Istat | Istituto Nazionale di Statistica

Rome, 01-02 December 2025

4rd WORKSHOP ON METHODOLOGIES FOR OFFICIAL STATISTICS

ARTIC:
Automatic Random error Treatment and Imputation for Categorical variables.

Filippini R., Toti S.

Istat | DCME - DIREZIONE CENTRALE PER LA METODOLOGIA E IL DISEGNO DEI PROCESSI STATISTICI **(Methodology Department)**

Outline

- Introduction
- ARTIC process flow
  - Phase I: Checking the data
  - Phase II: Localizing the erroneous variables
  - Phase III: Imputation
- Case study
- Conclusions

Introduction

- The identification and treatment of non-sampling errors is a fundamental step in the production of Official Statistics, aimed at ensuring higher accuracy and quality of the resulting estimates.
- Traditionally, the Editing and Imputation (E&I) process represents one of the most time-consuming phases.
- In the context of Surveys, developing generalized tools that can be easily adapted to different questionnaires offers valuable support throughout the entire process.

To this end, a standardized procedure in R, called **ARTIC** (*Automatic Random error Treatment and Imputation for Categorical variables*), has been developed for the detection and automatic correction of random errors for categorical variables, following the Fellegi–Holt paradigm.

Fellegi-Holt paradigm

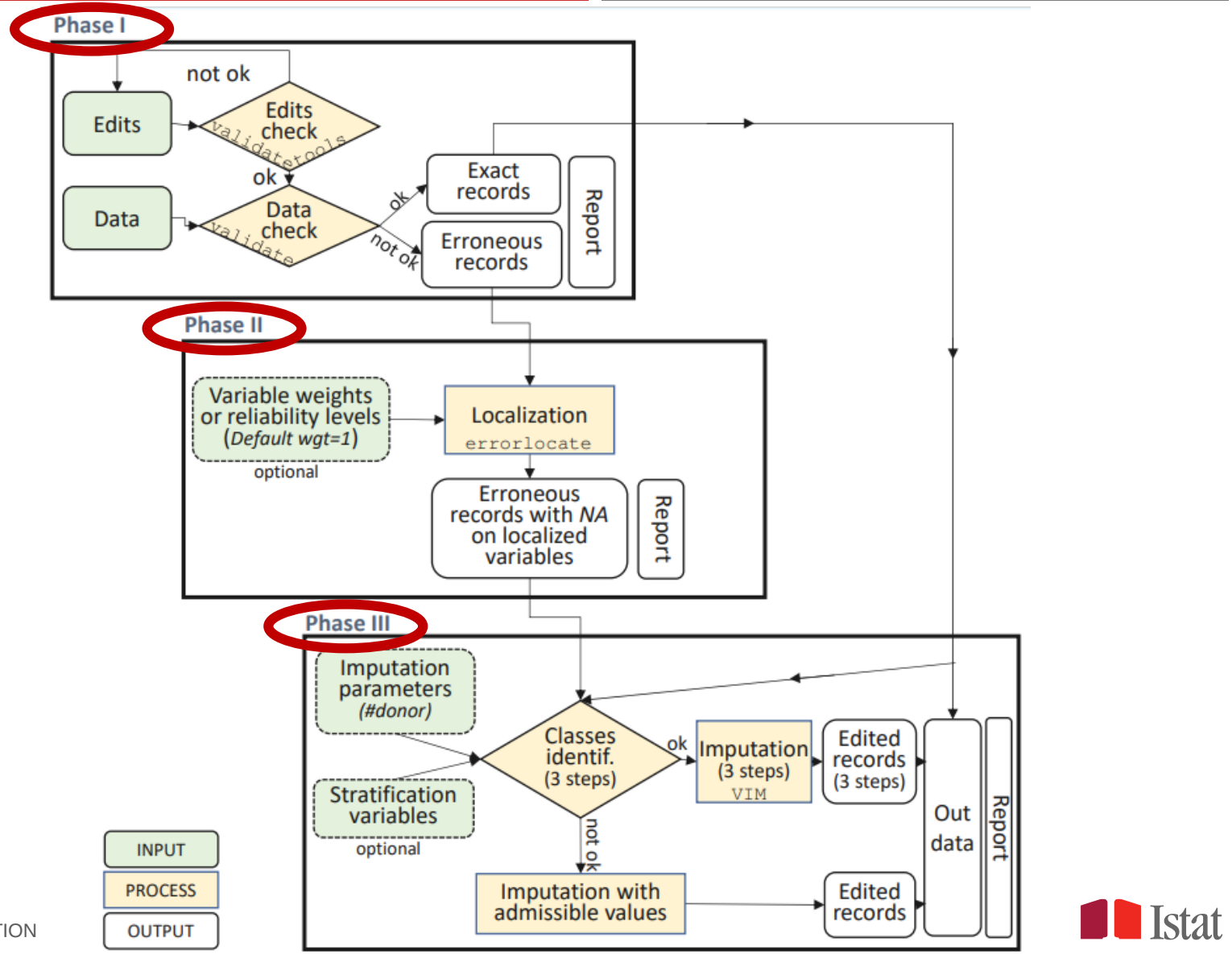
- It defines a formal framework for detecting and correcting data errors.
- It is based on three sequential phases:
  - **Checking** the data (editing) to identify units affected by errors.
    The process involves verifying each field to ensure that it contains a valid entry (validity rules), and examining specific combinations of fields to ensure that the entries are mutually consistent (consistency rules)
  - **Localizing** the erroneous variables within each unit.
    The data in each record should be made to satisfy all edits by changing the fewest possible items (minimum change criterion)
  - **Imputing** the identified errors.
    Imputation under edit constraints is implemented through a hot-deck imputation: fields identified as erroneous are replaced with values randomly drawn from donor records that have successfully passed all relevant edit checks. This allows to preserve the statistical distribution of the correct data.

    The construction of the donor pool is a crucial step in the imputation process

(Fellegi, I. and D. Holt. 1976. “On the use of data collection process information for the treatment of unit nonresponse through weight adjustment”. Survey Methodology, Volume 71: 17–35)

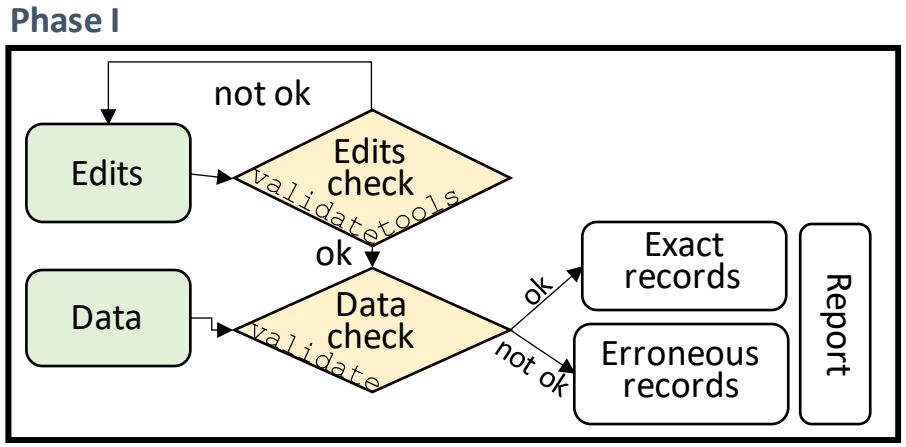
ARTIC: process flow

- Each phase of the process is performed in a semi-automated manner
- The interaction with the user is managed through shiny apps.
- using:
  - the R libraries `validate`, `validatetools`, `errorlocate`, `VIM`
  - and a set of *ad hoc* functions
- At the end of each phase, outputs are produced and ARTIC can be stopped. This ensures phase modularity.

