

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (v): Editing based on results (post-editing)

**TESTING THE USE OF ADMINISTRATIVE DATA TO EDIT THE 2009 AGRICULTURE
CENSUS**

Supporting Paper

Prepared by D. Lorca, National Statistical Institute, Spain

Abstract

Every ten years the Spanish National Statistical Institute conducts an Agriculture Census. We collect several variables from the farmers by personal interview. Different editing approaches are applied to the complex set of collected data. For the 2009 Agriculture Census we are testing the use of administrative data to edit census data focusing on selective editing. A score function is calculated to determine and prioritize the suspect units to be manually reviewed due to their discrepancy with the administrative data and their significant weight at the final estimates. The performance of this procedure is tested using data from the previous census and the Farm Structure Survey.

Key words: Agriculture Census, administrative data, selective editing

I. INTRODUCTION

1. The Spanish National Statistical Institute (NSI) following EU Regulations carries out every ten years an Agriculture Census. We collect several variables from the farmers by personal interview. Different editing approaches are applied to the complex set of collected data.

2. In an agriculture census the editing procedures suppose a large amount of time and resources dedicated to them. The large number of questionnaires collected by many interviewers, during a short time, about very different kind of holdings can have quite errors to amend in editing. To understand the quantitative aspect of editing the 1999 census collected more than 1,800,000 farms and involved about 6,000 interviewers during three months.

3. The editing is carried out in different phases of the Census. In the data collection phase simple checks are applied using build-in edits in a Computer Assistance Personal Interview (CAPI) system. Next, selective editing is applied to determine the units that will be manually reviewed. The other ones will be automatically editing using the DIA (version extended to quantitative data) system. Finally, macroediting procedure is carried out before publishing the results.

4. For the 2009 Agriculture Census we are testing the use of administrative data to edit census data using a selective editing procedure. A score function is calculated to determine and prioritise the suspect units to be manually reviewed due their discrepancy with the administrative data and their significant weight in the final estimates. The performance of this procedure is tested using data from the previous census and the Farm Structure Survey (FSS) carried out by NSI every two years.

5. This paper is organized as follows. Section II describes the selective editing methodology applied. Section III presents the study case carried out to test the performance of this approach and finally some conclusions are given.

II. SELECTIVE EDITING METHODOLOGY

6. The selective editing procedure calculates a measurement (global score) for each reported unit to distinguish those ones that exceeding a certain threshold (cut-off score) will be manually reviewed. Firstly a local score function is applied for each interest variable. Then, local scores are combined to obtain a global score at the level of reported unit.

7. The local score measures the magnitude of change between the reported and administrative value. Using simple expansion estimators as follows

$$\hat{X} = \sum_{i=1}^n w_i X_i$$

where \hat{X} is the total estimate of the X variable, w_i is the sample weight associated to the unit i, n is the number of the sample units and X_i is the X variable value for the unit i, the local score associated to the variable X and the unit i is given by

$$S_i = w_i |X_i - \hat{X}_i^a|$$

where X_i is the reported value and \hat{X}_i^a is the administrative value. In our pilot study we use data from probabilistic sample. When we will carry out the census $w_i=1$.

8. The local scores are scaled to a common basis to allow comparison across units, domain or items and combined them to obtain a global score. We scale them dividing by the total estimate as follows

$$LS_i = \frac{S_i}{\hat{X}_a}$$

where \hat{X}_a is the total estimate of variable X calculated from administrative data. As we have the administrative data before carrying out the census we use them in order to calculate LS_i since the first holding is collected. The global score is defined as the maximum absolute value of the local scores. Thus, we ensure that a significant discrepancy for any variable will be revised.

9. Once global scores are determined we choose the thresholds or cut-off scores using the simulation study approach by Lawrence and Mackenzie (2000). We calculate the absolute pseudo-bias of Latouche and Berthelot (1992) given by

$$PB_p = 100 \frac{|\hat{X}_p - \hat{X}_a|}{\hat{X}_a}$$

where \hat{X}_p is the total estimate of variable X obtained by replacing all reported values with a global score larger than the pre-determined threshold (the percentile p) by their administrative values and leaving reported values in place for the others and \hat{X}_a is the corresponding total calculated from administrative data.

10. We calculate the absolute pseudo-bias for several values of p and choose thresholds trying to balance between a low pseudo-bias for each interest variable and a low recontact rates calculated as the number of units with a global score larger than the pre-determined threshold (percentile p) divided by the total number of units.

