

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 4 of the provisional agenda

EVALUATING EDITING PROCEDURES: THE SIMULATION APPROACH

Submitted by the Italian National Institute of Statistics ¹

¹ Prepared by Giulio Barcaroli and Leandro D'Aurizio.

Abstract

The problem of evaluating a given editing procedure can be solved by adopting basically two different approaches: by re-doing more carefully all or some steps of the process (starting from the interviews, or limiting it to the data entry and/or to the check), or by simulating sets of “true” and “raw” data: in any case, clean data are compared to true data. The first approach is conceptually preferable, as it is more adherent to the real situation, but it is often expensive in terms of resources (even if it could be limited to a subset of units). The second one strongly depends on the models that are chosen to generate true and raw data, that can be far from representing the real world. In any case, under both approaches, a set of indicators are required to assess the capability of the editing procedure to detect errors and to correct them by imputing the true values. In this paper, some indications related to the problem of the preparation of test data sets are given, with particular regard to the case of a simulation approach, and a set of indicators is proposed in order to evaluate in a correct and complete way the performance of the editing procedure. A number of studies report experiences in the field, that make use of some methods rather than others (Granquist 1997), (Stefanowicz 1997). This paper is a proposal of a standard methodology of evaluation that can be applied in a variety of situations.

1. INTRODUCTION

1. The impact of nonsampling errors on the final estimates of a survey is usually non negligible, and often of the same magnitude, if not greater, of sampling errors. The purpose of an editing procedure is to find and correct as many errors as possible, in order to limit this impact². The problems related to the optimal design, implementation and tuning of an edit procedure lead to the following question: “how is it possible to evaluate an editing procedure?”, in other words, what methods and indicators one could use to know if a given procedure is satisfactory or not?

2. To evaluate the performance of a procedure, i.e. its desired capability to detect errors and correct them, together with its undesired effect of introducing new errors, it is necessary to ensure the availability of at least three data sets:

- a) the set of “true” data, with a perfect correspondence to the situation in the real world;
- b) the set of “raw” data, that can be seen from a double point of view: as the data that are available after the phases of data collection and data entry, or as the true data with errors in them;
- c) the set of “clean” data, the result of the application of the editing procedure to raw data.

For each variable, by comparing true and raw data we can know what values are wrong, and by comparing true and clean data we can determine how many wrong values have been detected and successfully replaced by true values.

3. At the basis of any evaluation process is the availability of true data: this is the subject of Chapter III, investigating the two opposite approaches of survey replication and data simulation, with a special attention to the latter. Under the second approach, the problem of how generating raw data starting from true data, i.e. what error generation model to adopt, has to be considered:

² At least in case of *automatic* procedures: when procedures are interactive approaches like macroediting or selective editing can be adopted and the target is no more to find and correct as many errors as possible, but only *relevant* errors, i.e. errors with strong impact on final estimates

a proposal regarding this aspect is contained in Chapter IV. Finally, in Chapter V a set of indicators to assess the quality of an editing procedure are proposed and discussed.

