Selective editing procedure for the Community Survey on the Structure of Agricultural Holdings

Orietta Luzi, Giovanni Seri, Roberta Varriale*
Istat
Via Balbo 16
Rome, Italy
*varriale@istat.it

DOI: 10.1481/icasVII.2016.g45c

ABSTRACT

The annual survey on the Structure of Agricultural Holdings (SAH) collects information on agricultural areas for cultivation type, type and quantity of livestock, agricultural production, structure and amount of labour involved in the holding. In the present work we describe and evaluate the results of the selective editing strategy used to identify the influential values of the SAH main numeric continuous variables (reference year 2013).

Keywords: selective editing, influential error, agricultural census, data modelling

1. Introduction

The survey on Structure of Agricultural Holdings (SAH below) collects information on Italian farms about land by type of cultivation, type and amount of livestock, type of production, structure and amount of family and not family labour (ISTAT, 2001). For the first post-census survey (reference year 2013), a new procedure for checking and correcting data has been designed, which integrates different methods and tools for the detection and treatment of different types of non sampling errors in the data. In particular, this work focuses on the description of the selective editing strategy designed to detect the possibly influential errors for the main continuous numeric variables from the survey (farms agricultural surfaces, livestock and some productions).

Within each data editing and correction strategy involving companies (in this case, farms), it is usual to investigate which are the units potentially affected by errors having significant impact (and therefore potential biasing effect) on the final estimates of the main target variables (see for example EDIMBUS, 2007, or MEMOBUST, 2014). One class of particularly effective methods in this area is known as selective editing (Latouche *et al.*, 1992), in which the units potentially affected by influential errors are selected to ensure a more severe control (generally manual/interactive revision) of their nature/source aiming at reducing their potential biasing effect on the final estimates. In the case of the SAH, the selective editing method adopted for the identification of influential errors is based on the latent class contamination models proposed by Di Zio and Guarnera (2013) and implemented in the generalized package Selemix (Guarnera *et al.*, 2013).

The work is structured as follows. Section 2 contains a description of the main features of the overall data editing and correction procedure implemented for the SAH numeric variables. Section 3 contains a brief description of the selective editing methodology and the Selemix software. Section 4 shows the results of applying the method to the most important survey variables of the SAH 2013. In Section 5 the results and efficiency of selective editing process are evaluated. Section 6 contains some concluding remarks and future work perspectives.

2. SAH survey: the data editing and correction procedure

The SAH 2013 responds to EU Regulation No. 1166/2008 of the European Parliament and of the Council of 19 November 2008. It aims to systematically produce statistics on the structure of agricultural holdings and on agricultural production methods. For numeric continuous variables, the final estimates are totals at National and regional level. The theoretical sample of the SAH 2013 includes 42,723[1] companies (D'Orazio, 2013), with a reference population of 1,138,214 companies active in the Country (about 70% of the 1,620,844 recorded at 2010 6th Census of Agriculture). A distinctive feature of this edition of the SAH is that this is the first survey conducted after the 6th General Census of Agriculture, which has made it possible to use the census as an auxiliary source, in order to make more efficient the survey process and more accurate the final results.

The survey was conducted by using web electronic questionnaires. Multiple factors contribute to the complexity of the data editing and correction strategy: the high number of observed variables, their types (both categorical and continuous), the complex relationships (both structural and statistical/mathematical) existing between the variables. With regard to non sampling errors, it is necessary to point out that some of them (for example duplication, skip items, balance) are already identified at the data entry stage, where some consistency checks are carried out at the time of entering information. This strategy ensures more accurate data with respect to a basic set of controls. However, the incompleteness of the controls means that collected data may still remain in situations of non-acceptability or inconsistency, to be properly treated at the subsequent data editing and imputation (E&I) stage.

As known, an error that occurs during collecting or data entry can have both systematic or random nature: these errors can be identified when they entail logical/statistical/mathematical inconsistencies, anomalous values, influential values on the target estimates. In order to identify the various types of error that are potentially present in the target variables (surfaces, livestock and farm productions), a complex procedure of E&I has been carried out. Its main steps are:

[1] In order to meet the Regulation requirements for some domain of interest the number of sample units was increased adding 2.030 small units outside the target population of the SAH: this kind of units are out of the purpose of this paper. The total number of sample units contributing to the estimates keeping into account also events like farms merge or split is 44552.

