Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO

Application of the MissForest algorithm for imputing income variables in the Survey on Income and Living Conditions

BIANCHI Blandine, KILCHMANN Daniel

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

UNECE Expert Meeting on Statistical Data Editing

Vienna - October 8, 2024

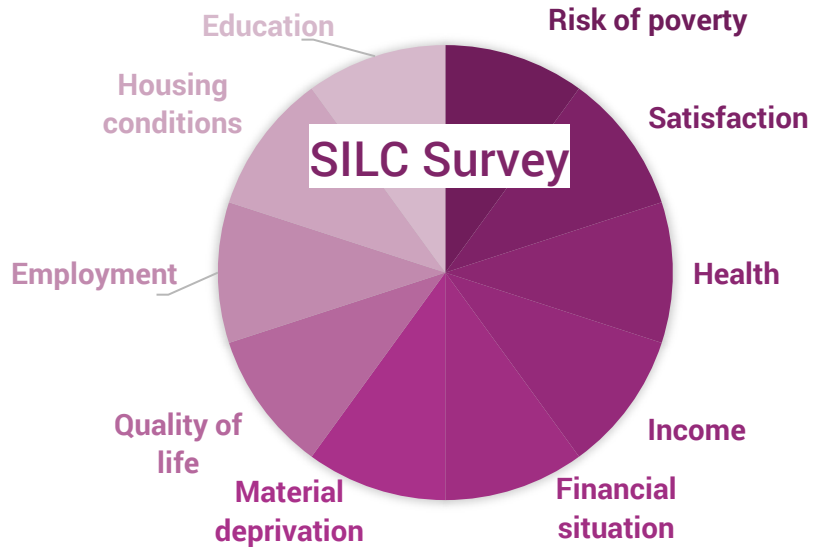
Missing data

All surveys are affected by potential bias due to non-random non-response.

Dealing with missing data typically involves different strategies depending on the type of non-response:

- **Complete non-response** is usually treated by **reweighing**.
- **Partial non-response** is usually treated by **imputation.**

Partial non-response occurs when at least one question has been answered. With imputation, missing and incoherent values are replaced by new values.



Imputation of partial non-response

- We applied **MissForest algorithm** to impute missing values due to **partial non-response** in the SILC (Income variables).
  - MissForest initially imputes all missing values using the mean (or the mode for categorical variables).
  - For each variable with missing values, it fits a **random forest** on the observed part and then predicts the missing values.
  - This process of training and predicting is repeated iteratively until a stopping criterion is met.

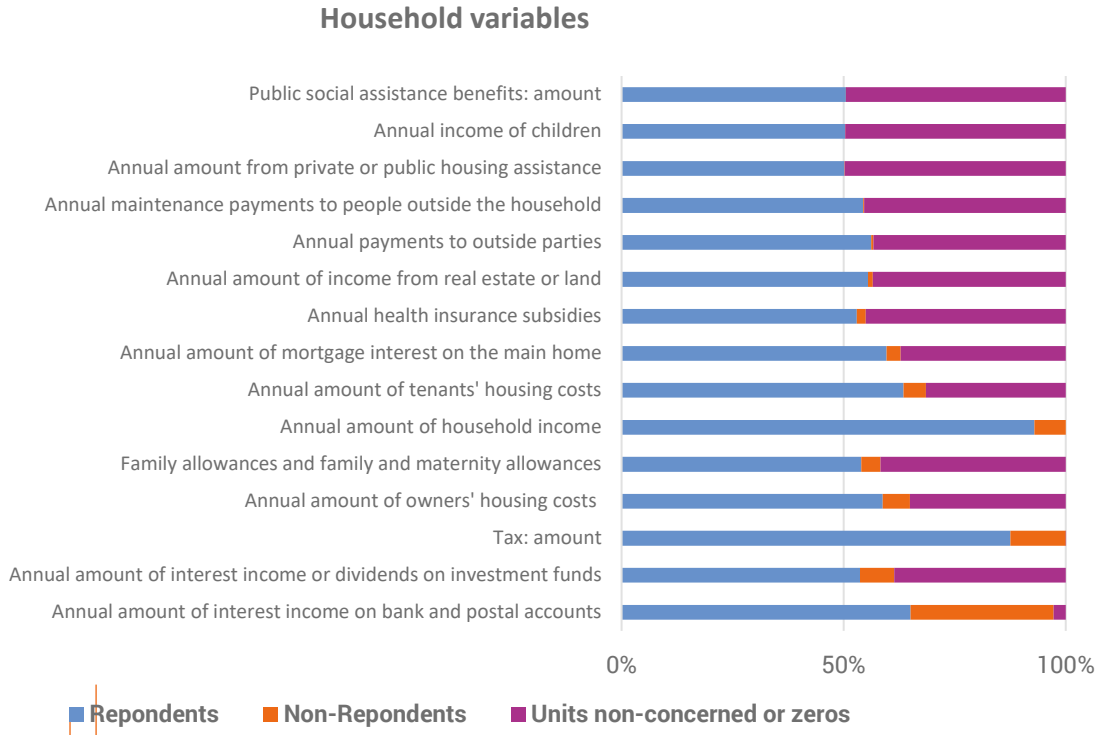
We proceeded in **four stages**. At each stage individual income variables were imputed using household and socio-demographic auxiliary variables:

