Improving the effectiveness of a probabilistic editing strategy for business data

Di Zio M., Guarnera U., Luzi O.

ISTAT – Italian National Statistical Institute
Roma, Via Cesare Balbo 16

1. Introduction

Data collected in statistical surveys are generally affected by different kinds of non-sampling errors. In Statistical Offices, procedures and tools are used at the post data capturing stage for identifying and eliminating errors that generally contaminate the collected data. Such procedures generally use sets of statistical or logical criteria (*edits*) expressing known relationships among the surveyed variables. These criteria are specifically designed for dealing with the different types of errors occurring during the overall survey process. A general distinction among errors refers to *stochastic* and *non-stochastic* errors, depending on their random or not random origin. Particularly in business surveys, an important further distinction has to be made between *influential* and *not-influential* errors, depending on their effects on the final survey estimates.

The identification of *influential* errors and other kinds of relevant errors (like unit measure errors and other systematic errors having a potential high impact on published figures) is generally performed in preliminary data editing steps in order to control the data variability at the publication level. Selective editing (Latouche *et al.*, 1995) and macro editing (Granquist, 1992) are useful approaches in this context.

Probabilistic algorithms like the Fellegi and Holt method (Fellegi *et al.*, 1976) are specifically designed for identifying stochastic non influential errors in statistical survey data taking into account coherence constraints (*edits*) among the investigated phenomena. Generally, for any given unit do not satisfying a given (sub-) set of edits, the Fellegi and Holt methodology (FH in the following) allows the identification of the minimum number of fields to be changed in order to make the units passes all edits. A commonly used classification of edits makes a distinction between *hard* edits, pointing out *fatal* errors (e.g. certainly erroneous relations among data), and *soft* (or *query*) edits, identifying suspicious but not necessarily unacceptable data relations. When soft edits are used in probabilistic algorithms as if they were *hard* ones they can produce the misclassification as fatal errors of some amounts of correct data (e.g. representative outliers). In fact, in these algorithms errors due to the failure of any edit are always considered as *fatal* errors, regardless of the nature of the failed edit. Furthermore, users generally apply to data more edits than necessary (e.g. either useless edits in terms of their capability of pointing out 'true' errors or edits that do not highlight unacceptable situations). This problem could affect the effectiveness of the editing process, e.g. by determining *over-editing* and/or high percentages of acceptable data erroneously classified as unacceptable.

From the above discussion it is evident that also in the editing phase, and particularly during the automatic data editing process, additional errors can be introduced among data because of an inaccurate design of edits. A very crucial problem in this context is then the rationalization and improvement of the edit rules used in the automatic error localization process. Granquist (1995; 1996) underlines the importance of the accuracy in the definition of edits, and in particular of query edits, e.g. by eliminating unnecessary edits, by focusing on edits that do not identify influential errors, by improving the query edits bounds, by monitoring the impact of editing on data.

The availability of generalized software for automatic editing allows at present the application of different editing techniques at low costs and time. Among others, the FH approach is at present available in a number of generalized packages (for both categorical and continuous data) developed in different National Statistical Offices: for continuous variables see Winkler *et al.* (1997a), Kovar *et al.* (1988), Kovar *et al.* (1996); De Waal (1996), Todaro (1999); for categorical variables see Riccini *et al.* (1995), Garcia Rubio (1990), Winkler *et al.* (1997b). The main advantages related to the use of these packages are the reduction of the editing costs, the standardization and the full documentation of the automatic editing step and the minimisation of the number of modifications of the original data under some general conditions. The FH approach is useful for detecting errors originated by random mechanisms, so their use is theoretically not appropriate for localizing other types of errors (e.g. systematic errors).

In the paper we describe a strategy for monitoring and improving the effectiveness of editing rules (in particular, ratio edits) to be used in the FH methodology for continuous business data. The adopted FH algorithm is that implemented in the *Generalised Edit and Imputation Software* (GEIS) (Kovar *et al.*, 1988; Cotton, 1991). In the proposed strategy, we concentrated on the definition and analysis of query edits. To this aim, the Hidiroglou and Berthelot algorithm (Hidiroglou et al, 1986) was used in combination to *Exploratory Data Analysis* techniques (*EDA* in the following). The Hidiroglou and Berthelot algorithm (HB in the following) has been originally proposed for defining acceptance bounds of longitudinal ratios (ratios between current and historical values of a same variable). In the strategy proposed in this paper the algorithm is used to find bounds for query edits having the form of ratios between related variables. EDA (Tukey, 1977; Ellfors *et al.*, 2000) represents a powerful approach for analysing statistical data structures and relations supported by appropriate graphical representations.

Taking into account that the FH algorithm works in the best way in presence of only random errors, the study has been performed in order to verify to what extent the algorithm itself can be adapted for dealing with any kind of errors.

The research activities and the results presented in the paper have been produced in the context of the EUREDIT project (www.cs.york.ac.uk/euredit/).

The paper is structured as follows. In paragraph 2 the overall strategy and the methods used for improving the effectiveness of edit sets including query edits in an automatic context are described. Paragraph 3 contains a description of an application of that approach to a sample of the U.K. Annual Business Inquiry data.

