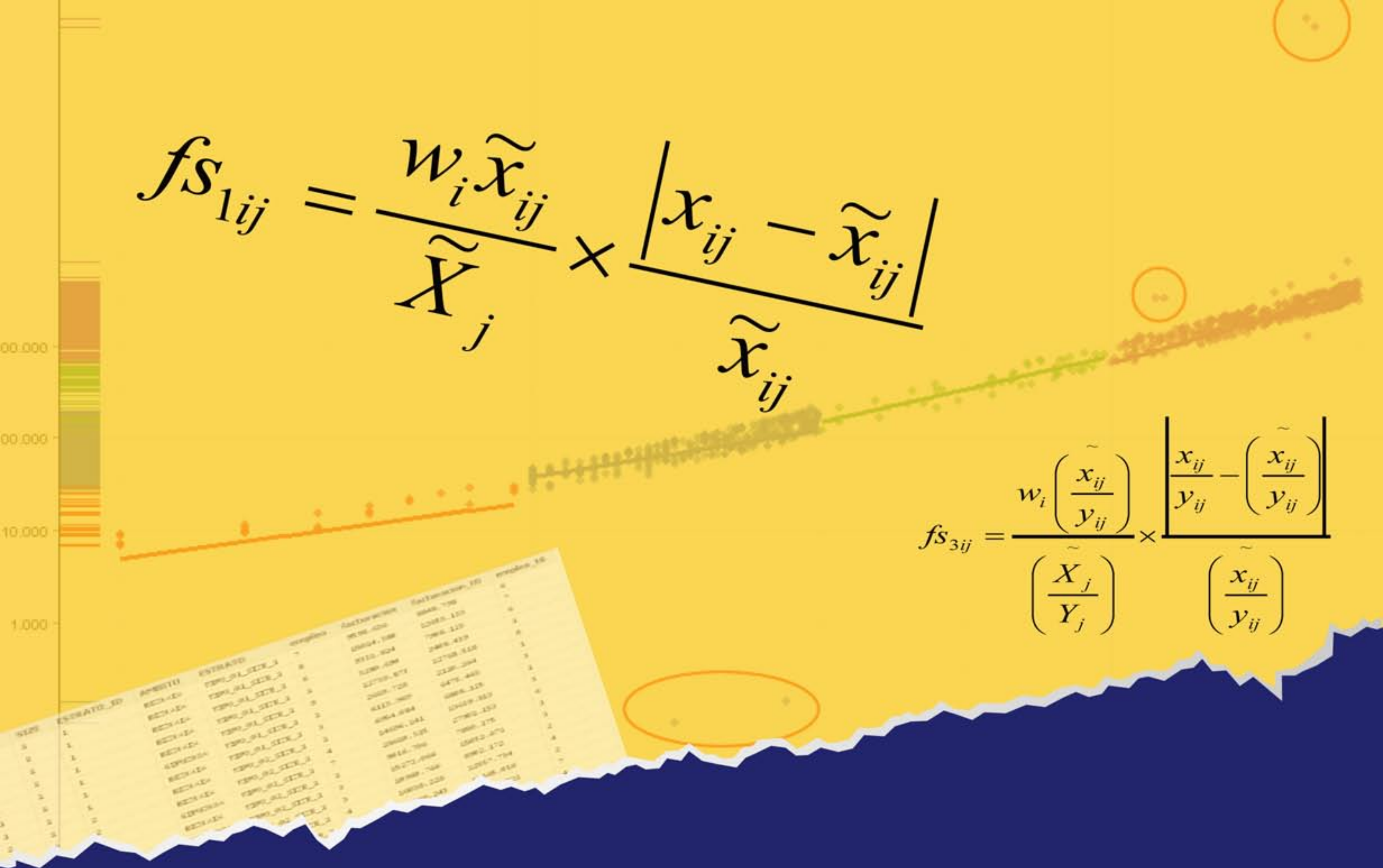

$fs_{1ij} = \frac{w_i \widetilde{x}_{ij}}{\widetilde{X}_j} \times \frac{\left| x_{ij} - \widetilde{x}_{ij} \right|}{\widetilde{x}_{ij}}$
$fs_{3ij} = \frac{w_i \widetilde{\left(\frac{x_{ij}}{y_{ij}}\right)}}{\widetilde{\left(\frac{X_j}{Y_j}\right)}} \times \frac{\left| \frac{x_{ij}}{y_{ij}} - \widetilde{\left(\frac{x_{ij}}{y_{ij}}\right)} \right|}{\widetilde{\left(\frac{x_{ij}}{y_{ij}}\right)}}$
00.000
00.000
10.000
1.000

Selective editing

2014

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

Organismo Autónomo del
EUSKO JAURLARITZA
GOBIERNO VASCO

Edition:
EUSTAT
Euskal Estatistika Erakundea
Basque Statistics Institute

Publication:
EUSTAT
Euskal Estatistila Erakundea
Basque Statistics Institute
Donostia-San Sebastián 1
01010 Vitoria-Gasteiz

First Edition
I/2015

Printing and Bounding:
Printing and Reprography Service. Basque Government.

ISBN: 978-84-7749-481-2

Legal Deposit: VI-793/2014

SELECTIVE EDITING

Imanol Montoya Arroniz

imanolmontoya@gmail.com

Traducido por: Ofilingua SL

Eustat

EUSKAL ESTATISTIKA ERAKUNDEA
BASQUE STATISTICS INSTITUTE

Donostia-San Sebastián, 1
01010   VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.eus
www.eustat.eus

Introduction

Data editing is one of the most time-consuming and expensive parts of the statistical processing procedure. Therefore, it is essential for statistical offices to have efficient methods for data editing.

The objective is to study and to apply different selective editing techniques. Selective editing helps selecting those errors whose correction has a significant influence on the published results, thereby reducing costs and delivery times.

This document is divided into several chapters. First the methodology is developed. Then a simulation study is done for studying previously developed methodology. Next chapter contains an explanation about the SAS macros that have been prepared for selective database editing. Later it presents a real example, specifically the new Basque Statistics Institute Services Statistic operation, where the methodology proposed for selective database editing has been applied. Finally, some conclusions are shown about the efficacy and usefulness of this methodology.

Vitoria-Gasteiz, December 2014

JOSU IRADI ARRIETA

General Director

Contents

. . . / . . .

Chapter

1

1. Introduction

The content contained in this Technical Notebook is the result of work undertaken during the training and research scholarship on statistical and mathematical methodologies, for the topic of selective database editing, granted in 2012 by the Basque Statistics Institute / Euskal Estatistika Erakundea.

This document is divided into the following chapters:

The second chapter contains an introduction and mentions the objectives that the publication of this technical log book have established.

Chapter three develops the micro-selection methodology, defending the "score" function, showing different types of functions, the strategies used to construct them, how they can be combined in a global score function and finally a threshold is established.

Chapter four develops the macro-selection methodology, differentiating the aggregated method and the distribution method.

In chapter five there is a simulation study for studying previously developed methodology. It explains how the simulated database was created, how the score functions were calculated and shows the results obtained.

Chapter six contains a brief explanation about the SAS macros that have been prepared for selective database editing.

Chapter seven presents a real example, specifically the new Basque Statistics Institute Services Statistic operation, where the methodology proposed for selective database editing has been applied.

Finally, some conclusions are shown about the efficacy and usefulness of this methodology.

I would like to thank everyone in the Methodology, Innovation and R&D Area and, in general, the helpfulness of all Eustat personnel.

KEYWORDS: Selective editing, local score function, global score, micro-selection, macro-selection.

Chapter

2

2. Introduction to selective editing

One of the parts that takes the most time and is the most expensive in the process of improving data quality is manual or interactive data editing. In the past, records were often edited manually, with the consequent cost in personnel and time. In recent decades, the effect of this manual data editing has been researched and it has been shown that the number of records to be edited manually can be greatly reduced, since for many of the records, manual editing has an insignificant influence on the estimators of the main parameters of interest.

Studies such as those of (Granquist, 1995), (Granquist and Kovar, 1997) and (Hoogland, 2000), among others, showed that it is generally not necessary to correct all the errors in order to obtain a viable figure for the parameter of interest. It is sufficient to correct the errors with the greatest influence.

The following graph shows the decrease in the influence that successively correcting the less important errors has on estimating the parameter of interest, in this case the number of employees (Hoogland, 2000).

