WP. 28
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (iv): New and emerging methods

AUTOMATIC EDITING OF STATISTICAL DATA

A NEW VERSION OF DIA SOFTWARE

Prepared by J.M. Gómez-Alonso, National Statistical Institute, Spain[1]

Abstract

The DIA system is a generalized software, developed by the Spanish National Statistical Institute (NSI), for automatic editing and imputation of qualitative data. It is based on the Fellegi-Holt methodology treating not only the random errors but also the systematic errors. We now have developed an heuristic algorithm to allow to extend the DIA system to continuous and integer data. This algorithm solves the error localisation (EL) problem, this is, to determine the minimum number of variables to impute without having to calculate the complete set of edits. It avoids to have to break-down the set of explicit edits (the edits given by the subject-matter experts). Some examples using logical, arithmetic and mixed edits are presented to illustrate it. The treatment of the systematic errors is a DIA specific feature. We introduces a new concept , Deterministic Imputation Edit , that permit to modify this treatment eliminating the re-imputations that some times can happen when the edits expressing random errors and edits expressing systematic errors overlap.

Key words: logical edit, arithmetic edit, mixed edit, error localisation, systematic errors, statistical data editing, deterministic imputation edit, empty variable.

I. INTRODUCTION

1. This paper aims to help solve a range of real problems that emerge when implementing algorithms based on the theory developed by Fellegi-Holt (FH) to optimise automatic editing of statistical data. We shall describe the problems and use a number of examples to illustrate the proposed method of solution. We shall use the terminology and concepts established by FH, and only add new concepts where be necessary to clarify a proposed solution.

2. FH's assumption of the availability of a Complete Set of Edits (CSE) to solve the Error Localitation (EL) problem, this is, to find the minimum number of variables to impute, though entirely acceptable in theory, as shown by FH's theorems, is often unusable in practice, because real cases of imputation emerge characterised by the following circumstances: the experts on the subject establish a Set of Explicit Edits (SEE) comprising a large number of edits, the number of (active) participating variables in the definition of the SEE edits is also very large and the variables are closely inter-related. In cases like these, the number of possible combinations of edits, which must be checked for generating an Essentially New Edit (ENE), increases exponentially. In order to avoid this exponential growth as far as possible, we need to develop complex filters to reduce drastically the number of verifiable combinations of edits so that the CSE may be generated within a reasonable time.

---

[1] Prepared by Gómez-Alonso, J.M and presented by Lorca, D (jmgomez@ine.es , mdlorca@ine.es)
The author is grateful for the useful comments provided by Ton de Waal.

3. In highly complex situations encountered in practice, the unfeasibility of generating the CSE in a reasonable time leads to two undesirable consequences:

- The experts break down the SEE into several partial subsets, and use each subset to generate the respective Set of Implicit Edits (SIE). Hence several CSEs emerge, instead of the single CSE stipulated by theory. The breakdown of the SEE is not a trivial task. As there is no method for doing automatically, the experts have to conduct laborious tests of successive breakdowns in an attempt to minimise the number of CSEs.
- The process of imputation–given that there are several CSEs–is carried out in several successive steps. On each step, one or more variables that were imputable in previous steps are declared non-imputable (fixed). Having fixed variables violates one of the core principles of the system posited by FH: the principle of "minimal change=maximum preservation". This means that the minimum set (MS) of variables should be imputed so as to preserve the original data to the maximum extent possible.

4. A natural way to solve the EL problem and determinate the MS of variables to impute while scrupulously observing the core principles of the FH model is to work not with the CSE but with the named CSE_R^o, this is the SEE plus, if it is necessary, a small SIE that it is specially required to impute the erroneous record R^0. When a original record R^o has to be imputed due to fail an edit, we determine whether the SEE is adequate for imputation while observing the principle of least possible change. If so, the successful imputation of R^o does not require the derivation of any implicit edit through the SEE edits (this is what most often happens in practice). Otherwise, we need to generate the SIE required by R^o (SIE_R^o) dynamically to obtain the relevant CSE_R^o=SEE+(SIE_R^o). This enables us to impute R^o in a way similar to how it would have been done using the CSE.

5. This paper is organised as follows. Section II presents some notation. Section 3 describes the method to determine the MS with CSE unavoilable generating the SIE_R^o.It is illustrated by three examples in the cases of logical, arithmetic and mixed edits. Section 4 illustrates the way to approach imputation when the observed errors are not random but systematics, while respecting the core principles of the FH model by using a new concept: Deterministic Imputation Edit (DIE). Finally, some conclusions are given.