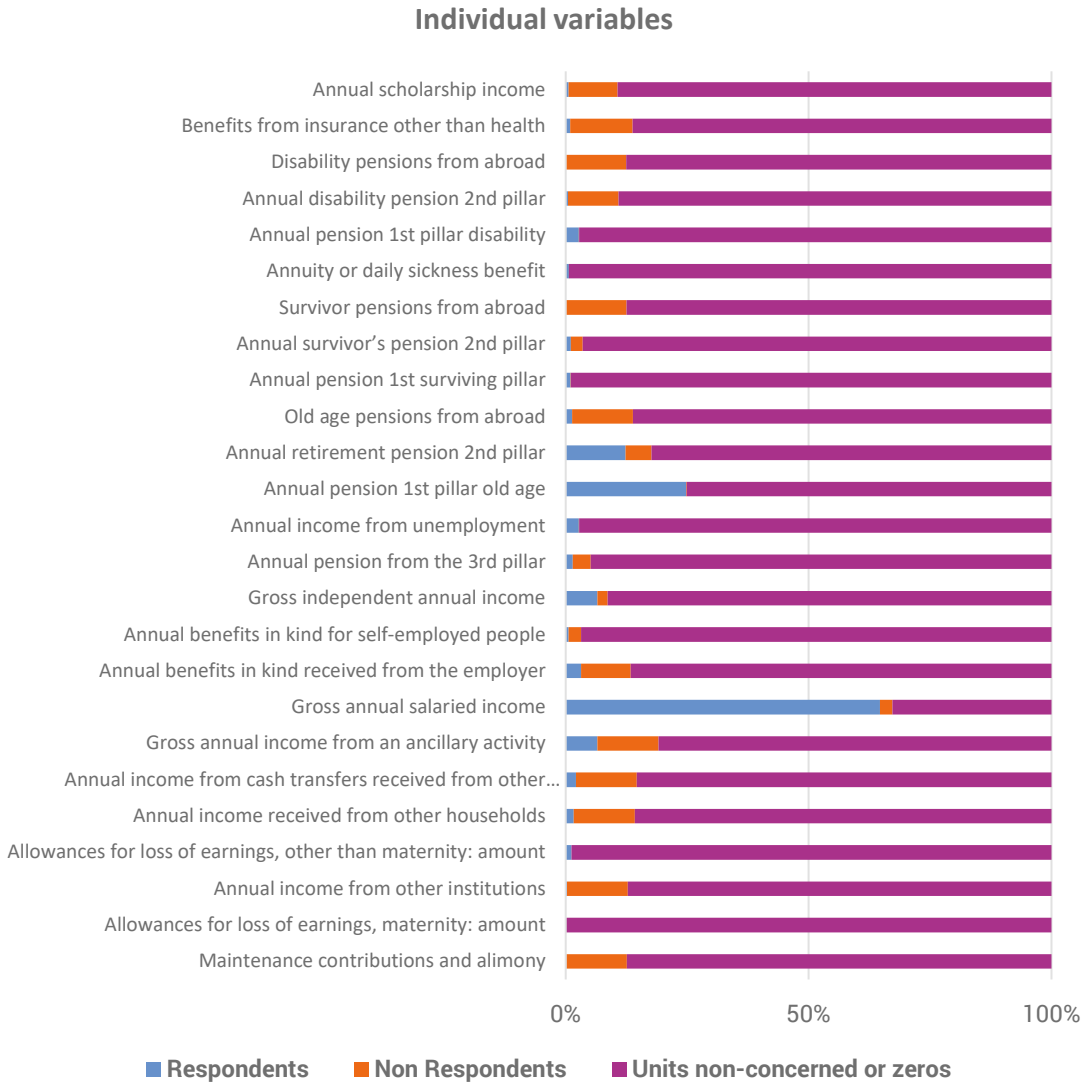
1. **Imputation of unconcernedness** (filter) of all variables together.
2. **Imputation of zeros** of all variables together.
3. **Imputation of positive values** of all variables together.
4. In this step, the **variables gross annual salaried income and annual old-age pension 2nd pillar are re-imputed** separately, as these variables contain a slightly higher number of missing values expecting a positive value.

