IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD FOR OUTLIER DETECTION IN CATEGORICAL MULTIVARIATE SURVEYS[1]

CARLOS LÓPEZ
Ingenieros Consultores Asociados
Cerro Largo 1321, Montevideo, Uruguay

Abstract: The detection of errors and outliers is an important step in data processing, specially those errors arising from the data entry operations because they are of the entire responsability of the data processing staff. The duplicate performance method is commonly used as an attempt to detect such type of errors. It implies typically typing twice the same data without any special precedence. If the errors are uniformly distributed among individuals, retyping a fraction of the total will also remove typically the same fraction of the errors. A new method which is able to improve that procedure by sorting the records putting first the most unlikely ones is presented. The ability of the present methodology has been tested by a Monte Carlo simulation, using an existing database of categorical answers of housing characteristics in Uruguay. At first, it has been randomly contaminated, and after that, the proposed procedure applied. The results show that if a partial retyping is done following the proposed order about 50% of the errors can be removed while keeping the retyping effort between 4 and 14% of the dataset, while to attain a similar result with the standard methodology 50% (on average) of the database should be processed. The new ordering is based upon the unrotated Principal Component Analysis (PCA) transformation of the previously coded data. No special shape of the multivariate distribution function is assumed or required.

Some keywords: Data checking, Census data management, Outlier detection, Principal Component analysis, Categorical data

I.- Introduction

A recurring problem in the creation or maintenance of a large computerized data base is the correctness of the information entering the base. If high volumes of data are involved, then data entry operation tends to be carried out by less qualified personnel, and verification is less extensive. Thus, action is required to maintain the base's integrity, and the fact that large volumes of machine-readable material are involved suggests that, as far as possible, this screening action should be automated. Clearly typing errors is not the single source of errors existing at the machine level; however on principle they can be kept under control.
There are many classical examples of typing errors, even from the early days of computer development. A classic one is described by Coale *et al.* (1962) who reported an error in the 1950 U.S. census figures that resulted when a small fraction of computer cards were

[1] Published in the Journal of the Italian Statistical Society, 1996, **5**, 2, 211-228

punched one column to the right of the proper position, so that an unusually high number of 14-year-old widows was reported. Only after discovering the reason for that error, they were able to anticipate errors in the age-distribution of Indians individuals. Even though the total amount of erroneous records was fairly small (below 1/100 of one per cent) certainly rare categories might be greatly affected by such spurious cases. Notice that all the fields in the record have values within their own feasible range.

A general procedure for locating typing errors in a data entry process is the duplicate performance method. If a data typing operation is performed twice, independently, and if the results are compared by a method that can be assumed error free (such as a computer program comparing files after data entry), and if all the disagreements are corrected, then the only errors remaining in the data set are those where both staff members were in error. If the ratio of disagreements to total items is low, then the individual error rates of both persons are low, and the probability of joint errors (the product of the probabilities of individual errors) is lower still (Strayhorn (1990)). The method is extremely simple, and it applies for any kind of data, both quantitative or categorical. Despite its simplicity, it has some desired properties: the probability of locating an error is independent of the error itself, so trivial errors will be corrected as well as subtle ones. This will help in keeping the statistical properties of the database. It is also independent of the *order* the retyping is performed, so in principle, if only a fraction of the dataset is retyped, typically the same fraction of the errors will be corrected. Another advantage is that the procedure does not require a large database, so it can be applied also to small ones.

The literature about editing survey data is considerable, but somewhat scarce regarding quality control of categorical data. Fellegi *et al.* (1976) presented a methodology specifically suitable for qualitative or categorical data. It is based upon the existence of rules which relate the different fields in each record. Such rules should be given by experts, and express the judgment of them that certain combinations of values or code values in different fields are unacceptable. If a particular record does not satisfy one or more of those rules, the field (or fields) that contribute to them are rejected or modified in order to attain a feasible record. Notice that this procedure relies on the existence of explicit rules (and experts behind them) and requires some manipulation of the rules before application. No experimental results are presented in the paper.

Paradice *et. al*. (1991) presented a methodology for controlling *incoming* data to a database. Their approach focuses in minimizing the time a wrong record stays in the system, basically by limiting its chances to pass some logical tests created by experts, and tailored for the particular application. Not all the attributes of a record are important for all applications, so new tests may be required for different users of the same data. For the applications the authors are involved in, individual records should be handled also individually and not "in aggregate" so errors will have significant effect for one particular record, but possible not for the whole database. The paper also gives a performance evaluator for the overall error diagnostic procedure, which gives an enterprise measure of success. They claim this benchmark gives a clear measure for evaluating current verification procedures and proposed changes. Even though we could not apply this methodology for an already existent database, or even one that is created in a single task (a national census, for example) it will give us the chance to qualify the procedures used in a continuously updated process (like economical data).