II. NOTATION

6. We denote the original record by R^0 and the imputed record by R^*. Let CNF the edit set in which a record R cannot fail and CF the edit set in which a record R can fail. Let AV(X) the edit set in which the variable X is active and FAV(X) the edit set whose failure is only avoidable by the value imputed to variable X. An edit e belongs to the FAV(X) set if it meets the following conditions:

- $e \in CF$
- $e \in AV(X)$.
- $e \in NAV(\alpha)$ where $NAV(\alpha)$ is the set of edits (not empty) in which the set of variables α is not active. A variable Z belongs to the set of variables α if $Z \in MS$ and is imputed after imputation of variable X (where X is the variable currently being imputed)

7. If after imputing the variable X there are no variables pending imputation, α will be an empty set. In this case the "$e \in NAV(\alpha)$" condition will be ignored. The edits in the $NAV(\alpha)$ set are caracterized by not to include any variable to impute after the imputation of the variable X. Therefore, if after imputation of X, the record R* fails any edits of FAV(X), the variables to be imputed later will not avoid its failure. This means that the imputed register R* would continue failling at least one of the explicit edits.

III. DETERMINATION OF MS WITH CSE UNAVAILABLE

8. The MS determination is based on knowledge of failed edit set (F) and the value of the named Index of Suspicion (IS). The selection of MS is an iterative process that it is repeated until the variables in MS cover off all failed edits. One variable is selected in each reiteration r. On starting reiteration 1, the set of failed edits pending of covering (FPC) matches with the F set. All active variables in a failed edit

are regarded as "suspicious". A suspicious variable is a "candidate" for inclusion in the MS if it is not yet included in the MS and it is active in an edit in the FPC set. The selected variable at each reiteration r covers off (is active in) a given number of failed edits denote by FC. The FPC set is updated when the reiteration is completed, so FPC(r)=FPC(r-1)–FC(r).

9. In each reiteration r, we determine the IS (r,i), Index of Suspicion corresponding to variable i in reiteration r, and we select the candidate variable with greater IS. Thus, the following data are taken into account:
-FPCC(i) note the number of edits in the FPC set which include (in which it is active) the variable i.
-W(i) note the weight allocated to variable i by the experts. A greater weigh corresponds to a greater degree of suspicion, that is, a lower reliability of the variable.
-P(i)=FPCC(i)/ACT(i) note the proportion of failed edits (pending of covering) which include the variable i, with respect to total edits which include the variable i.

10. Firstly, the IS(r,i) is associated to FPCC(i). If there are several variables with equal FPCC, then that with the greater W(i) is selected, unless that FPCC takes the value 1. In this case, the system randomly selects a variable to avoid systematically imputing the most suspicious variable when R^o fails one edit only. If there are several variables with equal FPCC and equal W, then that with the greatest P(i) is selected. And finally, if several variables are of equal value in the three data, the system randomly selects one variable. Thus, the selection of the candidate variable with greater IS(i) to be included in the MS, preserves the principle of least change. The iterative process terminates when the variables in the MS cover off all failed edits, that is, when FPC=∅. Once determinate the MS of variables to impute, it calculates the set of admissible values to impute. If it is not possible to obtain admissible values to impute the MS variables, it is necessary to generate an implicit edit and to return to starting the process of the MS determination. The process is finalized when all MS variables can be imputed. The implicit edits generated constitute the SIE_R^0.

A. Logical edits

11. The following table sets out the data we shall use to illustrate the determination of MS generating the SIE_R^o set, after converting the relevant explicit edits into binary strings. A bit 0 appears in their not problematic values.The observed values in the original record R^o, expressed by the respective bit 1, determine the failure of two of the four edits from the SEE. We also detail the active variables in each edit.

	VARIABLES/VALID VALUES														FAILED EDIT	ACTIVE VARIABLES
	A			B				C				D				
EDIT	1	2	3	1	2	3	4	1	2	3	4	1	2	3	(F)	
1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	F	A B
2	0	1	1	1	1	1	1	0	1	1	0	1	1	1	-	A C
3	1	0	1	1	1	1	1	0	0	1	1	0	1	0	-	A C D
4	1	1	1	1	1	1	0	1	1	1	1	0	0	1	F	**B D**
R^o		***1***				***1***					***1***			***1***	-------	------

12. We start the iterative process to determine the MS of imputable variables. On starting the reiteration r=1, FPC(1)=F=(1,4). It supposes that the weights W(i) are equal for all the variables, then to select the candidate variable for inclusion in the MS among the active variables in the FPC(1) set, we calculate the IS(1,i) associated a each variable:

- variable A: FPCC(A)=1; P(A)=1/3
- variable B: FPCC(B)=2; P(B)=2/2
- variable D: FPCC(D)=1; P(D)=1/2

The system selects the variable B because is the most suspect. The variable B covers off the edits 1 and 4, which are eliminated. On starting the reiteration r=2 we obtain FPC(2) is an empty set. Therefore, MS=(B).