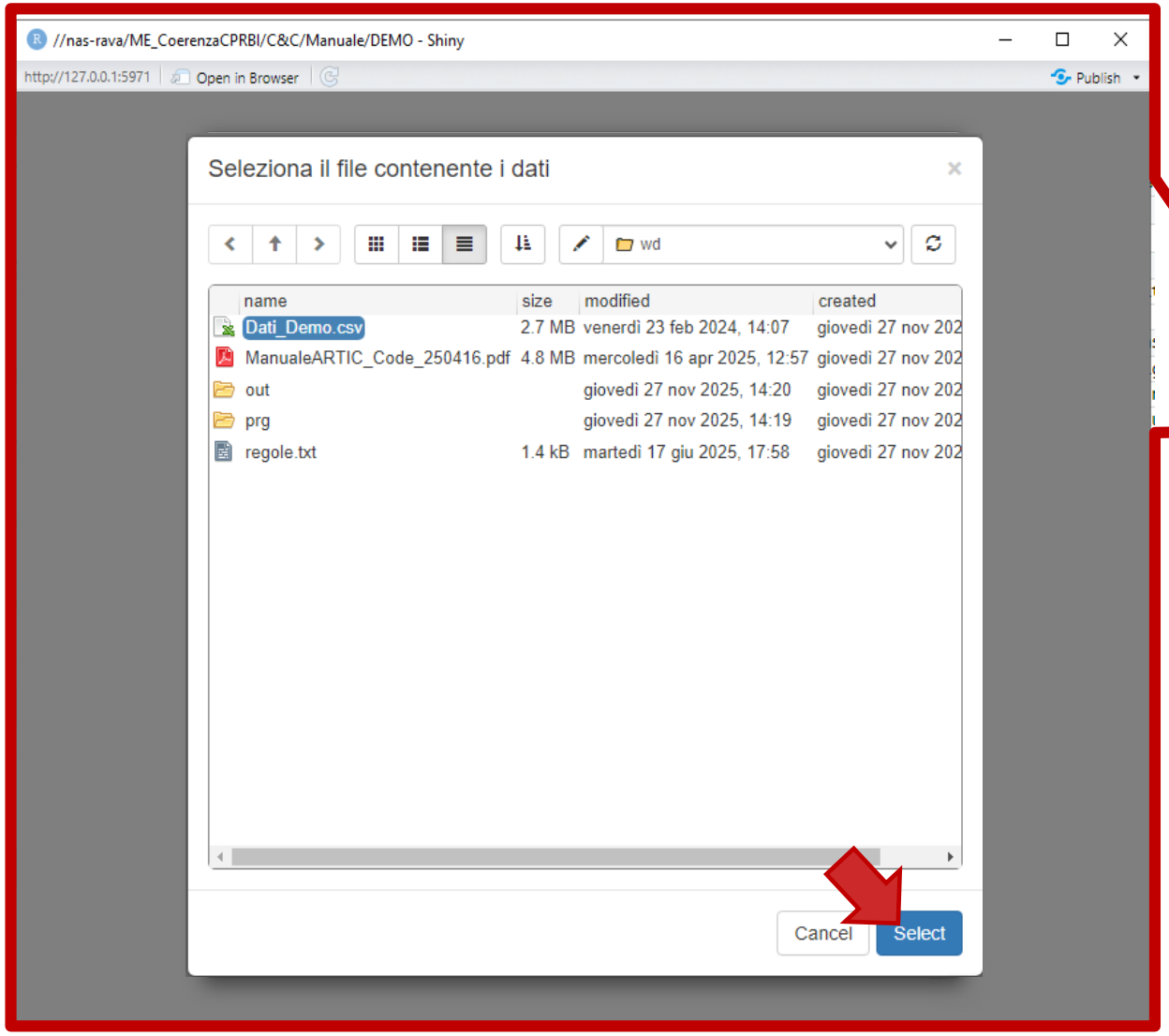


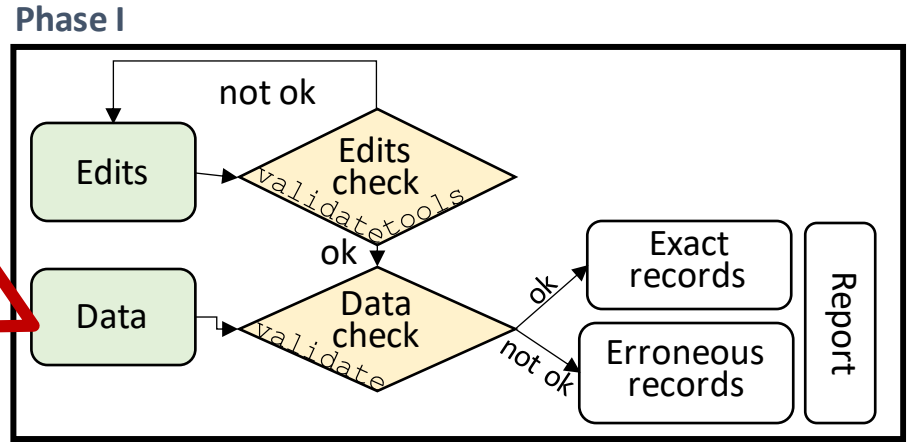
Phase I: Checking the data

- Error detection and validation are based on a consistency plan developed with subject-matter experts
- All rules must be verifiable at the individual record level and must not depend on the values of other records
- Records that violate at least one rule are deemed erroneous
- Data check is performed using functions from the `validate` library. Edits check is performed using functions from the `validatetools` library