2. The methodological approach

When using soft (or *query*) edits in probabilistic editing algorithms the main risk is the potential misclassification of correct data as erroneous. Particularly in business surveys, a generally high number of query edits having the form of *ratios*[1] is used in combination with fatal edits in order to identify statistically or mathematically non-consistent data. The use of these rules for units containing only non fatal errors (for example, unusual or anomalous but acceptable relations among variables) can be dangerous because of various reasons:

- the imputation of values of these units determines the replacement of true data with erroneous ones, i.e. the introduction of additional errors among data;
- the imputation of values of these units generally produces modifications of the variability of original data (for example, in case of continuous data, because of the truncation of the distribution tails).

For these reasons, when adopting automatic probabilistic algorithms for data editing, it is crucial to properly define the set of consistency rules in order to guarantee an acceptable trade off between correctly detected and resolved errors and risk of potential data misclassifications. In particular, query edits are to be limited and carefully defined: often the introduction of few query edits or the

[1] a *ratio check* has the form $a<X/Y<b$, were X and Y are related variables.

slight modification of query edits bounds may produce significant effects on data in terms of correct data classification.

In this section we propose a strategy for analysing and improving the effectiveness of query edits for continuous business survey data. This strategy consists of several combined analyses, performed by using different techniques in an integrated manner. The applicability of the proposed approach implies that the editing procedure for the current survey is defined and tuned by using a data set (called *test data set*) having the same characteristics of current data but for which the subject-matter expert has information on both the errors location and the corresponding 'true' data. This situation can be either artificially obtained by simulating errors among edited data, or approximated by using data from a previous survey repetition for which both the raw and the final edited data are available.

For any given initial set of editing rules, the proposed strategy consists of the following main analyses:

- identifying ineffective query edits;
- defining new soft edits;
- defining new bounds for the ineffective edits.

The first analysis aims at identifying soft edits that potentially produce data misclassification. The second analysis is based on the fact that the FH algorithm is effective when each variable to be analysed is involved in many hard rules, or, in other words, when the edit rules are strongly dependent one each other ("well connected set of edit").

In this experimental situation, the given error localization procedure can be tuned on test data by measuring and analysing its capability of correctly classifying data, through appropriate quality indicators (Chambers, 2000).

Once methods have been applied and tuned on test data, the obtained results can be used as a starting point for building the actual editing strategy. The underlying assumptions here are that: 1) the surveyed phenomena maintain similar behaviour, distributions and relations from one period to the next one, and 2) the error mechanisms are similar in the test and in the actual survey data.

The proposed strategy consists of the following main phases:

1 <u>on test survey data</u>:

  a) analysis of the original set of query edits in order to identify possibly ineffective rules;

  b) graphical analysis of relations between items in order to both assess the significance of ineffective query edits and possibly identify new query edits;

  c) by using statistical methods, identification of appropriate bounds for the new query edits and determination of new most effective bounds for the originally ineffective rules;

  d) by iterating the application of the FH algorithm using the different potential sets of edits, identification of the *optimal* set of rules based on the analysis of some performance indicators (in particular, the *probability of correct data classification*);

2) <u>on the actual raw data</u>:

  a) (<u>only if different items are surveyed in actual data with respect to the test ones</u>) graphical analysis of the relations between items in order to possibly identify new soft edits;

  b) by using statistical methods, determination of bounds for both the new query edits and the originally ineffective ones taking into account the results obtained in step 1.c).

Step 1.d) is possible because of the knowledge of the true data and errors in the test survey data. In this situation it is in fact possible to measure the effects on data of the edit rules (Chambers, 2000), and to analyse the changes in these effects due to changes of the editing strategy.

The methodologies and approaches used in each step are described in the following subsections.

2.1. Analysis on test data: identification of ineffective query edits

As already mentioned, the use of soft rules as if they were hard ones may cause anomalous but acceptable records are targeted as erroneous. This risk can be estimated by means of different measures:

1. the *failure rates* of query edits;
2. the *probabilities of correct data classification* associated with the applied edit rules (Chambers, 2000), based on contingency tables reporting the frequencies of actual vs predicted error status. These probabilities are based on the classification of each value of a given variable *Y* observed on *n* sampling units with respect to its status before the editing process (erroneous or true value) and its status after the editing process (suspicious or acceptable value). Therefore for each variable *Y* we have the following cross-classification table (Table 1):

Table 1 – Cross classification table of original vs edited values

	S = 0	**S = 1**
E = 0	n_a	n_b
E =1	n_c	n_d

where E=0 if a value is true and 1 otherwise, S=0 if the value is classified as acceptable by the error localisation algorithm and 1 otherwise, na+ nb+ nc+ nd = n. It is obvious that high frequencies in the table diagonal indicate good performance of the error localisation in terms of correct data classification. The probabilities of wrong data classification can be easy obtained as:

$\alpha = n_c/(n_c + n_d)$ (not detected errors out of the total number of errors);

$\beta = n_b/(n_a + n_b)$ (true data classified as erroneous out of the total number of true data).

2.2. Analysis on test data: identification of potential new query edits

It is well known that the effectiveness of the FH algorithm increases when variables are involved in a well connected set of hard rules: a typical example is represented by a double contingency table with row and column totals, or crossed ratio edits between couple of strictly related variables.

