Application of the MissForest algorithm for imputation in the Survey on Income and Living Conditions

Blandine Bianchi

Federal Statistical Office FSO / Data Science, AI and Statistical Methods/ Statistical Methods

UNECE Conference of European Statisticians

Expert meeting on Statistical Data Editing, October 5, 2022

Content

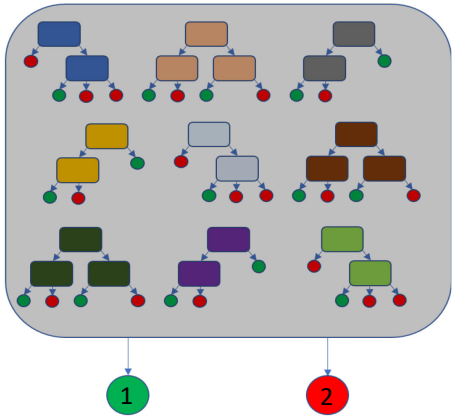
1. Introduction

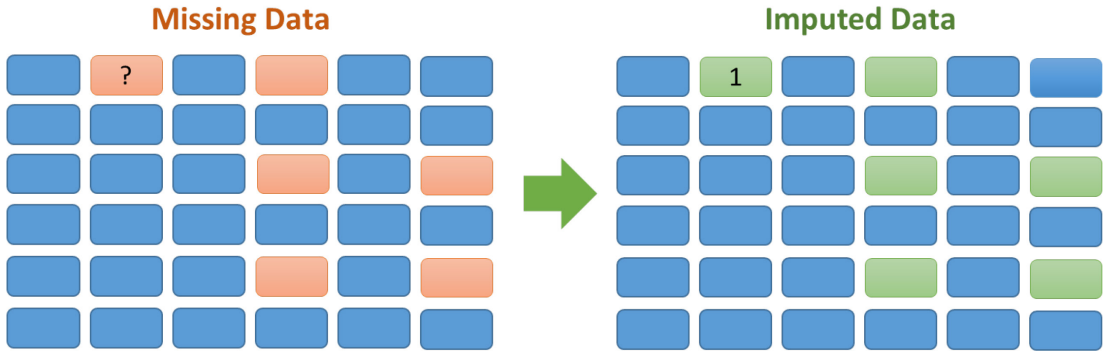
- The survey on Income and Living Conditions (SILC) is a yearly household survey.

- **Non-response** is a key issue and an **appropriate treatment is crucial to improve the quality of the survey** and **increase the reliability of the published results**.

2. MissForest Algorithm

- **Non-parametric imputation based on Random Forests** (Breiman 2001)
- **Can handle** continuous and categorical data simultaneously, **without making any assumptions** about the **structure** and **distribution of the data.**
- Sequential **imputation procedure** starting with the variable with fewest missing values.
- Random forest is trained on the available observations and used to make predictions for missing values.
- The **accuracy** of the predictions can improve using **auxiliary variables**.





Stekhoven and Bühlmann 2012

3. Data: SILC 2019

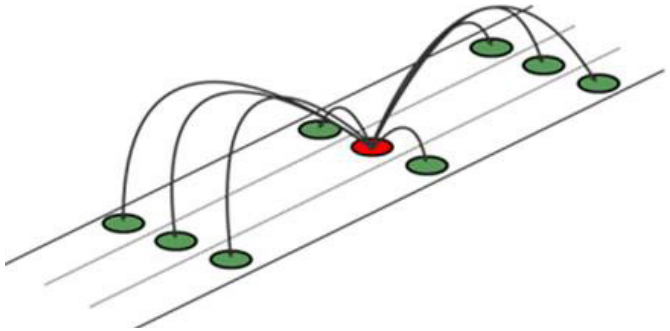
List of person and household variables concerning material and social deprivation to be imputed, taking mostly the values 1, 2, 3* :