Data

SILC2020 dataset:

- Survey 8 156 households and 15 177 individuals.
- 25 individual variables.
- 15 household variables.





Data preparation and simulation

Complementing the dataset with auxiliary and boundary variables:

- 91 auxiliary (socio-demographic) variables.
- 11 boundary variables.

Simulation:

- Simulation of 17% non-response based on the initial distribution of missing values to assess imputation accuracy.
  - **Analysis of the non-response mechanism** to deduce a non-response model.
  - **Simulation of missing values on the respondents**.
  - Random selection without replacement of a non-response class, proportional to the probabilities of belonging to the non-response classes obtained through a logistic regression.
- Imputation of individual income variables using household and socio-demographic auxiliary variables.
  - **Imputation without/with the boundary variables** to quantify the improvement.
  - **Assessment of the quality of the imputations** comparing the real values with the imputed ones.

Results & Validation

Impact of auxiliary and boundary variables

- ***Accuracy***: Common occurrences of non-concerned units.
- ***RAE:*** If imputed and original values are both ≥ 0, the Relative Absolute Error is calculated.
- ***RAE_adj:*** provides a clearer estimate of the imputation error, since it takes into account the number of missing values imputed positively and the number of positive values reported.

- Here are the results for the gross annual salaried income and the gross annual self-employed income :

		Imputation Error			Avg nb of units		
Individual Variable	**With Boundary Variables**	**Accuracy**	**RAE**	**RAE_adj**	**Avg of imputed units**	**Imputed units positively**	**Imputed units as non-concerned**
Gross annual salaried income	no	95.56%	0.4290	0.0115	308	205	103
Gross annual salaried income	yes	96.67%	0.0348	0.0009	315	214	102
Gross annual self-employed income	no	95.86%	0.8206	0.0179	247	12	236
Gross annual self-employed income	yes	98.93%	0.0792	0.0017	249	19	230