Phase I: Checking the data

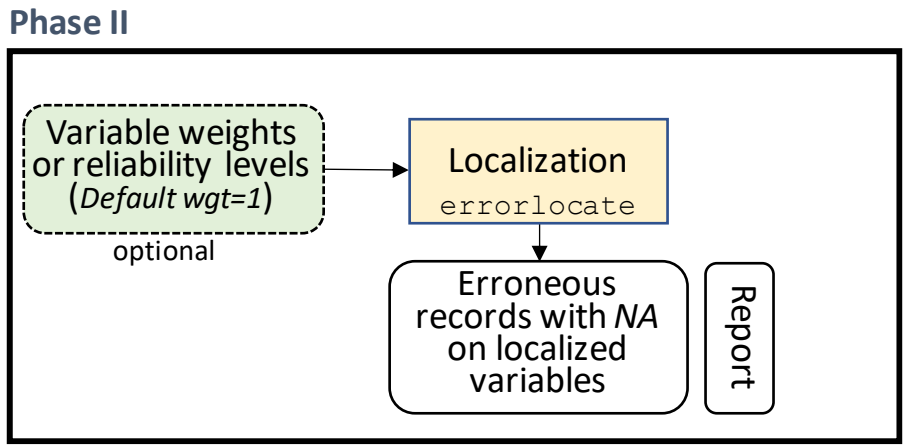




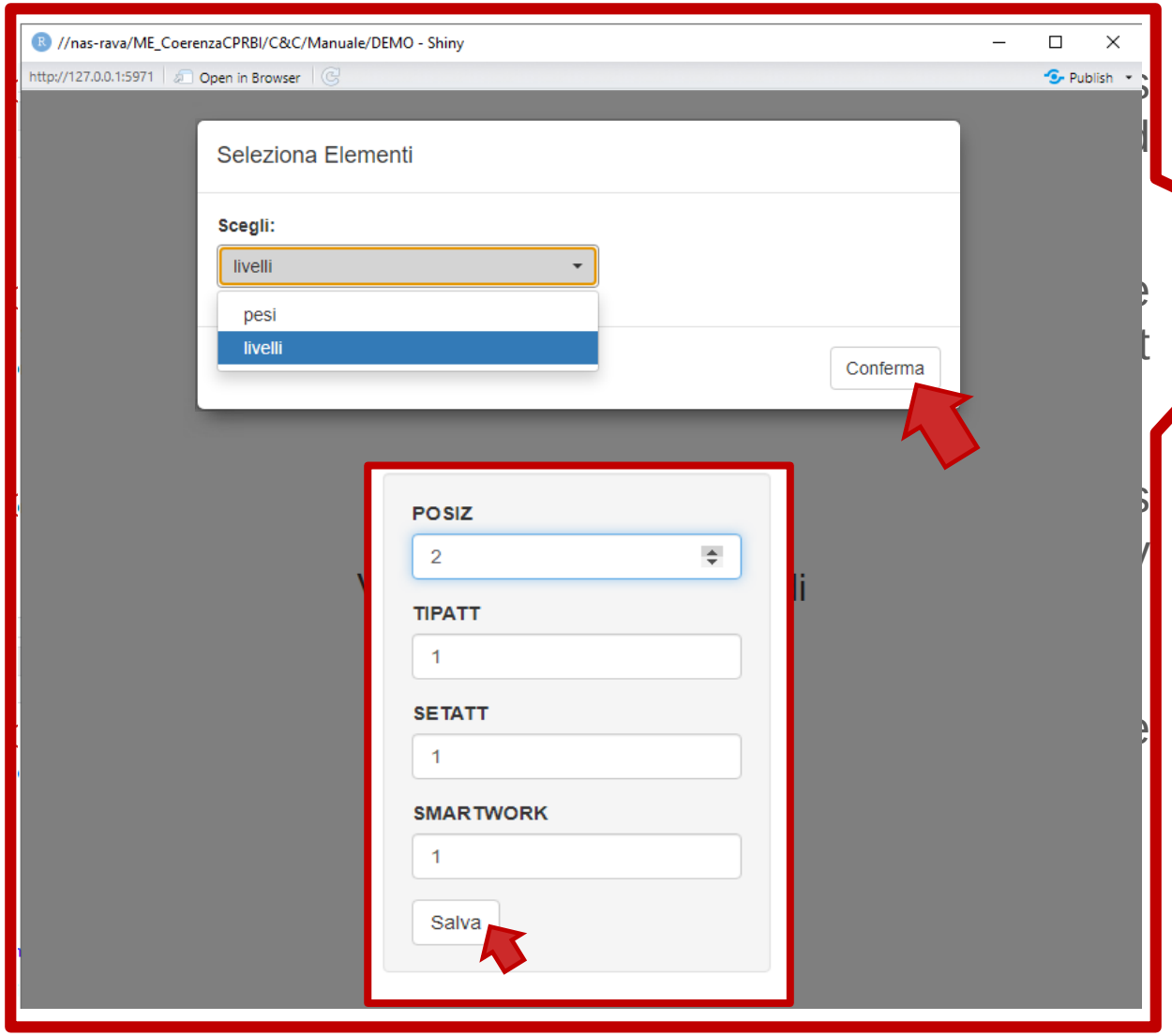
- The interaction with the user is managed through shiny apps.

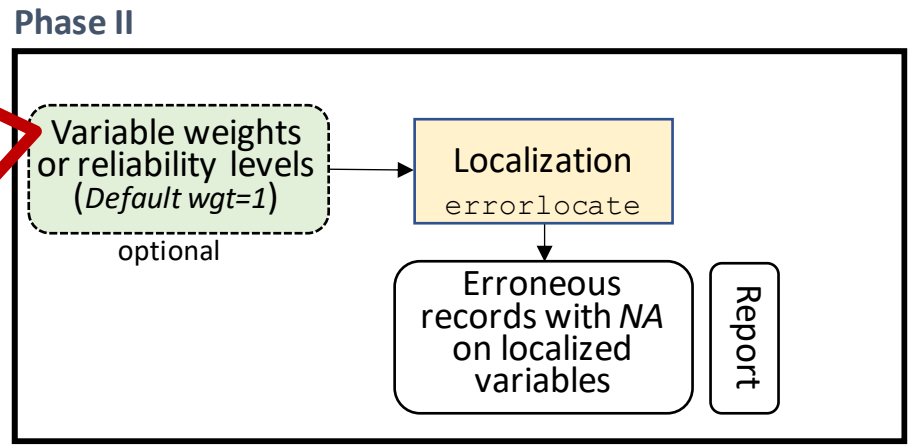
Phase II: Localizing the erroneous variables

- The minimum set of variables to be imputed consists of the subset of variables that “cover” all the violated rules and whose size is minimal.
- A weight can be assigned to each variable to guide localization based on its a priori reliability from expert knowledge
- Weights can be specified either as absolute values (*pesi*) or as relative measures reflecting the hierarchy (*livelli*) of variable reliability
- Functions from the `errorlocate` library are employed to solve the minimization problem.



Phase II: Localizing the erroneous variables



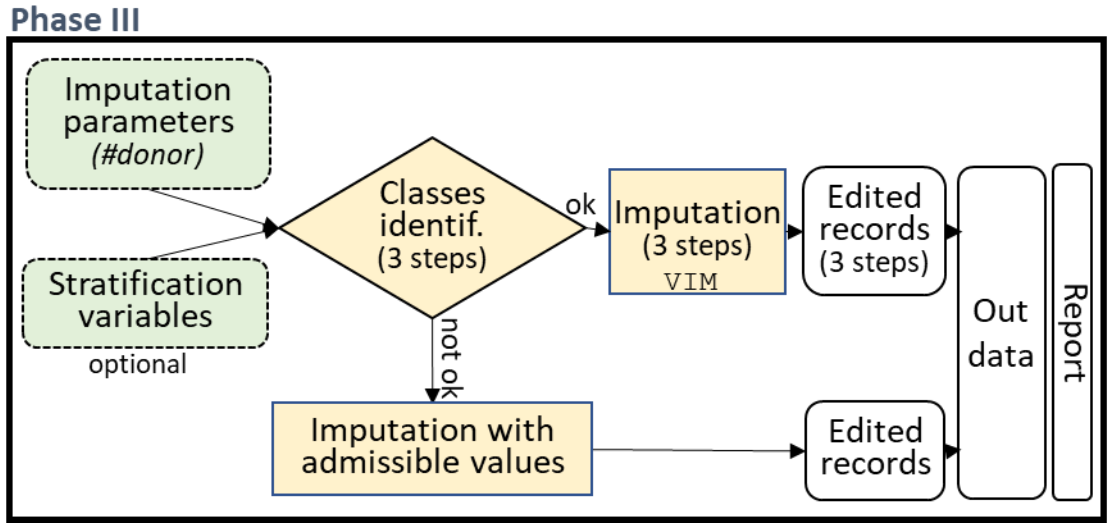


- The interaction with the user is managed through shiny apps.