		List of variables to impute for the analysis of material and social deprivation	Total	Total Applicable	Units with missing values	Missing rate to be imputed [%]
Person Variable (P)	**PD020**	Replace worn-out clothes by some new (not second-hand) ones	16662	13861	2562	18.48%
	PD030	Two pairs of properly fitting shoes (including a pair of all-weather shoes)	16662	13861	2553	18.42%
	PD050	Get-together with friends/family (relatives) for a drink/meal at least once a month	16662	13861	2563	18.49%
	PD060	Regularly participate in a leisure activity such as sport, cinema, concert	16662	13861	2565	18.51%
	PD070	Spend a small amount of money each week on yourself	16662	13861	2559	18.46%
	PD080	Internet connection for personal use at home	16662	13861	2556	18.44%
Household Variable (HH)	**HD080**	Replacing worn-out furniture	7341	7341	29	0.40%
	HS011	Arrears on mortgage or rent payments	7341	6867	26	0.00%
	HS021	Arrears on utility bills	7341	7341	13	0.18%
	HS031	Arrears on hire purchase instalments or other loan payments	7341	1354	12	0.01%
	HS040	Capacity to afford paying for one week annual holiday away from home	7341	7341	16	0.22%
	HS050	Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day	7341	7341	7	0.10%
	HS060	Capacity to face unexpected financial expenses	7341	7341	28	0.38%
	HS110	Do you have a car?	7341	7341	3	0.04%
	HH050	Ability to keep home adequately warm	7341	7341	10	0.14%

* (1) yes, I can afford it, (2) no, I cannot afford it, (3) no, I cannot afford it, but for non-financial reasons.

4. Simulations

- Evaluation MissForest imputation of the person variables of material and social deprivation:
  1. Subset of the data without missing values (82%)
  2. Joint simulation of random missing values (18%)
  3. MissForest imputation
  4. The imputed values are compared with the original values to quantify the error.
- Inclusion of auxiliary variables to improve the imputation
  - Household variables (imputed)
  - Socio-demographic variables
  - Historical variables

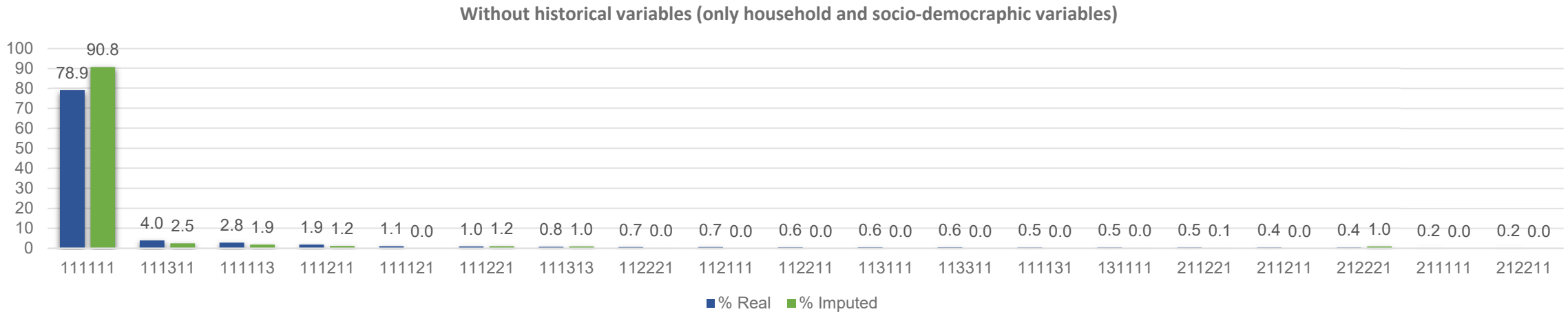


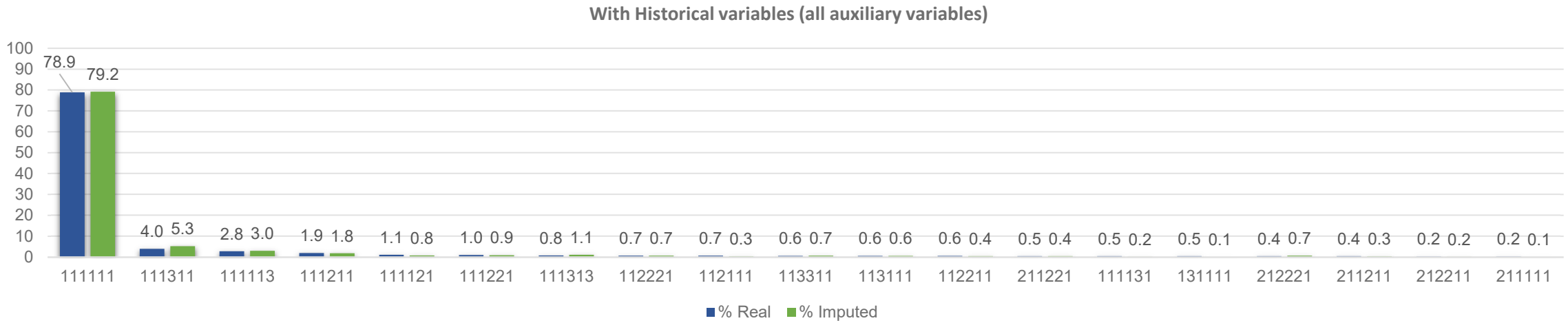
5. Validation (Imputation Error)