Often some item is involved in too many constrains with respect to the other ones so that the FH algorithm, based on the minimum change criterion, doesn't work in a *balanced* way. In these cases, the behaviour of the algorithm can be influenced either by balancing the set of edits (for example by adding further edits to the initial ones or by eliminating not effective rules) or by modifying the *reliability weights*[2] associated to variables. The latter strategy is generally used to take into account in the editing algorithm of the different level of reliability of each analysed variable.

In this phase of the data analysis, our objective is to identify new query edits that potentially improve the effectiveness of the error detection algorithm without increasing the over-editing risk.

[2] Although the error localisation algorithm determines for each erroneous unit the *minimum number of fields to be changed* (*solution of minimum cardinality*), the user can exert some influence on its choice through the utilisation of *weights*. Weights are assigned to variables depending on the user believe about variable reliability in an edit group. If weights are attached to variables, the criterion used to determine which fields should be imputed is still based on the minimum change criterion. However, in this case, the cardinality of the solution is given by the sum of the weights of fields involved in the solution instead of the number of fields.

To this aim we propose an approach based on the joint analysis of following elements: data distributions and relations, failure rates of edits, classification probabilities α and β defined in section 2.1.

Marginal and joint data distributions are graphically explored by using the EDA tools available in the SAS Insight module. EDA has shown itself to be an important methodology in the context of data editing, in the analysis of data, and for identifying outliers and inliers. Des Jardins *et al.* (2000) analyse the main advantages of using EDA for exploring data relations and editing statistical data and show how graphical methods can be easily applied to discover features that conventional methods could not highlight. Further, EDA is particularly helpful when data relations vary markedly in different data clusters (e.g., when using ratios, the use of graphical representations allows to highlight unpredictable changes of relations among items in different strata).

New software packages make it straightforward to perform several graphical analyses. These tools often allow the application of even sophisticated methods in a simple and quick manner. Some generalised tools implementing graphical approaches for exploring data have been developed (see for example Esposito *et al.*, 1994; Houston *et al.*, 1993). In this context, a powerful tool is represented by the *SAS Insight* module available in the SAS software. Many data representations and statistical analyses can be performed by jointly using a high number of tools, the simplest are scatters and box plots, regressions, analysis of residuals and so on. Furthermore, data transformations can be easily performed allowing the inspection of data relationships in the most appropriate scale.

In our application, possible new query edits have been identified by analysing the marginal and two-dimensional data distributions through scatter and box plots, using in some cases data transformations in order to better investigate some specific surveyed phenomena. Evidence of strong linear relations among pair of logically related items suggested testing the use of these relations for building new ratio edits.

By analysing marginal distributions of selected variables, univariate query edits are also defined in order to check that values of a given item are inside an appropriate acceptance region. An important reason for using univariate query edits in the FH algorithm is increasing the capability of the algorithm itself of detecting particular kinds of errors (like consistently reported unit measure errors or other kinds of non representative outliers that cannot be identified by using ratios).

2.3. Analysis on test data: find bounds for the original and the new query edits

Once the new potential query edits (ratios and univariate ones) have been identified, their optimal acceptance bounds are to be determined. In the application we developed a procedure based on the Hidiroglou and Berthelot algorithm (Hidiroglou *et al.*, 1986) for identifying acceptance bounds for ratio and univariate[3] edits. We also developed an algorithm for calibrating the Hidiroglou and Berthelot (HB in the following) parameters when historical data are available for a sample of units that differs from the one currently available.

2.3.1. Determining acceptance bounds for ratio edits

Given two related variables Y_j and Y_k observed on a given sample s, we want to determine the acceptance bounds of the distribution on s of the ratio $R = Y_j/Y_k$. To this aim, we use the following algorithm based on the HB method:

1. symmetry the distribution of R through the following transformation:

[3] Given the marginal distribution of a variable Y observed on a sample s, an univariate edit corresponds to an acceptance region determined on the Y distribution on s.

$e_i=1-(r_{median}/r_i)$ if $r_i < r_{median}$ (and in this case results $e_i < 0$);
$e_i=(r_i/r_{median})-1$ if $r_i \geq r_{median}$ (and in this case results $e_i > 0$),

where $r_i = y_{ji}/y_{ki}$ is the value of *R* in the unit *i* and r_{median} is the median of the *R* distribution.

2. Define the lower (*L*) and upper (*U*) acceptance bounds as:
$L = e_{inf} = e_{median} - C \times d_{Q1}$
$U = e_{sup} = e_{median} + C \times d_{Q3}$

where:

- $d_{Q1} = MAX\{ e_{median} - e_{Q1}, A \times e_{median}\}$ and $d_{Q3} = MAX\{e_{Q3} - e_{median}, A \times e_{median}\}$
- e_{Q1}, e_{median}, e_{Q3} are respectively the first quartile, the median and the third quartile of the e_i distribution;
- A is a suitable positive number introduced in order to avoid the detection of too many outliers when the e_i are concentrated around their median;
- C is a parameter used for calibrating the acceptance region width.

