WP. 22
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (iv): Macro-editing

USING SELECTIVE EDITING COMBINED WITH AN AUTOMATIC SYSTEM IN THE FSS OF SPAIN

Invited paper

Prepared by D. Lorca (mdlorca@ine.es), National Statistical Institute, Spain

I. INTRODUCTION

1. Traditionally, the editing system of the agricultural and enterprise surveys, carried out at the Spanish National Statistics Institute (NSI) consists of a micro-editing system using tailored-made programs for each of the surveys. The subject matter experts specify the edits and the processing department makes tailored-made programs to detect edit failures. In this traditional procedure, these edit failures are manually reviewed. This process spends much time and resources of the whole survey.

2. In this paper, we describe a new integrated editing and imputation process that combines the selective editing and the automatic system (Banff) applied at the Spanish Farm Structure Survey (FSS).

3. This integrated process consists of three phases: 1) Initial editing prior to selective editing, 2) Selective editing procedure, 3) Automatic system process. The initial editing consists of the consistence controls that are established in the data collection phase carried out by interviewers. For selective editing, score functions are built to determine and prioritize the survey suspect units to be reviewed manually due to their significant weight on the final estimates. The automatic system process is carried out using the generalized system Banff, developed by Statistics Canada.

4. The different types of data that the FSS collects, such as the utilised agricultural land, cultivated land by kind of crop, types of livestock, the structure and the amount of farm employment, machinery and equipment, contribute to the complexity of the integrated editing process strategy.

5. This paper is organized as follows. Section II describes the main characteristics of the FSS and the initial editing prior to selective editing that is carried out in the data collection

phase. In section III, it presents the selective editing process applied to several key variables. A macroediting approach combined with the selective editing is also described. In section IV some results are presented. In section V further research is discussed. Finally, some remarks are given.

II. THE FSS AND THE INITIAL EDITING PRIOR TO SELECTIVE EDITING

6. The Spanish NSI, following EU Regulations, carries out, every ten years, the Agrarian Census and every two years the FSS. The main goal of the FSS is to evaluate the Spanish agricultural situation and follow up the structural development of farms as well as obtaining comparable results among all the European Union Member States.

7. The FSS sample consists of a farm panel drawn from the last Agrarian Census. The sample design is a single stage design with stratification of the farms according to geographical area (region), type of farming (TF) and size.

8. The type of farming (TF) classifies the farms according to the proportion of gross margin of each activity regarding the farm's total gross margin. Activity is understood as each type of crop or type of livestock worked in the region. The farm size is defined taking into account 4 key variables: Used Agricultural Land (UAL), Cultivated Land (CL), Animal Units (AU) and Annual Labour Units (ALU).

9. FSS estimates the total of the interest variables using Horvitz-Thompson estimators as follows:

$$\hat{X}_{hj} = \sum_{i=1}^{n_h} F_h \, X_{hji} \qquad [1]$$

where $\hat{X}_{hj}$ is the total estimate of the jth variable in stratum h, F_h is the sample weight for the stratum h, n_h is the sample size in stratum h and X_{hji} denotes the jth variable value for the sampled unit i in stratum h.

10. Data Collection and the initial editing are carried out by interviewers in the NSI's provincial offices. In this phase, all fatal errors are corrected. Most of these fatal errors come from balance edits.

III. SELECTIVE EDITING PROCEDURE

11. The goal of the selective editing procedure is to select the survey units with suspicious values that may have a significant effect on survey estimates. The selected units are manually reviewed. This technique limits the number of re-contacts of the suspicious units.

12. The selection of survey units is carried out using a score function. The score function allocates every unit response, which is possibly in error, a number (score), which measures the relative importance of that error on the survey estimates.

13. The key variables chosen to the selective editing procedure are: Used Agricultural Land (UAL), Cultivated Land (CL), Woody Crops (WCs), Olive Grove (OG), Vineyard (VY), Animal

Units (AU) and Annual Labour Units (ALU). These variables are of very different nature. It implies different treatments in the selective editing procedure.