Apart from the methods specially devised for categorical data, we want to mention some of the methods available for quantitative data, since we will adapt some of them for the former

case. Typically the authors rely on assumptions about the data distribution. For example, Little *et. al*. (1987) presented a methodology based upon multivariate normality of the data. They used a log-transformed population, and look for linear relationships between the new variables. Using the squared Mahalanobis distance as an estimator, the author analyzed its sampled distribution exploring graphically the departure from a transformed chi-squared distribution. All instances that renders values that are "far" in some sense to the theoretically assumed behavior are flagged and edited by experts. They also extend their methodology for incomplete datasets, limiting for each individual the Malahanobis distance to the available data.
A related approach has been presented by Hawkins (1974) based upon Principal Component Analysis (PCA) of the data. Instead of using the Mahalanobis distance, he proposed to use other statistics which are intended to be more sensitive and to have better performance when compared with standard statistics (χ^2, etc). However, some problems arise while calculating the eigenvalues of the covariance matrix in real data. The existence of outliers may affect its values, so more robust procedures should be preferred, and not all the data can be regarded as normally distributed.
López *et al*. (1994) presented a methodology that overcomes some of these drawbacks. Instead of using the distribution of a single number like the Mahalanobis distance or the Hawkins´s statistic for flagging an instance, they proposes to use k independent tests applied over the projections of the given data on the eigenvector´s basis. No fitting with any distribution is required. The rest of the paper is devoted to show a connection of quantitative data procedures to categorical ones, and to present some simulated results.
The work is organized in nine sections. Section **I Introduction**, has discussed some work representing the state-of-the-art on the subject. Section **II Motivation and assumptions** introduces the main ideas. Section **III Experimental test design** describes the simulation carried out to examine the performance of the method with a particular dataset. Section **IV Methodology** describes the steps required to apply the procedure. Section **V Results** summarizes the success by means of some performance indicators and finally section **VI Discussion** compares the results and analyzes advantages and drawbacks, while section **VII** states the **Conclusions**. **Acknowledgments** and **References** are included as sections **VIII** and **IX**.

II.- Motivation and assumptions

For the sake of simplicity, we will assume hereinafter that by typing twice a record all errors are removed. This will help us in simplifying some arguments, and the reader will easily notice that this not a key hypothesis.
We mentioned before that the duplicate performance method ability is independent of the order the records are retyped. If we assume that the wrong individuals are uniformly distributed in the population, retyping a fraction of the dataset will most likely correct the same fraction of errors. This paper is devoted to find a reordering in the data, designed to put first the records that are prone to hold some errors, so partial retyping will eliminate more errors than without any reordering.
To do so, we will try to locate outliers in the dataset. What is meant by outlier in categorical data may differ from the concept for real-valued data. It is also assumed here that the dataset has passed successfully some trivial logical tests, which pointed out for example,

more than one mark in mutually exclusive answers, or similar things. Also all the coded values are within their prescribed ranges. These logical tests are very crude, and certainly should not be confused with the edits designed by experts in the particular data (Fellegi *et. al*. (1976)). It should be regarded more as a computer specification for the data, rather than a quality control procedure.
So we will consider only the problem of selecting a specific record (a single survey in the example) on the basis that there is something in the answers that make them unusual. Such record should be retyped. Notice that this procedure will diminish the variability in the data, because "feasible" errors are prone to be ignored.
In a real processing environment, if the record is still unusual it will be carefully analyzed by a trained specialist, which may found (or not) reasons to reject or modify some answers in the particular record. This fact will not be considered here, but the methodology is in fact devoted to give the specialist a smaller selected set, with higher probability of holding true errors.
It should be stressed that errors arising from the the typing stage is one among others sources of errors; however they are important in the sense that they can be kept under control. Significant errors can be introduced in earlier stages (like the coding of non-categorical answers) which cannot be controlled by the duplicate performance method, but can be handled by the procedure to be presented below.
In categorical data, the codification procedure usually generates for each question a set of feasible values. For technical reasons, those values are frequently coded as integers, but the integer value itself is meaningless. In order to manage categorical data with PCA, one should translate such integers in a way that the results do not depend upon :

a) changing the order of the alternatives in the question
b) changing the integer codes

It will also be assumed that all the answers have the same relative importance. The technique to be presented, was designed to be applied for processing the 1996 National Census of Population and Housing in Uruguay (population ~3 million, houses ~200.000) to check only categorical answers. The data was not be typed, but scanned and processed via automatic recognition routines, handling handwritten text, number and marks. Even though automatic recognition of marks are known to be very reliable, it is intended to flag and check dubious data while keeping the manual typing effort low.

III.- Experimental test design

A Monte Carlo simulation is performed, modifying the answers of a subset of the raw data collected and processed during the 1985 National Census in Uruguay, and testing the ability of the methodology to locate them. The subset chosen reports housing characteristics in the Flores region and has been typed twice. Only private houses cases, without missing values were considered. The final set has 4963 events, but to diminish computer time requirements, the simulations were carried out over only 2500 individuals.
The dataset is claimed to be typed twice, and the original records are not available. This fact makes it difficult to properly model a pattern of "rule" for real errors, so only reasonable assumptions could be made.