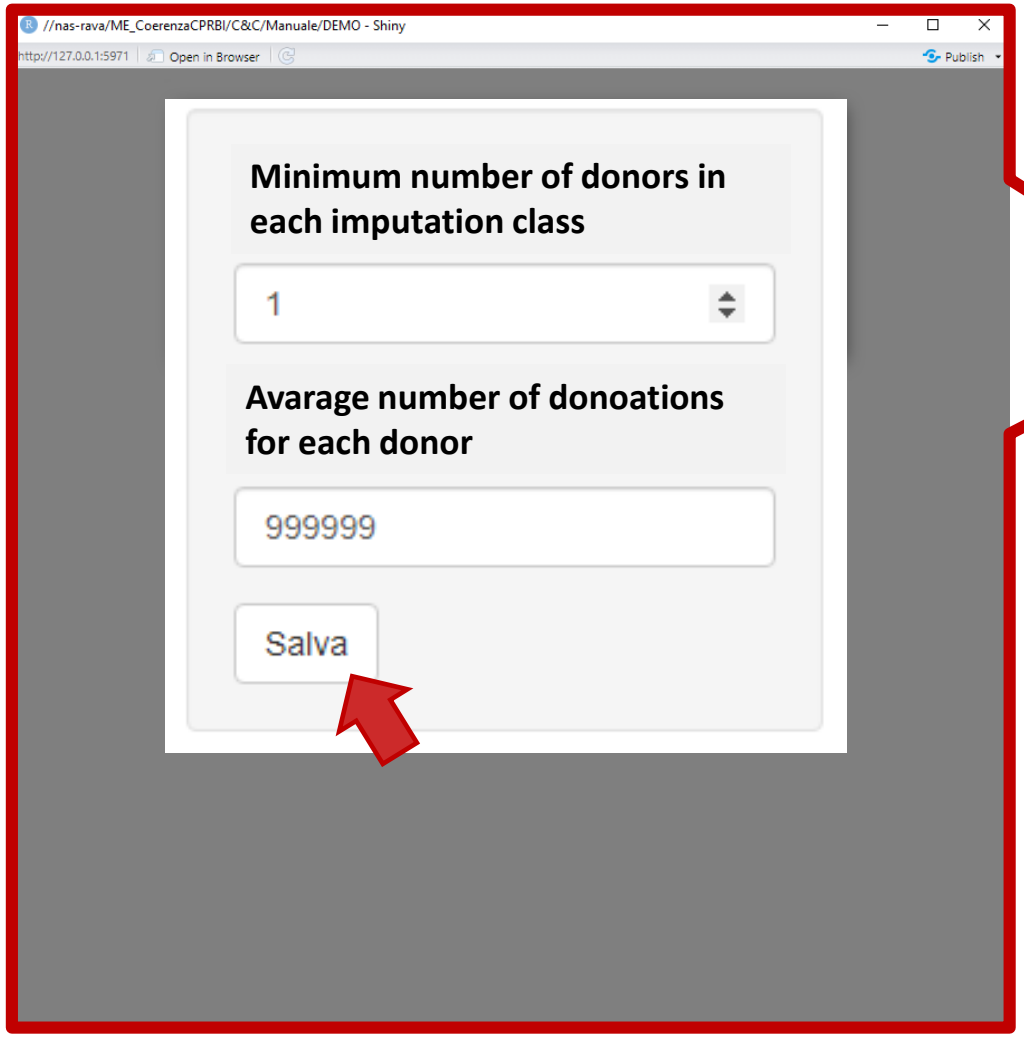
Phase III: Imputation

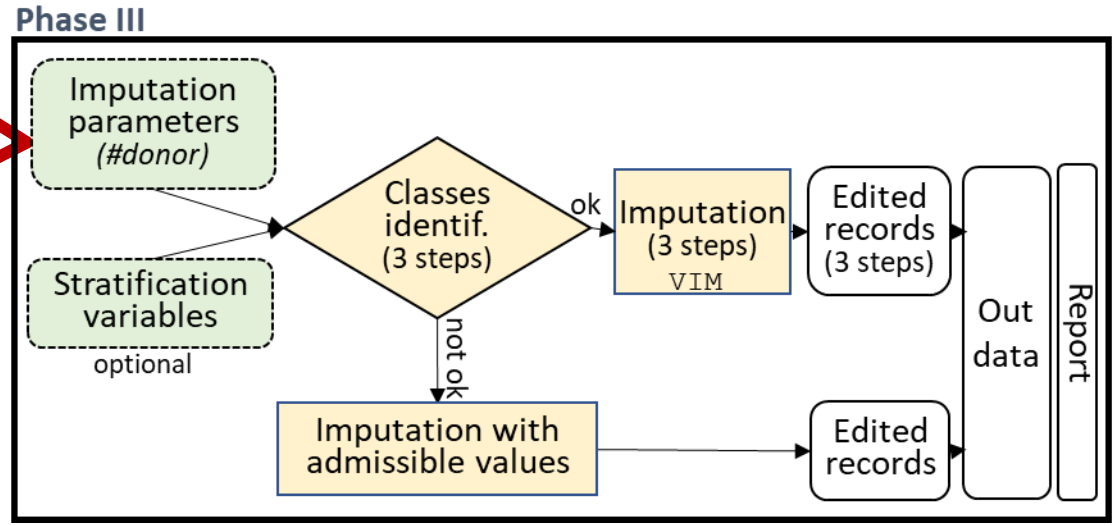
- Errors are replaced with admissible values that restore record consistency
- Record validity is ensured through the definition of appropriate donor pools
  - donors are randomly selected among error-free records
  - whenever feasible, erroneous variables are imputed using the same donor

  ➡ this approach ensures the preservation of marginal and, possibly, joint distributions
- User-defined stratification variables can be used to partition the donor pool into strata
- `hotdeck` function from the `VIM` library is used