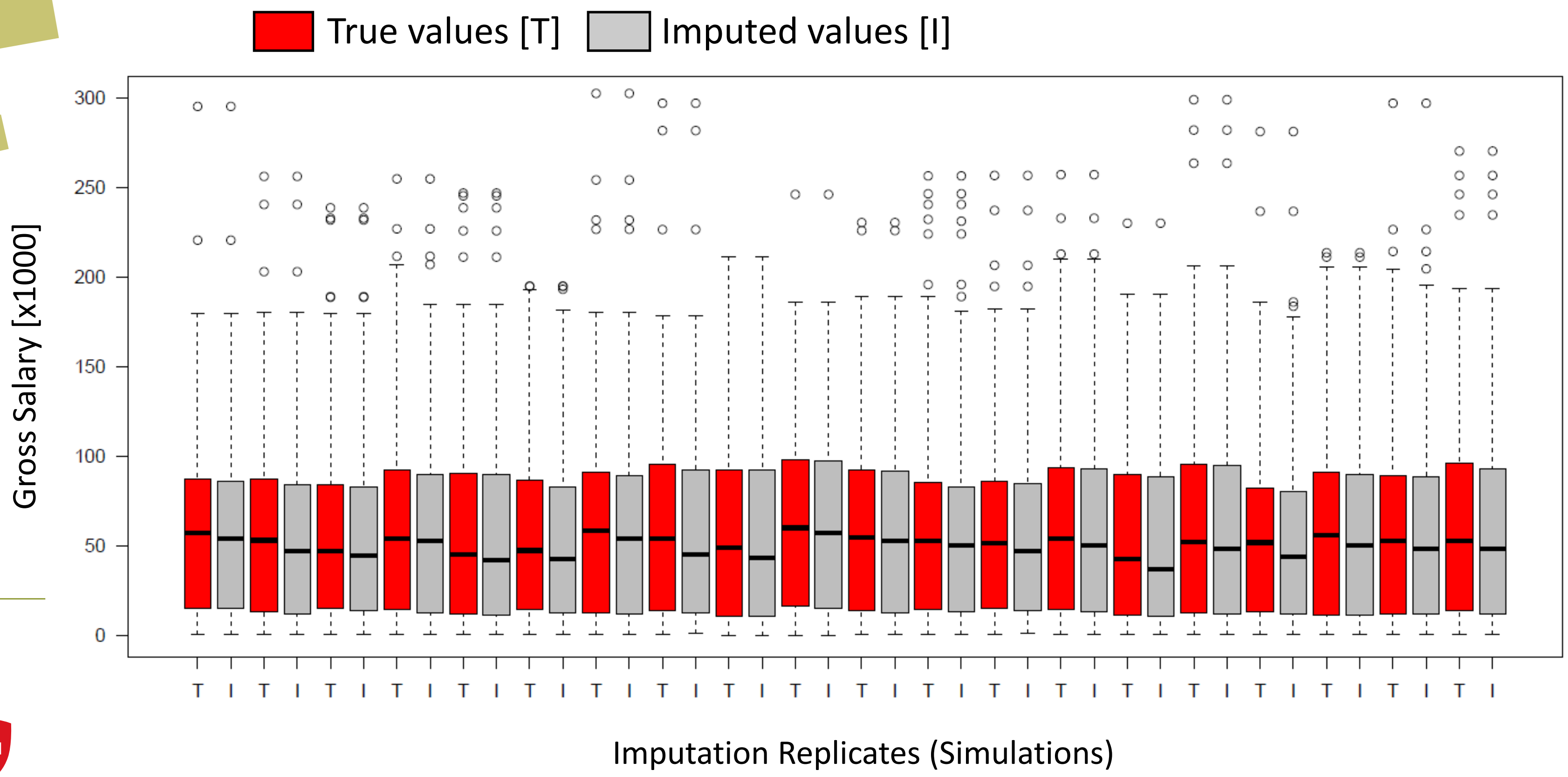
Impact of auxiliary and boundary variables

- The simulation showed that the imputation error of 'unconcerned' was low across all imputed variables ranging from 0 to 0.16 and the accuracy ranging from 89.6% to ~100% has been judged quite good.

- The overall trimmed mean accuracy between imputation without boundary variables and imputation using these variables increased from 97.47% to 98.12%, while the relative absolute error, RAE, decreased from 0.35 to 0.23.