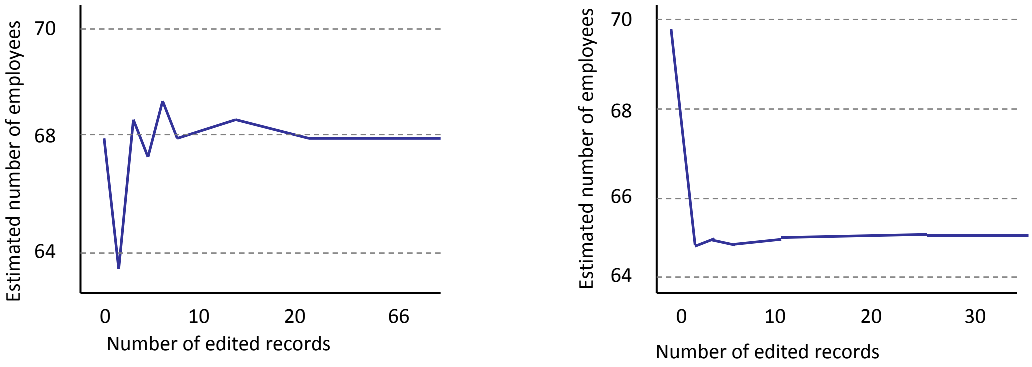


Figure 1. Estimated number of employees based on the number of edited records, which have been ordered depending on their influence on the final estimator. The graph on the left is based on construction company records, while the graph on the right is based on civil engineering companies (Hoogland, 2000).

Selective editing is the strategy by which only the errors whose correction significantly influences the results to be published are edited, thus reducing costs and delivery times.

There are different methods that enable selection of the records to be edited in a database. When they are applied in the first stages of data capture, even if the data capture is not completed, they are known as micro-selection methods. In general, these methods are applied individually to each record and are based on data from previous periods or estimations of homogenous subgroups. On the other hand, macro-selection methods are designed to be used when almost all of the data is available. These methods use the information from all the data available to detect influential values.

This technical notebook describes the different methods that enable more influential records to be selected for editing. It is mainly based on chapter 6 "*Selective Editing*" of the book "*Handbook of Statistical Data Editing and Imputation"* (de Wall, Pannekoek and Scholtus, 2011), in the technical notebooks published by the Holland National Statistics Institute (Hoogland, van der Loo, Pannekoek and Scholtus, 2011) and (de Wall, 2008) and in the European project recommendations (EUREDIT Project, 2004) and (EDIMBUS, 2007).

Chapter

3

3. Micro-selection

The main idea of micro-selection is to be able to select the records to be edited before the data capture has been completed.

This chapter explains the score function, the most frequent ways of constructing it, other strategies for constructing the function, how to combine it when calculating a global value and how to determine the threshold value that enables selection of the records to be edited.

The score function

The score function is the main instrument used in the micro-selection process when editing records. This function assigns a score to each record for each variable analysed. This scoring provides an indication of the expected effect on the parameter to be estimated if it is edited. Records with a high score are those selected first for editing.

It is known as the score function local to the function that measures the editing influence of a specific variable in a record. This local score function often has two components: risk and influence. Risk covers the size and the probability of a potential error, while influence covers the impact of the record on the estimation of the study parameter. Local scores are defined as the product of these two components,

$$s_{ij} = F_{ij} \times R_{ij} = influence_{ij} \times risk_{ij}$$

where s_{ij} is the score function for the record *i* in the variable *j*. The component of influence is generally measured as the relative contribution of the anticipated or expected value over the total estimator. The risk component is generally measured by comparing the raw value with respect to an anticipated or expected value. Small deviations between both values imply that there is no reason to assume that there is an error, while large deviations are an indication that there may be an error.

The global score is a function that combines the local scores to create a measurement for the entire record.

$$S_i = f(s_{i1},\dots,s_{iJ})$$

The micro-selection methods are applied with no need for the data capture to have been finished. Once the record is available, a global score is obtained and compared with a previously determined threshold value. If the score surpasses this threshold, the record will be designated as not plausible. These are the records that enter the branch of records that must be edited.

Formally, this selection is based on the plausibility indicator defined as:

$$Plausibili\,tyindicato\,r_i = \begin{cases} 1 \quad (plausible) & if \;\; S_i \le C, \\ 0 \; (not \;\; plausible) & inothercas\,es \end{cases}$$

where C is the threshold value.

The micro-selection strategy can be summarised in the following three steps:

- Calculate the local score functions for the main variables of interest, using anticipated or expected values based on data from previous periods or in homogeneous subgroups as a reference.
- Determine a function that combines these local scores in a global score.
- Determine the threshold values for the global values that select the records to be edited.

Types of score functions

o **Basic score functions for totals**

A score function must quantify the effect of editing the record in the estimator of interest. Therefore x_{ij} is the value of the variable x_j in the record i. If the estimator of interest is the total, this can be defined as:

$$\hat{X}_j = \sum_{i \in D} w_i \hat{x}_{ij}$$

where D is the set of data and i the records. The weights w_i are corrected by the unequal probabilities of inclusion and/or no response. The $\hat{x}_{ij}$ are the data once edited. This implies that certain records of the raw data, x_{ij}, have passed through the editing process and have been corrected. For the majority of the records, x_{ij} is considered correct and equal to $\hat{x}_{ij}$. Therefore, the effect (additive) on the total to be edited in a single record can be defined as the difference between the total estimated with or without the edited record i. The estimated total without editing the record i is $\hat{X}_j - w_i(\hat{x}_{ij} - x_{ij}) = \hat{X}_j^{(-i)}$, and therefore the difference can be expressed as:

$$d_i(\hat{X}_j) = \hat{X}_j^{(-i)} - \hat{X}_j = w_i(\hat{x}_{ij} - x_{ij})$$

The difference $d_i(\hat{X}_j)$ depends on an unknown corrected value $\hat{x}_{ij}$ and therefore cannot be calculated. A score is based on an approximation to this unknown value $\hat{x}_{ij}$, $\tilde{x}_{ij}$, as an expected value. Normally these expected values are:

- Edited values of the same record from previous periods in the same survey, multiplied by an estimation of the evolution between the two time periods.

- The value of a similar variable from the same record but obtained from a different source of data.

- The average or median of the variable of interest of a homogenous subgroup of similar records from a previous period.

The difference $d_i(\hat{X}_j)$ also depends on weights w_i. This is because these weights are corrected by the unequal probability of inclusion, but also the non-response, which is unknown up until the final capture of the data. The approximation used in these cases uses the "weights of the design", v_i which are only corrected by the unequal probability of inclusion.

Using these approximations, the effect of editing the record i can be quantified by the score function:

$$s_{ij} = v_i \left| x_{ij} - \tilde{x}_{ij} \right| = v_i \tilde{x}_{ij} \times \frac{\left| x_{ij} - \tilde{x}_{ij} \right|}{\tilde{x}_{ij}} = F_{ij} \times R_{ij} = influence_{ij} \times risk_{ij}$$

This score function, therefore, can be understood as the product between an influence factor and a risk factor. The risk factor is a relative measure of the difference between the raw and the expected value $R_{ij} = \left| x_{ij} - \tilde{x}_{ij} \right| / \tilde{x}_{ij}$. Large differences indicate that the value may be erroneous. The influence factor, $v_i \tilde{x}_{ij}$, is the record's contribution to the estimated total.

Multiplying the risk by the influence provides a measure of the effect that editing the record would have on the estimated total. Large values would indicate that the record may contain an influential error and it may be worth checking it. Small values on the other hand, indicate that the records may not contain influential errors and it is therefore not entirely necessary to meticulously edit them.

For non-negative variables, such as the majority of the economic surveys, the risk factor can also be based on the ratio between the raw value and the expected value, instead of the absolute difference between these values.

$$(x_{ij} - \tilde{x}_{ij}) / \tilde{x}_{ij} = \frac{x_{ij}}{\tilde{x}_{ij}} - 1$$

This way the risk is expressed as a ratio between the raw value and the expressed value, and -1 is added to ensure that the risk is zero when the two values are the same. In any case, this expression still does not cover that the large and small differences in the ratio indicate deviations with the expected value. In order to correct for this factor, the following risk function is based on the ratio:

$$R_{ij} = \max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1$$

This function ensures that multiplicative increments of equal value, whether upwards or downwards, provide the same score. Multiplying this risk by the factor of influence provides an alternative score function to the additive:

$$s_{ij} = v_i \tilde{x}_{ij} \times (\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1)$$

Finally, a scaled version of the score function is normally used, replacing the factor of influence F_{ij} by the relative influence $F_{ij} / \sum_i F_{ij}$, where

$$\sum_i F_{ij} = \sum_i v_i \tilde{x}_{ij} = \tilde{X}_j$$

Therefore the scaled value obtained is the original value divided by an estimation of the total based on expected values. Scaling the value enables this value to be independent of the size and unit of the variable studied. This is useful when several score functions will be combined to generate a global value.

In summary, two local scaled score functions for totals, one additive and another multiplicative, are, respectively:

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{\left|x_{ij} - \tilde{x}_{ij}\right|}{\tilde{x}_{ij}} \quad \text{and} \quad s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times (\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1)$$

- **Models for expected values**

In general, an expected value is a function of auxiliary variables and coefficients:

$$\tilde{x}_{ij} = f(\hat{\mu}_1, \ldots \hat{\mu}_K, z_{i1}, \ldots z_{iK})$$

These auxiliary variables are often obtained from the current database or, as is more often the case, from previous records or surveys that have already been edited.