14. The selective editing procedure is applied to regions, in each stratum, formed by the type of farming (TF), with the condition that the total estimate of the analyzed variable in the stratum is larger than 3% out of total estimate of the analyzed variable in the region.

A. Selective editing: crop and employment variables

15. The relative stability over time of selected crop variables implies that anomalous variations, from the previous year to the current one, can be a sign of data errors. For this reason, we determine the units with anomalous and significant variations of the selected crop variables. This process is applied to the permanent units, i.e. units belonging to the sample at time t and t-1.

16. The selective editing procedure is built up in the following steps:
i) In each stratum, we obtain the units with anomalous variation with respect to the previous period of the analyzed variables, using the Hidiroglou-Berthelot (1986) method of outlier detection.
ii) The units for manual editing are selected among the outliers identified previously having a significant weight on the population total estimates using a score function.

17. Within each stratum h, we determine the units with anomalous variations with respect to the previous period for the analyzed variables, using the SAS procedure (PROC OUTLIER) of the generalized system Banff. This SAS procedure applies the Hidiroglou-Berthelot method of outlier detection to ratio: $r_{hji} = \frac{X_{hji}^{t}}{X_{hji}^{t-1}}$ where X_{hji}^{t} and X_{hji}^{t-1} denote the variable value j for the ith sampled unit at time t and t-1 respectively. This method finds outliers on both tails of the distribution (positive and negative changes).

18. The "PROC OUTLIER" of Banff needs two parameters specified by the user: (a) a parameter that controls the importance associated with the magnitude of the data and (b) a parameter that determines the length of the acceptance interval. The former is set to 1. Thus, the larger the size of the weighted unit, the smaller the percent change it allows from one period to the next. According our simulation studies the latter one is set to 5.

19. In the next step, we select the units among the outliers obtained previously for manual editing. For each unit i and variable j, we calculate the following scaled local score function (Latouche and Berthelot 1992):

$$\Delta_{hji} = \frac{F_{h}^{t-1} \left| X_{hji}^{t} - X_{hji}^{t-1} \right|}{\hat{X}_{hj}^{t-1}} \qquad [3]$$

where

$$\hat{X}_{hj}^{t-1} = \sum_{i:1}^{n_h^{t-1}} \hat{F}_h^{t-1} X_{hji}^{t-1} \qquad \hat{F}_h^{t-1} = \frac{\hat{N}_h^{t-1}}{n_h^{t-1}}$$

where $\hat{N}_h^{t-1}$ is the population size estimate in stratum h at time t-1.

20. Within each stratum, the Δ_{hji} values are sorted in descending order. We determine a threshold value, Δ_{hj}^c, that depends on the Δ_{hji} empirical distribution and the number of recontacts that it aims to achieve. Then, outliers from the previous step with $\Delta_{hji} > \Delta_{hj}^c$ are selected for manual editing.

21 The threshold value Δ_{hj}^c is determined through the following formula (Lawrence and Mckenzie 2000):

$$a_{hj} = \sqrt{\frac{3k}{n_h}} SE(\hat{X}_{hj}) \qquad [4]$$

where a_{hj} is the threshold value of the jth variable in stratum h, $SE(\hat{X}_{hj})$ is the standard error of the jth variable in stratum h, n_h is the sample size in stratum h and k is a value such as :

$$E(bias^2(\hat{X}_{hj})) \leq kVar(\hat{X}_{hj})$$

where $bias^2(\hat{X}_{hj})$ denotes the bias in the survey estimate of not editing a set of survey units. Using the above formula ensures that the bias due to not editing some of the survey units is less than k% of the variance of the estimate. The value of k is set to 10%.

22. For the employment variable, the selective editing procedure is similar to the previous one. But in this case, it takes into account the variation rate of the employment number in agriculture obtained through the Force labour Survey (FLS).

23. The selective editing procedure is applied to the ALU variable. One ALU is equivalent to the work carried out by one person on a full-time basis over one year.