	Error Imputation without boundaries			Error Imputation with boundaries		
Individual Variable Stat	Accuracy	RAE (Relative Absolute Error)	RAE_adj	Accuracy	RAE (Relative Absolute Error)	RAE_adj
Max	100.00%	2.584	0.082	99.99%	10.871	0.158
Min	87.49%	0.095	0.000	89.58%	0.035	0.000
Mean	97.47%	0.429	0.008	98.12%	0.647	0.011
Trimmed mean	97.47%	0.350	0.006	98.12%	0.229	0.005

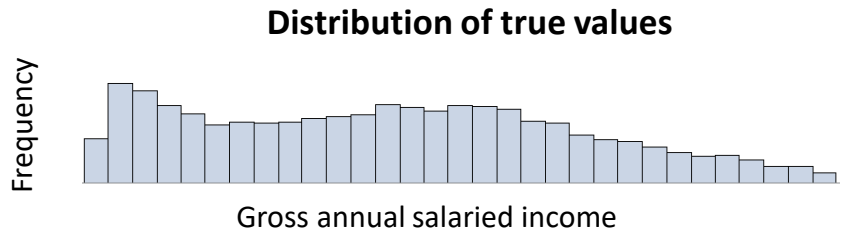
Evaluation of imputation accuracy through simulations

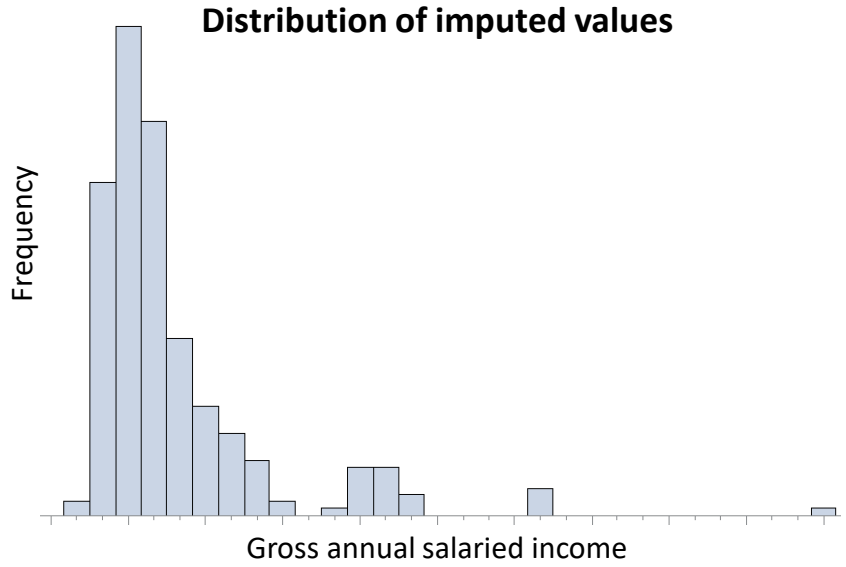


- The distributions of imputed and original values are very close.
- The boundaries conditions are set to the percentile 0.95 to ensure that extremes values do not skew the results and provide a more robust measure of the lower and central tendency.

Impact analysis

- The imputed distribution differs substantially from the distribution of known values.
- This highlights the potential bias due to non-response and the importance of using auxiliary and boundary variables for imputation.





Plausibilisation

- With MissForest there is no guarantee that the imputed values lie within the boundary variables, even if these are used as auxiliary variables.