4. In fig.1 a general flow of the evaluation process, based on simulation techniques, is reported.

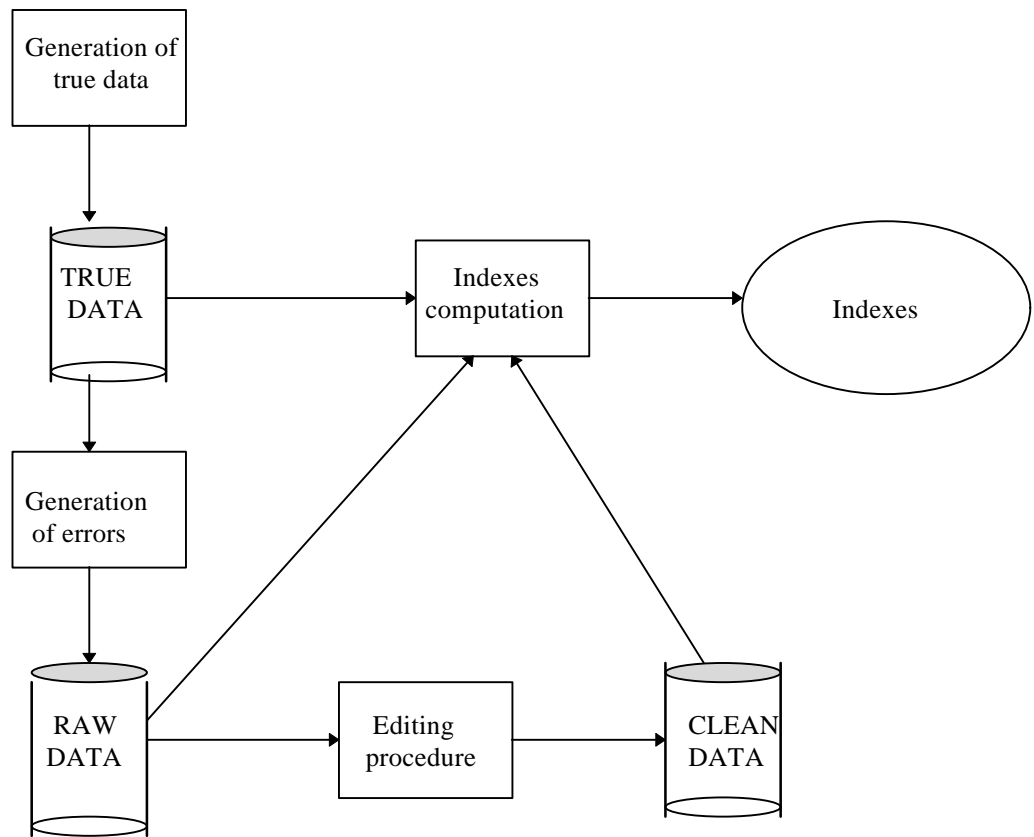


Fig.1 - General flow of the evaluation process based on simulation

II. PRODUCTION OF “TRUE” DATA

5. True data are essential to understand, for each variable, what values are to be considered wrong and, in case of imputation, if assigned values can be considered correct, or at least, not far from true values.

6. How to obtain a set of true data? As errors can arise mainly during the phases of compilation of the questionnaires, and of their memorisation (data entry), it is possible to organise a repetition of these activities under better conditions (Lessler, Kalsbeek 1995). For instance, it is possible to employ a set of interviewers chosen among the best, or supervisors can be called to make the interviews. Furthermore, interviews can be based on reconciliation of current answers that are different from the answers given by the respondent during the “normal” survey. As for data entry operations, it is possible to adopt the double typing, in order to limit keying errors. Again, it is possible to adopt different survey techniques, such as CAPI, that allow to obtain better quality by applying checks during the interview.

7. The general philosophy of this kind of approach is to do again these activities minimising the insurgence of errors by investing more resources and/or time. This is subject to two main limits:

- a) errors are still present in data obtained in this way: we can only say that their amount is lower than in normal occurrences of the survey (in best situations they are negligible);
- b) it is not always possible to afford such activities, as they can have a high cost both for the statistical institute (in terms of resources to be employed) and for the respondents (in terms of response burden): anyway, this cost can be reduced by limiting the investigation to a subset of the respondents.

8. For these reasons, sometimes it is preferable to obtain true data in a different way:

- a) by linking information coming from other sources if available (for instance data from other surveys, or administrative data, if these sources and the survey share common characters and involve the same statistical units and the same period of time (Poulsen 1997); This solution is possible if these sources exist, their data are available, and their quality is reasonably high;
- b) by considering as true data the result of the application to raw data of very careful editing operations (performed by expert personnel and based on the follow-up of respondents); This option is subject to the same limits due to higher cost of the previous approach, though in a less dramatic way, but on the other side a greater amount of errors in data is to be expected.
- c) by creating true data in an artificial way. This solution is the less expensive, but its quality depends strictly on the model that you assume to generate data. The more the generated data can represent the real situation, the more the evaluation based on this simulation approach can be assumed to be correct: the same problem, as we will see, concerns the generation of artificial raw data, i.e. the introduction of errors in true data.
- d) by considering as true data the result of the application to raw data of the same editing procedure that has to be evaluated. This is a common practice (Garcia, Peirats, 1994), the most straightforward and the less expensive, but it is clear that in this way the resulting set is by no means the set of *true* data (if it were so, there would be no need to evaluate the procedure), but rather just a *starting* set.

9. Every data set is characterised by a set of variables, each of them with a given definition domain: in case of qualitative variables, the domain is a finite set of values, in case of quantitative variables it is a range of numbers. The simplest way to generate artificial data is to obtain the Cartesian product of the domains (in case of qualitative variables), as a first step, and then to eliminate the combinations of value that are logically impossible. This solution does not take into account the distribution of real cases in the target population.