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

In order to obtain a contaminated set, a prescribed number of records were chosen at first and then a random number generator choose a fixed number of questions (out of 20) to modify. For each of them, the existing answer was changed to a different value, but still belonging to its feasible set (assuring that they were different with the original one). That was considered a suitable choice for modeling "true" errors. The *total number of contaminated records* were fixed as 10, 5, 3 and 1.5% of the subset of 2500 individuals. The figures to be presented correspond to the 3% case, which implies 75 wrong cases.

IV.- Methodology

In this section, all steps required for processing a categorical dataset are described. Given the data, the corresponding question list and the feasible options, the user should eliminate those fields which are *a priori* uncorrelated with the others. Typical examples for survey data are all the information related with the zip code, city code, address, etc. Also numerical quantitative data should not be considered (for example: age, size of the building, etc.) except if a categorization is applied.

The dataset is usually available in table format, one individual per row, and one question per column. In order to have a numerically useful representation, we will binarize the dataset, creating a new table containing only 1 or 0. This also make the data homogeneous (dimensionless). In order to binarize the dataset, one may think on a multiple choice sheet. For any particular question, there are room to choose between some (maybe mutually exclusive) alternatives. Instead of coding a single number for the answer, we may equally store all the alternatives, putting a 1 or 0 if the option is true or not. In other terms, each column of the original table expands to as many columns as alternatives in the question, allowing only 0 or 1 as an answer. After repeated for all questions, the data are transformed into binary format, and the covariance matrix can be calculated.

Since the methodology to be applied relies upon exploiting the empirical relationships between the answers, all the questions that are weakly correlated with the other data will not be considered by this procedure. In early stages of the work it has been found that also "almost trivial" answers were a source of problems, because they behave like uncorrelated answers. For example, in the test dataset more than 96% of the population has direct connection with the electrical power supply. So the corresponding answer has trivially nothing to do with the others answers. That was also the case for the questions "do you have a freezer?", "do you have telephone at home" and others which almost always have been answered "no" in this particular dataset. So, if more than 95% of the population answers are the same for any binary option, the option will be removed for the final test. A second criteria was applied trying to eliminate uncorrelated answers. If the off-diagonal elements of the correlation matrix are very close to 0, the corresponding option is also removed. The threshold has been chosen as 10 times the machine ε. (defined as the largest number which satisfies $1+\varepsilon=1$ in finite precision arithmetic). Those questions were removed before applying the outlier detection process. The final dataset has 20 questions, with 69 alternatives (options).

To highlight unusual records, a PCA derived method is being proposed. PCA is a well known methodology that transform the original (mutually correlated) data in another uncorrelated but equivalent presentation. Usually such transformation is performed in order to reduce the dimensionality of the problem. Only the first Principal Components (PC) are

retained, and most of the variance in the original set is explained through them. The remaining PC are usually neglected.

Hawkins (1974) pointed out that those neglected PC may serve as outlier detectors. PCA transforms the covariance matrix Σ to diagonal form, so $E\Sigma E^T = \Lambda$. Any instance of the data X_i is also transformed to $W_i = E(X_i - \mu)$, being μ its sampled mean value. Obviously the elements w_{ij} are a linear combination of the components of X_i. The $w_j(X_i) = w_{ij}$ components are mutually uncorrelated, and have variance λ_j (the associated eigenvalue). The PCA residual test statistic is defined by Hawkins as

$$T_2(X_i) = \sum_{j=1}^{k} \frac{1}{\lambda_j} w_{ij}^2$$

being k limited for only *some* of the terms in the vector w_i. When the summation takes place over *all* the terms, this statistic equals the Mahalanobis distance, defined as

$$\Delta_i^2 = (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) = (X_i - \mu)^T (E^T \Lambda E)^{-1} (X_i - \mu) =$$

$$(X_i - \mu)^T E^T \Lambda^{-1} E (X_i - \mu) = W_i^T \Lambda^{-1} W_i = \sum_{j=1}^{n} \frac{1}{\lambda_j} w_{ij}^2$$

Hawkins proposed to flag any instance i that renders values for $T_2(X_i)$ inside a so called *outlier region* (Davies *et. al*. (1993)). López *et. al*. (1994) applied a closely related procedure also based upon PCA, to handle daily rain datasets. They proposed to flag an instance when for any one $j \in [K_1, K_2]$, the projection $w_j(X_i) \notin [LB_j, UB_j]$ being LB_j and UB_j lower and upper bounds which define the non-outlier region for projection $w_j(X_i)$. Those limits are derived from the distribution of $w_j(X)$. The eigenvalues themselves are not required as well as any specific distribution for the data.

This paper follows almost the same idea, but since we are now working with categorical data some details need to be discussed.

It should be pointed out that, even in numerical datasets, usually the mean value and the Principal Components are real vector values, and so are the projections of the dataset on the PC, which are called here scores. That holds even if the data are integer or even binary numbers. For example, in a rain dataset, all values are integer and positive, but the scores are real, i.e., they belong to a different number category. When considering categorical binary answers a similar situation arises. However, even real, the possible values are limited due to a combinatorial problem. We are implicitly requiring that this finite number is a large number (in the experiments, 2^{69}-1) and the reason is presented below.