3. Express the acceptance bounds (r_{inf}, r_{sup}) of the original distribution through the following back-transformation:

$$r_{inf} = r_{median}/(1-e_{inf})$$
$$r_{sup} = r_{median} \times (1+e_{sup}).$$

A central role in determining the acceptance limits for a given ratio is played by the C parameter. Roughly, C is a calibration parameter measuring the size of the acceptance region. We tried to develop an algorithm to "estimate" C from data. In particular, we implemented a generalized procedure calibrated on historical data that can be applied to current data, making the assumption that in the two considered periods the variable distribution as well as the error mechanism generating outliers are similar.

In this procedure we exploit the availability of both true and raw values for historical data. In fact, in this case, for each algorithm parameters setting we are able to build a 2x2 contingency table T containing the cross frequencies of original status (erroneous, not erroneous) vs post-editing status (suspicious, acceptable). It is obvious that the higher are the correct classification frequencies, the better is the quality of an editing method. Since in general the two correct classification frequencies (erroneous data classified as suspicious and not erroneous data classified as acceptable) cannot be simultaneously maximized, a “best” contingency table can be found only in subjective way: for example if it is believed that to classify as suspicious a correct value is more dangerous than to accept an erroneous value, the two erroneous classification frequencies are “weighted” in different way.

<u>Algorithm applied to historical data</u>

1. Initialize parameters A (A=0.05 as generally suggested), C, and a “step” parameter S; the initial values C_0 and S_0 depend on ratio distribution characteristics.
2. Repeat the following steps several times for different values of S:
   - iterate the revised HB method for K different values C_k, where $C_k = C_{k-1} + S$ and K is chosen so that the resulting acceptance region will include the most extreme not erroneous value;
   - for each C_k analyze the correspondent 2x2 contingency table T_k and identify $T_k = T^*$ “optimum”.
3. Determine the S* generating the same T* a "high" number of times.

Algorithm applied to current data

1. For k=1,…K, detect outliers through revised HB method with $C_k = C_{k-1} + S^*$.
2. Choose the "optimum" C* among C_k producing the same number of detected outliers a "high" number of times.

2.3.2. Determining acceptance bounds for univariate edits

The determination of the acceptance bounds for the marginal distribution of a given variable Y (univariate edit) has been performed by following the same procedure used for ratio edits. In this case, the previous algorithms are directly applied on the marginal distribution of Y.

2.4. Analysis on test data: identifying the 'best' set of edits

The selection of the final editing strategy is performed by analysing the results obtained by applying on test survey data different sub-sets of constraints (chosen among the original edits and the new ones) and analysing the corresponding results. The adopted quality indicators correspond to the probabilities α and β defined in section 2.1. The *optimal* set of edits is obviously the one that produces the best trade-off between α and β. A more accurate evaluation should take into account also the relative importance of detected/undetected errors: further analyses are requested to this aim.

3. The experiment

In this paragraph we illustrate the application of the strategy described in the previous section and the corresponding results.

Data used for the experiment are subsets of a sample of the U.K. *Annual Business Inquiry* inquiry (*ABI*), provided by the U.K. Official National Statistics in the context of the EUREDIT project. Only one Economic Sector has been considered in the application, while two years data (1997 and 1998) have been used. Both true and artificially contaminated data have been made available for the year 1997: these data have been used as *test* ones. The test sample consists of 6,099 records, while 6,233 observations are to be analyzed for 1998. Each record corresponds either to a *long* or to a *short* form. The data sets contain responses to selected questions from the ABI: there are some differences in the definition and treatment of some items between *long* and *short* forms and between years. Out of the 33 collected variables in *long* forms, 27 ones have been artificially perturbed and need to be edited. Out of the 17 items reported in *short* forms, 11 need to be edited. Because of the need of working within homogenous sub-sets of data, the ABI samples have been stratified by form type and class of registered turnover (*large* businesses have registered turnover greater than 1 million of pounds). Therefore, the overall data processing has been split in several sub-analyses, each referring to a different stratum.

The original ABI set of edits consists of 25 rules (*hard* and *soft*). Unfortunately, being most of the variables involved in not more than two edits, the initial set of edits doesn't form a well connected 'grid' of constraints among items. Furthermore, most of the edits are soft with too narrow acceptance regions. For these reasons, starting from the initial set of edits, we tried to improve the effectiveness of the editing process by introducing new edits and by redefining the acceptance region of some of them. In particular we have introduced both ratio and univariate edits. Univariate edits have been defined only for the most important employment variables, originally involved in too few edits. In this way the main variables are involved in a larger number of edits and the system is more effective in identifying errors.

3.1. Analysis of original query edits

The analysis of the original edits has been performed by using the measures introduced in section 2.1. The failure rates of the original query edits range from a minimum of 1.22% to a maximum of 21.7% for long forms, while in the subset of short forms they range from a minimum of 3% to a maximum of 27.6%. Furthermore, it has to be noted that the original set of ABI rules is quite *poor*, in the sense that edits do not form a well connected 'grid' of constraints among variables: each variable is in fact involved in one or at maximum two hard edits, and the most important survey variables (e.g. TURNOVER, NUMBER OF EMPLOYEES) are involved only in few soft edits. As a result, by using the original set of (hard + soft) edits in the FH probabilistic algorithm, it does not work in the optimal conditions. This is directly indicated by the probabilities α and β of correct/erroneous data classification computed after the application on the test data of the FH algorithm using the original set of rules. In the first part of tables 1 and 2 the α and β probabilities obtained using the original edits are reported for the main survey variables TOTAL EMPLOYEES COSTS, EMPLOYEES WAGES AND SALARIES, NUMBER OF EMPLOYEES, TURNOVER, TOTAL PURCHASES separately for long and short forms.