An expected value that is based on auxiliary variables may be the estimated value of the average or median of the variable of interest in a specific subgroup. For example, in economic records a subgroup can define the type of industry and its size.

When there is an auxiliary variable that is highly correlated with the variable of interest, the variable of interest is often divided between the auxiliary variable and this ratio compared with the expected value for this ratio. For example, supposing that the number of employees is the auxiliary variable and turnover is the variable of interest, the ratio would be turnover per employee. Turnover may vary greatly between different establishments, even in the same type of industry, while the ratio by employee often varies much less.

When the ratio between two variables is used in the score functions, x_{ij} and $\tilde{x}_{ij}$ are replaced in the risk factor by the raw value of the ratio and by the expected value, $\frac{x_{ij}}{y_{ij}}$ and $\left(\frac{\tilde{x}_{ij}}{y_{ij}}\right)$. Once again, the expected value for the ratio may be the average or the median in a previous period, as it may belong to a homogenous subgroup.

In general, the models used in practice for the expected values are often not very sophisticated. Therefore their predictions are not entirely accurate. Even so, these predictions are often useful since the objective of micro-selection is to correctly select the records sent for editing, and not to obtain an accurate prediction of the records (Lawrence and McKenzie, 2000).

o **Score functions with longitudinal data**

In surveys or records that are captured every certain period of time, the values from previous periods are usually used as auxiliary values. The following formula shows the risk component that uses values from previous periods as expected values, based on the ratio, and was proposed by (Hidiroglou and Berthelot, 1986):

$$R_{ij} = \max\left(\frac{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}}\right)}, \frac{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}}\right)}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}\right) - 1$$

Where $x_{ij,t}$ is the value of the variable of interest x_j for the record *i* in the current period *t* and $\hat{x}_{ij,t-1}$ the value corresponding to the same unit in the previous period once edited. As the expected value for the change the median of the changes of all the records is proposed, although for this, all data must be captured. An alternative for this case is to use the average of the changes in previous periods, for example between *t* -2 and *t* -1, but only when it is assumed that this change will be similar to *t* -1 y *t.*

When calculating the influence, the information from the previous period can also be taken into account:

$$F_{ij,t} = \left[\max\left(x_{ij,t}, \hat{x}_{ij,t-1}\right)\right]^{c}$$

With $0 \le c \le 1$. In this formula *c* serves to control the importance of the parameter of influence. For example, one study (Latouche and Berthelot, 1992) estimated a *c* value of 0.5 as reasonable in its data. On the other hand, the maximum function guarantees that an error in $x_{ij,t}$, even when it is an infra-estimation, has an influence as at least the same as its edited version from the previous period $\hat{x}_{ij,t-1}$.

Other strategies for constructing the score function

The different ways that have been looked at for calculating the score functions are based on the deviation between the raw value and the expected value. These types of functions are the most used in national statistics institutes. Even so, other types of strategies have been proposed, although they still do not have acceptance as the "traditional" strategy.

- **Parametric models for data with errors**

An alternative is to specify a parametric model that takes possible errors in the data into account. This model assumes that the data with errors and the data without errors come from different distributions. This strategy has been proposed by (Ghosh-Dastidar and Schafer, 2006), (Di Zio, Guarnera and Luzi, 2008) and (Bellisai et al., 2009). These authors assume that the correct data comes from normal distribution with average μ and variance σ^2 and that the incorrect data comes from a normal distribution with the same average but with a variance inflated by a factor of *c* > 1. These assumptions provide a contaminated normal model, which has density as the function:

$$f_x = \pi N(\mu, \sigma^2) + (1-\pi) N(\mu, c\sigma^2)$$

where the probability π is the proportion of data with no errors. Using this model, the conditional probability can be estimated, $\hat{\pi}_i$, from which a record is free from error, given its observed value: $\Pr(x_i = x_i^* \mid x_i)$. This probability, conditional in the data observed, is known as the *a posteriori* probability. Values lower than an appropriate cut-off point are considered atypical values and are sent for editing.

The previous model can be extended enabling the presence of missed values, logarithmically transforming the variable to make it more symmetric and so that the assumption of normality is more realistic. It can also be extended using co-variables that enable the average value μ to vary.

- **Strategy associated with the validation "edits"**

Another strategy proposed by (Hedlin, 2003) is to evaluate at which point in a record the validation "edits" have failed. In other words, how many "edits" do not meet the criteria and by how much they do not meet the criteria. The idea of this strategy is that the influential errors violate several of the "edits" or that the same failure would be a considerable amount. At the end of the study, (Hedlin, 2003) showed that the strategy of using the score function provided better results than the strategy associated to the validation "edits".

- **Prediction model strategy**

This strategy proposes constructing a model that relates the presence and size of influential errors on the analysis variable with other predictive variables in the same record. This strategy requires training data that contain original raw data as well as edited data. Using this training data, the influence of editing each record on the total estimation can be calculated.

Once the influence of each record is known, the "error probability" can be predicted π classifying each record in a variable of, for example, as (Van Lancen, 2002) did, 6 categories: the first category is that in which the records did not contain errors, the others contain 20% of the records with errors, and the last type has more influential errors. The probabilities assigned to each category π: 0; 0.2; 0.4; 0.6; 0.8; and 1. A logistical regression model including predictive variables and using the training data can be used to predict this probability. Once the model's parameters have been calculated, they are used in the actual data and the probability of containing an influential error is estimated for each record.

This strategy has not shown to be superior to the strategy based on the score function calculations (de Wall, Pannekoek and Scholtus, 2011).

Global score function

In order to be able to select an entire record and thus edit it, a value is required that combines the information from the different score functions. This value is known as the global score. This score must reflect the importance of entirely editing the record. In order to be able to combine the different local scores, it is important that the local score functions are measured on comparable scales. To do this, these local values are usually scaled dividing them by their total or their expected total.

The most common options for combining the local score functions, previously scaled, are:

- The sum of the local score functions (Latouche and Berthelot, 1992):

$$S_i = \sum_{j=1}^{J} s_{ij}$$

- The maximum of the local score functions (Lawrence and McKenzie, 2000):

$$S_i = \max\left(s_{ij}\right)$$

- A proposal that addresses both of the above (Hedlin, 2008):

$$S_i^{(\alpha)} = \left(\sum_{j=1}^{J} s_{ij}\right)^{1/\alpha}$$

Where $S_i^{(\alpha)}$ is the global value based on the parameter α , s_{ij} is the *j*-th value of the local score function, *J* is the number of local values.

The disadvantage of the first way of combining the local functions is that records with many but moderate deviations have priority over records with few but significant deviations. In the second case, the advantage with respect to the previous case is that the records in which there are important deviations are prioritised. Even so, this option is not able to discriminate records with a single large local score with respect to others with many large local scores. In the latter option, it is α that determines the influence of the local values in the global value. A $\alpha = 1$value implies the first option, the sum of the score functions. The value of α next to ∞gives the second option as a result, the maximum of the local score functions.

Another option is to select a specific weight for each variable in the global value depending on the importance given to it. This weight may be assigned by experts and may vary, for example, between 0, 1, 10 and 100.

Set a threshold

The ultimate objective of a global score function is to select the records that must subsequently be edited. If the editing can wait until all the data has been captured, editing can be stopped when the parameters of interest do not change substantially. This method of editing usually takes a lot of time if the quantity of data and variables is large. In order to be able to begin editing data in the capture phase, a decision must be made based on the score of each record, with no need to compare it with other records. To do this, a threshold is set by which if the global score function of a record surpasses this value, the record must be sent to be edited.

A simulation study is normally undertaken to determine the threshold in which the effect of different threshold values is researched. In other words, the effect of editing more or fewer records on the parameters of interest. This simulation study uses original raw data and this same data edited manually.

For the simulation study, the records are ordered based on their value in the global score function. Next, the first *p*% of records are selected for manual editing. This is done by replacing these records with the data from the edited database. The subset of the *p*% of edited records is known as E_p. These steps are repeated for a range of *p* values. Next, the parameter of interest is estimated based on this new database with *p*% of records edited and is compared with the estimated parameter with the completely edited database. The absolute value of the relative difference between these estimators is known as the absolute pseudo-bias:

$$AB_j(p) = \frac{1}{\hat{X}_j} \left| \sum_{i \notin E_p} w_i \left(x_{ij} - \hat{x}_{ij} \right) \right|$$

This value is known as the absolute pseudo-bias because if only errors are edited, there would be a real bias due to not editing all the records. But as it is not certain that the edited data are the correct data, this bias is an approximation to the truth, and therefore a pseudo-bias.

