

3-6 October 2022, (virtual)

EXPERT MEETING ON STATISTICAL DATA EDITING

# THE SCIA SYSTEM IMPLEMENTING THE FELLEGI-HOLT METHODOLOGY COMPARED TO THE RECENT R PACKAGES

MARIA TERESA BUGLIELLI, ROMINA FILIPPINI, **SIMONA ROSATI**

Istat | DCME

# Outline

---

- Introduction
- The SCIA system
- R for editing and imputation
- The experimentation
- Results
- Conclusions and future developments

# Introduction


---

- In 90s, at Istat, the SCIA system was developed to implement the Fellegi-Holt methodology
- Only qualitative variables can be handled correctly by the system and an update of the system is required
- But the programming languages used are too obsolete to proceed with a new release
- Before planning a new project, we investigated among the most recent R packages for E&I in order to find an alternative option
- Performance between R and SCIA were compared

# The SCIA system

---

Open source software developed at Istat for automatic editing and imputation of **random errors in qualitative variables**. It implements the **Fellegi-Holt** methodology based on the minimum change principle (Fellegi and Holt, 1976)

- 
- **Check** the data with respect to explicit edit rules
  - **Localization** of erroneous or missing data: identify the minimum number of variables to be modified to restore consistency
  - **Correction/imputation** of erroneous and missing data: preserving the marginal and joint distributions of the original variables
- 
- Derivation of the **implied edit rules**
  - Ensuring the final **“correctness”** of records

# R for editing and imputation

---

## Packages used:

- **validate**: provides functions to formulate validation rules written as positive logical formulas, to confront data and analyze or visualize the results
- **validatetools**: a set of functions for finding redundancies or contradictions between the rules formulated with validate
- **errorlocate**: allows to localize errors given a set of rules according to the Fellegi-Holt algorithm
- **VIM**: provides the functions to impute missing values as kNN and hotdeck

# The experimentation

---

- R and SCIA were both applied to the section of the questionnaire that collected information on the farm manager and on the other gainful activities (OGA) directly related to the farm (section D).
- The questionnaire section under study consisted mainly of qualitative variables (**43 variables were treated**).
- The number of **explicit edits** specified by the experts was **119** (for **R**) and **83** (for **SCIA**) including field validation rules. The different number of edits was due to the specific syntax required by the two software. Failures expressed by the edits were conceptually identical.
- As far as **SCIA** is concerned, the **complete set of edits**, containing also the essential implied edits logically derived from the explicit edits, amounted to **104** edit rules in all.
  - More than 1 million of records were submitted to the E&I process.
  - The performance of the two software was compared in terms of edit failures and localized errors

# Results: Edit Failures

Consistency rule	N. failures
if (ATT_CON == 2) TEMPO_ATT_CONN <= 0	70,191
if (TEMPO_ATT_EXTRA < 3) SETT_AGR + SETT_EXTRA_AGR >=1 and SETT_AGR + SETT_EXTRA_AGR <4	979
if (ATT_CONR + ATT_CONQ < 4 and (ATT_CONR >0 or ATT_CONQ>0)) ORE_TERZATT >= 1	224
if (ATT_CON ==1) PERC_ATT_CON_REN >= 1 PERC_ATT_CON_REN <= 100	74
if (ATT_CONQ == 1 ) ORE_TERZATT >= 0 ORE_TERZATT <= 99999	62
if (ATT_CONR == 1 ) ORE_TERZATT >= 0 ORE_TERZATT <= 99999	17
if (CAPO_AZ == 2) CAPO_REL >=1 CAPO_REL <=5	3
if (ATT_CON ==1) ATT_CON_REN %in% c(A,...,Z)	1

- Data check carried out through R and SCIA returns similar results, meaning that the two set of rules are essentially the same
- Only eights edits failed
- The highest number of failures concerned the “Time dedicated to related activities” (TEMPO\_ATT\_CONN) that takes values greater than zero when no related activities (ATT\_CON) are declared.
- SCIA: only logical edits can be specified, hence additional variables were needed to correctly specify the highlighted edits in the table

# Results: Error Localization

1/2

Number of erroneous values per variable (excluding field errors)

Variable	SCIA	test R1	test R2	R1 - SCIA	R2-SCIA
TEMPO_ATT_CONN	70,198	70,198	70,198	0	0
SETT_AGR plus SETT_EXTRA_AGR	908	938	935	30	27
ORE_TERZATT	146	188	178	42	32
ATT_CONQ plus ATT_CONR	80	38	48	-42	-32
PERC_ATT_CON_REN	74	74	37	0	-37
TEMPO_ATT_EXTR	71	52	52	-19	-19
CAPO_AZ	3	3	3	0	0
ATT_CON	0	0	37	0	37
Total	71,480	71,491	71,488	11	8

- Two tests were carried out with R
  - **test R1: fixity weights** were specified for those more reliable variables that could not be changed
  - **test R2:** weights were not specified

- **R with weights**
  - minor differences were found from SCIA
  - localization process with R produced a consistent solution with the set of the specified edits
- **R without weights**
  - unexpected outcome: the variable ATT\_CON was localized as an error, i.e. if it is modified by imputation, one of the explicit edits will failed
  - Is it a lack in the error localization algorithm, which did not consider the entire set of explicit edits?
- **SCIA**
  - SCIA found a consistent solution in any case

# Conclusions and future developments

---

- Different performance between R and SCIA, mainly due to whether or not the complete set of edits is considered to localize the errors
- The SCIA system provides an effective and efficient one-step solution
- In R, **errorlocate** function may take long time
- R approach may take more than one step to arrive at an acceptable solution
- R offers more flexible tools than SCIA for managing large and complex data sets, and it allows mixed types of variables to be treated jointly
- R packages are extensible; other functionalities or features can be improved or developed
- **More investment should be made in using R in the E&I process**

# thank you

MARIA TERESA BUGLIELLI | [bugliell@istat.it](mailto:bugliell@istat.it)

ROMINA FILIPPINI | [filippini@istat.it](mailto:filippini@istat.it)

**SIMONA ROSATI** | [sirosati@istat.it](mailto:sirosati@istat.it)