13. The next step is to eliminate the edits that cannot fail. An edit e belongs to the CNF set if it meets some of following conditions :

- no variables in MS are actives in the edit e .
- some variable included in the edit , but not included in the MS , has a value avoiding the failure of edit e.

Thus CNF=(2,3).We now impute the variable B and for this we calculate the FAV(B) set. An edit e belongs to the FAV(B) set if $e \in CF=(1\ 4)$ and $e \in AV(B)=(1\ 4)$. The third condition $e \in NAV(\alpha)$ is ignored becuase α is an empty set (B is the only imputable variable). Then FAV(B)=(1 4). If the set FAV(B) set has several edits, the set of admissible values for imputing the variable B is produced by the result of performing on the edits of FAV(B) the operation union ($\cup$) in the columns of bits corresponding to the variable B. Thus $\cup B(1\ 4)=(1\ 1\ 1\ 1)$. If–as in this case–the resulting union vector has no bit 0, we term B is an empty variable, because there is no value of B which avoids the failure of all edits in the FAV(B) set. The variable B is empty because the current set of edits is incomplete with respect to the record R^o. Therefore, we now execute a 'call' to the generating procedure, in order to generate a new edit on the basis of, precisely and exclusively, the contributing edits in set FAV(B) and the generating variable B. The generating procedure used is the one specified by FH: intersection of the FAV(B) edits in all variables except the empty variable B and union of the FAV(B) edits in the variable B.The result is an ENE 5 denote as (1,4,B):

No. of Edit	Contributing edits and Generating field	VARIABLES														FAILED (F)	ACTIVE VARIABLES
		A			B				C				D				
		1	2	3	1	2	3	4	1	2	3	4	1	2	3		
5	(1,4,B)	1	1	0	1	1	1	1	1	1	1	1	0	0	1	F	**A D**

14. Before updating the set of available edits, the generating procedure verifies that the new binary string is really an ENE. The generation of the new edit 5 involves to obtain an updated version of the MS, based on the current set of edits. Therefore, the above steps are repeated.To determinate the new MS, we calculate the IS(1,i) associated to each candidate variable i. The system selects the variable B. It covers off the edits 1 and 4 which are eliminated. On starting the reiteration r=2, we find FPC=(5). Amongst active variables in edit 5, the system selects the most suspect D. In the next reiteration r=3 we find FPC is an empty set. Therefore, MS=(B D). The edits that cannot fail are CNF=(2 3). The observed values in the variables outside of MS ($C^o = 4$ and $A^o = 2$) avoid their failure. To impute the first variable of MS, the variable B, we calculate FAV(B). An edit e belongs to the FAV(B) set if $e \in CF=(1\ 4\ 5)$, $e \in AV(B)=(1\ 4)$ and $e \in NAV(D)=(1,2)$. The result of the intersection of these sets is FAV(B)=(1). There is only a edit in FAV(B). Thus, the set of admisible values for the variable B is given by the 0 bits in edit 1.

15. The binary string for the variable B in edit 1 is (0 0 1 1). It is assumed that the imputation module decides to impute the value B*=1 (which avoids the failure of edit 1), and it then imputes the following variable of the MS, the variable D. For it, we have FAV(D)=(4 5). The set of admissible values for imputing the variable D is given by the union of edits in the columns of bits corresponding to the variable D. Thus, $\cup$ D(4 5)=(0 0 1). It is assumed that the imputation module decides to impute the value D*=2 (which avoids the failures of edits 4 y 5). Since D*=2 deactivates edits (4 5), CF is an empty set, i.e., the pair of imputations B*=1; D*=2 and the original values A^o=2 and C^o=4 ensure that register R* will not fail any edit. Successful, the imputation requires the generation of only one implicit edit, the SIE_R^o set is only formed for the edit 5. The principle of least change has been preserved. It has not been necessary to generate the CSE that contains 9 edits.