3.2.Identifying potential new query edits

As discussed in section 2.2, the identification of new rules has been performed by a graphical analysis of the marginal and joint distributions of the considered variables. Data distributions have been explored within specific domains where appropriate. Data transformations (e.g. in logarithmic scale) have been performed when useful.

As an example of graphical representation and analysis of data relations and errors, in figures 1 and 2 multiple scatter plots of transformed variables (logarithm of TURNOVER, REGISTERED TURNOVER and TOTAL PURCHASES) are shown. Data refer to a given data domain, i.e. long forms having the registered turnover less than £1,000,000 (*small* long forms). The plots reported in figure 1 have been obtained from the *true* test data, while in figure 2 the same plots refer to the *raw* test data. As it can be seen from both plots, strong relationships exist between the three analysed items, thus confirming the opportunity of including the corresponding ratios in the editing strategy.

Furthermore, figure 2 brings to light the existence of markedly separate clouds of observations.

These clusters could include both acceptable but unusual data relations (*representative outliers*) and erroneous data. For example, by observing the *l_turnover* (logarithm of TURNOVER) vs *l_turnreg* (logarithm of the REGISTERED TURNOVER) scatter plot it is evident as a same relation between these two items exists in the two clusters, but it is also probable that the upper cloud corresponds to observations affected by consistent systematic errors due to a wrong measure unit.

Our aim is to design the edits in order to make the FH algorithm identify also this kind of errors, taking under control the probability that original true data are classified as unacceptable.

3.3. Find bounds for both original and new query edits

The bounds of query edits were found by following the procedure illustrated in sections 2.3.1 (for ratios) and 2.3.2 (for univariate edits).

3.4. Identifying the 'best' set of edits

The optimal set of edits has been determined based on the following procedure:

1) out of the original set of edits, select all the hard ones (*starting set of edits*);
2) taking into account the percentages of true data failing the original query edits, include in the *starting set of edits* the original soft edits not requiring the revision of their bounds;

3) include in the *starting set of edits* the original edits judged as ineffective with bounds revised as described in previous sections;

4) select the *optimal set of edits* by analysing the α and β probabilities produced by using different sub-sets of query edits defined in steps 2 and 3 in association to the hard ones.

Figure 1 – Multiple scatter plots of logarithms of Total Turnover (l_turnover), Registered Turnover (l_turnreg) and Total of Purchases (l_purtot) for the Small-Long Forms stratum on the *true* 1997 survey data.

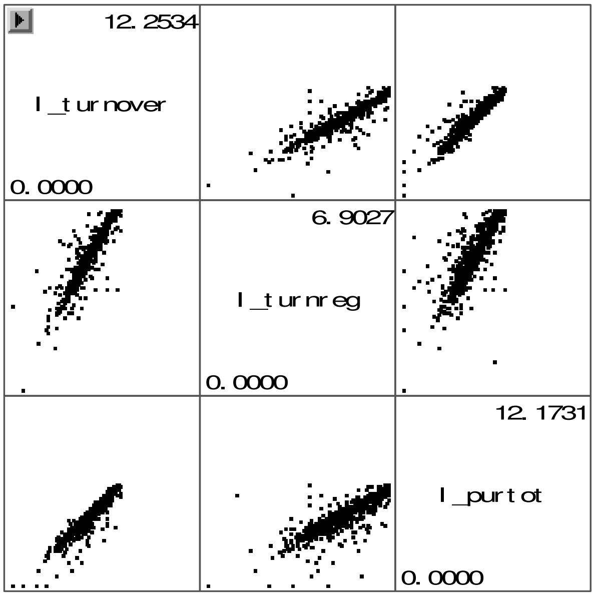
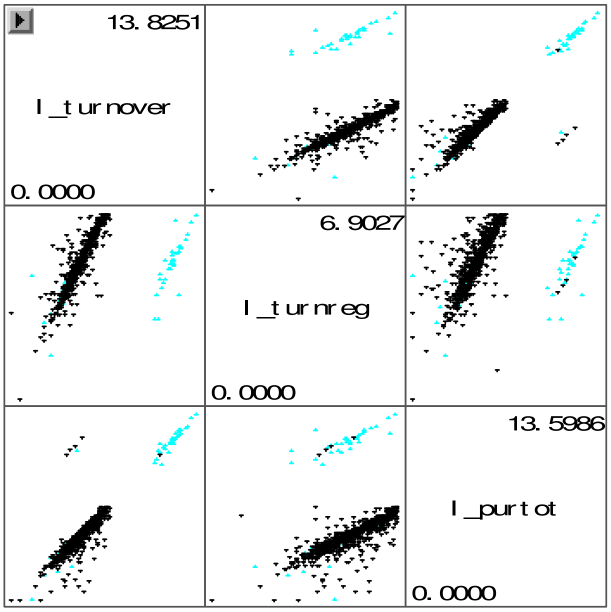


Figure 2 – Multiple scatter plots of logarithms of Total Turnover (l_turnover), Registered Turnover (l_turnreg) and Total of Purchases (l_purtot) for the Small-Long Forms stratum on the *raw* 1997 survey data



It is important to stress the fact that, in the latter step, different combinations of variable weights have been tested for any given set of edits.