Once the data (without missing values!) are binarized and presented in table (or matrix) format, the PCA can be performed straightforwardly. Principal components can be derived as the eigenvectors of the covariance matrix (Lebart *et. al*. (1977)). Let's call E the square nxn matrix whose columns are the eigenvectors, which satisfy $E\Sigma E^T = \Lambda$, being Σ the sample's covariance matrix, and Λ a diagonal matrix which holds the (sorted in ascending order) eigenvalues. "n" is not the number of controlled questions but the sum of all the

options within them. It is assumed that the population is big enough to represent properly the true covariance with the sample´s covariance matrix.
Other subtle requirement should be stated: the procedure will not be of use if the number of options for the answers is low, because the distributions won´t look like those of continuous data. Notice that the real numbers w_{ij} are not arbitrary because they arise from a finite number of possible answers.
Anyway, since the matrix is range-defective due to the logical interrelationships between mutually exclusive answers, there will be some zero eigenvalues. This makes a slight difference with the situation for quantitative data (Hawkins (1974), López *et. al*. (1994)) where the Σ matrix is positive definite. The matrix of scores is defined here as:

$$W = E(X - \mu)$$

being X the binary data (one row for each record) and μ the arithmetic mean (among columns) of the matrix X. Matrix W has the same dimensions as matrix X, and its column-wise mean is zero. This is a linear transformation of the original data, and so each element w_{ij} depends directly upon *all* the elements x_{im}, where m ranges from 1 to n. This is an important fact, because the discrete distribution of the linear combination is completely different from the dichotomic one for x_{ij}, as it is shown in fig. 1.

Two facts should be remarked:

a) the sampled probability density function looks like the one of an ordinary continuous variable, even though it is based on a linear combination of dichotomic terms.

b) Its shape is different depending on the index of the score, following the same behavior noticed for scores derived from continuous variables, being more symmetrical as the index increases.

That's why we claim that the same procedures reported there could be used from now on. Once the sample distribution is created, confidence limits can be calculated. These values will define the outlier region (Davies *et. al*. (1993)) but without assuming any particular distribution shape. Why do we claim that this is the outlier region?. Fig. 2 shows the sampled probability distribution function for the given database of some of the scores and the arrows point to two values: those marked with an "o" correspond to the original answers for a particular record; those marked with an "x" are related to the same record, but now contaminated by modifying one of the answers. In this particular case, it was imposed that the house is equipped both with a color and a black and white TV set, while originally it has only black and white. Notice that the effect is important mostly in the "weakest" scores (i.e. those associated with the lower eigenvalues of matrix Σ) and that the ones associated with the "strongest" ones are only minimally modified. The proper limit between the "weakest" and the "strongest" is to be determined, and some guidance is given below.