10. The optimal way to generate a representative set of true data, however, is to consider a density function of the general multivariate distribution, considering jointly all the variables in the data set. In real cases, however, it is rather difficult to define and estimate such a function. An acceptable solution is to consider many density functions, one for each variable. A density function can be *unconditioned*, i.e. the probabilities do not depend on the values assumed by other variables, or *conditioned*, in the opposite case. The latter solution is preferable when a significant relation among a subset of variables does exist, and the generation process is hierarchical, as values are assigned first to variables with unconditioned functions.

11. Let us consider the different situations concerning qualitative and quantitative variables. As for qualitative variables, in case of conditioned functions, the probability related to a single value of a given variable depends on the values assumed by a subset of other variables. In a first step, values are randomly generated for variables that are characterised by a simple probability function, and then the same process can be applied to the others that have conditioned functions. A given variable can be conditioned by one or more variables that in their turn can be conditioned by other variables: in complex cases, the generation sequence can be determined by considering a graph representing relations among variables. As a constraint the graph must be acyclic. An interesting example of creation of artificial data in the case of qualitative variables is in (Nordbotten, 1995).

12. In case of quantitative variables, use can be made of regression functions to take into account statistical relations among subsets of variables. For each regression function, values have to be assigned randomly to independent variables, characterised by a simple density function, and then proceed to compute the value of the dependent variable (it is also possible to add a random component in the regression function). Also in this case a predictive variable can be in its turn a dependent variable in a regression function, so it is necessary to determine the sequence of generation with care. An experience of quantitative data generation is reported in (Verboon, Schulte Nordholt 1997).

IV. GENERATION OF ERRORS

13. Once the set of true data is available, no regard of how they have been produced (repetition of the survey, other sources, result of an editing procedure, artificial data), it is possible to obtain raw data by introducing artificially generated errors. This choice has less meaning under the first three approaches (as in this case we do already have real raw data), but is mandatory if we do not have raw data corresponding to true data, as in the case these latter have been artificially generated.

14. It is crucial, for the correctness of the evaluation of the editing procedure, to be able to consider artificial errors as representative of real ones, both from a quantitative (incidence of non sampling errors in data) and a qualitative (type of errors) point of view. The definition of the error generation model (i.e. the assumptions on the mechanisms that in the real survey determine the insurgency of errors) should be based on a careful consideration of the organisation of the survey and on *ad hoc* quality studies.

15. On the basis of our experience, we propose to consider a model taking into account two different components of non sampling error: errors occurring in the phase of the compilation of the questionnaire, and errors occurring in the data entry phase. In both components we assume the presence of stochastic errors, but we also admit the possibility that systematic errors are present, due to imperfections in the data collection system. The main difference between the two components is that the first one involves data as *variables*, while the second one involves data as *bytes*, which has important implications concerning the mechanism of error production.

16. According to these assumptions, the generation of errors is made record by record and is characterised by the following sequence:

- a) generation of stochastic and systematic errors corresponding to the phase of questionnaires compilation and involving any interested variable as a whole: item non responses,

misplacement errors (the answer to a question given to another question), misunderstanding of measure units (for instance, units instead of thousands), interchange errors (characterised by an error in the mark of a category for qualitative variables, or by a change of digits for quantitative variables), errors dependent on structural misunderstanding of routing conditions, or caused by a systematic tendency to under or over-report the answers to a given question;

- b) generation of stochastic errors representing those occurring during the data entry phase, involving bytes one by one: typically, keying errors producing interchange of values (in a different sense from the interchange errors seen in the compilation phase, because now these errors can produce also out-of-domain values for qualitative variables).

17. In the implementation of the program, for any type of errors it is necessary to define:

- the incidence of the error, i.e. the probability of occurrence;
- the mechanism of generation, i.e. the set of rules to apply in order to determine the new value to assign to the involved variable.

The mechanism of errors generation is based on a Monte Carlo approach (Kleijnen, Van Groenendaal, 1992).

18. In the following, we give examples concerning the generation mechanism for any type of error. We refer to SAS language statements; in particular, the function

RANUNI (L,U)

generates randomly a number belonging to the interval L-U, where the density function is uniform.

The generic variable to be perturbed is named X; the single byte to be modified is named B.