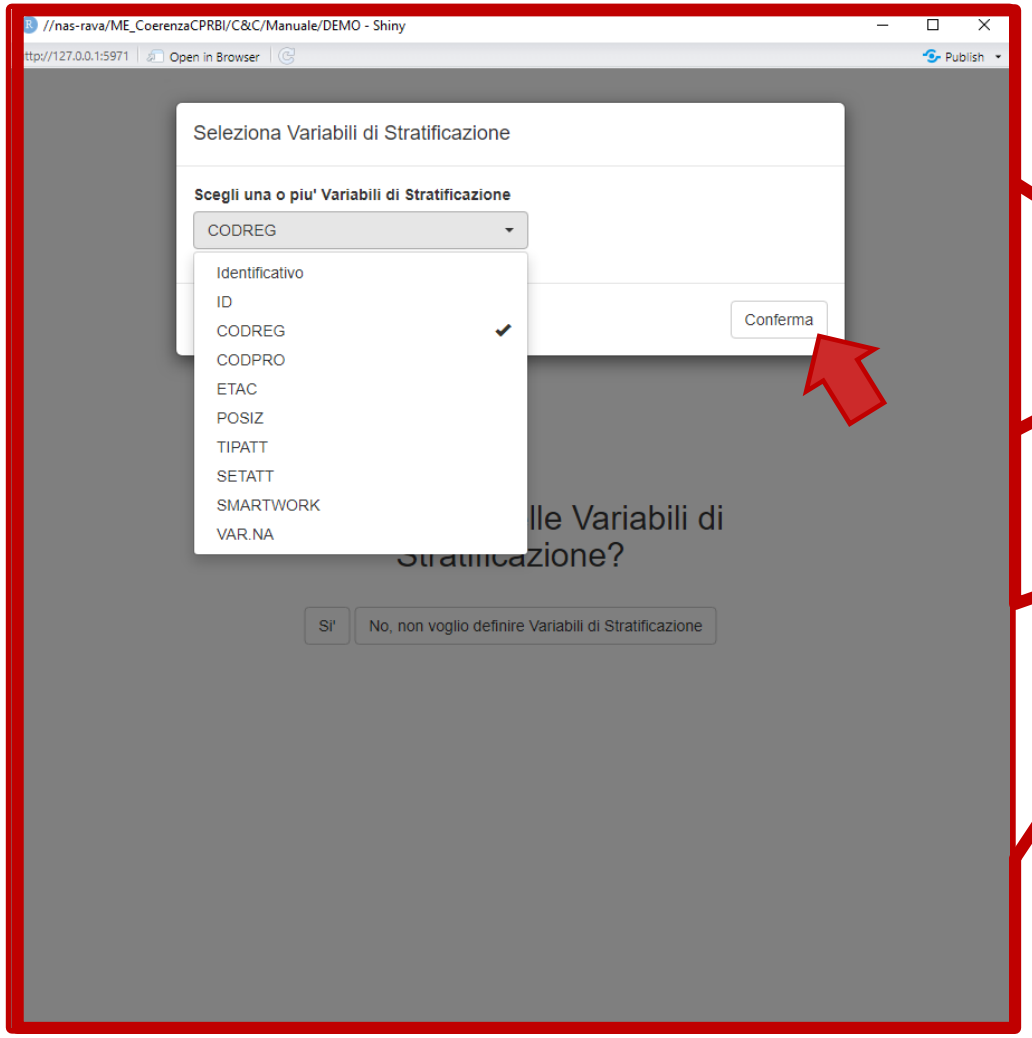
Phase III: Imputation

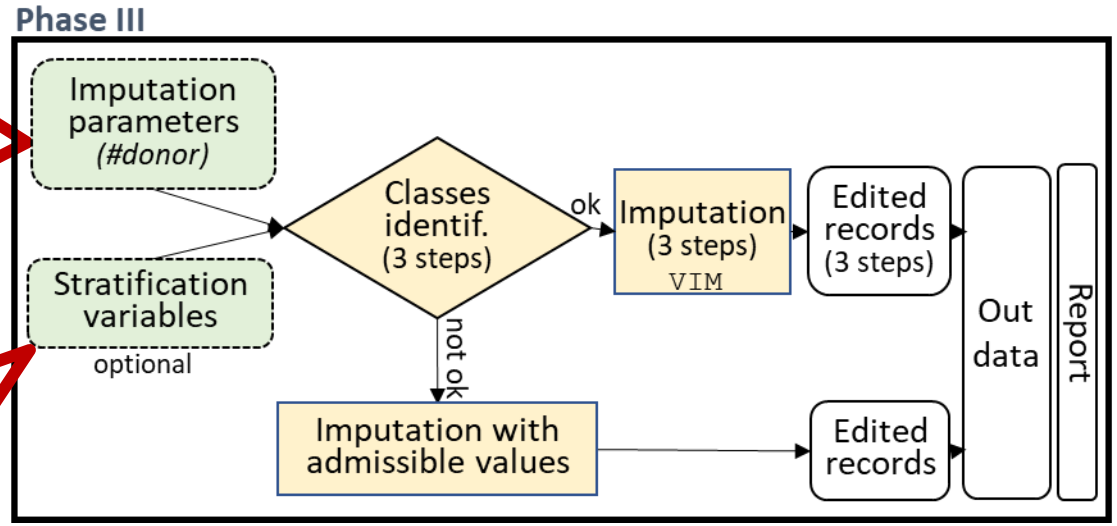




- The interaction with the user is managed through shiny apps.