For each type of form and for each variable of interest, in tables 2 and 3 we report the results obtained by using the original and the *optimal set of edits*. In each table, the cross-classification table associated to each variable has the form of Table 1 (see section 2.1): rows indicate the status of the true variable values (E=0 → not erroneous value; E=1 → erroneous value), while columns indicate the status of the edited variable values (S=0 → value classified as not erroneous; S=1 → value classified as erroneous). Frequencies of cells out of the main diagonal correspond to the α and β misclassification probabilities.

Table 2 – Long forms - Cross-classification of values by variable and set of edits

	Original edits				**Revised edits**			
FIELD		**S=0**	**S=1**	*Total*		**S=0**	**S=1**	*Total*
TURNOVER	**E=0**	1,261 **(89.18)**	153 **(10.82)**	1,414	**E=0**	1,364 **(96.74)**	46 **(3.26)**	1,410
	E=1	3 **(4.48)**	64 **(95.52)**	67	**E=1**	7 **(10.45)**	60 **(89.55)**	67
	Total	1,264	217	1,481	**Total**	1,371	106	1,477
WAGES OF EMPLOYEES	**E=0**	1,379 **(99.78)**	3 **(0.22)**	1,382	**E=0**	1,370 **(99.42)**	8 **(0.58)**	1,378
	E=1	86 **(86.87)**	13 **(13.13)**	99	**E=1**	42 **(42.42)**	57 **(57.58)**	99
	Total	1,465	16	1,481	**Total**	1,412	65	1,477
TOTAL EMPLOYEES COSTS	**E=0**	1,356 **(99.41)**	8 **(0.59)**	1,364	**E=0**	1,357 **(99.78)**	3 **(0.22)**	1,360
	E=1	106 **(90.6)**	11 **(9.4)**	117	**E=1**	64 **(54.70)**	53 **(45.30)**	117
	Total	1,462	19	1,481	**Total**	1,421	56	1,477
NUMBER OF EMPLOYEES	**E=0**	965 **(66.37)**	489 **(33.63)**	1,454	**E=0**	1,430 **(97.54)**	36 **(2.46)**	1,466
	E=1	8 **(29.63)**	19 **(70.37)**	27	**E=1**	9 **(81.82)**	2 **(18.18)**	11
	Total	973	508	1,481	**Total**	1,439	38	1,477
TOTAL PURCHASES	**E=0**	1,244 **(93.6)**	85 **(6.4)**	1,329	**E=0**	1,306 **(98.27)**	23 **(1.73)**	1,329
	E=1	136 **(89.47)**	16 **(10.53)**	152	**E=1**	80 **(54.05)**	68 **(45.95)**	148
	Total	1,380	101	1,481	**Total**	1,386	91	1,477

In case of long forms, for example, for the TURNOVER the final set of rules minimises the probability of introducing new errors in the data (in fact, the probability β of classifying original true data as errors decreases from 10.82% to 3.26%). This high improvement has a negative impact on the edits capability of recognising errors (in fact the probability of classifying true errors as acceptable increases from the original 4.48% to 10.45%).

A similar result has been obtained for the NUMBER OF EMPLOYEES, but in this case the reduction of the β probability is higher than in the previous case (it decreases from 33.63% to 2.46%). Also in this case this result is associated to a lower capability of the algorithm to identify true errors (α increases from 29.63% to 81.82%).

Different effects correspond to variables WAGES OF EMPLOYEES and TOTAL EMPLOYEES COSTS: in both cases the capability of the error localisation algorithm of correctly classifying true data is high and similar. In these cases, the probabilities of correctly classifying true errors as unacceptable increase for both items (e.g., the percentage of detected true errors for TOTAL EMPLOYEES COSTS increases from 9.4% to 45.3%).

A mixed situation characterises the TOTAL PURCHASES: in this case, both the α and β probabilities of erroneously classifying data decrease when using the revised set of edits.

Similar considerations can be performed for short forms (Table 3).

Table 3 – Short forms - Cross-classification of values by variable and set of edits