16. In our example R^o fails 50 % of the edits in SEE. In real situations the failure rate is not naturally so high as in the example. The difference between sizes of the CSE (when it could be generated) and the CSE_R^o is normally very large. In real, highly complex statistical files (size of SEE=1500 edits; number of variables=146; total number of valid values=3521), the average number of edits generated by erroneous records (invalid or inconsistent values) was 3.2 implicit edits.

B. Arithmetic edits

17. This example, proposed by De Waal, T [3], shows the problem which may arise when trying to impute an integer type variable. It is assumed that a correct record R must meet the following conditions, where T denote the turnover of an enterprise, B its benefits, C its costs and E the number of employees. The variables T, B and C are continuous variables and E is an integer variable:

Validity edits	Consistency edits
Turnover T >= 0	Benefit B=T-C
Cost C >= 0	T<=2C
Employees E >= 0	10C<=11T
	T<=550E
	320E<=C

18. Transforming the consistency edits into normalized edit gives the following set of explicit edits:

Normalized consistency edits	
1. B+C-T>0	4. 10C-11T>0
2. –B-C+T>0	5. T-550E>0
3. T-2C>0	6. 320E-C>0

19. Let us consider an original record R^0 with values B=2020, C=3040, E=5 and T=5060. We here consider an auxiliary variable A that always takes the value 1. Expressing with c(i,j) the coefficient of the variable j in the edit i and M1 the coefficient matrix we have $P=M1*R^0$ product vector whose elements with a value greater than zero will indicate the failed edits by R^0. Therefore, we have the following in this example:

TABLE 1

EDITS(i)	MATRIX M1					Active Variables	Type of inequality TI(i)
	B	C	E	T	A		
1	1	1	0	-1	0	BCT	>
2	-1	-1	0	1	0	BCT	>
3	0	-2	0	1	0	CT	>
4	0	10	0	-11	0	CT	>
5	0	0	-550	1	0	ET	>
6	0	-1	320	0	0	CE	>

TABLE 2

EDITS (i)	p(i, j) = c(i, j).R°(j)					P=Σ	FAILURE: F (F → P>0)	Active Variables
	B	C	E	T	A			
1	2020	3040	0	-5060	0	0	-	BCT
2	-2020	-3040	0	5060	0	0	-	BCT
3	0	-6080	0	5060	0	-1020	-	CT
4	0	30400	0	-55660	0	-25260	-	CT
5	0	0	-2750	5060	0	2310	F	ET
6	0	-3040	1600	0	0	-1440	-	CE

20. To determinate the MS of imputable variables, we start the iterative process r=1 with FPC(1)=F=(5).To select the candidate variable for inclusion in the MS among the active variables in the FPC(1) set, we calculate the IS(1,i) associated a each variable and the system selects the variable E because is the most suspect. The edits that cannot fail are CNF=(1 2 3 4). To impute the variable E, we have FAV(E)=(5 6). A value of E is adequate as an imputation if it avoids failure of all edits in the FAV(E) set. Given that the variables B, C and T are now considered to fixed (they do not belong to the MS), the values to E that avoid failures of the edits in FAV(E) set must meet the following:

$E5 \rightarrow T-550E \leq 0 \rightarrow E \geq T/550$

$E6 \rightarrow 320E-C \leq 0 \rightarrow E \leq C/320$

21. The admissible lower value (ALV) to impute the variable E is T/550= 5060/550=9.2 and the admissible upper value (AUV) is C/320=3040/320=9.5. Then, the interval of admissible values (IAV) for the variable E is (9.2;9.5). Given that an admissible imputation for E must be a integer, we say that the variable E is empty. In cases such as this a new edit is generated of a different way to usual: if AUV-ALV<1 it would be an error/edit. Then C/320-T/550-1<0 would be a not normalized edit and finally the

edit 7=(5,6,E,*)=32T-55C+17600>0 would be a normalized and reduced edit. The asterisk in the edit 7, derived by the contributing edits 5, 6, and generating variable E, means that the edit was generated a different way to usual. The edit 7 dominates the derived edit 32V-55>0 which is obtained the usual way. Varying the set of edits could change the MS of imputable variables. We now have FPC(1)=F=(5,7). The variable T is the most suspect. The edits that cannot fail are CNF=(6). To impute the variable T, we have FAV(T)=(1 2 3 4 5 7). A value of T is adequate as an imputation if it meets the following conditions:

E1:B+C–T≤0→T≥B+C→T≥5060	E4:10C-11T≤0→T≥2763.63
E2:-B-C+T≤0→T≤B+C→T≤5060	E5:-550E+T≤0→T≤2750
E3:-2C+T≤0→T≤6080	E7:-55C+32T+17600≤0→T≤4675

Interval of Admissible Values: IAV=∅

22. The variable T is empty because it does not meet the conditions required, thereby causing a call to the generating procedure (Module Generating Edits for Rº, MGERº) in order to generate a new edit on the basis of the generating variable T and the contributing edit pair that meets the following conditions: 1) the edits of the pair are non-associated edits, 2) the coefficients of pair c(i, T) and c(i', T) have a different sign, 3) at least Rº fails one of the edits, 4) the derived edit is failed by Rº and 5) generate a non-redundant edit.