Phase III: Imputation

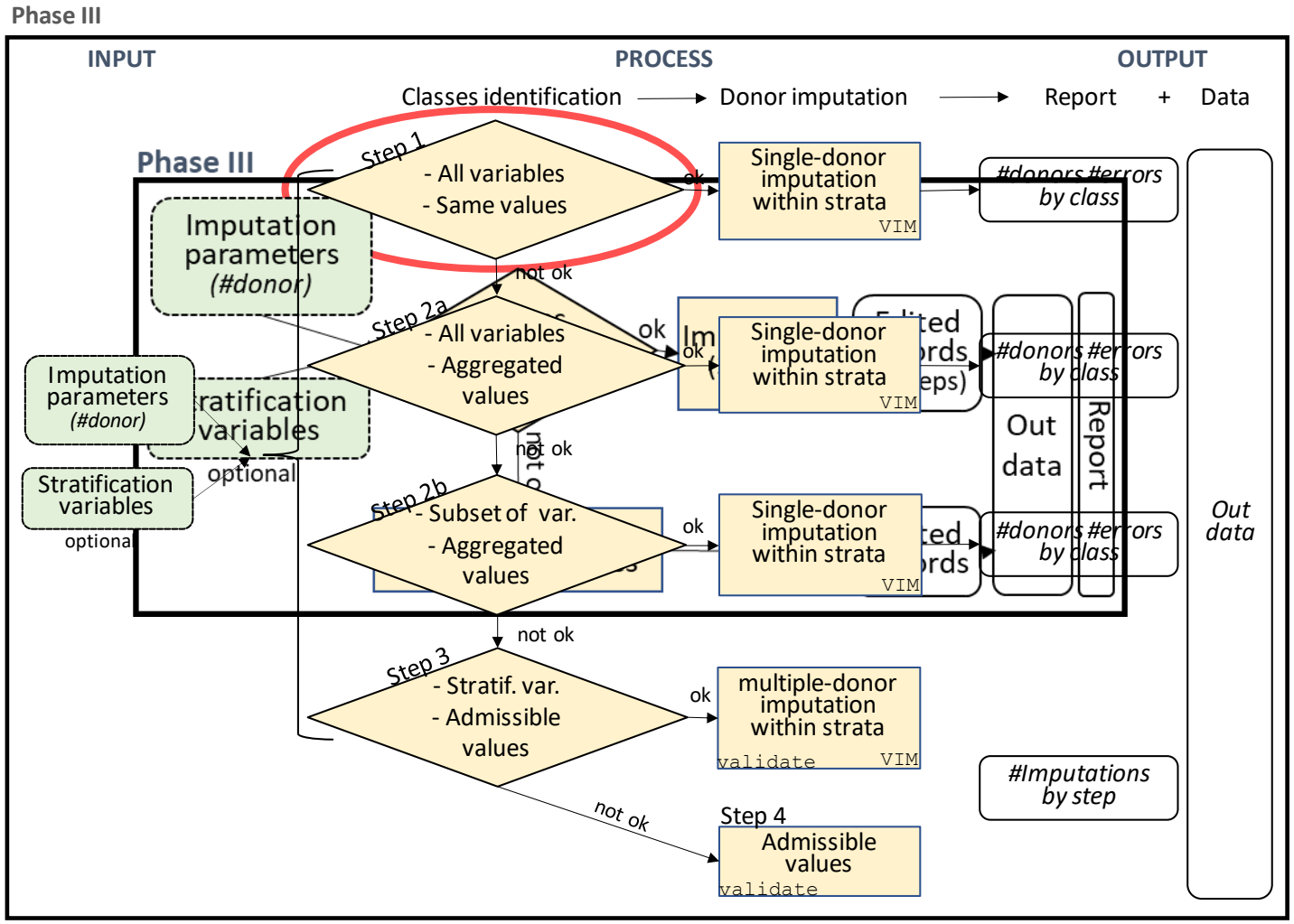




- The interaction with the user is managed through shiny apps.

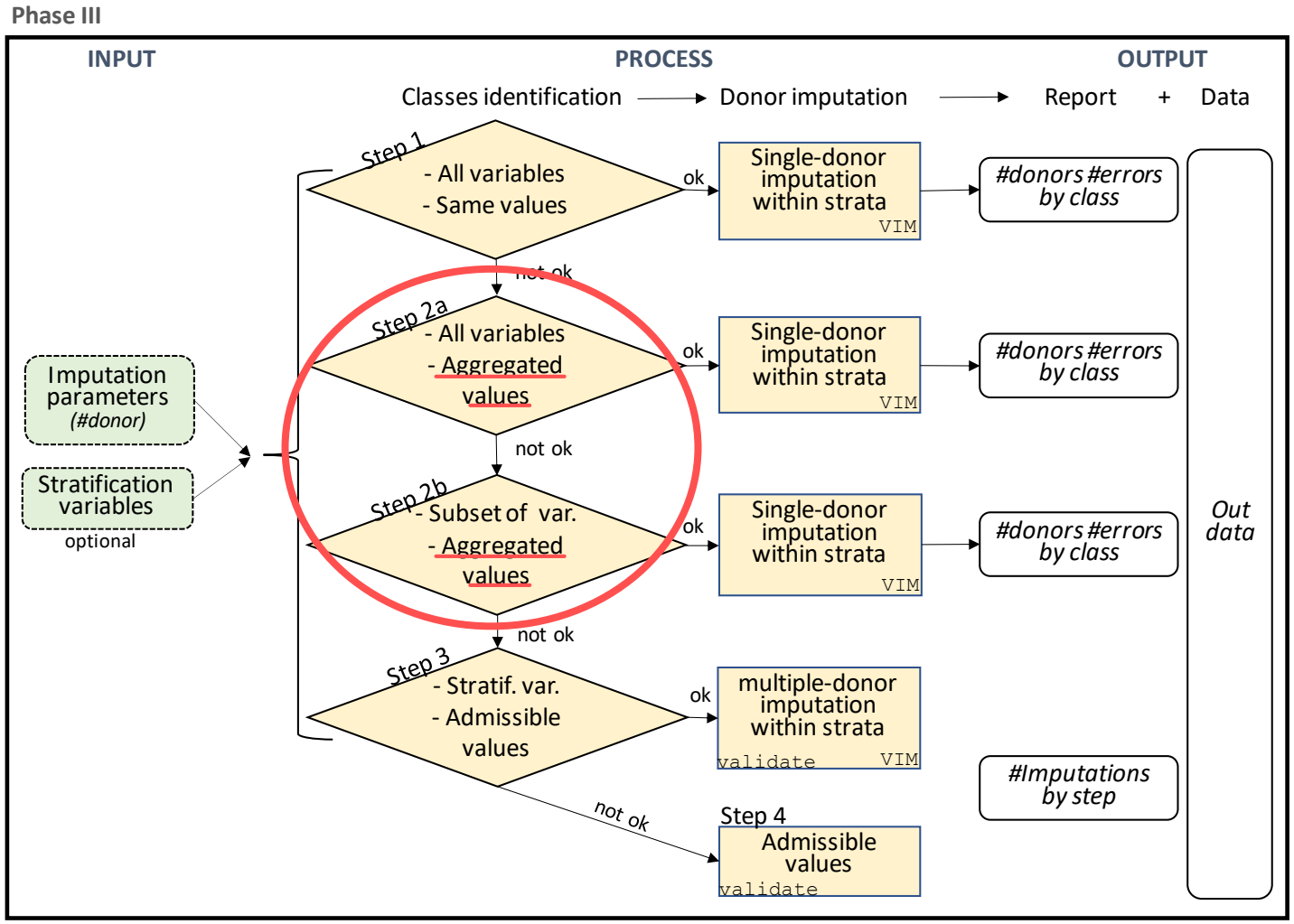
Phase III: Imputation Step 1

- All variables
  - Imputation classes are defined on the basis of all variables involved in the rules (but not subject to imputation), as well as any user-defined stratification variables.
- Same values
  - For each erroneous record, a donor is searched within the subset of individuals having the same values for all other variables involved in the set of consistency rules.
- Single donor
  - Missing values in the record are imputed using the same donor record.