The incidence of every error type is given by a real number belonging to the interval 0-1.

19. **Item non responses** (incidence: I1)

```
P= RANUNI (0,1);
IF P < I1 THEN DO;
    X = . ;
END;
```

20. **Misplacement errors** (incidence: I2)

If X is close to another variable Y in the sequence of questions, Y is of the same type of X (both quantitative or both qualitative), and the possibility of misplacement is admitted, then:

```
P = RANUNI (0,1);
IF P < I2 THEN DO;
    Z = X;
    X = Y;
    Y = Z;
END;
```

21. **Loss or addition of zeroes** (incidence: I3)

If X is a quantitative variable, and its rightmost digits are zeroes, it can be perturbed by multiplying or dividing by zero:

```
P = RANUNI (0,1);
IF P < I3 THEN DO;
    P1 = RANUNI (0,1);
    IF P1 < 0.5 THEN X = X * 10;
    ELSE X = X / 10;
END;
```

22. **Interchange errors** (incidence: I4)

If X is a qualitative variable, and $S = \{1,2,\dots,n\}$ is the set of its possible values:

```
P = RANUNI (0,1);
IF P < I4 THEN DO;
    X = INT(RANUNI (1,n));
END;
```

If X is a quantitative variable, with k digits:

```
P = RANUNI (0,1);
IF P < I4 THEN DO;
    K1 = INT(RANUNI(1,k);
    K2 = INT(RANUNI(1,k);
    ALFA_X = X;
    A = SUBSTR(ALFA_X, K1, 1);
    SUBSTR(ALFA_X, K1, 1) = SUBSTR(ALFA_X, K2, 1);
    SUBSTR(ALFA_X, K2, 1) = A
END;
```

23. **Routing errors** (incidence: I5)

If an answer in question X is conditioned by values assumed by another variable Y, we have two cases.

If X is to be non missing when $Y \in \{\dots\}$, then:

```
IF Y IN (...) THEN DO;
    P = RANUNI (0,1);
    IF P < I5 THEN DO;
        X = . ;
    END
END;
```

If X is to be missing when $Y \notin \{\dots\}$, then:

```
IF Y NOT IN (...) THEN DO;
    P = RANUNI (0,1);
    IF P < I5 THEN DO;
```

```

        X = INT(RANUNI (1,n)) ;
    END;
END;

```

24. Under/over-report errors

If X is a quantitative variable affected by under-reporting, i.e. answers are lower than the real situation, with a non ignorable systematic mechanism (for instance: the greater the income, the greater the under-report), then:

```

X = X - X * ( RANUNI (0, X / MAX(X) ) );

```

25. Keying errors (incidence: I6)

B is a single byte in the record. Then:

```

P=RANUNI (0,1);
IF P < I6 THEN DO;
    B = INT( RANUNI (1,9));
END;

```

IV. INDICATORS FOR THE EVALUATION

26. We want to evaluate the performance of an editing procedure for all the variables observed on a set of units. The process introduced here is to be performed for each variable of interest.

A file of true values is available, labelled as F1, containing a set of V records, each one corresponding to one unit. We will indicate with $|V|$ the number of elements of a set V, from now on.

An alteration mechanism is used to get a file of altered data, labelled as F2: it is composed by the sets V_1 and V_2 , containing unaltered and altered records respectively.

We define as *alteration index* the following ratio:

$$p = \frac{|V_2|}{|V|}$$

27. Different ratios of altered records are useful to test the editing procedure under increasingly difficult situations, so it is recommended to make different assumptions concerning the incidences of errors, from optimistic to pessimistic, in order to cover a thorough range of operational conditions.

28. An editing procedure is a sequence of logical steps, after which every available record is defined to be true or wrong and a correction can be tried only in the latter case.

We want to assess whether the editing procedure performs correctly. Let us assume that the procedure is based on two steps: wrong records are located in the first one and the outcome is a file containing some records with an error flag. This file is labelled as F3 and it is composed by the following sets:

V_3 = set of flagged records

V_4 = set of not flagged records

29. These two sets can be decomposed:

$V_3 = V_5 \cup V_6$, where $V_5 \subset V_1$ and $V_6 \subset V_2$

$V_4 = V_7 \cup V_8$, where $V_7 \subset V_2$ and $V_8 \subset V_1$

It must moreover be taken into account that $V_1 = V_5 \cup V_8$ and $V_2 = V_6 \cup V_7$.