- Thus instead of imputing the bounded variable directly, we imputed a quotient variable defined as:

$$q_{imp} = \frac{(x_{imp} - x_{low})}{(x_{up} - x_{low})}$$

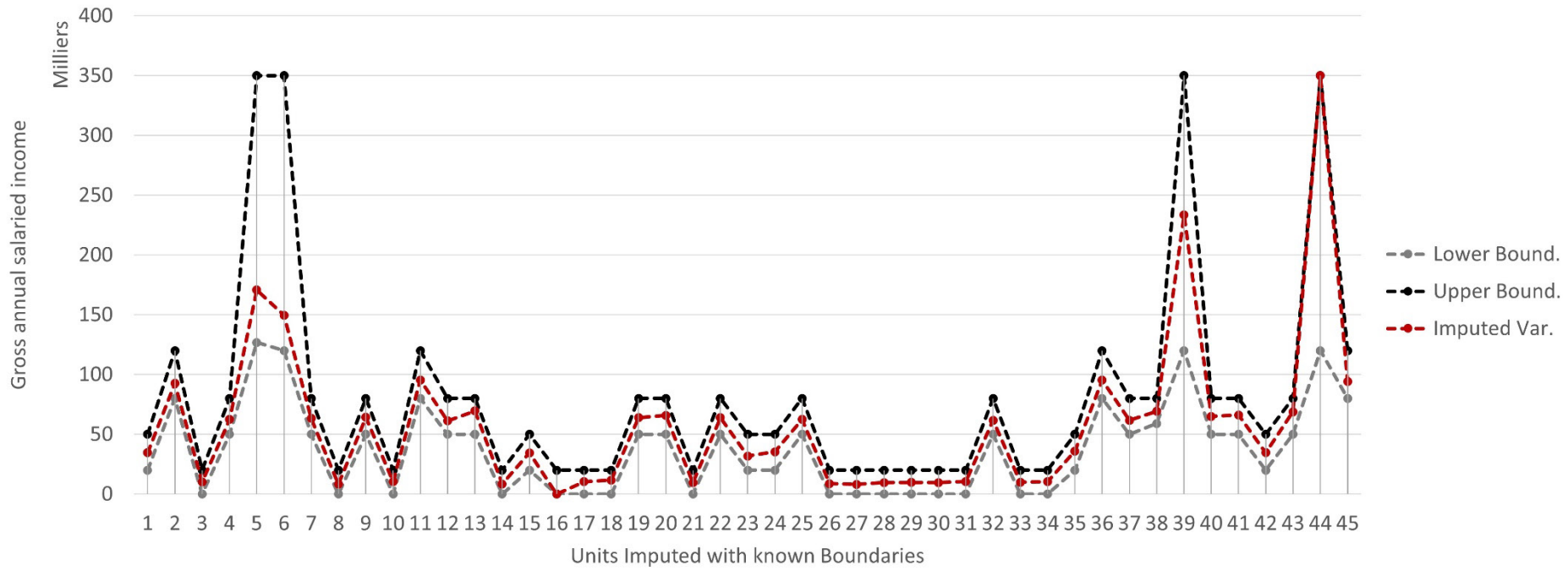
from which we obtain the variable of interest:

$$x_{imp} = q_{imp}(x_{up} - x_{low}) + x_{low}$$

- If the unit in question has boundaries, then the imputed value must be within those boundaries. If the imputation is outside the limits, the variable of interest takes the value of the closest limit. This is done by a post-imputation procedure.

Plausibilisation

- Imputation of the real missing values for the gross annual salaried income, with known boundaries.