24. The difference with the previous procedure is the use of auxiliary information to estimate the expected amended value. The local score function is now the following:

$$\Delta_{hji} = \frac{F_h^{t-1} \left| X_{hji}^t - trX_{hji}^{t-1} \right|}{tr\hat{X}_{hj}^{t-1}} \qquad [5]$$

where tr is the bi-annual variation rate of the agriculture employment number in the region r and $\hat{X}_{hj}^t$ is the total estimate of the ALU variable in stratum h at time t. The PROC OUTLIER is also applied using this auxiliary information.

B. Selective editing: livestock variables

25 In the case of livestock variables, an anomalous variation with respect to the previous period for the analyzed variable does not always imply data errors. The FSS collects the existing livestock in the farm on the day of the interview. Thus, a farm can have a strong livestock variation depending on the interview date. For this reason, the selective editing procedure for livestock is different to the rest of variables.

26. The selective editing procedure is built up in the following steps:
i) Units that fail some of the edits, which are specified in the traditional approach, are selected as suspicious units.
ii) For each suspicious unit or edit failure, an estimate of the expected amended response of AU variable is calculated.
iii) We determine, among the suspicious units detected at the previous step, those units with a significant weight on the total estimate of the AU variable.

27. The edits specified in the traditional approach are of type: $y_{hji} < c_{hj}$ where y_{hji} is the jth variable for the unit i in stratum h and c_{hj} is a constant determined by the historical empirical distributions. We obtain the units that fail at least one edit.

28. The importance of a suspicious unit is measured through the total estimate of the AU variable. Livestock data are expressed in AUs which are obtained by applying a coefficient to each species and type in order to group different species in one common unit.

29. An estimate of the expected amended response of AU variable is calculated using the upper boundaries of the acceptable region defined for the edits. These maximum values are expressed as AUs. Then, $e_{hji}=x_{hji}-x'_{hji}$ determines the magnitude of failure for the unit i where x_{hji} denotes the AU value reported by respondent i within stratum h and x'_{hi} is the expected amended response for the respondent i for AU variable.

30. For each suspicious unit, we calculate a score using the following scaled local score function:

$$\Delta_{hji} = \frac{F_h^{t-1} e_{hji}}{\hat{X}_{hj}^{t-1}} \qquad [6]$$

where $\hat{X}_{hj}^{t-1}$ denotes the total estimate of the AU variable in stratum h at time t-1.

31 The threshold value is calculated using the formula (4) as in the previous cases. Within each stratum, the Δ_{hji} values are sorted in descending order. Then, the suspicious units with $\Delta_{hji} > \Delta_{hj}^{c}$ are selected for manual editing

C. Global score function

32. The local score functions, calculated for each jth variable, are combined to give a global score function for unit, which is used for prioritisation of the entire survey unit. A global score function is defined as the maximum value of the local scores:

$$G\Delta_{hi}=\max_j(\Delta_{hji})$$

where j index indicates the selected key variables.

33. Then, a unit is selected to manual revision if at least one of their local score is over their local threshold value.

D. Macroediting and selective editing approach

34. A macroediting approach is combined with the above selective procedure. In first place, a selection of the strata with the largest variation with respect to the previous period of the analyzed variables is carried out. After, the steps of selective editing procedure are applied only to the farms of the selected strata.

35. For each stratum h and variable j, we calculate the following expression:

$$\Delta_{hj}=\frac{\left|\hat{X}_{hj}^{t}-\hat{X}_{hj}^{t-1}\right|}{\sum_{h:i}^{l}\left|\hat{X}_{hj}^{t}-\hat{X}_{hj}^{t-1}\right|} \qquad h:1\ldots l \qquad [2]$$

where l is the stratum number, $\hat{X}_{hj}^{t}$ y $\hat{X}_{hj}^{t-1}$ are the population total estimates of jth variable in stratum h at time t y t-1 respectively.