So V_5 are the records correctly deemed to be wrong, whereas V_7 are the ones erroneously judged to be correct. These two sets are the first sources of failure for the procedure.

30. We can now define the following three indices:

a) **wrong value identification capability**

$$C_I = \frac{|V_6|}{|V_2|} \in [0,1]$$

is the index that quantifies the ability of the procedure to identify correctly the errors (wrong values) in raw data.

b) **wrong value identification error**

$$E_I = \frac{|V_5|}{|V_1|} \in [0,1]$$

is the index that quantifies the failures of the procedure when it identifies true data as wrong (it can be considered as an *error of the first type*).

c) **correct value identification error**

$$E_{II} = \frac{|V_7|}{|V_2|} \in [0,1]$$

is the index that quantifies the failures of the procedure when it identifies wrong data as true (it can be considered as an *error of the second type*).

31. The three terms are not independent. This can be easily shown:

$$E_{II} = \frac{|V_7|}{|V_2|} = \frac{|V_2| - |V_6|}{|V_2|} = 1 - C_I$$

Index C_I , together with errors of first and second type E_I and E_{II} , is useful to assess the performance of the procedure in the phase of errors detection.

32. The editing procedure imputes all the flagged records it is able to impute in the second step, generating a clean file, labelled as F4.

33. The four files we have defined so far contain the same set of records. The value of the only variable we are studying can change in the following ways:

- it can be altered by the alteration mechanism, passing from F1 to F2;
- an error flag can be added, passing from F2 to F3;

– a flagged value can be modified, passing from F3 to F4.

34. F4 is composed by two sets of records:

V_9 = set of imputed records

V_{10} = set of non imputed records

We impose a consistency constraint on our editing procedure, meaning that it can impute only previously flagged records: $V_9 \subset V_3$

We can decompose V_9 and V_{10} too:

$V_9 = V_{11} \cup V_{12}$, where $V_{11} \subset V_5 \subset V_1$ and $V_{12} \subset V_6 \subset V_2$

$V_{10} = V_{13} \cup V_{14}$, where $V_{13} \subset V_3$ and $V_{14} = V_4 = V_7 \cup V_8$

So V_{11} represents the set of true records that are imputed, as a direct aftermath of the fact that E_1 is greater than zero.

35. Let us introduce the *operational efficiency index* of the editing procedure:

$$I_{\text{EFF}} = \frac{|V_9|}{|V_3|} \in [0,1]$$

this index quantifies the capability of the procedure to impute flagged records.

It follows that V_{10} is formed both by records which must not be imputed (V_4) and by records which should be imputed but are not (V_{13}), because of inner faults of the imputing procedure, expressed through $I_{\text{EFF}} < 1$.

We suppose that I_{EFF} is uniform over all subsets of V_3 and V_9 . This uniformity can be expressed as:

$$\forall V_3^1 \subset V_3, \text{ if } V_9^1 = V_3^1 \cap V_9 \Rightarrow \frac{|V_9^1|}{|V_3^1|} = \frac{|V_9|}{|V_3|} = I_{\text{EFF}}$$

This is a reasonable assumption: we require that the imputation procedure behaviour, as far as the number of imputed records is concerned, is not dependent on the particular set of flagged records.

36. F4 can alternatively be seen as formed by:

- really true records (V_{ES}),
- really wrong records (V_{ERR}).

Let us have a closer look at how these two sets are structured. Having this aim in mind, we decompose V_9 again, by writing:

$$V_{12} = V_{15} \cup V_{16} ,$$

where V_{15} and V_{16} are respectively the sets of correctly and incorrectly imputed records, of all the ones correctly deemed to be wrong. So, a new source of inconsistency arise for the editing procedure: a significant percentage of wrong and flagged records not properly edited.

37. It is possible to further decompose V_{10} too, by writing:

$$V_{13} = V_{17} \cup V_{18}, \text{ where } V_{17} \subset V_6 \subset V_2 \text{ and } V_{18} \subset V_5 \subset V_1.$$

The set V_{17} is an error source directly coming from $I_{EFF} < 1$, whereas V_{18} is a strange outcome from both E_I and I_{EFF} being not ideal. Because V_{18} is formed by true records, erroneously judged wrong, but left unimputed, it is not a further source of error.