23. Given the formal analogy existing between the generation processes of logical and arithmetic edits, it would be logical that the contributing edits were from the FAV(T) set which, in this case, would be six edits. However, several tests show that, generally, it is more efficient to base on pairs of contributing edits rather than on FAV(T) edits. For each variable, the MGERº knows which pair of contributing edits was the final one in the generation of a derived edit failed by Rº, in such a way that the next time the MGERº is invoked the process continues from the right point. In our case, the first edit pair that, on the basis of the generating variable T, allows the generation of an edit which meets the required conditions is the pair (1, 5) from which a new edit is obtained: Edit 8=(1,5,T):B+C–550E>0. Varying the set of edits could change the MS.We now have FPC(1)=F=(5 7 8). The first variable selected by the IS to include in MS is E. The variable E covers off the edits 5 and 8 which are eliminated. On starting the reiteration r=2 we obtain FPC(2)=(7). Amongst active variables (C,T) in edit 7, the system selects the most suspect T. In the next reiteration r=3 we find FPC is an empty set. Therefore MS=(E T).The edits that cannot fail are CNF=(∅). To impute the first variable of MS, the variable E, we have FAV(E)=(6 8). A value of E is adequate as an imputation if it meets the conditions:

E6:-C+320E≤0→E≤C/320→E≤9.5
E8:B+C-550E≤0→E≥(B+C)/550→E≥9.2

24. Given that the admissible imputation for variable E has to meet the conditions 9.2≤E≤9.5, the variable E is empty. Just as before a new edit is generated in a different way to usual. It is considered that an error occurs if the following happens: VSA–VIA<1, it is : C/320-(B+C)/550-1<0. Once the edit is normalized and reduced, we obtain: edit 9=(6,8,E*)=32B–23C+17600. The MS is of new calculated. We have FPC(1)=F=(5 7 8 9). The first variable selected by the IS to include in MS is C. The variable C covers off the edits 7, 8 and 9 which are eliminated. On starting the reiteration r=2 we obtain FPC(2)=(5). Amongst active variables (E,T) in edit 5, the system selects the most suspect E. In the next reiteration r=3 we find FPC is an empty set. Therefore MS=(C E).The edits that cannot fail are CNF=(∅). To impute the first variable of MS, the variable C, we have FAV(C) =(6 8 9). A value of C is adequate as an imputation if it meets the conditions:

E6:–C+320E≤0→C≥1600
E8:B+C–550E≤0→C≤730 → incompatible with the previous condition .

25. The variable C is empty as it does not meet the required conditions thereby causing a call to the MGERº to generate a derived edit based on the corresponding edit pair, which is pair (1 7).The new derived edit is: edit 10=(1,7,C)=55B-23T+17600>0. The MS is of new calculated. We have FPC(1)=F=(5 7 8 9 10). The first variable selected by the IS to include in MS is B. The variable B covers off the edits 8, 9 and 10 which are eliminated. On starting the reiteration r=2 we obtain FPC(2)=(5 7). The only variable which covers off the two failed ones pending coverage is the variable T. Hence, the system decides that MS=(B T).The edits that cannot fail are CNF=(6). To impute the first variable of MS, the variable B, we have FAV(B)=(8 9). A value of B is adequate as an imputation if it meets the conditions:

E8:B+C–550E≤0→B≤550*5-1000→B≤ - 290
E9:32B–23C+17600≤0→B≤(23*3040-17600)/32→B≤1635

26. It is assumed that the imputation module decides to impute the value B*= - 400 (which avoids failures of edits 8 and 9), and it then imputes the following variable of the MS, the variable T. We now have CNF=(6 8 9) and FAV(T)=(1 2 3 4 5 7 10). A value of T is adequate as an imputation if it meets the following conditions:

E1/E2: $T=B+C \rightarrow T=2640$

E3: $T-2C\leq 0 \rightarrow T\leq 6080$

E4: $10C-11T\leq 0 \rightarrow T\geq 2763.63$ → incompatible with the first condition .