1 error localization in the phase of identification of the statistical units and/or coverage (errors in identification codes, merge/split of farms compared to the census, membership to the target population,…). This phase will be referred to as pre-editing;

2 identification of possibly systematic errors (in particular unit measure errors) on the basis of a deterministic approach, in order to apply proper automatic corrections;

3 identification of influential errors, with a presumably random nature, through selective editing, for subsequent manual/interactive review;

4 identification and automatic correction of not influential errors, with a presumably random nature, on the basis of a probabilistic approach. In particular, a data-driven type method, implemented in software Diesis (Bruni *et al.*, 2002), has been used.

In the following, the focus will be on the second and third phases of the procedure.

3. The selective editing approach for the identification of influential errors

Let X be a continuous variable and θ a parameter to be estimated on the X's distribution. We define as influential with respect to θ those X values which are potentially affected by a measurement error and have a possibly biasing effect on the θ estimate. Given their statistical relevance, these values need an accurate verification, which is generally performed by manual/interactive revision by well-trained expert clerks. In order to reduce costs and time of such data treatments, they have to be limited to the most critical units. All the residual errors could be eliminated by using less costly automatic data editing procedures. The main objective of selective editing is actually to limit the most demanding activities of data editing (manual reviewing, re-contact, etc.) to those cases where the expected benefit in terms of costs reduction is highest (Lawrence *et al.*, 2000). A score function (Latouche *et al.*, 1992) is used to rank the observed units, so that observations with the highest score are supposed to contain the most potentially influential errors on the target estimates (in the case of the SAH, the variables' totals). A specific value for the score function (threshold) is set in advance, and the m units with score above the threshold are selected for manual editing, where m depends on the expected estimates accuracy. The latter generally corresponds to the allowed residual error on the estimates computed on the not edited data. Since usually the estimates of interest (e.g. totals) involve several variables, a different score function is computed for each variable (local score) and a unique global score is obtained by suitably combining local scores. The selective editing approach allows to reduce not only the overall cost of interactive editing, but also the over-editing effect (that is the resources spent to manually check non sampling errors having low biasing effect on the target parameters estimates).

Typically, the score functions are based on the comparison of the observed values of the target variable with the corresponding predicted values obtained through some (explicit or implicit) model, taking into account the possible sampling weights. A key issue with this approach is that errors tend to be identified with residuals with respect to the assumed model, so that it is difficult to separate the natural variability of the investigated phenomenon from the extra variability due to the presence of errors in data. Moreover, it is not obvious how to define the threshold determining how many units have to be selected for interactive editing. A recent approach (Di Zio *et al.*, 2013 and 2008) tries to overcome this difficulty by explicitly modelling both true (i.e. not contaminated) data and error mechanism. In particular, in order to capture the intermittent nature of the model, a two-component mixture model is used, where the components are naturally associated with error-free and contaminated data, respectively. In other words, errors are assumed to affect only a subset of

data in such a way that each unit in the dataset is corrupted by an error with an (unknown) a priori probability. Based on the appropriate conditional distribution, for each observed value the corresponding predicted "true" value is obtained, and the error component is determined as the difference between the observed and the true variable values. In this way, the number of units to be selected for manual revision can be directly associated to the required accuracy for the target estimates. It has to be underlined that the estimation of contamination models can be performed taking into account an appropriate stratification of the population units, that can be different by that one underlying the estimation domains (reference data partitions at the influential errors identification step). In practice, often the estimation domains correspond to a more detailed data partition with respect to the data stratification adopted at the contamination model estimation stage: this depends on the fact that in order to obtain "robust" and reliable estimates, a suitably large number of sampling units is needed in each strata. The described method is implemented in the R package SeleMix (SELEctive editing via MIXture models) available on the website http://www.R-project.org (Guarnera *et al.*, 2013). This package includes functions for the estimation of the model parameters via EM algorithm, computation of prediction of true values conditional on observed values, prioritization of units for interactive editing according to a user-specified threshold. If the analysis is performed on sample surveys, in order to select the most influential errors SeleMix uses the sampling weights. Missing values in the response contaminated variables are allowed. In this case SeleMix can also be used as a tool for (robust) imputation of incomplete data. The covariates included in the model are supposed to be error-free and not affected by non-response, therefore the efficiency of this approach also depends on the reliability of the available auxiliary information.