36. In each region, the Δ_{hj} values are sorted in descending order. We determine a threshold value, Δ_j^* and strata with $\Delta_{hj} > \Delta_j^*$ are selected to find in them the units with anomalous variations. This threshold value is set to 3%.

IV. RESULTS

37. In our study we use the data from 3690 farms. We compare the results obtained for the following editing procedures:(A) Traditional microediting approach, (B) Selective editing procedure and (C) Macroediting and selective editing approach.

38. The performance of these procedures is analised by the rate of number of edited units and corrected units. Table 1 presents these rates. In Table 2 we calculate the rate of change of the total estimate for the CL variable obtained from the B and C procedures with respect to A procedure. In Table 3 we show the 95% confident interval calculated for the total estimate of CL variable obtained from A procedure and the total estimates of CL variable obtained from B and C procedures.

Table 1

Procedures	A	B	C
Rate of edited units(%)	21.5	9.0	4.8
Rate of corrected units(%)	3.9	7.2	9.1

Table 2: Change rates of total estimate for the CL variable (%)

Change rate of (B) over (A)	Change rate of (C) over (A)
0.8	1.1

Table 3

95% Confident interval	Total estimate from B procedure	Total estimate from C procedure
(72657,2; 86031,5)	78770,72	78471,56

It can be seen that significant reductions of edited units could be attained using both selective editing and macroediting/selective approach. Moreover, the hit rate is improved. The differences obtained on the final estimates for CL variable from the procedures considered are not significant.

V. FURTHER RESEARCH

39. Banff will be applied to the rest of units that have not been edited in the selective editing procedure.

40. Due to such a different nature of the data collected by FSS, we will build up three edit groups: crop, livestock and employment. For each of these edit groups, we will run the procedure of the error localization (PROC ERRORLOC). This procedure identifies the fields that must be changed in each record in error in order to minimize the number of fields requiring imputation, following the Fellegi-Holt methodology. Different methods of imputation will be tested.

VI. FINAL REMARKS

41. Integrating the PROC OUTLIER of Banff to detect suspicious values and a score function to select units for manual editing has been useful in the Spanish FSS. Reduction in cost and processing time would be attained using this approach. It also would help to reduce response burden from carrying out less number of recontacts. We need more research in order to use Banff to edit units with a low risk of seriously affecting survey estimates.

References

1) Belcher R. (2003): "Application of Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Survey", in: Proceeding of Survey Methods, SSC Annual Meeting, June 2003.
2) Farwell F. (2004): "The General Application of Significance Editing to Economic Collections", Research paper, Australian Bureau of Statistics.

3) Granquist L. (1992): “A Review of Methods for Rationalizing the Editing of Survey Data”, in: United Nations Statistical Commission and Economic Commission for Europe, Statistical Data Editing Methods and Techniques, Vol. 1.
4) Hedlin, D. (2002): “Score Functions to Reduce Business Survey Editing at the ONS” in: UN/ECE Work Session on Statistical Data Editing, Helsinki, Finland, 27-29 May 2002.
5) Hidiroglou, M.A. and Berthelot J.M. (1986): “Statistical Editing and Imputation for Periodic Business Surveys”, Survey Methodology, 12:73-83.
6) Latouche M. and Berthelot J.M. (1992):”Use of Score Function to Prioritize and Limit Recontacts in Editing Business Surveys”, Journal of Official Statistics, Vol. 8, No 3, pp. 389-400.
7) Lawrence D. and McKenzie R. (2000): “The General Application of Significance Editing”, Journal of Official Statistics, Vol. 16, No. 3, pp. 243-253.
8) Luzi O. And Pallara A. (1999): “Combining Macroediting and Selective Editing to Detect Influential Observations in Cross-Sectional Survey Data”, in: UN/ECE Work Session on Statistical Data Editing, Rome, Italy, 2-4 June 1999.
9) Statistics Canada (2005): “Functional Description of the Banff System for Edit and Imputation”, Generalized System Methods Section, Business Survey Methods Division, June 2005.