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

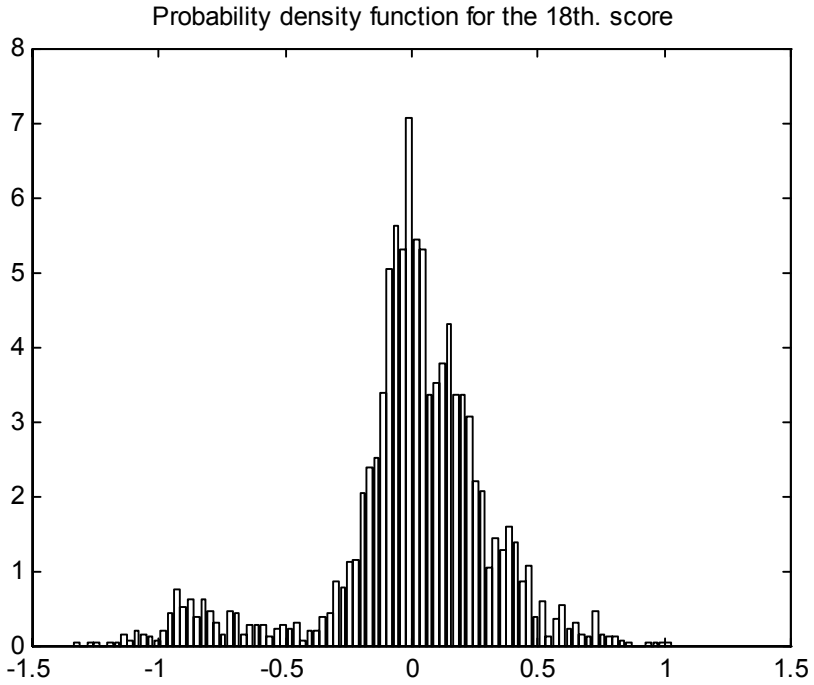

Figure 1 Example of the distribution of the 18th. score

A known fact is that each of those Principal Components associated with low eigenvalues have significant weights only with few variables in the original data. That implies that controlling the outlier region of one or two weak scores protects only some of the variables, which may be unadvisable. Those Principal Components associated with larger eigenvalues are typically insensitive regarding outliers, as it can be seen in fig. 2, so they should be avoided for our purposes. Summing up, neither too few or too many scores should be checked, and the appropriate number is a matter that is not uniquely solved in the literature.
Some rule of thumb suggest to neglect those terms whose associated eigenvalues are over a previously defined threshold. Hawkins (1974) suggests a more refined criteria, which chooses the limit in order to protect all the variables by proper inspection of the elements of matrix E. He did not formalize the criteria, so we will propose some objective one. The rows of matrix E are related with the original variables. Assume that K_1 is the index of the first non-zero eigenvalue, and K_2 is another integer index to be determined, ($K_1 < K_2$ because we assumed that the eigenvalues are sorted). In order to assure that at least once the variable X_j significantly affects some score, at least one of the eigenvectors with index ranging from K_1 to K_2 (i.e. columns in matrix E) should have a non negligible weight. The weights are the elements e_{jk} of row j of matrix E while $k \in [K_1,K_2]$, and they should be considered in absolute value for this purpose. The limit for negligible-not negligible is based upon a threshold value. If any abs(e_{jk}) is larger that such threshold for some $k \in [K_1,K_2]$, the variable is said to be *protected.* The threshold value cannot be chosen as a

fixed constant like 0.17, because (due to normalization) the e_{jk} are related with the size n of matrix E. So the proper threshold should take this fact into account. Since a mathematically valid eigenvector could be $(1,1,1,\ldots,1,1)/\sqrt{n}$, we choose as a threshold value a multiple of $1/\sqrt{n}$, now independent of size n itself. In the simulations the chosen multiple was 0.15, and the resulting range $[K_1,K_2]$ was [21,45].

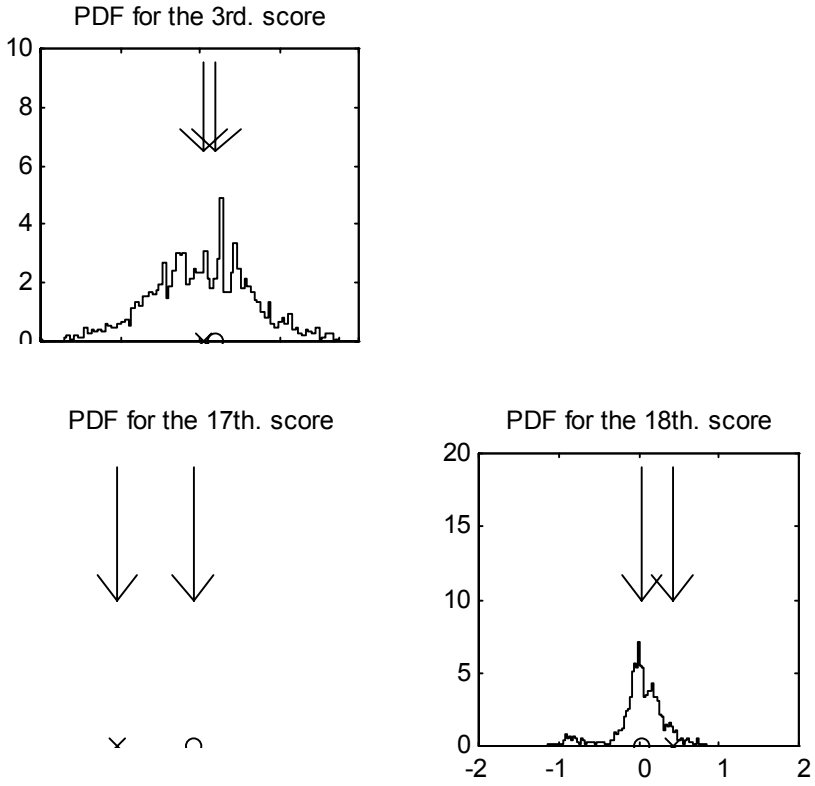

Figure 2 Example of the effect of a single outlier in a particular record

Once the limits K_1 and K_2 are defined, the sampled probability density function can be created for each score, and limits for the outlier region arise for each k, $k \in [K_1,K_2]$. The procedure is now straightforward, and it implies:

a) for each k-th score, look for records with values a_{ik} in the corresponding $k^{\underline{th}}$ outlier region, $k \in [K_1,K_2]$

b) once those records (rows) are retyped (and maybe modified or not), they can be included back in population X and new values for μ, E and outlier regions are calculated.

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

The procedure is iterative, and some stop criteria should be given. In each step, the dataset is classified in two categories. The first one holds the records which are likely to have an error, and the second one holds the ones accepted. When such a decision is made, it is certainly possible to reject good quality as poor or classify nondefective items as defectives; then, the associated error is called Type I. When a decision is made to accept poor quality as good (classify defective items as nondefectives), the error is called Type II (Minton (1969)).

We will denote as *number of contaminated records found* the successfully identified records which belong to the candidates set. That set is suggested by the algorithm, and its size (the *number of candidates analyzed*) depends strongly on the parameters, as will be pointed out later. Its quotient is an estimate of the complement of the Type I error:

$$CE_I = 1 - E_I = \frac{(\text{number of contaminated surveys found})}{(\text{number of candidates analized})}$$

and it measures the rate of success looking from the point of view of the reviewing process. It should be noted that in a production environment CE_I can be measured by the end user without knowing the total number of errors (i.e., without retyping twice the whole dataset).