4. The selective editing procedure for the SAH

The aim of the selective editing approach implemented for the 2013 SAH survey was the identification of the potentially influential errors affecting a subset of key survey variables on firm surface (*Total Agricultural Surface, Utilized Agricultural Surface, Actually Irrigated Surface*) and livestock (*Total Number of Cattles, Total Number of Equines, Total Number of Cattle "Bufale", Total Number of Sheep, Total Number of Goats, Total Number of Pigs, Total Number of Rabbits*), Production of milk. For surface and livestock variables, auxiliary information from the 6th Census of Agriculture (reference year 2011) is available. However, no census information is available on milk produced by dairy cows, sheep and goats. Out of the 42.723 firms included in the SAH theoretical sample, the selective editing models have been applied to the subset of 38.330 respondents which resulted correctly identified (also in a longitudinal perspective) at the *pre-editing* stage[2]. Based on exploratory data analyses and the direct support of area experts, we defined: the contamination model at regional level, the approach to be adopted in order to predict the "true" values (model-based values vs census variable values), and the threshold on the target estimates accuracy. In particular, the choice of the contamination model for each single target variable as well as the specific auxiliary information to be used have been mainly based on the amount of available observed information and on the data characteristics (e.g., high rates of zero values in variables' distributions).

Before applying the selective editing strategy, some preliminary analysis and editing activities have been performed on the data in order to identify and possibly eliminate the systematic and/or measurement errors affecting the target variables. Different latent-class regression models have been developed for the different variables subject to selective editing (*Total and Utilized*

[2] Responding firms consistent with respect to Census units, e.g. splitted units are properly reconstructed in order to match with the Census generating units.

Agricultural Surface, Livestock, Produced Milk, Total Irrigated Surface), using as auxiliary information, when appropriate, the corresponding items observed at the Census of Agriculture and possible related survey variables. As some of the target variables (i.e., *Total and Utilized Agricultural Surface*) are observed in more than one item in the survey questionnaire, all the available information is simultaneously used in the corresponding estimated models. As mentioned, depending on the considered variables, the auxiliary information from the 2010 Census of Agriculture perform differently in terms of accuracy of predictions: i.e., the census provides good auxiliary information for the agricultural surfaces, on the contrary for variables related to the production of milk no reliable information is available from the census, therefore information from the survey itself has been used in order to exploit existing relations among variables for editing purposes. For each target variable, different stratifications are used at the model estimation and at the influential data identification steps, in order to guarantee the robustness of model estimates in each stratum and the efficient identification of influential errors at the most appropriate publication domains. Furthermore, different thresholds have been set for each target variable to properly tailor the balance between the expected estimates accuracy and the costs for manual revisions.

In Table 1, the number of observations identified as influential by variable and by region, is reported. As an example, 183 observations have been selected as influential for the variable Utilized Agricultural Surface (SAU2) and 129 ones for the variable Total Agricultural Surface (SAT2) (about 0.5% and 0.3% of the sample units, respectively): about 250 agricultural firms will have to be manually checked in this case, as there is an overlap of about 80 units in the two sets of potentially influential observations. In general, for the most part of the considered variables, the results highlight a not uniform distribution of potential influential errors in the Italian Regions: this suggests the need of most accurate preliminary data analyses aiming at identifying potential lacks or systematic sources of errors at the data collection and/or at the data entry stages.

5. Evaluation of the selective editing process

In this section the evaluation of the selective editing process with respect to the SAU2 and SAT2 variables is reported. Tables 2 and 3 show, respectively for SAU2 and SAT2, the number of: collected data (A), data corrected (imputed) through automatic or interactive editing procedures (B), data reported as influential (C) and, among them, corrected data using interactive editing (D), by Italian region. Moreover, Tables 2 and 3 show the following percentages:

- imputation rate (B/A): number of corrected information out of the number of collected data,
- rate of influential errors (C/A): number of data reported as influential out of the number of collected data,
- imputation rate by influential errors (D/B): number of data reported as influential and interactively corrected out of the total number of correct data,
- hit rate (D/C): number of data reported as influential and corrected by means of interactive editing out of the total number of data reported as influential.