The pseudo-bias of editing *p*% of records can also be interpreted as an estimator of the gain in the precision of the estimator if 1 – *p*% of the remaining records are edited. Therefore, if the pseudo-bias is calculated for a range of *p* values, an idea about the improvement to accuracy can be obtained based on *p*. At a certain point of *p* it may be decided that it is not worth the effort to continue editing records since there is not much improvement in the accuracy of the parameter of interest.

Finally, the simulation study is a way of being able to check the effectiveness of the selective editing process. It can be checked whether these records with higher values in the global score function in reality had influential errors or not, and conversely, whether records with small global values contained non-influential errors.

Chapter

4

4. Macro-selection

The main idea behind macro-selection is to select the records to edit once the data capture has finished or has almost finished.

Aggregate method

Once all the data has been captured, the main aggregates are calculated for the variables of interest. If these aggregates differ greatly from what is expected, for example based on data from previous periods, it will need to be checked.

There are many reasons the aggregates may differ from what is expected:

- There may be influential errors in the data.
- There may have been problems with the weights used in the design.
- There may have been unexpected variations that are real.

One example of a score function at a macro level is:

$$S_j = X_j - \tilde{X}_j$$

where X_j is the aggregated estimator for the variable x_j based on the unedited data and $\tilde{X}_j$ an expected value for this aggregated estimator.

The relative difference between the aggregate and its expected data can also be calculated:

$$S_j = \frac{X_j - \tilde{X}_j}{\tilde{X}_j}$$

In some cases, it is often more efficient to use ratios between aggregates than aggregates separately:

$$S_j = \frac{X_j}{X_k} - \widetilde{\left(\frac{X_j}{X_k}\right)}$$

Or in relative terms:

$$S_j = \frac{\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k}\right)}{\left(\frac{\tilde{X}_j}{X_k}\right)}$$

An example of this case may be the ratio between the total turnover and total costs for a type of industry, or the total salary among the total number of employees for a type of industry.

One way that the variance of the aggregate can be controlled is to divide the difference in the aggregates or in the ratios by its relative standard deviation:

$$S_j = \frac{X_j - \tilde{X}_j}{d.e.\left(X_j - \tilde{X}_j\right)} \quad \text{and} \quad S_j = \frac{\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k}\right)}{d.e.\left(\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{X_k}\right)\right)}$$

This is the same as what occurs in micro-selection; the differences between the aggregates or the ratios can be expressed in an additive or multiplicative manner.

Once a suspicious aggregate has been detected for the data, the aggregates will need to be investigated at a lower level, for example the aggregate for a certain type of industry. At the end of this process, the records must be checked for possible influential errors.

The main differences between micro-selection and the application of these techniques once the data is captured are:

1. The actual data can be used as a source of information for the expected values. For example, the median of a homogenous group of the actual data can be used instead of the data from the previous period. As the current database has not yet been edited it is important to use the medians because they are robust estimations in the presence of atypical values.

2. It is not necessary to use an approximation of the weights with the weights of the design v_i since the final weights will be available w_i when calculating the estimators.

3. It is not necessary to calculate a threshold beforehand, since the scores of the score functions provide an order by which the records will be edited, and how much the final estimator changes can be controlled. The edit can be stopped when it is considered that the improvement in the estimator's precision is insignificant.

Distribution method

The purpose of the distribution method is to identify the values that do not seem to fit well with the distribution observed. This is done using graphic tools and statistical measurements. The atypical values, the most suspicious records, are checked. If an atypical value shows that it is an incorrect and influential value it is corrected.

The method is applied to quantitative variables and a certain normality or asymmetry in the data is often assumed. If it is not the case, some kind of transformation is usually applied to the data.

One robust measurement commonly used to detect atypical values is based on the median $x_{ij} - median(x_{ij})$. This is similar to using a score function in which the expected value is the median. In order to be able to compare deviations with respect to the median in different groups, this difference is often standardised by the median of its absolute values, for the records of each group.

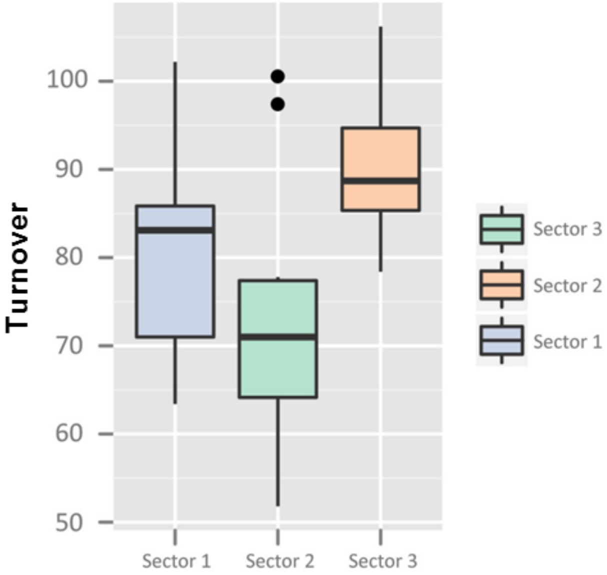
$$o_{ij,c} = \left| x_{ij} - med(x_{ij,c}) \right| / (1{,}4826 \times DAM_c(x_{ij,c}))$$

where $med(x_{ij,c})$ is the median for the records in group *c* and $DAM_c(x_{ij,c})$is the absolute deviation of the median for these records given by

$$DAM_c(x_{ij,c}) = med(\left| x_{ij} - med(x_{ij,c}) \right|)$$

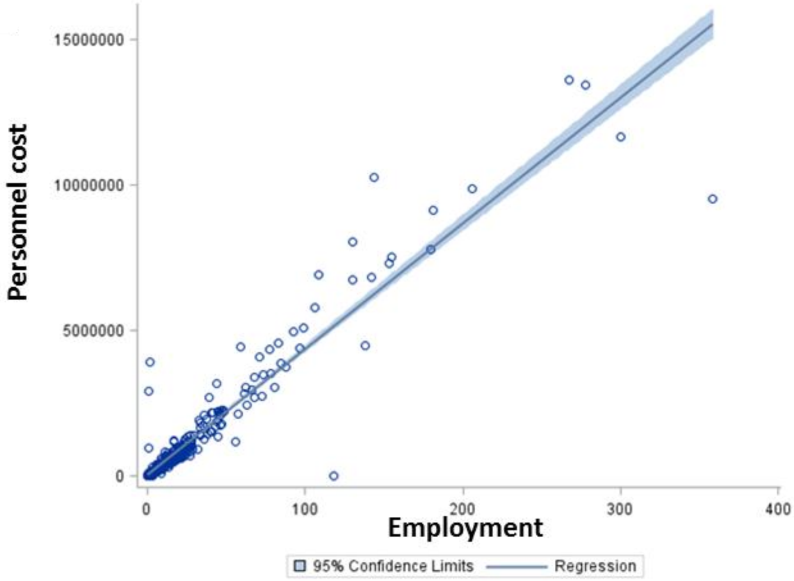
For a normal distribution $1{,}4826 \times DAM$ this is an estimator consistent with the standard deviation. Other robust measures used for detecting the outliers are the Winsor averages and the truncated averages. Non-robust dispersion measures such as variance or standard deviation can also be used.

Graphics are commonly used as the boxplots for representing the deviations with respect to the median. On the one hand these graphs show a box that contains 50% of the records, lines that normally limit 1.5 times the interquartile range with respect to the first and third quartile. Values beyond these lines are considered atypical values.



This graph shows, for three sectors, the distribution of turnover in the sector's establishments in thousands of euros. As can be seen, in *Sector 2* there are two atypical values.

Another graph technique often used for detecting atypical variables is the *scatterplot*. As opposed to other box plots, the scatterplot is often used when comparing the distribution of two continuous variables.



This graph shows the relationship between employment and the personnel cost in the "Machinery & Equipment" sector.

Chapter

5

5. Simulation

The main objective of this chapter is to evaluate how different score functions behave by generating several types of errors with a simulated database.