Another important number is the probability associated with a purely random choice, i.e. without using any rule in selecting the candidate set. As long as the procedure goes forward, an *accepted* set is created. The Type II error associated is defined by the quotient

$$E_{II} = \frac{(\text{number of contaminated surveys not found})}{(\text{total number of "classified as acceptable" surveys})}$$

This quotient can be expressed in more rigorous terms, as:

$$E_{II} = \frac{\left(\begin{array}{c}\text{initial number of} \\ \text{contaminated surveys}\end{array} - \begin{array}{c}\text{accumulated number of} \\ \text{contaminated surveys found}\end{array}\right)}{(\text{total number of surveys} - \text{accumulated number of candidates analized})}$$

The E_{II} value also measures the probability to locate an error in the *acceptable* dataset with any blind (or random) procedure like the standard duplicate performance. Instead of presenting the evolution of the E_{II} index, a clearer measure of success is used, and it was defined as

$$\eta_2 = \frac{(\text{accumulated number of contaminated surveys found})}{(\text{initial number of contaminated surveys})}$$

This statistic monotonically increases from step to step, and it is bounded by 100 %, which implies that all the contaminated values have been located. It will also allow to compare directly the improvement over the standard duplicate performance method.

V.- Results

The calculations were carried out for 1, 2 and 3 wrong answers per record. Figure 3 shows the results for the first three steps in terms of the ratio η_2 for 100 replications of the experiment. The best results arise for a marginal value of 0.10%, where the methodology were able to locate 25% of the original errors (and in some cases, nearly 50%) *in a single step* of the procedure. The case of 0.01% looks very striking, because it represents two different bimodal, bell shaped distributions. This behavior is connected to some extent to the small number of records involved, as it will be shown later, and the picture is still incomplete because it does not show the effort involved in each step. Again, the value chosen for the marginal value is not crucial.

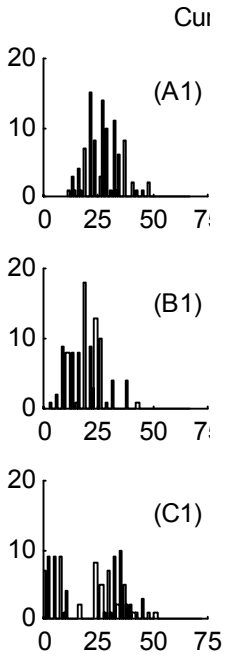

Figure 3 Distribution of the accumulated fraction of the total errors located up to the first three steps. Plots derived after 100 experiments, modifying 3% of the records with 2 errors each.

Figure 4 is itself a global summary of the behavior of the method. The x-axis is the fraction of the total dataset retyped, while the y-axis represent η_2, the fraction of the total errors found. We should emphasize that the continuous line indicates the locus of the theoretical evolution of the standard (blind) duplicate performance method, i.e.: by typing the x% of

the whole dataset, the same x% of the errors were removed (notice that the line goes through the (20%,20%) point). For *any* choice of the marginal value, the methodology proves to be better than the standard duplicate method, and since the behavior was very similar, only the case of 0.10% is shown. The dotted line is the best you can attain: retype first only those records that have errors. In the figure while retyping only 5% of the original data (x-axis) we can locate an amount of the original errors ranging from 25-60%, and when retyping 10%, 40-75% can be located.

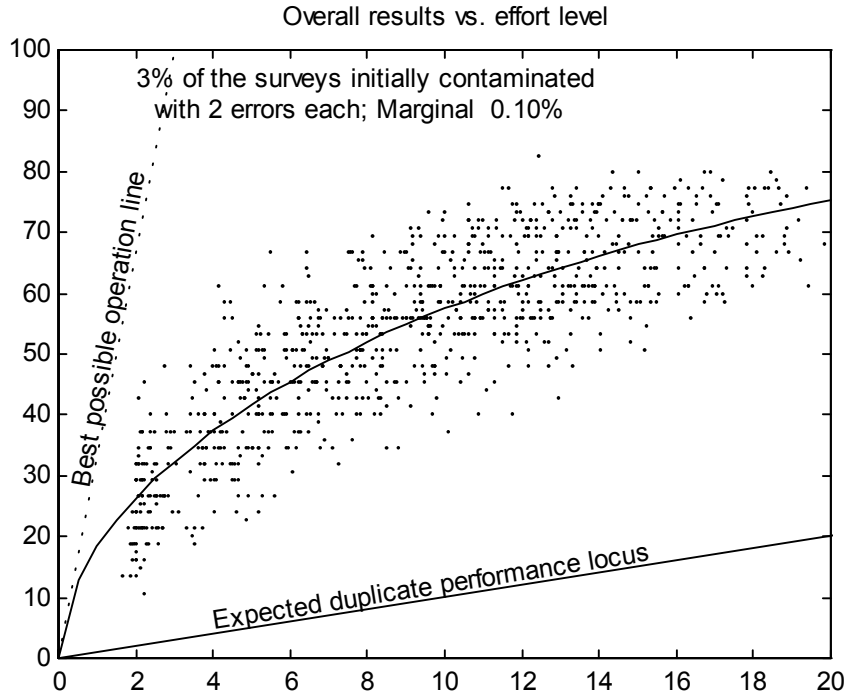

Figure 4 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3% of the records with 2 errors each.