We can write:

$$V_{20} \equiv V_8, \quad V_{ES} = V_{20} \cup V_{15} \cup V_{18}, \text{ where } V_{18} \text{ is a not desirable component of the set.}$$

Being necessarily $V_{19} \equiv V_7 \subset V_{ERR}$, it then follows: $V_{ERR} = V_{11} \cup V_{16} \cup V_{17} \cup V_{19}$.

38. Finally we introduce two more indices.

a) correction capability

$$C_{II} = \frac{|V_{15}|}{|V_2|} \in [0,1]$$

quantifies the ability of the procedure to assign the true value to the variable during the phase of imputation.

b) correction error

$$E_{III} = \frac{|V_{16}|}{|V_2|} \in [0,1]$$

quantifies the opposite, i.e. the number of records imputed with a value different from the true one.

E_{III} can be derived from previously introduced terms:

$$E_{III} = \frac{|V_{16}|}{|V_2|} = \frac{I_{EFF}|V_6| - |V_{15}|}{|V_2|} = I_{EFF} C_I - C_{II},$$

where we have assumed the uniformity of I_{EFF} .

39. Given $|V_1|$ and $|V_2|$, the number of elements of all the sets composing V_{ERR} and V_{ES} can be derived using only C_I , C_{II} , E_I and I_{EFF} .

The tables 1 and 2 show the results:

Table 1

V_{ERR}	$0 < I_{EFF} < 1$	$I_{EFF} = 1$
V_{11}	$I_{EFF} E_I V_1 $	$E_I V_1 $
V_{16}	$I_{EFF} (C_I - C_{II}) V_2 $	$(C_I - C_{II}) V_2 $
V_{17}	$(1 - I_{EFF}) C_I V_2 $	0
V_7	$(1 - C_I) V_2 $	$(1 - C_I) V_2 $

Table 2

V_{ES}	$0 < I_{EFF} < 1$	$I_{EFF} = 1$
V_{15}	$C_{II} V_1 $	$C_{II} V_1 $
V_8	$(1 - E_I) V_1 $	$(1 - E_I) V_1 $
V_{18}	$(1 - I_{EFF}) E_I V_1 $	0

Results are shown even for the particular case $I_{EFF} = 1$, because an auxiliary imputation method is often added to an automatic editing procedure with $I_{EFF} < 1$, in order to get $I_{EFF} = 1$.

The results obtained under $I_{EFF} < 1$ are particularly useful when the editing procedure core, composed by the totally automatic parts, must be appraised. Many editing systems are composed by a main automatic method, designed to remove the greatest part of errors, whereas their residual part is left to ancillary methods, expressly devised to act when the main method fails.

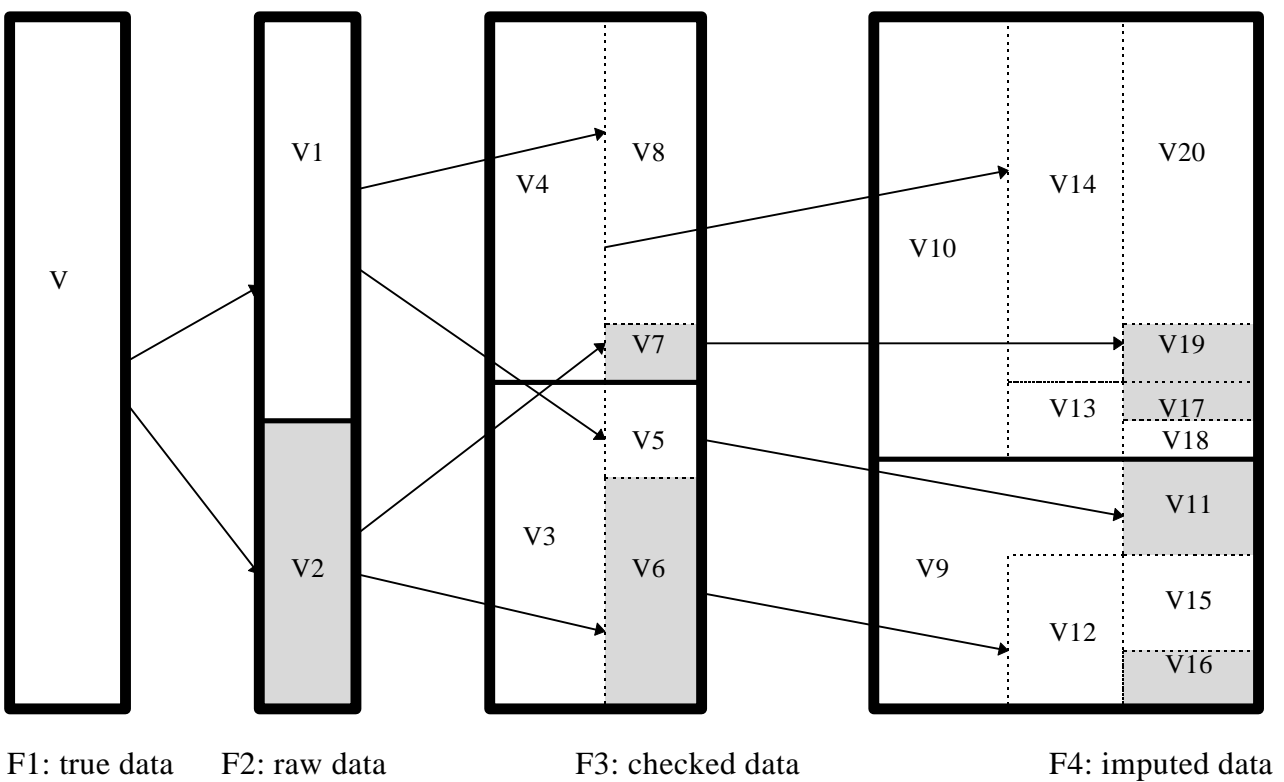
Whenever the term I_{EFF} is used, its uniformity has been exploited.