Table 2 shows that out of the 44,552 SAU2 observed values in the final dataset, 97.18% is not corrected. Complementarily, the imputation rate is 2.82%, while the percentage of cases reported as 'influential' is 0.41% (183 units). Out of these units, only 20 (hit rate = 10.93%) were classified as errors and corrected. For SAT2 (see Table 3), similarly, the 96.53% of the observations is not subject to any correction. The proportion of units reported as 'influential' is 12.29% (129 cases). Out of these, only 24 (hit rate = 18.60%) were classified as errors and corrected.

Table 1: *Number of influential observations, by Region per and target variable*

Region	SAU2	SAT2	CATTLE	EQUINS	CATTLE "BUFALA"	SHEEP	GOATS	PIGS	RABBITS	COW MILK	"BUFALA" MILK	SHEEP MILK	GOAT MILK	IRR1	IRR2	Total (a)
Piemonte	12	4	9	4	2	17	10	**80**	4	7	0	0	0	0	0	144
Valle D'Aosta	17	12	1	1	0	3	0	0	0	0	0	1	3	5	5	32
Lombardia	13	4	0	5	0	6	0	**68**	5	5	0	1	0	0	0	104
Veneto	5	4	3	5	1	4	11	28	8	2	1	0	0	**240**	**431**	586
Friuli-Venezia Giulia	4	3	8	1	1	3	8	12	2	0	0	3	0	0	**77**	114
Liguria	13	4	5	0	0	2	11	0	0	0	0	1	1	**115**	**90**	164
Emilia-Romagna	3	3	0	1	0	2	8	32	1	2	0	1	0	0	44	94
Toscana	5	9	**44**	9	3	**67**	11	10	4	4	0	**25**	1	0	0	171
Umbria	18	9	19	2	1	7	1	10	3	3	1	1	0	0	0	61
Marche	12	5	18	1	0	30	6	1	5	2	0	3	1	54	50	162
Lazio	10	5	**88**	**18**	8	29	6	5	2	5	3	17	0	0	0	178
Abruzzi	14	13	8	3	0	3	7	8	1	0	0	6	0	1	32	80
Molise	13	17	26	3	3	31	4	11	0	4	0	2	0	39	23	130
Campania	3	2	11	1	16	20	8	8	3	8	4	0	1	0	1	78
Puglia	13	12	11	4	1	29	14	6	2	0	0	7	1	**202**	72	274
Basilicata	9	6	0	6	3	**57**	**38**	8	2	0	0	**32**	3	31	7	171
Calabria	4	4	**69**	2	3	**110**	**96**	21	1	0	1	9	0	0	0	280
Sicilia	5	4	**170**	**24**	2	**58**	**71**	2	3	4	0	**60**	**12**	34	0	406
Sardegna	3	1	3	3	1	0	0	37	1	0	0	1	0	81	0	129
Trentino-Alto Adige-Trento	1	1	0	0	0	4	1	0	0	0	0	0	0	0	0	6
Trentino-Alto Adige-Bolzano	6	7	1	0	0	6	3	2	1	0	0	0	1	1	0	22
Total	183	129	494	93	45	488	314	349	48	46	10	170	24	803	832	3386

(a) The Total does not represent the row sum but the number of distinct farms for which at least one influential value has been identified.

Table 2: *Number of observations, by regions. SAU2 variable*

Region	Total (A)	Total Corrected (B)	Influential (C)	Influential Corrected (D)	B/A (%)	C/A (%)	D/B (%)	D/C (%)
Piemonte	2236	82	12	1	3.67	0.54	1.22	8.33
Vale D'Aosta	230	21	17	4	9.13	7.39	19.05	23.53
Lombardia	2090	96	13	1	4.59	0.62	1.04	7.69
Veneto	2877	39	5	0	1.36	0.17	0.00	0.00
Friuli-Venezia Giulia	1152	66	4	1	5.73	0.35	1.52	25.00
Liguria	649	39	13	2	6.01	2.00	5.13	15.38
Emilia-Romagna	2698	82	3	1	3.04	0.11	1.22	33.33
Toscana	2400	14	5	0	0.58	0.21	0.00	0.00
Umbria	1117	32	18	1	2.86	1.61	3.13	5.56
Marche	1716	91	12	1	5.30	0.70	1.10	8.33
Lazio	3231	116	10	1	3.59	0.31	0.86	10.00
Abruzzi	2342	13	14	0	0.56	0.60	0.00	0.00
Molise	974	26	13	2	2.67	1.33	7.69	15.38
Campania	2847	38	3	0	1.33	0.11	0.00	0.00
Puglia	3081	49	13	1	1.59	0.42	2.04	7.69
Basilicata	1749	73	9	1	4.17	0.51	1.37	11.11
Calabria	4181	49	4	0	1.17	0.10	0.00	0.00
Sicilia	5432	240	5	2	4.42	0.09	0.83	40.00
Sardegna	2218	9	3	0	0.41	0.14	0.00	0.00
Trentino-Alto Adige-Trento	524	52	1	1	9.92	0.19	1.92	100.00
Trentino-Alto Adige-Bolzano	808	29	6	0	3.59	0.74	0.00	0.00
Total	44552	1256	183	20	2.82	0.41	1.59	10.93