Further work will show a degraded performance, because the "worst" errors have already been located. The limit goal of the procedure will be also the (100%,100%) point, because if all the data are checked we assume that all the errors will be removed. This procedure is intended to be applied for *partial* retyping.

The previous figures have presented the results with the records contaminated with 2 errors each. As expected, with 3 or more errors per record the results will be better while with 1 error they will be worse. For the 3 errors per record case, figure 5 show that after retyping 10% of the database, 50-85% of the errors have been corrected. For the case of a single error per record, fig. 6 shows that after retyping 10% of the database, only 25-50% of the errors have been corrected at most. The reader may notice that the cloud do not show any point over 16%; that's because we limit ourselves to 10 steps in the procedure. Even in this difficult case, the method is typically 4 times better than the blind retyping.

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

Some results regarding the initial number of erroneous records (not presented) show that the behavior of the best fit curve is almost independent of such value, but the dispersion is lower for larger initial number of erroneous records. This fact is a very desirable property, because poor quality dataset can be handled without loosing performance.

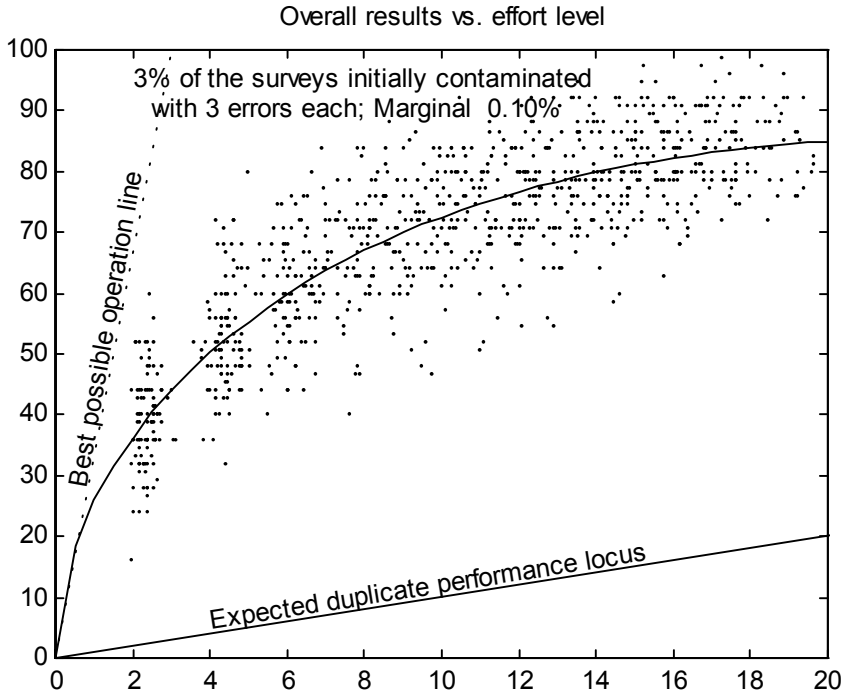

Figure 5 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3% of the records with 3 errors each.

VI.- Discussion

Comparing the use of logical edits against the present methodology, some clear differences arise. The methodology proposed in this work does not require any expert, since the "rules" (if any) are embedded in the population. Even the dichotomic answers (like marital status, sex, etc.) which are mutually exclusive, are handled gracefully, and need not to be analyzed separately. Moreover, when the population is updated using mostly the same questions, but with changes in some of them, all related rules should be revised. If a question is ambiguous, the rule can be wrong, while the proposed methodology probably will flag the answers as "uncorrelated" and will remove automatically from the feasible set.

Since the mere retyping is a completely "blind" methodology, it will locate equally well errors in "unusual" as well as "typical" individuals, keeping the variability of the dataset, while both the proposed methodology and the logical edits are oriented toward flagging

only those errors which make a particular individual "unusual". Then they will diminish the variability of the dataset.
However, it should be admitted that the application either of logical rules or mere retyping do not require a large population of individuals, while this methodology implicitly does. Other limitation of the reported methodology is that not all the questions can be controlled, either because of almost trivial answer or low correlation with other answers. Moreover, it cannot handle individuals with missing values.