	Original edits				**Revised edits**			
FIELD		**S=0**	**S=1**	*Total*		**S=0**	**S=1**	**Total**
TURNOVER	**E=0**	3,916 (**88.96**)	486 (**11.04**)	4,402	**E=0**	4,348 (**98.77**)	54 (**1.23**)	4,402
	E=1	198 (**91.67**)	18 (**8.33**)	216	**E=1**	41 (**19.07**)	174 (**80.93**)	215
	Total	4,114	504	4,618	**Total**	4,389	228	4,617
TOTAL EMPLOYEES COSTS	**E=0**	3,388 (**77.67**)	974 (**22.33**)	4,362	**E=0**	4,361 (**99.98**)	1 (**0.02**)	4,362
	E=1	218 (**85.16**)	38 (**14.84**)	256	**E=1**	57 (**22.35**)	198 (**77.65**)	255
	Total	3,606	1,012	4,618	**Total**	4,418	199	4,617
NUMBER OF EMPLOYEES	**E=0**	3,574 (**78.64**)	971 (**21.36**)	4,545	**E=0**	4,532 (**99.74**)	12 (**0.26**)	4,544
	E=1	59 (**80.82**)	14 (**19.18**)	73	**E=1**	35 (**47.95**)	38 (**52.05**)	73
	Total	3,633	985	4,618	**Total**	4,567	50	4,617
TOTAL PURCHASES	**E=0**	4,027 (**97.91**)	86 (**2.09**)	4,113	**E=0**	4,078 (**99.15**)	35 (**0.85**)	4,113
	E=1	489 (**96.83**)	16 (**3.17**)	505	**E=1**	402 (**79.76**)	102 (**20.24**)	504
	Total	4,516	102	4,618	**Total**	4,480	137	4,617

From the previous results it results a general low capability of the editing strategy of correctly identifying true errors (all the *alpha* values are quite high). This fact depends on some main reasons. First of all, as already mentioned in previous sections, the FH algorithm works in an optimal way when variables are involved in many edit rules and the error mechanism is random. In case of ABI, most of the variables appear just once in the edits. Furthermore, since most of the edits are soft and the corresponding acceptance regions are too narrow, we had to enlarge them in order to avoid the classification as errors of acceptable data. Because of the poor knowledge of the investigated phenomena, we preferred approaching the problem by prioritizing the identification of very large errors, and minimizing the probability of misclassifying correct data. Since most of large errors in this survey correspond to the systematic error "variable values multiplied by a 1,000 factor", the correct way of dealing with them in a real context is to preliminary identifying them through appropriate techniques. On the other hand, since our main goal was to evaluate strengths and weaknesses of probabilistic editing, we tried to use the FH algorithm also for identifying this kind of error, even if it is a priori known that this approach is not suitable to this aim.

Results show that the variables involved in a higher number of edit rules are those with lower α values (TURNOVER, WAGES OF EMPLOYEES and TOTAL EMPLOYEES COSTS). It has also to be observed that β values are generally very low, as a consequence of the attention paid to the misclassification of acceptable data.

In any case, it is useful to note that a natural decrease in the quality of the editing process of 1998 compared with 1997 data is expected, since in the latter case we calibrated the procedure parameters knowing the true values. However since for some variables the decrease is quite remarkable, this seems to suggest further causes.

In such situations, a first analysis should be addressed in order to verify if the error mechanism in the two surveys can be considered the same. In fact, the approach of editing a survey through a strategy calibrated on a development dataset (a previous survey where original contaminated data and imputed data are available) is strongly based on the assumption that the error mechanism affecting data is basically the same. It is obvious that a direct comparison of the error mechanism cannot be made, nevertheless other indirect information might be useful: for instance the analysis of the frequency of the edit failures in the two years. It is also clear that a change in the error mechanism in just one variable may affect the editing performance also on the others, because all variables are connected by the overall grid formed by the edit rules.

As already mentioned in the paper, an important aspect to be further investigated relates to the *importance* of not identified errors, in terms of potential biasing effects on final figures. On the other hand, in our application we observed that the actual *optimal* editing strategy is able to identify the most part of biggest errors affecting test data (this result can be observed in Figure 2, where light points correspond to data classified as erroneous by the probabilistic error localisation algorithm). An acceptable percentage of errors are "lost" in the upper clusters, making us confident about the quality of results also in terms of preservation of aggregates and distributions. However, further investigations are needed concerning these aspects.

3.5. Identifying the 'best' set of edits for the current survey repetition

In order to build the editing strategy for the current survey data, the same approach adopted in the definition of the final editing procedure for test data has been followed. Similar criteria have been used for obtaining strata and for defining the edits and the FH parameters, exploiting the experience coming from the test phase. In other words, the editing strategy for current data has been designed by essentially reproducing the process of data and parameter definition followed for test data.

A similar approach and the same algorithms described in paragraphs 2.3.1 and 2.3.2 have been adopted for defining the bounds of query edit in each data domain. The final set of edits for current data has been defined by reproducing as much as possible the final structure of rules applied to test data.

4. Conclusions

In Official Statistics, large part of survey data are edited by using automatic procedures using pre-specified sets of rules checking for the coherence of the captured information. The potential drawbacks of using inaccurate or inappropriate edits in automatic editing are well known: for example, the increasing risk of introducing new errors among data, as a consequence of the misclassification of true data as errors, and the amplification of the risk of biasing effects on marginal and joint observed distributions. Particularly in business surveys, these problems can be more relevant because of the use of the so called query edits (in particular, ratio eidts), which bounds are to be carefully determined in order to limit the above mentioned risks.

In the paper, a strategy for analysing and improving the effectiveness of the automatic editing in business surveys when using probabilistic algorithms is illustrated. This strategy aims at analysing the query edits effects on data in order to improve their effectiveness, and uses test data for designing an *otpimal* editing strategy (in terms of probability of correct data classification). In the adopted approach Exploratory Data Analysis (EDA) and the Hidiroglou and Berthelot method are combined together.