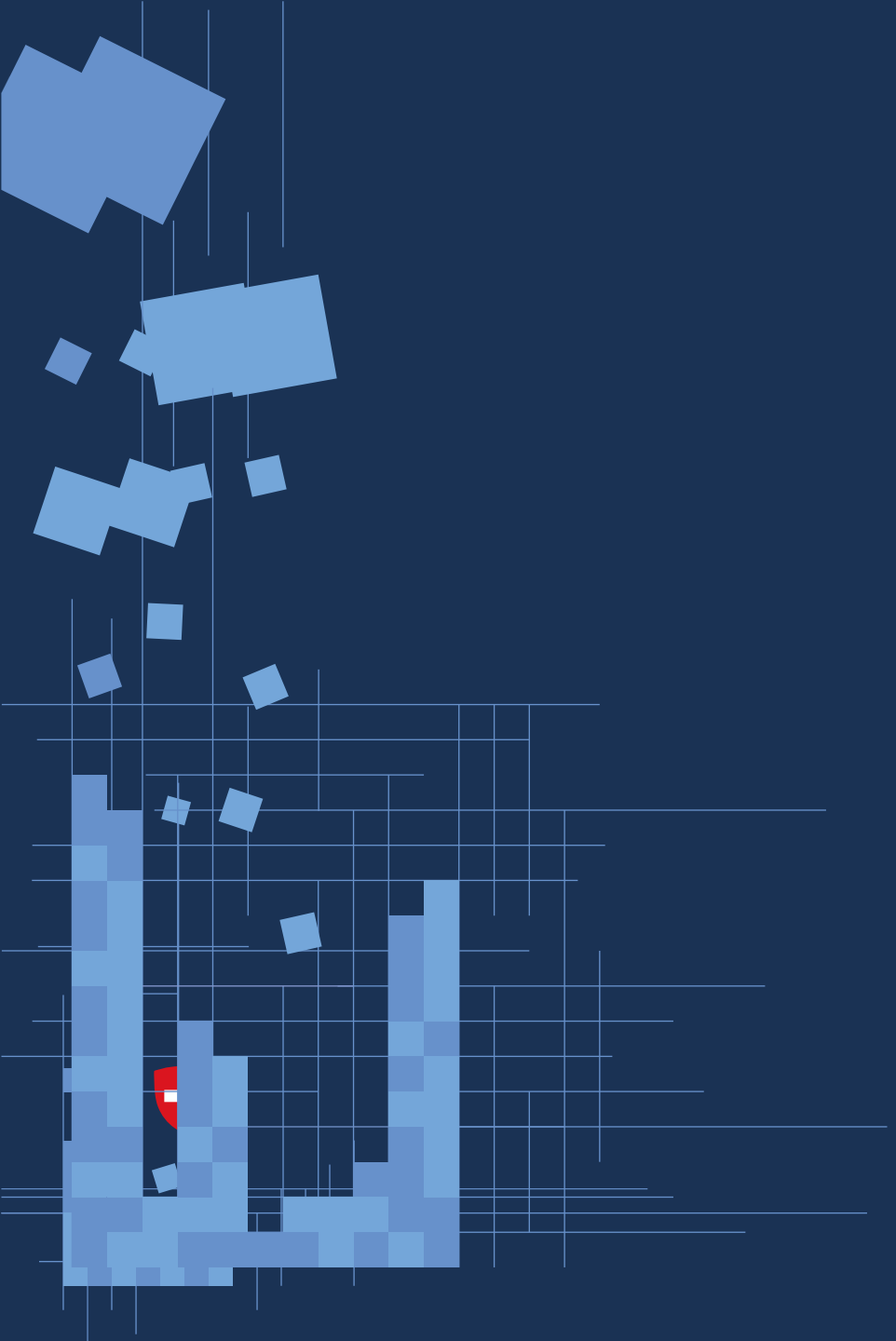
Conclusions

- **The MissForest algorithm can accurately impute the SILC2020 income variables of interest**.
- We **studied non-response** and **simulated missing values** to evaluate imputations without and with the use of boundary variables:
  - **The accuracy of the imputation increases substantially with the use of boundary variables: Relative Absolute Errors decrease up to 10-fold.**
  - We imputed individual income variables, using socio-demographic auxiliary variables and reported household variables.
  - Based on our simulations, the magnitude of the error is acceptable and is mainly related to the initial distribution of the variable of interest to be imputed: the fewer positive values there are in the variable of interest, the less likely it is that positive values will be imputed.
- The **plausibility check** to ensure that the imputed values lie within the boundaries variables.
- Impact analyses show that the distribution of imputed values can differ substantially from that of the reported ones.
- **We observed from both the simulation and the impact analysis that the most important variables in terms of salaries are those for which imputation error of positive values is amongst the lowest.**

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO



Thank you !

Questions ?