27. The variable T is empty as it does not meet the required conditions, thereby causing a call to the MGERº to generate a derived edit based on the corresponding edit pair, which is pair (4 5).The new derived edit is: edit11=(4,5,T)=C-605E>0. The MS is of new calculated. We have FPC(1)=F=(5 7 8 9 10 11) and the first variable selected by the IS to include in MS is C. The variable C covers off the edits 7, 8, 9 and 11 which are eliminated. On starting the reiteration r=2 we obtain FPC(2)=(5 10). The only variable which covers off the two failed ones pending coverage is the variable T. Hence, the system decides that MS=(C T).The edits that cannot fail are CNF=(∅). To impute the first variable of MS, the variable C, we have FAV(C)=(6 8 9 11). A value of C is adequate as an imputation if it meets the conditions:

E6: $-C+320E\leq 0 \rightarrow C\geq 1600$

E8: $B+C-550E\leq 0 \rightarrow C\leq 550*5-2020 \rightarrow C\leq 730$

E9: $32B-23C+17600\leq 0 \rightarrow C\geq (32*2020+17600)/23 \rightarrow C\geq 3575.65$ → incompatible with the previous condition .

28. The variable C is empty as it does not meet the required conditions thereby causing a call to the MGERº to generate a derived edit based on the corresponding edit pair, which is pair (2, 8).The new derived edit is: edit12=(2, 8, C)=-550E+T>0.The MS is of new calculated. We have FPC(1)=F=(5 7 8 9 10 11 12). The first variable selected by the IS to include in MS is E. The variable E covers off the edits 5, 8, 11 and 12 which are eliminated. On starting the reiteration r=2 we obtain FPC(2)=(7 9 10). Each variable B, C and T coverss off two of the three failed edits pending coverage. The variables B and T have a greater IS than variable C. Hence,the system decides that MS=(B E T).The edits that cannot fail are CNF=(∅). To impute the first variable of MS, the variable B, we have FAV(B)=(9).A value of B is adequate as an imputation if it meets the condition: E9: $32B-23C+17600\leq 0 \rightarrow B\leq (23*3040-17600)/32 \rightarrow B\leq 1635$. It is assumed that the imputation module decides to impute the value B*=1500 (which avoids failure of edit 9), and it then imputes the following variable of the MS, the E variable. We now have CNF=(9) and FAV(E)=(6 8 11).A value of E is adequate as an imputation if it meets the following conditions:

E6: $-C+320E\leq 0 \rightarrow E\leq 3040/320 \rightarrow E\leq 9.5$

E8: $B+C-550E\leq 0 \rightarrow E\geq (1500+3040)/550 \rightarrow E\geq 8.25$

E11: $C-605E\leq 0 \rightarrow E\geq C/605 \rightarrow E\geq 5.02$

29. As the value of variable E must be integer , the imputation module decides to impute the value E*=9, which avoids failure of edits (6 8 11), and it then imputes the following variable of the MS, the variable T.We now have CNF=(6 8 9 11) and FAV(T) =(1 2 3 4 5 7 10 12).A value of T is adequate as an imputation if it meets the following conditions:

E1/E2: $T=B+C \rightarrow T=4540$	E7: $-55C+32T+17600\leq 0 \rightarrow T\leq 4675$
E3: $T-2C\leq 0 \rightarrow T\leq 6080$	E10: $55B-23T\leq 0 \rightarrow T\geq (55*1500)/23 \rightarrow T\geq 3586.95$
E4: $10C-11T\leq 0 \rightarrow T\geq 2763.63$	E12: $55B-23T+17600<=0 \rightarrow T>=(55*1500+17600)/23 \rightarrow T>=4352.17$
E5: $T-550E\leq 0 \rightarrow T\leq 4950$	

30. The only admissible value is T*=4540.Then, we finally obtain the following imputed values: $B^{*}=1500$, $C^{0}=3040$, $E^{*}=9$ and $T^{*}=4540$ and SEE=(1 2 3 4 5 6) and SIE_Rº=(7 8 9 10 11 12).

C. Mixed edits

31. A mixed edit comprises two types of component: qualitative component (QC) and arithmetic component (AC). A mixed edit fails when both of its active (non-empty) components fail. As a study case of how to work with mixed edits, we shall assume that the arithmetic variables are continuous and positive, the experts determine four mixed edits, one arithmetic edit and one logical edit and the original record Rº has the following values: Aº=2, Bº=2, Cº=2, Xº=5, Yº=10, Zº=3.We obtain the following matrix of edits MX.