The two tables could be rewritten using $|V|$ instead of $|V_1|$ and $|V_2|$, using the relations:
 $|V_1| = (1-p)|V|$, $|V_2| = p|V|$.

We try now to summarise and visualise what we introduced until now.

In Figure 2, grey subsets correspond to the errors contained in the different files, starting from F2 (raw data), ending to F4 (clean data). In this latter file, grey subsets contain non imputed wrong values, or erroneously imputed true values, or non adequately imputed wrong values. The consideration of these subsets, and their relations, lead to the construction of the introduced indicators that allow us to assess the quality of the editing procedure.

Figure 2



Where:

- V : true data
- V1: not altered data
- V2: altered data
- V3: flagged data for imputation
- V4: not flagged data
- V5: true data flagged for imputation
- V6: altered data flagged for imputation
- V7: altered data not flagged for imputation
- V8: true data not flagged for imputation
- V9: imputed data
- V10: not imputed data
- V11: true data flagged for imputation and imputed

V12: altered data flagged for imputation and imputed
 V13: flagged data for imputation and not imputed
 V14: not flagged data for imputation and not imputed
 V15: altered data flagged for imputation and correctly imputed
 V16: altered data flagged for imputation and not correctly imputed
 V17: altered data flagged for imputation and not imputed
 V18: true data flagged for imputation and not imputed
 V19 \equiv V7
 V20 \equiv V8

REFERENCES

- GARCIA E., PEIRATS V. - Evaluation of Data Editing Procedures: Results of a Simulation Approach, *Statistical Data Editing Methods and Techniques Vol. I* Conference of European Statisticians, Statistical Standards and Studies, N.44, 1994, pp.52-68
- GRANQUIST L. - An Overview of Methods of Evaluating Data Editing procedures, *Statistical Data Editing Methods and Techniques Vol. II* Conference of European Statisticians, United Nations 1997
- KLEIJNEN J., VAN GROENENDAAL W. - Simulation. A Statistical Perspective, John Wiley 1992
- LESSLER J.T., KALSBECK W. D. - Nonsampling Error in Surveys, John Wiley 1995
- NORDBOTTEN S. - Editing Statistical Records by Neural Networks, *Working Paper N.40* Conference of European Statisticians, Work Session on Statistical Data Editing, Athens 1995
- POULSEN M. E. - Evaluating Data Editing Process Using Survey Data and Register Data, *Statistical Data Editing Methods and Techniques Vol. II* Conference of European Statisticians, United Nations 1997
- STEFANOWICZ B. - Selected Issues of Data Editing, *Statistical Data Editing Methods and Techniques Vol. II* Conference of European Statisticians, United Nations 1997
- VERBOON P., SCHULTE NORDHOLT E. - Simulation Experiments for Hot Deck Imputation, *Statistical Data Editing Methods and Techniques Vol. II* Conference of European Statisticians, United Nations 1997