Table 3: *Number of observations, by regions. SAT2 variable*

Region	Total (A)	Total Corrected (B)	Influential (C)	Influential Corrected (D)	B/A (%)	C/A (%)	D/B (%)	D/C (%)
Piemonte	2236	82	4	0	3.67	0.18	0.00	0.00
Valle D'Aosta	230	21	12	4	9.13	5.22	19.05	33.33
Lombardia	2090	95	4	1	4.55	0.19	1.05	25.00
Veneto	2877	38	4	0	1.32	0.14	0.00	0.00
Friuli-Venezia Giulia	1152	62	3	0	5.38	0.26	0.00	0.00
Liguria	649	32	4	2	4.93	0.62	6.25	50.00
Emilia-Romagna	2698	91	3	0	3.37	0.11	0.00	0.00
Toscana	2400	24	9	1	1.00	0.38	4.17	11.11
Umbria	1117	50	9	2	4.48	0.81	4.00	22.22
Marche	1716	97	5	1	5.65	0.29	1.03	20.00
Lazio	3231	145	5	1	4.49	0.15	0.69	20.00
Abruzzi	2342	22	13	1	0.94	0.56	4.55	7.69
Molise	974	48	17	5	4.93	1.75	10.42	29.41
Campania	2847	50	2	0	1.76	0.07	0.00	0.00
Puglia	3081	65	12	3	2.11	0.39	4.62	25.00
Basilicata	1749	92	6	2	5.26	0.34	2.17	33.33
Calabria	4181	160	4	0	3.83	0.10	0.00	0.00
Sicilia	5432	270	4	0	4.97	0.07	0.00	0.00
Sardegna	2218	18	1	0	0.81	0.05	0.00	0.00
Trentino-Alto Adige-Trento	524	58	1	0	11.07	0.19	0.00	0.00
Trentino-Alto Adige-Bolzano	808	27	7	1	3.34	0.87	3.70	14.29
Total	44552	1547	129	24	3.47	0.29	1.55	18.60

The differences between SAU2 and SAT2 variables are higher for the total number of corrections, than for the number of corrected influential cases. The hit rate values (10.93% and 18.60% for SAU2 and SAT2, respectively) can be interpreted as a measure of 'efficiency' of the procedure adopted in identifying, among the influential cases, actual errors. The low values of the index may be due to the fact that in the estimated models it was considered as auxiliary variable the information (measured in the census) of three years before the survey, so with low predictive power, and that this information could have quality defects. The imputation rate for influential errors exceeds 10% only in Valle d'Aosta and Molise, probably because of the low number of total cases. In Trento, there is only one reported unit as a possible influential outlier.

As a further element of evaluation, Table 4 shows, respectively for variable SAU2 and SAT2, the relative differences between the estimates calculated on the following sets of data: before the E&I phase (Raw), after the step of interactive revision of the influential errors (Semi final), and at the end of the entire process of E&I (Final). The estimates are computed using sample weights adjusted for non-response. In other words, Table 4 shows the impact on estimates obtained using Raw data of the corrections made on the only influential data and of the total corrections, respectively. The impact is measured on total estimates and, therefore, there could be possible compensations for different sign of the data corrections. Out of the (National) Total estimate, the corrections involve a change of less than one percent of the estimate using raw data.

Table 4: *Relative differences between estimates computed using different data, SAU2 and SAT2 variables*

Region	(Semi final-raw)/Raw (%)	(Final-Raw)/Raw (%)
Total SAU2	0.03	0.98
Total SAT2	-0.04	0.09

To better evaluate the impact of the corrections on the data reported as influential, Table 5 shows the sum and the average of the absolute differences between the data flagged as influential and corrected and the relative raw data in comparison with the same measures of the absolute differences between the corrected data but not reported as influential and the raw data. The corrections on the influential data are larger than those on non-influential data: 1.23 times for SAU2 variable and 2.8 times for the SAT2 variable. This underlines the capacity of the selective editing procedure to identify the most influential errors on the final estimates.

