value of HMR- NEW



HMR: Example of a Decision Tree



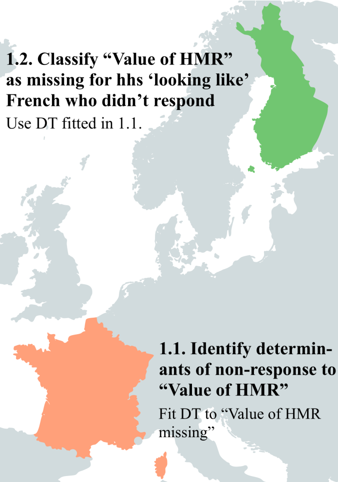
HMR: Decision Tree with arbitrary splits



Classification error: 22.54%

HMR example: steps 1 to 3 - NEW

- 1.1. and 1.2.



- Steps 2 and 3: K-Nearest-Neighbor

HMR Example: Results

	I	II	III	IV	V	VI
	Admin: FI	**Survey: FR**			**ML Imputation**	
	H owners	H owners			Training	Imputed
			Answered	*Imputed**		
N	8,526	8,477	1,051	1,776	500**	1,776
μ	216	315	366	253	202	282
p_{50}	180	225	250	167	193	231
σ	144	340	387	330	155	168

Mean, median, stdev rounded to nearest thousand Euro

*Responded: "No answer" or "Don't know"; **Targeted (classification error tolerance: 32%)

- Does FR imputation underestimate HMR? (IV vs. VI)
- Does ML imputation inherit moments of FI? (V,I vs. III)

Conclusion

Conclusion

- I propose an imputation procedure for missing survey items
- It aims to share benefits of country specific quasi-admin data

- Work to be done
  - Check robustness of training dataset with respect to
    - three approaches for step 1
    - using other quasi-admin country data
    - tolerance of classification error
  - Account for country-specific item distributions
    - Transform quasi-admin distribution?
    - Adjust imputation algorithm?
  - Are admin data always a better measure for HFCS variables?

THANKS for your attention

I am grateful for comments and suggestions

Johannes.Fleck@eui.eu

Non-response in survey data

- Survey observations are either complete or **missing**
- Types of **missing:** item non-response vs. unit non-response

ID	Head Age	Income	Real Assets	Financial Assets	Classification
1	34	100,000	233,000	64,000	*Complete*
2	21	12,000	0		**Missing** *(Item non-response)*
3	57		459,231		**Missing** *(Item non-response)*
4					**Missing** *(Unit non-response)*
5	78				**Missing** *(Item non-response)*
6	66	45,230	120,000	330,000	*Complete*
7	47	78,000	450,000	0	*Complete*
...	...	...	...	...	...
N	39	60,000			**Missing** *(Item non-response)*

Return

ML for imputation

Table: Imputation Algorithms - literature examples (TBC)

INPUTS		**OUTPUTS** categorical	continuous
complete	continuous	*Decision Trees, Random Forest*	*Fuzzy K-means*
	categorical	*Singular Value Decomposition*	
	mixed	*Logistic Regression*	
missing items	continuous		
	categorical		*Nearest Neighbor*
	mixed	*Neural Networks*	

Return