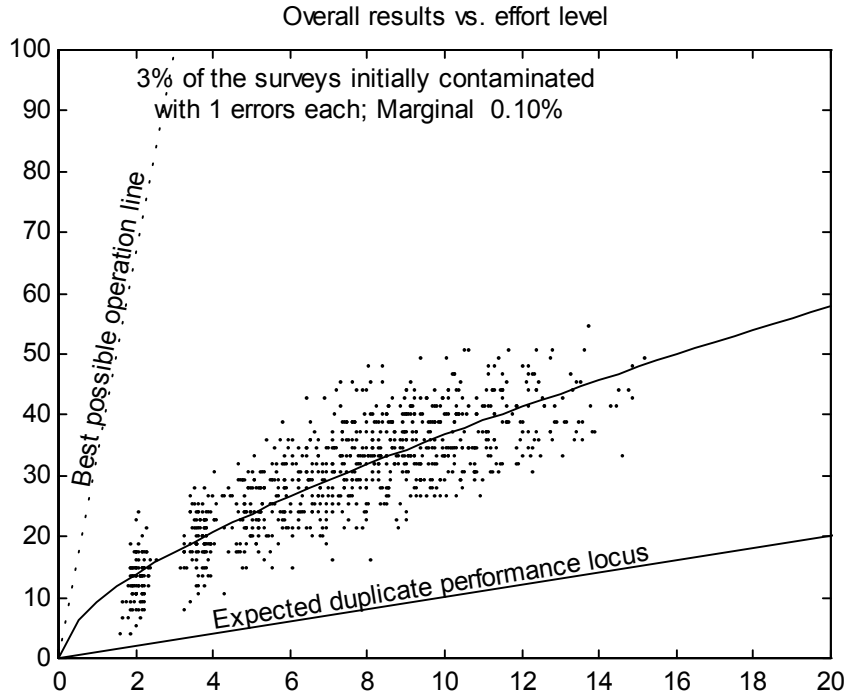

Figure 6 Evolution of the remaining errors against the retyping effort for the suggested depuration order and the blind retyping. Plots derived after 100 experiments, modifying 3% of the records with 1 errors each.

The numerical procedure is quite simple. It requires first to transform all answers to a "check box" format, so only ones or zeros will be admitted as answer. Then, the covariance matrix is constructed and its eigenvectors calculated, and a new table of projections (scores) of the original individuals over the eigenvectors is created. By analyzing the eigenvectors, a critical set of the scores is chosen in order to calculate for each an outlier region. Every individual with at least one of its scores lying on those region should be retyped. All the procedure can be automated. Once calculated the eigenvectors and the critical set, it can be applied even during the first typing process, allowing for near real time quality control.
The sensitivity to some parameters have been tested during the work, and for others not. Among the first, the margin (related with the number of individuals to be retyped in each step) was only weakly significant. The methodology for selecting the principal components to check seems feasible, but no further tests have been done.

For perfectly uncorrelated answers as a limiting case, the procedure is equivalent to looking for answers with low probability, which is also a feasible procedure.

IX.- Conclusions

The problem of quality control of categorical data is treated with a methodology derived from statistical procedures for quantitative data. Two other alternatives can be analyzed, the duplicate performance method and the use of logical edits. The first is very simple and popular, and requires typing again the same dataset. Its ability in locating errors for a given typing effort is known to be low. The use of logical edits strongly rely on the existence of an expert, which should prepare a set of rules, expressed in terms of logical relationships between the answers. When any of them is not met, the record is flagged as unusual, and retyping is performed. Here an alternative is proposed in order to reorder carefully what should be retyped.
Some limitations of this procedure are: a somewhat large (yet undefined) population is required as well as a minimum number of options for the answers, it cannot handle missing data, and depending on the inherent characteristics of the population, some answers or options for answers are not checked. The users for a methodology like this are still those which are either collecting or using the raw data; we are not giving any tool to check derived statistics (like averages in a region, etc.).

X.- Acknowledgments

The author is indebted to the authorities of the National Bureau of Statistics of Uruguay, which allowed access to the raw data used in this work, and to Prof. Friedrich Quiel, Royal Institute of Technology, Stockholm, who provided helpful discussion and suggestions while preparing the manuscript.

XI.-References

Coale, A.J. and Stephan, F.F (1962) The case of the Indians and the teen-age widows, *Journal of the American Statistical Association*, Vol. 57, No. 298, 338-347

Davies, L. and Gather, U. (1993) The identification of multiple outliers. *Journal of the American Statistical Association,* Vol. 88, No. 423, 782-801

Fellegi, P. and Holt, D. (1976) A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, Vol. 71, No. 353, 17-35

Minton, G. (1969) Inspection and correction error in data processing. *Journal of the American Statistical Association*, Vol. 64, No. 328, 1256-1275

Hawkins, D. M. (1974) The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, Vol. 69, No. 346, 340 -344

Little, R. J. A. and Smith, P. J. (1987) Editing and Imputation for quantitative survey data. *Journal of the American Statistical Association*, Vol. 82, No. 397, 58-68

López, C.; González, E. and Goyret, J. (1994) Análisis por componentes principales de datos pluviométricos: a) aplicación a la detección de datos anómalos (in spanish) Estadística, Vol. 6, No. 146-147, 55-83. (English version available at http://www.fing.edu.uy/cecal/reports/rep92_1/papera.html)

Paradice, D.B. and Fuerst, W.L. (1991) An MIS data quality methodology based on optimal error detection. *Journal of Information Systems*, Vol. 5, No. 1, 48-66

IMPROVEMENTS OVER THE DUPLICATE PERFORMANCE METHOD

Strayhorn, J. M. (1990) Estimating the errors remaining in a data set: techniques for quality control, *The American Statistician*, Vol. 44, No. 1, 14-18