Table 5: *Sum and average of the absolute differences, SAU2 and SAT2 variables*

Variabile	Somma	N	Media	A Media/B Media % C Media/D Media %
A. SAU2 |Influent corrected– Raw|	402342	20	20117	
B. SAU2 |Not Influent corrected – Raw|	20179311	1236	16326	123
C. SAT2 |Influent corrected – Raw|	2104088	24	87670	
D. SAT2 |Not Influent corrected – Raw|	47493823	1523	31184	281

6. Conclusions

The aim of the selective editing approach implemented for the 2013 SAH survey was the identification of the potentially influential errors affecting the estimates of a subset of key survey variables to be subjected to a clerical review. For the 2013 SAH, the set of variables considered for

the selective editing procedure includes total firm surfaces and livestock, and the auxiliary information is the same set of variables observed at the 2010 6th Census of Agriculture.

The obtained results differ when dealing with surfaces and livestock. Indeed, the variables relative to the total surface of a firm are more stable over the time and generally grant a relevant number of observations. This allowed to apply the selective editing procedure to quite detailed domains of estimates given by the combination of the variables (Region and OTE). The number of influential units to be subject to clerical checks results to be ‘adequate according to the available resources. As far as the livestock is concerned, observed figures changes easily over time. Consequently, the effectiveness of the 2010 6th Census of Agriculture as a source of auxiliary information resulted lower for livestock variables than for surfaces variables. This has probably led to identify a number of 'influential' cases higher than expected and/or desired for the purpose of containment of the interactive control, despite a careful definition of diversified thresholds for different animal types. Also the information on milk production, naturally depending from livestock numbers, has been affected by this instability. Moreover, the concentration of influential cases in some regions for some variables led to hypothesize the presence of systematic factors on the results. Further experiments are therefore needed in order to check on the one hand the effectiveness of the models and of the covariates used for this type of variables, on the other hand the existence of additional and more updated sources of auxiliary information. The assessment of the effects of selective editing procedure was focused on SAU2 and SAT2 variables. The share of cases reported as influential and actually corresponding to errors proved to be quite low. Nevertheless, the average number of identified errors is significantly higher for the cases reported to be influential than for the overall average of correct observations.

Overall, the selective editing process has produced satisfactory results, but its efficacy was lower than expected probably because of both the nature of the data and the absence of auxiliary information with a suitable quality. In subsequent editions of the survey, the use of alternative auxiliary information (other administrative sources, data from the current survey in a longitudinal perspective) will be evaluated. As far as the selective editing procedure is concerned, both a multivariate strategy for modeling data and the use of different thresholds of the estimates of accuracy on individual domains will be analysed. Finally, the results of the present work, will be used to plan the future survey edition with respect to the various phases of the entire process (for example, data recording) to identify and prevent any systematic error causes on the main variables.

REFERENCES

Bruni R., Reale A., Torelli R.. (2002) DIESIS: a New Software System for Editing and Imputation. *Proceedings SIS2002*, Milan.

Di Zio M., Guarnera U. (2013) Contamination Model for Selective Editing. *Journal of Official Statistics*, Vol. 26, n. 4, pp . 539-556

Di Zio M., Guarnera U., Luzi O. (2008) Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. *UN/ECE Work Session on Statistical Data Editing*. Vienna 21-23 Aprile. http://www.unece.org/stats/documents/2008.04.sde.htm.

D’Orazio M. 2013. *Il campione per l’indagine sulla struttura delle aziende agricole*. Doc. interno 2013.

Guarnera, U. e M.T. Buglielli. (2013) SeleMix: an R Package for Selective Editing. http://cran.r-project.org/web/packages/SeleMix/vignettes/SeleMix-vignette.pdf

ISTAT. 2001. *Struttura e Produzioni delle Aziende Agricole – Anno 1999 - Italia*. Coll. Informazioni ISTAT.

Latouche M., Berthelot J.M. (1992) Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8, n.3, 389-400.

Lawrence D., McKenzie R. (2000) The General Application of Significance Editing. *Journal of Official Statistics*, 16, n. 3, 243-253.