Edit	QUALITATIVE VARIABLES											ARITHMETIC VARIABLES				F=failed N=not failed I=inactive			Active variables in edit
	A			B				C				X	Y	Z	**C T E**				
	1	2	3	1	2	3	4	1	2	3	4	2	0	-1	0	Q	A	E	
1	1	1	1	1	1	1	1	1	1	1	1	2	0	-1	0	I	**F**	**F**	**X Z**
2	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0	F	**I**	**F**	**A C**
3	1	1	1	0	1	1	0	0	0	1	1	0	0	4	-10	N	**F**	**N**	**B C Z**
4	1	0	1	1	1	1	0	1	1	1	1	-1	2	4	0	N	F	N	A B X Y Z
5	1	1	0	1	1	1	0	0	1	1	0	0	3	0	-5	F	F	F	A B C Y
6	1	1	1	0	0	1	1	1	1	1	0	5	0	-2	2	N	F	N	B C X Z
Rº		*1*			*1*				*1*			*5*	*10*	*3*	1	----------			

32. For mixed edits, the determination of the MS of imputable variables is a process similar to the one for strict logical and arithmetic edits. So it is based on the determination of the failed edits (F) and the IS associated to each candidate variable for inclusion in the MS.Therefore, on the basis of the current data, we have FPC(1)=F=(1 2 5). In the first reiteration, the selected variable by the IS to include in the MS is A. The variable A covers off the edits 2 and 5, which are eliminated. In the second reiteration, we obtain FPC(2)=(1). Amongst active variables (X, Z) in edit 1, the system selects the most suspect, the variable X. Thus, MS=(A X). To impute the variable A, we have CNF=(3 6). The failure is avoided by the fixed values C^o=2 and B^o=2. We have FAV(A)=(2 5). The union in A of edits 2 and 5 is the bit vector: (1 1 1). Given that the union vector is an unitary vector, the variable A is empty. Thus, it is necessary to generate a new edit.

33. To generate mixed edits, we distinguish two cases:
Case 1: the empty variable is of type qualitative: To generate the QC we have as generating variable, the empty variable, as contributing edits, the edits from FAV and as generating system, the system used only to operate with qualitative edits. i.e., union for the generating variable and intersection for the rest of them. To generate the AC we have that it does not exist the generating variable, as contributing edits, the edits from FAV and as generating system, AC=ΣAC(i) where AC(i) is the AC of the i-essimo mixed edit from FAV.
Case 2: the empty variable is of type arithmetic: To generate the QC, we have that it does not exist generating variable, as contributing edits, the edit pair determines by the $MGER^0$ and as generating system, the intersection of the contributing edits for all variables. To generate the AC we have as generating variable, the empty variable, as contributing edits, the edit pair determines by the $MGER^0$ and generating system, the used one only to operate with arithmetic edits. Thus, the QA and AC generated in our example would be as follow:

EDIT	QUALITATIVE VARIABLES											ARITHMETIC VARIABLES				F=failed N=not failed **I=inactive**			ACTIVE VARIABLE IN EDIT
	A			B				C				X	Y	Z	CTE				
	1	2	3	1	2	3	4	1	2	3	4	MATRIX OF COEFFICIENTS				**Q**	**A**	**E**	
7	1	1	1	1	1	1	0	0	1	1	0	0	3	0	-5	F	**F**	**F**	**B C Y**

34. Varying the set of edits could change the MS. We now have FPC(1)=F=(1 2 5 7). The variable selected by the IS is the C. The variable C covers off the edits 2, 5 and 7. In the second reiteration, we obtain FPC(2)=(1). Amongst active variables (X, Z) in edit 1, the system selects the most suspect, the variable X. Thus, MS=(C X). To impute the variable C, we have FAV(C)=(2 3 5 7). The union in C of edits 2, 3, 5 and 7 is the bit vector: (0 1 1 1). The imputation module decides to impute the value C*=1,

which avoids the failures of edits (2 3 5 7), and it then imputes the variable X. We now have the following sets of edits: CF=(1) and FAV(X)=(1). A value of X is adequate as an imputation if it meets the following condition:E1:$2X-Z\leq 0\rightarrow 2X-3\leq 0\rightarrow X\leq 1.5$. The imputed value is assumed to be X*=1. We therefore finally have an error-free record R*: $A^o=2$, $B^o=2$, C*=1, X*=1, $Y^o=10$, $Z^o=3$.