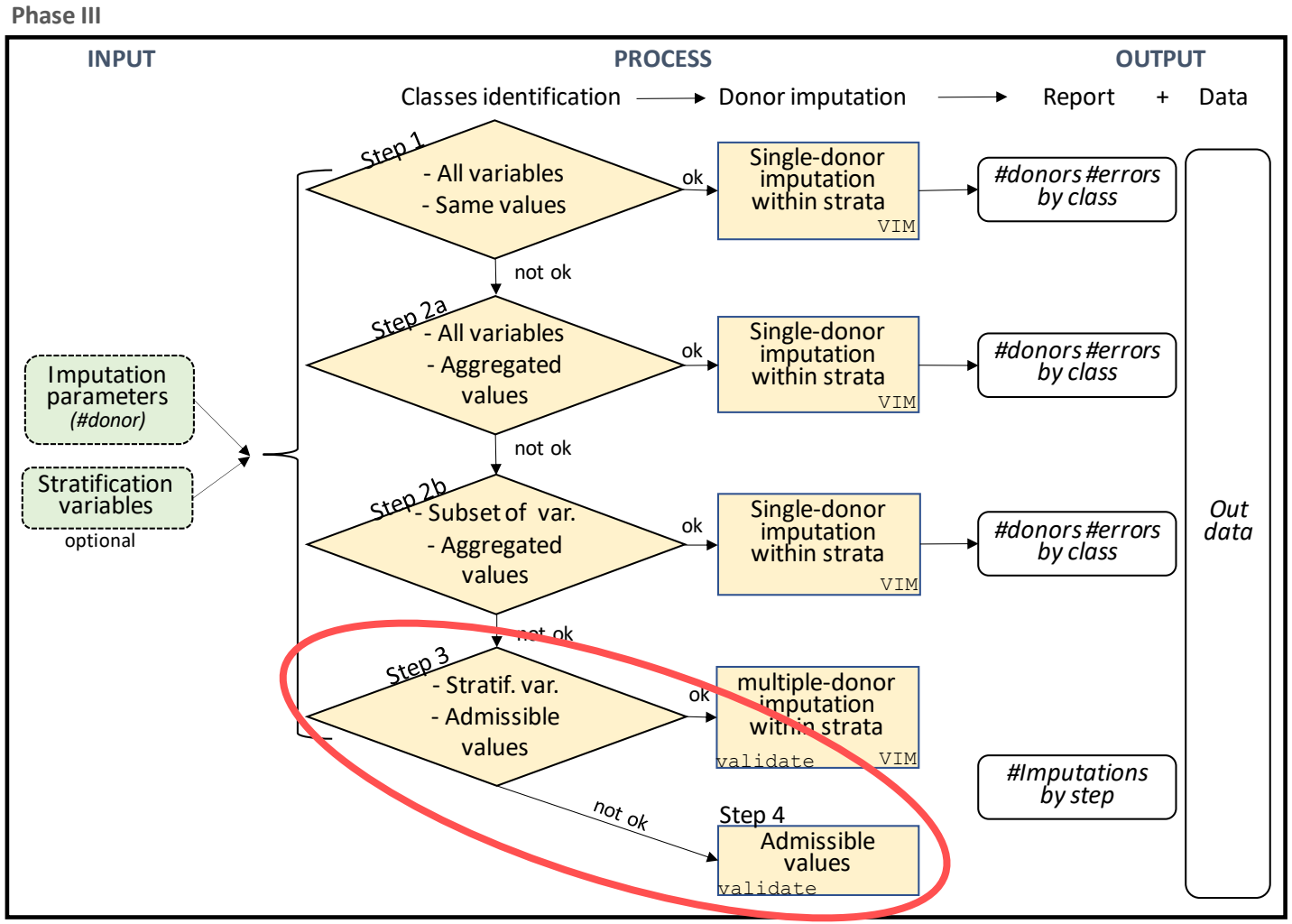
Phase III: Imputation Step 2

- Aggregated values
  - are defined from consistency rules: records in each group violate or satisfy the same rules
- Step 2a: All variables
  - Imputation classes are defined on the basis of all variables involved in the rules, and on stratification variables.
- Step 2b: Subset of variables
  - Imputation classes are defined on the basis of variables involved in the same rules as those to be imputed.
- Single donor
  - Missing values in the record are imputed using the same donor record.



Phase III: Imputation Step 3-4

- Admissible values
  - For each remaining record, admissible values are defined according to the consistency plan, considering the observed values.
- Step 3: Stratification variables
  - Donors with values within the admissible set are selected from the same stratum
- Step 4: No stratification variables
  - Donors with values within the admissible set are selected from the entire dataset
- Multiple donor
  - different donor may be selected for the same recipient record



Case study: Employment Characteristics in PC survey

The data and rule set are stored in a dedicated working directory (together with the program):

- consistency plan: 9 validity rules and 23 consistency rules
- variables: employment status (POSIZ), type of contract (TEMPP), type of occupation (TIPATT), economic activity sector (SETATT), and indicator for remote work (SMARTWORK).

At the end of each of the 3 phases:

- output files are automatically stored in the directory, named "out", within the working folder
- the user may decide whether to continue to the next phase or exit the procedure
- a confirmation message is displayed

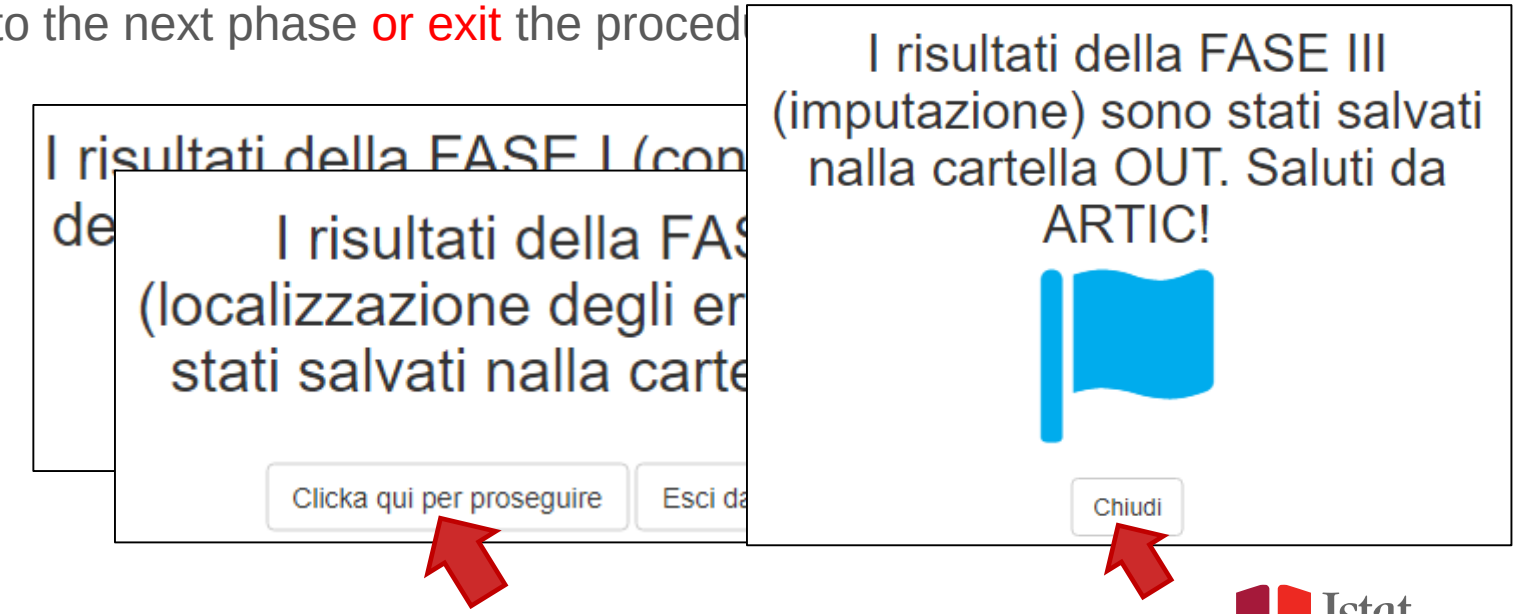
Case study: Employment Characteristics in PC survey

The data and rule set are stored in a dedicated working directory (together with the program):

- consistency plan: 9 validity rules and 23 consistency rules
- variables: employment status (POSIZ), type of contract (TEMPP), type of occupation (TIPATT), economic activity sector (SETATT), and indicator for remote work (SMARTWORK).