In the paper we concentrated on the probabilistic algorithms for editing continuous data proposed by Fellegi and Holt as implemented in the GEIS software.

From the analysis of the application results, all the main elements characterizing strength and weakness of an editing strategy based on the Fellegi and Holt approach have been highlighted. The main problem relates to the setting of the edit rules. This is not a simple task: in fact, if on one hand edits must form a grid of "well connected" rules, on the other hand they have to be thought in order to avoid the problem of over-editing. A similar trade-off problem arises when soft edits are introduced. Since the algorithm interprets soft edits as they were hard, the acceptance regions of each soft edit rule must be large enough in order to not cut the tails, but at the same time strict enough in order to find as many errors as possible. A further problem whith the Fellegi and Holt approach relates to its ability of dealing in the appropriate way with not random error mechanisms. The application has shown that even if the effectiveness of the probabilistic error localization results can be markedly improved by carefully designing the edit rules, the most appropriate way of dealing with non-sampling errors other than the random ones (e.g. representative and non representative outliers and systematic errors) remains their treatment before any automatic data processing phase.

References

Chambers R. (2000). Evaluation Criteria for Statistical Editing and Imputation, T001.05*, EUREDIT report*.

Cotton C. (1991). *Functional description of the generalized edit and imputation system*, Statistics Canada, Business Survey Methods Division, July 25.

De Waal T. (1996). CherryPi: a computer program for automatic edit and imputation, *UN Work Session of Statistical Data Editing*, 4-7 November 1996, Voorburg.

Des Jardins D. (1997). Experiences with introducing new graphical techniques for the analysis of Census data, *UN-ECE Conference, Work Session on Statistical Data Editing*, 14-17 October, Prague.

Des Jardins D., Winkler W. (2000). Design of Inlier and Outlier Edits for Business Surveys, *Proceedings of the Second International Conference on Establishment Surveys,* 17-21 June, Buffalo.

Ellfors C., Arvidson G., Granquist L., Nornerg A. (2000). Experiences of introducing exploratory data analysis methods at Statistics Sweden, *Proceedings of the Second International Conference on Establishment Surveys*, 17-21 June, Buffalo.

Esposito R., Lin D., Tideman K. (1994). ARIES: a Visual Path in the Investigation of Statistical Data, *Journal of Computational and Graphical Statistics,* Vol. 3, N° 2, 113-125.

EUREDIT official web-site: www.cs.york.ac.uk/euredit/

Fellegi I.P., Holt T.D. (1976). A systematic approach to edit and imputation, *Journal of the American Statistical Association,* vol.71, pp.17-35.

Garcia-Rubio E., Villan I. (1990). DIA System: Software for the Automatic Editing of Qualitative Data,. in *Proceedings of the Sixth Annual Research Conference of the Bureau of the of Census*, Washington, D.C.: U. S. Bureau of the Census.

Granquist L. (1992). A Review of methods for rationalizing the editing of survey data, *Statistical Data Editing Methods and Techniques*, United Nations, vol. I.

Granquist L. (1996). The new view on editing, *UN-ECE Conference, Work Session on Statistical Data Editing*, Voorburg (Netherlands).

Granquist L. (1995). Improving the Traditional Editing Process, in *Business Survey Methods*, eds. B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, and P.S. Kott, New York: Wiley, 385-401.

Hidiroglou, M A, Berthelot, J M (1986). Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, June 1986, vol.12, N.1, pp.73-83.

Houston G., Bruce (1993). GRED: Interactive Graphical Editing for Business Surveys, *Journal of Official Statistics*, Vol. 9, N° 1, pp. 81-90.

Kovar J.G., MacMillian J.H., Whitridge P. (1988). Overview and strategy for the generalized edit and imputation system, Statistics Canada, Methodology Branch, April 1988.

Kovar, J.G., Winkler, W.E., (1996). Editing Economic Data, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 81-87.

Latouche M., Berthelot J.M. (1992). Use of Score Functions to Prioritise and Limit Recontacts in Editing Business Surveys, *Journal of Official Statistics*, Vol. 8, N. 3, Part II.

Nordbotten S. (1965). The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means of Improving the Quality of Statistics, *Proceedings of the International Statistical Institute Meetings,* pp.417-441.

Riccini E., Silvestri F., Barcaroli G., Ceccarelli C., Luzi O., Manzari A. (1995). *La metodologia di editing e correzione per variabili qualitative implementata in SCIA*. Documento Istat.

Todaro T. (1999). Overview and evaluation of the AGGIES automated edit and imputation system, *U.N. Work Session of Statistical Data Editing*, 2-4- June, Rome.

Tukey J.W. (1977). *EDA: Exploratory Data Analysis,* Addison-Wesley, Massachusets.

Winkler W. E., Draper L.A. (1997a). The SPEER Edit System, *Statistical Data Editing (Volume 2), Methdos and Techniques*, United Nations.

Winkler W. E., Petkunas T. (1997b). The DISCRETE Edit System, in J. Kovar and L. Granquist (eds), *Statistical data Editing, Volume II, U.N. Economic Commission for Europe,* 56-62.