Personal variables Imputed with MissForest - Simulations		% Total Error		
		Only household variables	Household and socio-demographic variables	Household, socio-demographic and historical variables
PD020	Replace worn-out clothes by some new (not second-hand) ones	**24.2%**	**2.9%**	**1.2%**
PD030	Two pairs of properly fitting shoes (including a pair of all-weather shoes)	**4.8%**	**0.9%**	**0.9%**
PD050	Get-together with friends/family (relatives) for a drink/meal at least once a month	**26.1%**	**6.0%**	**2.8%**
PD060	Regularly participate in a leisure activity such as sport, cinema, concert	**26.4%**	**12.1%**	**5.7%**
PD070	Spend a small amount of money each week on yourself	**23.7%**	**6.2%**	**2.1%**
PD080	Internet connection for personal use at home	**17.7%**	**6.0%**	**0.7%**

With Total Error = $\frac{1}{n}\sum_k I(y_k \neq y_k^*)$

5. Validation (Distribution)

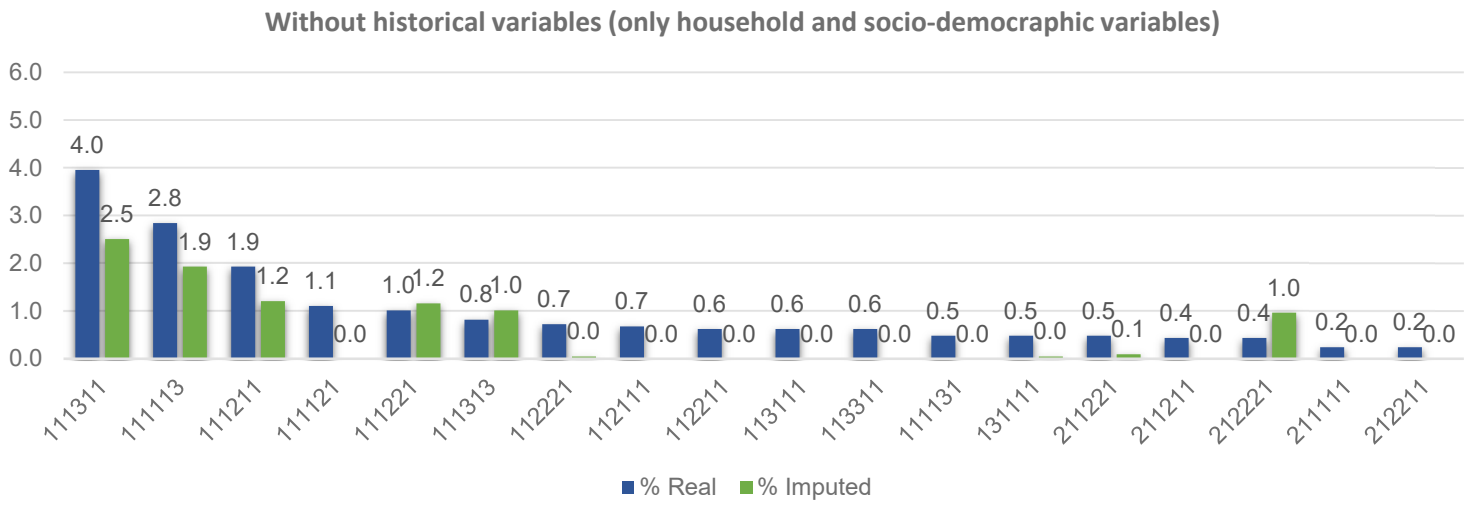


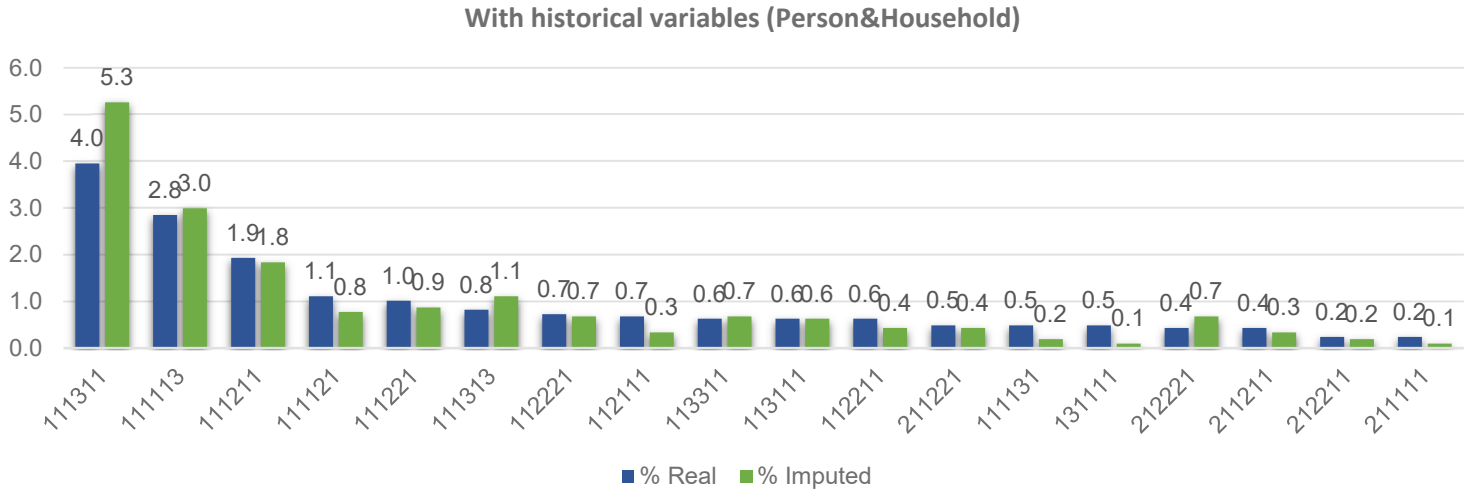


5. Validation (Distribution)

«111111»	No hist. var.
78.9%	%Real (original)
90.8%	Imputed

«111111»	With hist. var.
78.9%	%Real (original)
79.2%	Imputed

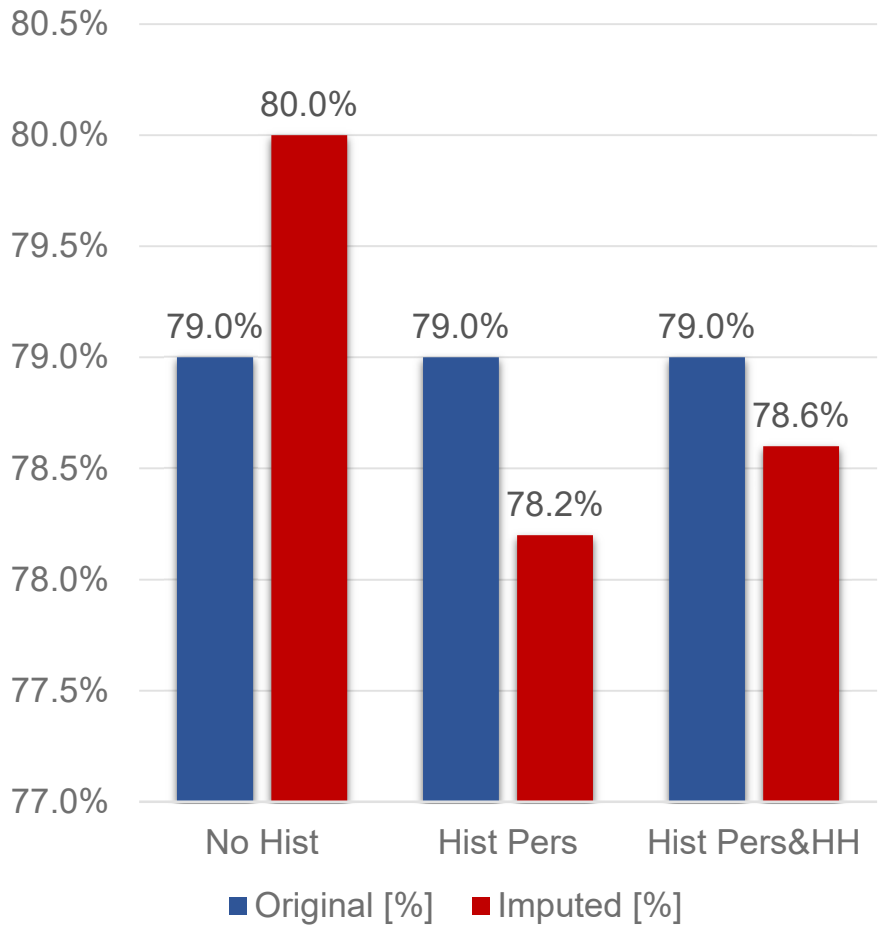




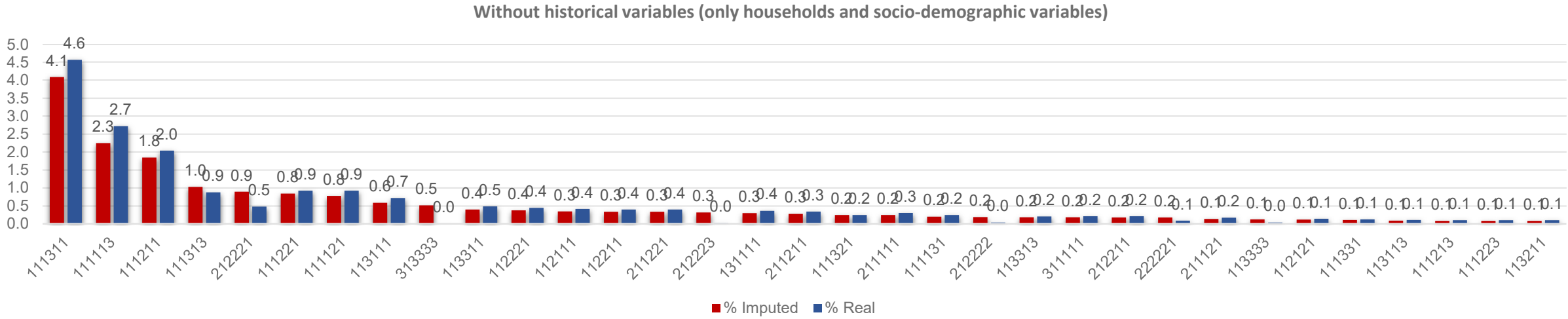
6. Imputation Impact Analysis (1)

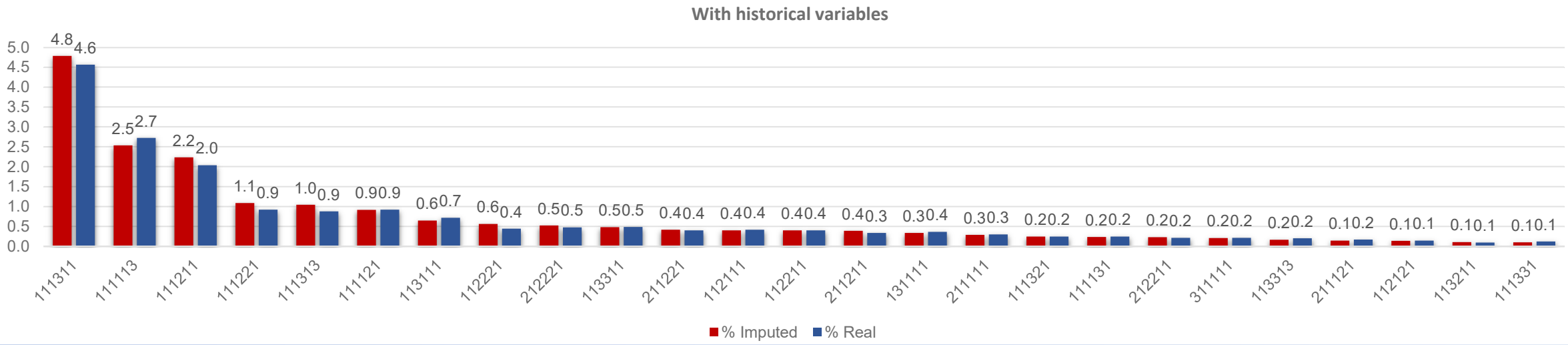
- Relative frequencies of original (blue) and imputed (red) values of class “111111”, corresponding to no social and material deprivation, by the amount of auxiliary variables used.

- The inclusion of historical variables resulted in a significant improvement of the imputations.



6. Imputation Impact Analysis (2)





7. Conclusions

- For the **personal material and social deprivation variables the non-response is quite high** leading to a important challenge to reduce the risk of bias in the results.
- **MissForest provides accurate imputation of the missing values** of material and social deprivation of the SILC.
- The errors range from 0.92% to 12.11% using **household and socio-demographic variables as auxiliary variables**. The use of **historical variables** strongly reduces the error to 0.9 - 5.7%.
- These findings still need confirmation from the Unit analyzing the data.

Thank you!

Questions?