Simulated database

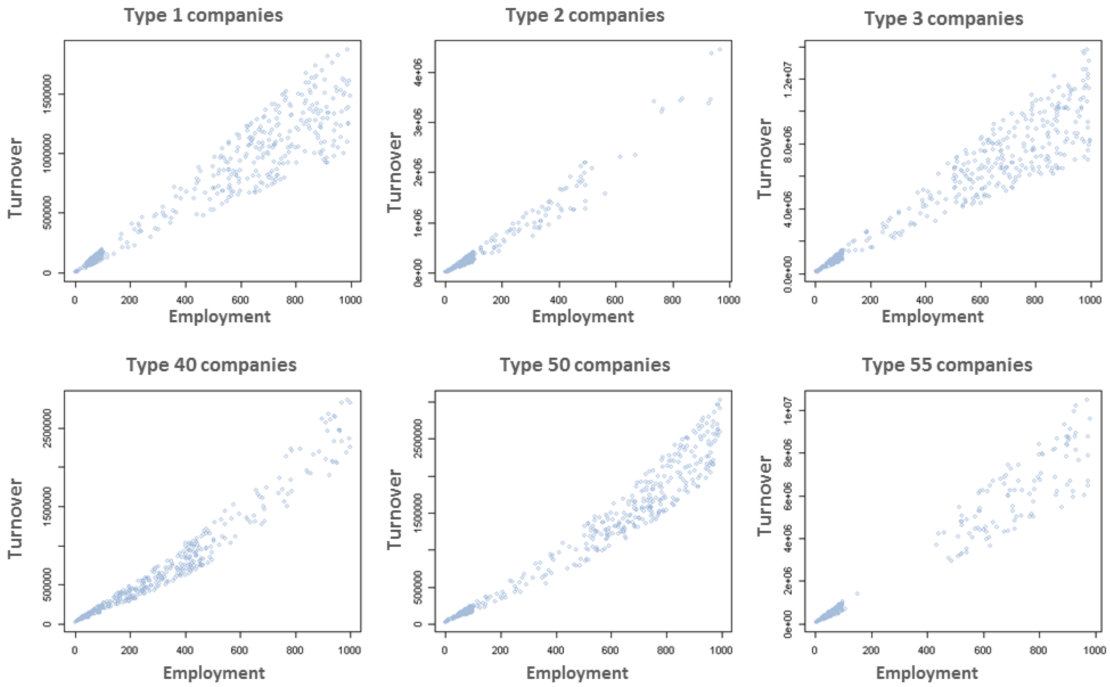
The database has been simulated to emulate the data of the Trade Registry. When generating the database, different characteristics have been taken into account to make it as real as possible.

Firstly, the data come from companies that have been classified according to their CNAE-2009 activity type and according to their size based on the number of employees. Each company has been assigned a historical territory with an approximate probability of being in Bizkaia of 0.50, 0.33 of being in Gipuzkoa and 0.17 of being in Araba/Álava. The number of companies in the simulated database was 36,719.

The turnover of each company was generated taking the type of activity and number of employees into account. For a larger number of employees the turnover would be higher, using lineal and quadratic functions. A certain variability or noise has been added to them. This variability in turnover would increase as the number of employees in the establishment increases.

In order to calculate some score functions it is essential to know the value from the previous period, prior values have also been generated for the variable turnover and number of employees. The same as when the turnover variable was generated, in this case noise or variability has been added, which is why the value of the previous period has a correlation with the current period plus a random noise.

The following graph shows the correlation between turnover and the number of employees in companies in six types of activities.



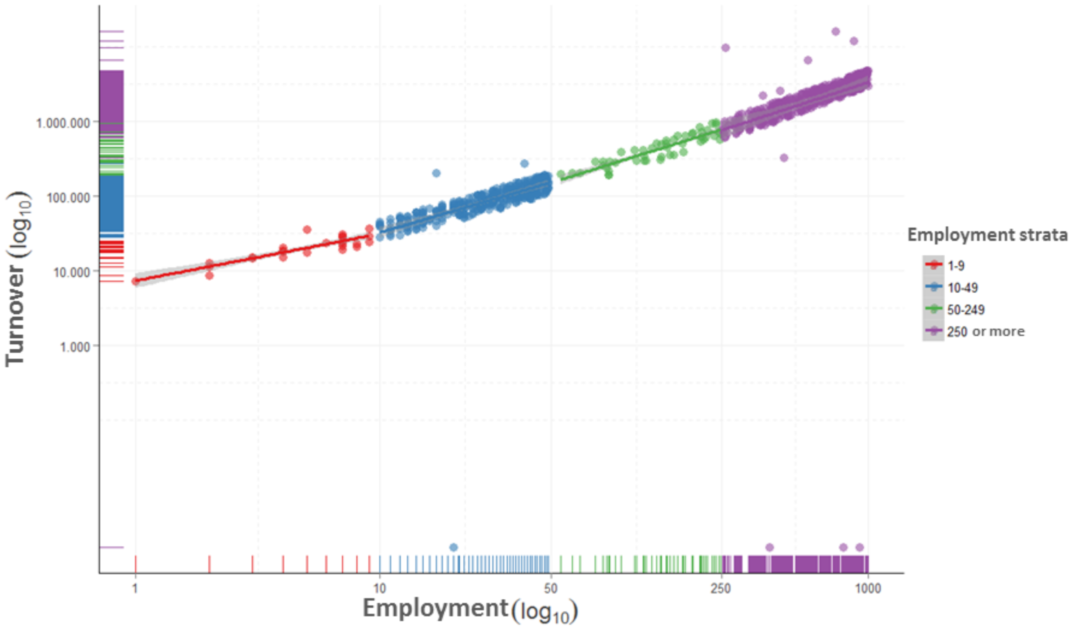
Generating different types of errors

Once the original database is generated, the next step was to add errors to the variable turnover. These errors must be as realistic as possible if the aim is to check how the score functions work when they are detected.

Firstly, 500 companies were randomly selected from the 36,719 in the database and a random error was added with a normal distribution with 0 average and a standard deviation equal to two times the original turnover. If there is a negative result the turnover would equal zero. This originates errors that may multiply turnover up to 4 or more times or that may mean that it is converted to 0.

Unit errors have also been added. Fifty companies whose turnover has been multiplied by 1,000 and another 50 companies whose turnover has been divided by 1,000 have randomly been selected.

The graph below shows the relationship between employment and turnover for all the companies belonging to one type of activity. The logarithmic scale has been used in order to be able to separate each stratum of employment well and shows the lineal regression line by employment tranches. “Anomalous” observations can be observed.



Score functions

These are the different score functions that have been calculated with the simulated database. For each company the score has been calculated with each of the functions.

Case 1: Turnover and what is expected from a homogenous group

Additive

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{|x_{ij} - \tilde{x}_{ij}|}{\tilde{x}_{ij}}$$

Multiplicative

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times (\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1)$$

Where,

x_{ij} → Turnover

$\tilde{x}_{ij}$ → Turnover in a homogenous group (same type of activity and number of employees)

Case 2: Ratio with the number of employees and expected number of employees from a homogenous group

Additive

$$s_{ij} = \frac{w_i \overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)}}{\overset{\sim}{\left(\frac{X_j}{Y_j}\right)}} \times \frac{\left| \frac{x_{ij}}{y_{ij}} - \overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)} \right|}{\overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)}}$$

Multiplicative

$$s_{ij} = \frac{w_i \overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)}}{\overset{\sim}{\left(\frac{X_j}{Y_j}\right)}} \times (\max\left(\frac{\frac{x_{ij}}{y_{ij}}}{\overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)}}, \frac{\overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)}}{\frac{x_{ij}}{y_{ij}}} \right) - 1)$$

Where,

$\frac{x_{ij}}{y_{ij}}$ → Turnover/Number of employees

$\overset{\sim}{\left(\frac{x_{ij}}{y_{ij}}\right)}$ → Turnover/Number of employees in a homogenous group (Same type of activity and number of employees)

Case 3: Ratio with the previous period and what is expected from a homogenous group

Additive

$$s_{ij} = \frac{w_i \overset{\sim}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}}{\overset{\sim}{\left(\frac{X_{j,t}}{\hat{X}_{j,t-1}}\right)}} \times \frac{\left| \left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right) - \overset{\sim}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)} \right|}{\overset{\sim}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}}$$

Multiplicative

$$s_{ij} = \frac{w_i \overset{\sim}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}}{\overset{\sim}{\left(\frac{X_{j,t}}{\hat{X}_{j,t-1}}\right)}} \times (\max\left(\frac{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}{\overset{\sim}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}}, \frac{\overset{\sim}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)}}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}\right)} \right) - 1)$$

Where,

$\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}$ → Turnover in the period t / Edited turnover in the period t - 1

$\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}}\right)$ → Turnover in the period t / Edited turnover in the period t - 1 in a homogenous group (same type of activity and number of employees)

Case 4: Influence using turnover and risk with the ratio with the number of employees.