IV. IMPUTATION IN EVENTS OF SYSTEMATIC ERRORS

35. We usually distinguish two types of non-sampling errors: random and systematic errors. The former ones can arise at any time affecting any variable and they are uniformly distributed. Systematic errors arise through inadequate understanding of the questions, concepts, definitions or instructions by the respondent or by the agents involved in the various phases of the statistical process or intentionally, introduced by the respondent to protect the privacy or for fear the information could be used for tax or policing purposes.

36. The pure methodology proposed by FH is wholly satisfactory for processing random errors. But the methodology is inadequate for editing systematic errors, such as those arising in the Building Census conducted in Spain in 1981 and described in bibliographical reference (2). Hence the Spanish NSI included in the DIA system a module aimed at processing systematic errors, which are removed on the basis of Rules of Deterministic Imputation (RDIs). A RDI is a rule combining detection and imputation in two parts, the conditional part, on the left of equal sign, which expresses the systematic error and the other part determines the applicable imputation if the conditional part is satisfied. An example of RDI is: $A(1)\cap B(2)\cap C(3)=B(\text{blank})$ where we assume that a systematic error arises if a record has the values A=1, B=2 and C=3 and if so, the most reasonable course of action is to impute the value 'blank' to the variable B.

37. The first member of the RDI is really the normalised form by FH to express an edit. This type of edits is called an Edit Derivated of RDI (EDR). Then, in the DIA system there is an EDITS_RDI analyser which reports and/or resolves conflicts between both types of edits. The input to the generating module of the CSE is SEE plus the set of EDR. At the initial stage, RDIs are executed according to imputations determined by experts. Later, imputations are decided and applied on the basis of the FH methodology. The gap between the two types of imputation can bring about reimputation, e.g., a variable X is imputed at the initial stage by a RDI, and at the second stage, the system determines its inclusion in the MS, such that the imputed value X finally differs from the value determined by the RDI. One way to avoid the reimputation would be to use a special kind of edit we shall call a Deterministic Imputation Edit (DIE), obtained from a RDI as follow.

38. A RDI encompasses two types of variables: First, one only variable appears in both members of the RDI. We shall generically name this variable Deterministic Imputation (DI) variable, because it is the only subject to deterministic imputation. In the first member of the RDI, the variable DI appears to express one of the conditions for a systematic error to be declared. In the second member, it determines the required imputation. Second, the rest of variables of the RDI, which appear only in its first member, express the rest of conditions for a systematic error to be declared. We shall name conditional variables.

39. The rules for converting a RDI into a DIE are:

1. The failure condition imposed in the RDI on a conditional variable is expressed in the same way in the DIE.
2. The failure condition imposed in the RDI on variable DI becomes, in the DIE, the failure condition imposed on a new variable IMA_DI, named the image of variable DI that it will be a copy of variable DI and it never will be imputed.
3. The complement ($\neg$) to the RDI imputation for variable DI becomes, in the DIE, the failure condition imposed on variable DI.

40. The system converts the above RDI into the following DIE:
A(1)∩IMA-B(2)∩C(3)∩B(¬blank). Hence the DIE resulting from the above conversion perfectly matches the required structure for an edit under the FH model. Then, as usual the system determines the set of failed edits, but, as there are DIEs, the system does the following:
- If a DIE fails, the respect DI variable is included in the MS. This means that the imputation required by the RDI associated with the failed DIE will always be preserved. In our example, the system would include the variable B in the MS and the imputation module can only impute the value 'blank', as required by the RDI.
- After completion of the imputation required by the failed DIEs, the usual process starts of imputing the variables in the MS to correct any failures arising in the SEE.
Given that the MS cannot contain repeated variables, it is impossible for a variable to be imputed twice, i.e., the possibility of reimputation disappears.

VI. CONCLUSIONS

41. The heuristic algorithm presented and illustrated by some examples in this paper solves the EL problem without having to calculate the CSE. This permits to extend the DIA system to quantitative data. It avoids having to break-down the SEE into several partial subsets reducing the number of imputations to carry out.

42. The DIE allows integrating the edits expressing systematic errors with the edits expressing random errors according to FH model and thus, we can apply the DIA system simultaneous to both type of errors avoiding possible re-imputations.

References

(1): De Waal, T(200x). Automatic Error Localisation for Categorical, Continuous and Integer Data. Statistics Netherlands.
(2): DIA (1994). Descripción del Sistema DIA. Instituto Nacional de Estadística (INE). Madrid.
(3): Fellegi, I.P., and Holt, D(1976). A Systematic Approach to Automatic Edit and Imputation, Journal of the American Statistical Association, V 71, pp 17-35.