At the end of each of the 3 phases:

- output files are automatically stored in the directory, named "out", within the working folder
- the user may decide whether to continue to the next phase or exit the proced
- a confirmation message is displayed



Case study: Data check

- Upon initialization, ARTIC performs a logical coherence check on the rule set

- The data-checking phase produces a comprehensive report summarizing the number of rule violations for each edit constraint

id	regola	N_esatti	N_errati	di.cui.inc oerenti	perc_ errati	perc_inc oerenti
V01	POSIZ %in% as.character(1:8)	823540	0	NA	0	NA
V02	TEMPP %in% as.character(1:2)	821684	1856	NA	0.23	NA
V03	TIPATT %in% as.character(1:10)	791827	31713	NA	3.85	NA
V04	SETATT %in% as.character(1:21)	816455	7085	NA	0.86	NA
V05	SMARTWORK %in% as.character(1:4)	815435	8105	NA	0.98	NA
V06	TITSTU %in% as.character(1:17)	823540	0	NA	0	NA
V07	CITTAD %in% as.character(1:3)	823540	0	NA	0	NA
V08	ETAC >=0	823540	0	NA	0	NA
V09	ETAC <=114	823540	0	NA	0	NA
V10	if (POSIZ==5)TITSTU%in%5:17	823540	0	0	0	0
V11	if (POSIZ==4) TIPATT==9	822679	861	0	0.1	0
V12	if (POSIZ==5) TIPATT %in% 1:9	820680	2860	22	0.35	0
V13	if (POSIZ==6) TIPATT %in% c(1:5,7:9)	818366	5174	1785	0.63	0.22
V14	if (POSIZ==7) TIPATT %in% 1:9	823266	274	6	0.03	0
V15	if (POSIZ==8) TIPATT %in% 1:6	822416	1124	621	0.14	0.08
V16	if (POSIZ %in% c(4,5)) SETATT %in% c(1:14,16:	822756	784	113	0.1	0.01
V17	if (POSIZ %in% c(6,7)) SETATT %in% c(1:14,16:	822383	1157	34	0.14	0
V18	if (POSIZ==8) SETATT %in% c(1:14,16:20)	823301	239	4	0.03	0
V19	if (TIPATT==8) TITSTU %in% 5:17	822539	1001	1001	0.12	0.12
V20	if (TIPATT==10) POSIZ==1	823418	122	122	0.01	0.01
V21	if (TIPATT==4) SETATT %in% c(1,3,6,7,9,15,18,1	823329	211	80	0.03	0.01
V22	if (TIPATT==5) SETATT %in% c(1,3:20)	822457	1083	27	0.13	0
V23	if (TIPATT==9) SETATT %in% c(1:19,21)	823330	210	19	0.03	0

Case study: Error localization

The dataset of erroneous records is automatically loaded

- Age and Citizenship, corrected in earlier stages, are specified as fixed
- Employment Status (POSIZ) is considered more reliable: it is assigned a reliability level of 2 (while the other variables retain a level of 1)

A synthetic report summarizing the localized erroneous variables is then produced

```
[[1]]
        VAR   ERR
5    TIPATT 35448
7 SMARTWORK  8105
6    SETATT  7349
4     TEMPP  1856
3     POSIZ     5
2      ETAC     0
8    TITSTU     0
9    CITTAD     0

[[2]]
  N_ERR N_REC
1     1 31971
2     2  8815
3     3  1054

[[3]]
                  VAR_NA  N_REC
1                      0 781700
2                  POSIZ      4
3          POSIZ&SETATT      1
4                 SETATT   5585
5       TEMPP&SMARTWORK    802
6 TEMPP&TIPATT&SMARTWORK   1054
7                 TIPATT  26382
8         TIPATT&SETATT   1763
9       TIPATT&SMARTWORK   6249
```


Istat

Case study: Imputation

The dataset (erroneous records + error-free records) is automatically loaded

- Region of residence and age class are selected as stratification variables
- Default parameters for the hot-deck imputation are retained (donor per stratum: 1; maximum n. of donations: 999,999)

➡ Most imputations are carried out in step 1

- No forced imputation

Tabella 1: Frequenza di passo

PASSO	FREQ	PERC
0	1850859	95.28
1	88389	4.55
2	3108	0.16
2b	13	0
3	146	0.01

Tabella 2: Frequenze di passo sui soli errati

PASSO	FREQ	PERC
1	88389	96.44
2	3108	3.39
2b	13	0.01
3	146	0.16

Case study: Quality of imputation

Imputation classes: subset of records within which donors are selected

- based on the stratification variables and the variables involved in the rule set.

Assists in the specification of imputation parameters

For each imputation step

donor per stratum

max # of donations

Var.di.Str	MOD.DOM	Freq. esatti	Freq. err	Perc.er .su.es	VAR.NA
CODREG CLETA POSIZ TIPATT ETAC	1_ 1_1_3_ 16	6	1	16.67	SETATT
CODREG CLETA POSIZ TIPATT ETAC	1_ 1_1_3_ 17	14	1	7.14	SETATT
CODREG CLETA POSIZ TIPATT ETAC	1_ 1_1_5_ 16	4	1	25	SETATT
CODREG CLETA POSIZ TIPATT ETAC	1_ 1_1_5_ 17	21	2	9.52	SETATT
CODREG CLETA POSIZ TIPATT ETAC	1_ 2_1_1_ 21	319	2	0.63	SETATT
CODREG CLETA POSIZ ETAC	20_ 2_1_ 18	66	1	1.52	ETATT&SMARTWORK
CODREG CLETA POSIZ ETAC	20_ 4_1_ 30	859	1	0.12	ETATT&SMARTWORK
CODREG CLETA POSIZ ETAC	20_ 6_2_ 43	22	1	4.55	ETATT&SMARTWORK
CODREG CLETA TIPATT SETATT ETAC	3_ 1_1_6_ 17	10	1	10	POSIZ
CODREG CLETA TIPATT SETATT ETAC	5_ 1_3_3_ 17	18	1	5.56	POSIZ
CODREG CLETA TIPATT SETATT ETAC	7_ 1_3_3_ 17	1	1	100	POSIZ
CODREG CLETA TIPATT SETATT ETAC	15_ 1_5_3_ 17	0	1	NA	POSIZ
CODREG CLETA ETAC	17_8_54	694	1	0.14	ETATT&SMARTWORK

Conclusions

- The procedure has been fully developed in R (free from licensing costs and easy to customize or extend).
- Is accessible to non-expert users of R: tasks such as data loading, parameter specification, and other interactions are managed through interactive web applications developed using the shiny library.
- Several sections of the code were parallelized using functions from the parallel package to enhance execution efficiency.
- Outputs are generated to provide diagnostic information and assess the quality of imputation
- The procedure is accompanied by a detailed manual (currently available in Italian)
- Provides a reproducible, extensible, and shareable framework

ROMINA FILIPPINI | filippini@istat.it

SIMONA TOTI | toti@istat.it

Istat | Istituto Nazionale di Statistica