Additive

$$s_{ij} = \frac{w_i \tilde{x}_{ij,t-1}}{\tilde{X}_{j,t-1}} \times \frac{\left| \frac{x_{ij,t}}{y_{ij,t}} - \overset{\sim}{\left(\frac{x_{ij,t-1}}{y_{ij,t-1}}\right)} \right|}{\overset{\sim}{\left(\frac{x_{ij,t-1}}{y_{ij,t-1}}\right)}}$$

Multiplicative

$$s_{ij} = \frac{w_i \tilde{x}_{ij,t-1}}{\tilde{X}_{j,t-1}} \times (\max\left(\frac{\frac{x_{ij,t}}{y_{ij,t}}}{\overset{\sim}{\left(\frac{x_{ij,t-1}}{y_{ij,t-1}}\right)}}, \frac{\overset{\sim}{\left(\frac{x_{ij,t-1}}{y_{ij,t-1}}\right)}}{\frac{x_{ij,t}}{y_{ij,t}}} \right) - 1)$$

Where,

$x_{ij,t}$ → Turnover in the period t

$\tilde{x}_{ij,t-1}$ → Turnover in the period t-1 edited

$\frac{x_{ij,t}}{y_{ij,t}}$ → Turnover/Number of employees in the period t

$\overset{\sim}{\left(\frac{x_{ij,t-1}}{y_{ij,t-1}}\right)}$ → Turnover/Number of employees in the period t-1 edited

Case 5: Turnover and estimation from the robust regression

Additive

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{\left| x_{ij} - \tilde{x}_{ij} \right|}{\tilde{x}_{ij}}$$

Multiplicative

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times (\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}} \right) - 1)$$

Where,

x_{ij} → Turnover

$\tilde{x}_{ij}$ → Estimation of turnover based on the number of employees and the company. Robust regression methods are used for each stratum.

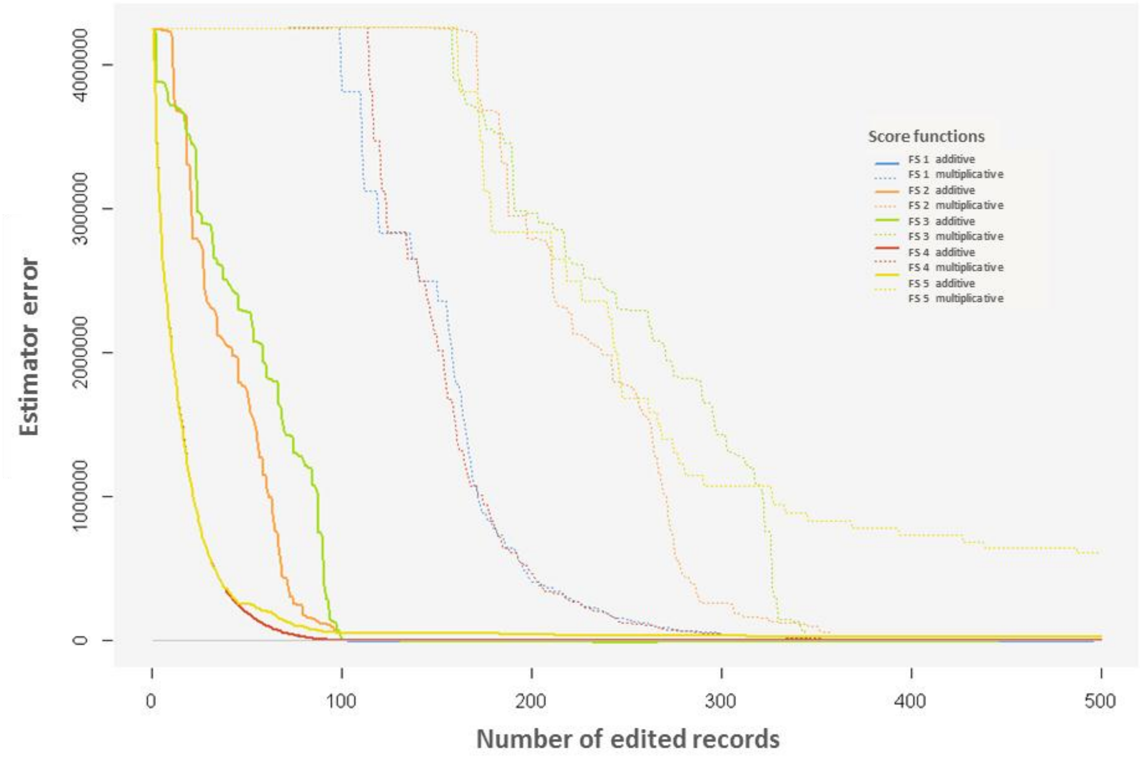
Results of the simulation

Once scoring with the score function is calculated for each company, these scores are ordered. The companies with the highest scoring are the first selected for editing.

The following graph shows how the error of the estimator evolves, in this case the average of the turnover, to the extent that companies are edited. This error is calculated as the difference between the average of turnover with no errors and the average of turnover with errors.

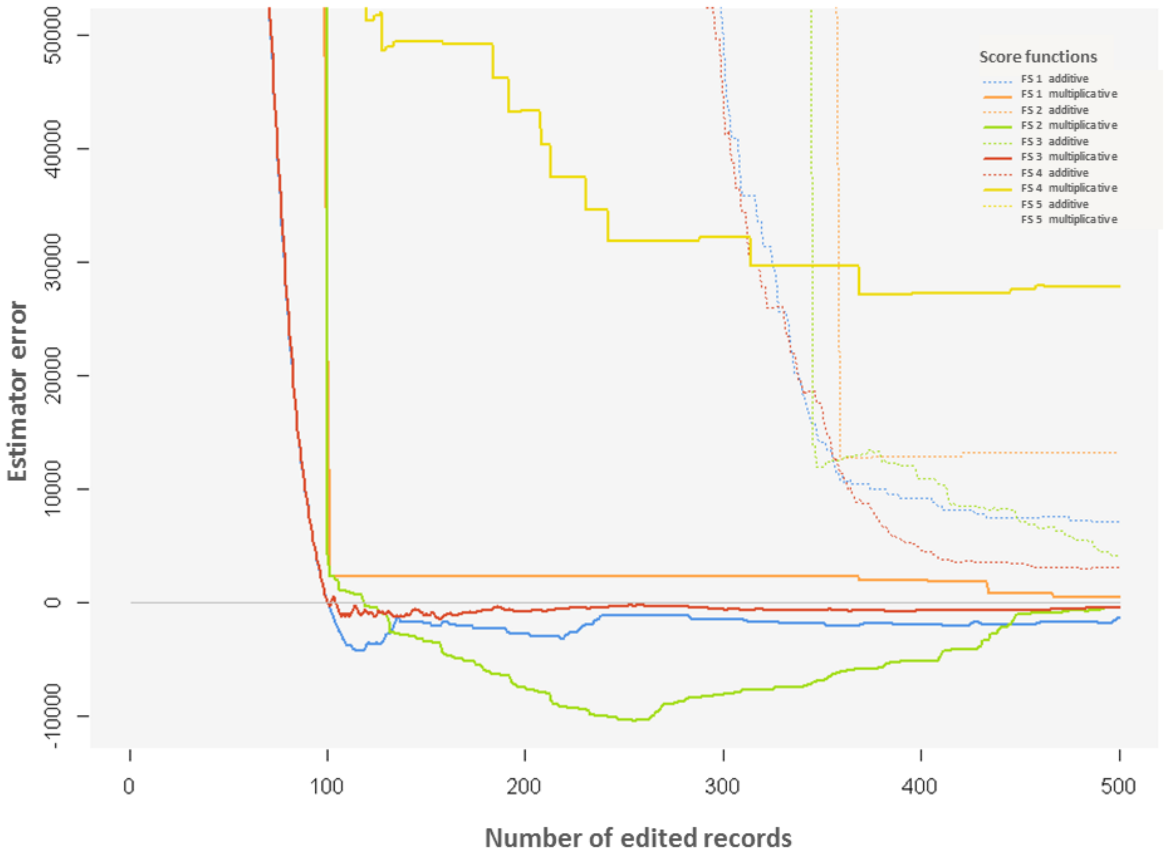
It can be seen that the score functions that use the multiplicative scale in the risk are the least efficient and take the longest to diminish the error in the estimator.

Among the additives, core functions 1, 4 seem to be the best. These are the functions whose term of “influence” takes expected contribution of turnover for each company over the total estimator into account.



A zoom has been made on the following graph on the error axis on the estimator in order to be able to compare how the support functions behave once the most important errors in turnover are eliminated.

Looking at the graph, the 4 additive score function is that which previously approached the null error and that which most closely maintained this value almost at all times. Therefore, the score function that behaves best in this simulation is the one whose term of “influence” takes the expected contribution of turnover of each company into account over the total estimator whose term of “risk” is calculated taking the turnover and number of employees ratio into account.



Chapter

6

6. SAS macro

The SAS macros that enable selective editing to be carried out on a database are presented below.

For further information see the “SAS application user manual: Selective Editing”.

Macro *FUNCION_SCORE*

The SAS *FUNCION_SCORE* macro enables three score functions on each record to be calculated.

- Type I score function: Taking the previous period and the influence from the record in the stratum defined as a reference.

- Type II score function: Taking the median value of the stratum defined and the influence of this record on the stratum as a reference.

- Type III score function: Taking the estimated value is through robust regression in the stratum defined and the influence of this record on the stratum as a reference.

- Type IV score function: Taking the median value and the interquartile range as a divider of the stratum defined and the influence of this record on the stratum as a reference.

Entry data

A SAS dataset is required that contains at least:

- The variable to be edited at the moment t.
- The variable to be edited of the previous moment t-1.
- Optional: A variable associated with the edit in order to be able to make a regression.
- Optional: Variables that identify the strata in which the median or the influence of each record for the score function are correctly calculated.

Syntax of the macro

This is a brief description of the necessary arguments:

- dataset = SAS dataset on which the selective editing is carried out.
- var = The variable to be edited at the moment t.

- var_ant = The variable to be edited of the previous moment t-1.
- var_reg = A variable associated with the editing in order to be able to make a regression.
- tipoIestrato = Stratum in which the influence of the record for the type I score function is calculated.
- tipoIIestrato = Stratum in which the median and the influence of the record for the type II score function is calculated.
- tipoIIIestrato = Stratum in which the robust regression is estimated and the influence of the record for the type III score function is calculated.
- tipoIVestrato = Stratum in which the median, the standard deviation and the influence of the record for the type IV score function is calculated.
- varpositiva = Whether or not the variable only takes positive values for its estimation in the regression must be defined. It takes the value “T” by default, which means it is positive. Its opposing value would be “F”.
- n_estrato = The minimum number of records in each stratum to be able to estimate the robust regression.
- fscore_1 = Gives the option of whether or not to calculate the type I score function. “T”= yes and “F” =no.
- fscore_2 = Gives the option of whether or not to calculate the type II score function. “T”= yes and “F” =no.
- fscore_3 = Gives the option of whether or not to calculate the type III score function. “T”= yes and “F” =no.
- fscore_4 = Gives the option of whether or not to calculate the type IV score function. “T”= yes and “F” =no.

FS_GLOBAL macro

The SAS *FS_GLOBAL* macro enables the global score function *fs_global* to be calculated, which combines the previously calculated local score functions.

Entry data

Requires an SAS dataset in which at least the local score functions for a variable have been calculated with the *FUNCION_SCORE* macro.

Syntax of the macro

This is a brief description of the necessary arguments:

- dataset = SAS dataset on which the selective editing is carried out.
- var = The variable to be edited at the moment t.

- vartext = Literal of the variable.
- max = Calculates the maximum of the three score functions. This value is “T” by default. To calculate the sum of the three score functions it is sufficient to specify max = “F”.
- w1 = Weight for the type I score function. By default 1.
- w2 = Weight for the type II score function. By default 1.
- w3 = Weight for the type III score function. By default 1.
- w4 = Weight for the type IV score function. By default 1.
- weights = weight of each of observation in the population. By default 1.

Chapter

7

7. Practical implementation in the Services Statistics operations

The new Services Statistics operations of the Basque Statistics Institute is created, on the one hand, with data from the questionnaire aimed directly at establishments and, on the other, with data from three administrative records: the Trade Registry, the Cooperatives Registry and the Associations and Foundations Registry.

The use of registry information has enabled better estimations to be obtained by having a larger quantity of information. However, having to work with such a large volume of information has involved a series of difficulties of a different nature. Especially with the Trade Registry, which is the source that provides the most data; specifically, in the 2012 Services Statistics information from 16,251 companies was used.

For this reason it is necessary to work with more efficient and reliable sources, as is the case with selective database editing.

Practical implementation of the selevtive editing

The selective editing macro has been applied to the database that combines the information from the questionnaire aimed directly at establishments as well as the information from the administrative records: Trade Registry, Cooperatives Registry and Associations and Foundations Registry.

The variables that have been used are Turnover, Value Added at Factor Cost and Personnel Cost, which are considered the main variables available and are also the variables that most correlated with the 'number of people employed' variable.

- **Editing the Net Turnover variable**

The following options were selected when applying the macro. Firstly, it is considered that, for this variable and in this case, the most relevant score functions are type II and III, which seek, respectively, records that are separated from the median of the stratum or from the estimated value of the robust regression, with employment being the adjustment variable.

Since there is no data for each establishment from the previous year, the type I score function was not calculated, which is the score function that takes the deviation with respect to the previous year into account. It was also opted not to calculate the type IV score function which is the one that most takes variability in each stratum into account.

The stratum in which it was decided to calculate the score functions was the combination of the CNAE to two digits together with its employment stratum. In order to be able to estimate the robust regression, a minimum number of 20 establishments was decided on in each record and the combination of global score functions was set as the maximum between them.

As an example, the following table shows the first 10 establishments from the output of the editing macro.

Table 7.1. Editing the Net Turnover variable: First 10 establishments.

Establishment	Company name	Employment	CNAE09	Employment Stratum	Net Turnover	Median	Estimate	fs global
a1	2. stratum value	3	5229	1	3.470.393	25.344		1,530
a2	3. regression value	1	7022	2	20.881.251	105.883	79.446	0,486
a3	3. regression value	2	5221	1	1.244.845	25.344		0,479
a4	3. regression value	1	7112	1	423.372	44.925	45.666	0,178
a5	2. stratum value	1	6910	1	308.531	41.729	40.776	0,128
a6	3. regression value	1	7111	1	284.882	44.925	45.666	0,111
a7	2. stratum value	1	6910	1	288.304	41.729	40.776	0,110
a8	3. regression value	1	6621	1	170.406	32.962		0,104
a9	3. regression value	2	7022	2	11.482.794	105.883	119.914	0,097
a10	3. regression value	2	7022	2	11.443.789	105.883	119.914	0,096

As can be seen in table 7.1., these establishments are greatly separated from the median of their stratum or from what is expected in their stratum with this number of employees. The Net Turnover in these establishments far exceeds that of establishments in the same sector and with similar employment, and adjusting by the number of employees also greatly exceeds what is expected.

In the cases in which the number of establishments in the stratum does not reach 20, no value is estimated for Net Turnover.

- **Editing the Net Turnover variable ratio by person**

In this section the Net Turnover ratio among the establishment's number of employees is edited, taking only the median of the stratum into account and assigning equal weights to all establishments.

Table 7.2. Editing the Net Turnover variable with different weights for each establishment: First 10 establishments.

Establishment	Company name	Employment	CNAE09	Employment Stratum	Net Turnover Ratio per person	Median	fs global
b1	2. stratum value	1	7022	2	**20.881.251**	65.193	0,213
b2	2. stratum value	3	5229	1	**1.156.798**	25.344	0,062
b3	2. stratum value	18	9001	4	**807.432**	45.749	0,031
b4	2. stratum value	2	5221	1	**622.423**	25.344	0,017
b5	2. stratum value	2	7022	2	**5.741.397**	65.193	0,016
b6	2. stratum value	2	7022	2	**5.721.895**	65.193	0,016
b7	2. stratum value	1	7112	2	**5.311.000**	70.454	0,014
b8	2. stratum value	108	9312	7	**438.467**	40.003	0,014
b9	2. stratum value	1	7219	2	**1.305.609**	65.612	0,010
b10	2. stratum value	1	6820	2	**4.266.072**	45.036	0,010

It can be seen that the Turnover Ratio by person in these establishments is very different to what occurs in its stratum, and therefore must be checked.

- **Editing the Value Added at Factor Costs variable**

In order to edit this variable the same options were selected as for the Net Turnover Amount. Firstly, since there is no data for each establishment from the previous year, the type I score function was not calculated, which is the score function that has the deviation with respect to the previous year. It was also opted not to calculate the type IV score function which is the one that most takes variability in each stratum into account.

It is similarly considered that the most relevant score functions for Value Added are type II and III, which seek, respectively, records that are separated from the median of the stratum or from the estimated value of the robust regression, with employment being the adjustment variable.

The stratum in which it was decided to calculate the score functions was the combination of the CNAE to two digits together with its employment stratum. In order to be able to estimate the robust regression, a minimum number of 20 establishments was decided on in each record and the combination of global score functions was set as the maximum between them.

The following table shows the first 10 establishments from the output of the editing macro.

Table 7.3. Editing the Value Added at Factor Costs: First 10 establishments.

Establish ment	Company name	Employ ment	CNAE09	Employ ment Stratu m	Value Added at Factor Costs	Median	Estimate	fs global
c1	3. regression value	1	6820	1	**537.319**	23.443	26.655	3,116
c2	2. stratum value	1	7112	1	**925.311**	27.002	25.698	2,466
c3	3. regression value	2	5221	1	**718.272**	17.668		1,428
c4	3. regression value	2	7022	2	**11.021.163**	63.658	76.697	0,271
c5	3. regression value	1	4725	1	**142.580**	8.187	7.633	0,255
c6	3. regression value	1	6910	1	**262.088**	30.000	29.363	0,173
c7	3. regression value	1	9001	1	**285.340**	26.355		0,172
c8	3. regression value	1	6910	1	**251.182**	30.000	29.363	0,158
c9	2. stratum value	1	7220	1	**60.703**	20.234		0,142
c10	3. regression value	1	5320	1	**44.390**	30.717		0,117

- ○ **Editing the Value Added at Factor Costs by person**

In this section the Value Added ratio between the establishment's number of employees is edited, taking only the median of the stratum into account and assigning equal weights to all establishments.

Table 7.4. Editing the Value Added at Factor Costs by person.

Establishment	Company name	Employm ent	CNAE09	Employ ment Stratu m	Value Added at Factor Costs by person	Median	fs global
d1	2. stratum value	2	7022	2	**5.510.582**	39.980	0,066
d2	2. stratum value	2	5221	1	**359.136**	16.517	0,043
d3	2. stratum value	1	7112	1	**925.311**	26.220	0,041
d4	2. stratum value	5	4742	3	**515.179**	9.097	0,034
d5	2. stratum value	1	6820	1	**537.319**	22.294	0,030
d6	2. stratum value	2	7022	2	**3.419.244**	39.980	0,025

d7	2. stratum value	108	9312	7	**407.720**	32.938	0,021
d8	2. stratum value	7	4764	3	**382.244**	9.097	0,019
d9	2. stratum value	50	5210	6	**842.255**	54.571	0,016
d10	2. stratum value	10	6820	4	**580.367**	69.618	0,016

The Value Added ratio observed by a person in these establishments is very different to what occurs in a stratum, and they are therefore checked to see if their value is correct or if it needs to be edited.

- **Editing Personnel Cost by person**

In this case it was correctly opted to use the Ratio between Personnel Cost and employment of each establishment.

Table 7.5. Editing Personnel Cost by person.

Establishment	Company name	Employment	CNAE09	Employment Stratum	Personnel Cost Ratio per person	Median	fs global
Establecimiento	2. stratum value	108	9312	7	**340,707**	30,095	0.078
e1	2. stratum value	23	9312	5	**340,707**	27,364	0.032
e2	2. stratum value	65	9312	6	**144,770**	22,734	0.025
e3	2. stratum value	32	9312	5	**267,197**	27,364	0.019
e4	2. stratum value	246	8220	7	**80,607**	19,478	0.017
e5	2. stratum value	30	9312	5	**218,031**	27,364	0.013
e6	2. stratum value	106	9312	7	**144,770**	30,095	0.012
e7	2. stratum value	6	9102	3	**79,146**	31,681	0.012
e8	2. stratum value	7	4764	3	**340,707**	27,338	0.011
e9	2. stratum value	108	9312	7	**340,707**	30,095	0.078

This table shows establishments where the Personnel Cost Ratio by employment greatly surpasses the median of the stratum.

Selective editing implementation summary

The macro programmed in Selective Editing SAS developed in the Institute has served to edit economic information obtained from different sources, after validating this information, integrating it and contrasting it with the Eustat Economic Activities Directory.

The main variables that have been used are Turnover, Value Added at Factor Costs and Personnel Cost, which are considered the main variables available and are also variables that are the most correlated with the 'number of people employed' variable.

The analysis that the macro provides has enabled establishments to be detected with extreme and influential values within the elevation strata (activity and employment stratum) taking its influence within these strata into account. Thus it has been possible to efficiently and reliably evaluate the data, especially taking into account the volume of data being worked with.

Chapter

8

8. Conclusions

Finally, this chapter presents a summary of the main conclusions of this work about selective database editing that has been undertaken throughout the training and research scholarship on statistical and mathematical methodologies.

Summary and conclusions about selective editing

Having efficient editing methods is essential for statistical entities given that one of the parts that takes the most time and is the most expensive in the process of improving data quality is manual or interactive data editing.

It has been shown that the number of records to edit can be largely reduced, given that for many records, manual editing has an insignificant influence on the estimators of the main parameters of interest.

In this context, a selection strategy is required that separates the records into two parts: one critical with the records that supposedly contain influential errors, and another with records whose editing is not expected to change the results to be published.

Selective editing is the strategy by which only the records whose correction significantly influences the results to be published are edited, thus reducing costs and delivery times.

The score function is the main instrument of selective editing. This function assigns a score to each record for each variable analysed. This scoring provides an indication of the expected effect on the parameter to be estimated if it is edited. Records with a high score are those selected first for editing.

Different types of score functions can also be calculated depending on what the expected reference value will be or, also, if several variables are being edited together. It is for this reason that in order to be able to select an entire record and thus edit it, a value is required that combines the information from the different score functions. This value is known as the global score. This score must reflect the importance of entirely editing the record.

This work has shown that the score function is a valid and efficient instrument for selecting anomalous and influential records. Firstly, and in a more theoretical framework, a simulated database is used in which the behaviour of different types of score functions can be observed when selecting the records to be edited.

Subsequently, these techniques are applied in a real framework, specifically in the new Basque Statistics Institute Statistical Operations. The macro programmed in SAS has served to edit economic information obtained from different sources, after validating this information, integrating it and contrasting it with the Eustat Economic Activities Directory.

It must be pointed out that the editing techniques described here as well as the programmed SAS macros can be applied and modified with relative ease on any type of database that needs to be edited. The parameters can be changed on the SAS macros so that a unique variable or several variables together can be used to edit. A selection can be made from among four score functions and they can be combined and different weights assigned to each one. Each score function can also be calculated on different strata. Additionally, the weight of each record can be different or the same, depending on what is of interest.

The SAS macros can be offered on the EUSTAT web site to institutions, statistical institutes or researchers interested in implementing them.

Finally, this technical notebook is mainly based on the methodology applied by the Holland National Institute of Statistics and what has been published in the technical notebooks (Hoogland, van der Loo, Pannekoek and Scholtus, 2011) and (de Wall, 2008) and in the model described in the European EDIMBUS project (Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys 2007).

Chapter

9

Bibliography

BELLISAI, D.; DI ZIO, M.;GUARNERA, U. AND LUZZI, 0. (2009)

A selective editing approach based on contamination models: An application to an ISTAT business survey. Working Paper No. 27, UN/ECE Work Session on Statistical Data Editing, Neuchatel.

DE WAAL, T. (2008)

An overview of statistical data editing. Statistics Netherlands, The Hague/Heerlen.

DE WAAL, T., PANNEKOEK, J. AND SCHOLTUS, S. (2011)

Handbook of statistical data editing and imputation. Wiley.

DI ZIO, M.; GUARNERA, U. AND LUZZI, 0. (2008)

Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. Working Paper No. 22, UN/ECE Work Session on Statistical Data Editing, Vienna

EUREDIT PROJECT (2004a)

Towards Effective Statistical Editing and Imputation Strategies. Findings of the Euredit Project, Volume 1. Available at: http://www.cs.york.ac.uk/euredit/results/results.html

EDIMBUS (2007)

Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. Manual prepared by ISTAT, Statistics Netherlands and SFSO.

GHOSH-DASTIDAR, B. AND SCHAFER, J. L. (2006)

Outlier Detection and Editing Procedures for Continuous Multivariate Data. Journal of Official Statistics 22, pp. 487-506.

GRANQUIST, L (1995)

Improving the Traditional Editing Process. Business Survey Methods. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, eds. John Wiley & Sons, New York, 385–401.

GRANQUIST, L. AND KOVAR (1997)

Editing of Survey Data: How Much Is Enough? Survey Measurement and Process Quality. L.E. Lyberg, P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin, eds. John Wiley & Sons, New York, pp. 415–435.

HEDLIN, D (2003)

Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics 19, 177-199

HEDLIN, D (2008)

Local and Global Score Functions in Selective Editing. Working Paper No. 31, UN/ECE Work Session on Statistical Data Editing, Vienna.

HIDIROGLOU, M. A. AND BERTHELOT, J. M. (1997)

Statistica Editing and Imputation for Periodic Business Surveys. Survey Methodology 12, pp. 73-78.

HOOGLAND, J (2002)

Selective editing by means of Plausibility Indicators. UNECE Work Session on Statistical Data Editing, Helsinki, working paper no. 33.

HOOGLAND, J; VAN DER LOO, M; PANNEKOEK, J. AND SCHOLTUS, S. (2011)

Data editing. Detection and correction of errors. Statistics Netherlands, The Hague/Heerlen.

LATOUCHE, M. AND BERTHELOT, J. M. (1992)

Use of a Score Function to prioritise and Limit Recontacts in Editing Business Surveys Data editing. Journal of Official Statistics 8, pp. 389-400

LAWARENCE, D. AND MCKENZIE, R. (2000)

The General Application of Significance Editing. Journal of Official Statistics 16, pp. 243-253

VAN LANGEN, S (2002)

Selective Editing by Using Logistic Regression. Report, Statistics Netherlands, Voorburg.

Organismo Autónomo del

EUSKO JAURLARITZA
GOBIERNO VASCO

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

www.eustat.eus