III. STUDY CASE

11. We use data from 1999 census, 2005 FSS and administrative registers. The FSS is a sample survey carried out every two years of approximately 50,000 farms drawn from the last census. Its main goal is to evaluate the Spanish agricultural trend between the censuses and follow up the structural development of farms, as well as obtained comparable results among the European Union Member States. FSS estimates the interest variables using simple expansion estimators.

12. The selected variable to test this approach is the area of olive grove planted. The number of common units between the 1999 census and administrative register is 253,038 units and between 2005 FSS and the administrative register is the 5,804 units.

13. Tables 1 and 2 show absolute pseudo-biases for the variable olive grove planted broken down by main geographical areas of Spain using the global score distribution percentiles for $p=95, 90, 85, 80, 75$ and 70 for census and FSS data respectively. The PB_{100} values correspond to the case where we only compute reported values. In tables 3 and 4 we present the total number of units and the number of recontacts associated to the percentiles for census and FSS data respectively. In both cases we observe a pattern quite similar. With a recontact rate of 5% the reduction in absolute magnitude of the pseudo-bias is much greater than in the rest of rates. In most regions, the increase of recontacts (more than 5%) does not compensate the reduction of pseudo-bias. Then, at least for most region we would use the global score distribution percentile $p=95$ threshold.

Table 1: Absolute pseudo-bias for the variable olive grove planted using data from 1999 Census

Region	PB_{100}	PB_{95}	PB_{90}	PB_{85}	PB_{80}	PB_{75}	PB_{70}
Andalucía	21.2	9.6	6.9	5.2	4.0	3.1	2.5
Aragón	33.9	16.0	10.9	7.9	6.0	4.4	3.2
C-León	14.0	7.8	5.5	3.7	2.4	0.7	0.4
C-Mancha	20.5	9.9	7.1	5.3	4.1	3.2	2.4
Cataluña	15.5	5.5	3.3	2.3	1.6	1.1	0.8
Valencia	13.1	3.7	2.3	1.3	0.9	0.6	0.5
Extremadura	12.2	5.6	4.5	3.7	3.1	2.6	2.1
Madrid	8.6	6.3	4.9	3.9	3.1	2.7	2.2
Murcia	23.5	7.0	4.0	2.5	1.1	0.6	0.1
Navarra	9.7	2.9	1.5	0.5	1.2	1.1	1.1
La Rioja	15.2	3.2	2.7	1.1	0.8	0.7	0.8

Table 2: Absolute pseudo-bias for the variable olive grove planted using data from 2005 FSS

Region	PB_{100}	PB_{95}	PB_{90}	PB_{85}	PB_{80}	PB_{75}	PB_{70}
Andalucía	19.6	9.8	6.7	3.8	2.6	1.6	1.2
Aragón	35.0	13.6	8.8	6.5	4.4	4.1	3.8
C-León	23.1	10.1	7.7	6.5	4.9	5.0	5.0
C-Mancha	16.6	8.2	5.4	4.7	3.3	2.6	1.9
Cataluña	14.0	4.4	1.2	1.3	0.4	0.7	1.0
Valencia	5.3	3.8	2.0	1.1	0.7	0.1	0.4
Extremadura	14.3	9.7	7.6	4.1	3.0	1.8	1.4
Madrid	11.0	10.3	7.0	5.0	3.7	3.7	3.7
Murcia	34.8	9.9	5.0	4.6	2.2	1.1	0.7
Navarra	35.3	1.8	1.6	3.5	3.4	2.6	1.0
La Rioja	30.8	2.9	5.6	1.1	3.2	0.6	1.8

Table 3: Recontacts associated to the percentiles p. 1999 Census

Region	Total sampled units	Recontacts					
		P=95	P=90	P=85	P=80	P=75	P=70
Andalucia	119511	5976	11953	17927	23918	29880	35939
Aragon	10974	549	1100	1647	2197	2744	3300
C-Leon	1794	92	180	270	359	453	544
C-Mancha	48716	2437	4886	7331	9765	12190	14651
Cataluña	15681	785	1572	2356	3140	3924	4705
Valencia	24449	1223	2445	3674	4906	6116	7343
Extremadura	20997	1050	2103	3155	4205	5259	6307
Madrid	2393	120	241	359	481	599	718
Murcia	4111	206	414	619	823	1028	1237
Navarra	2604	131	261	392	522	660	782
La Rioja	1579	79	158	240	316	396	475

Table 4: Recontacts associated to the percentiles p. 2005 FSS

Region	Total sampled units	Recontacts					
		P=95	P=90	P=85	P=80	P=75	P=70
Andalucia	1879	94	188	282	377	470	564
Aragon	361	19	37	55	73	91	109
C-Leon	242	13	25	37	49	61	73
C-Mancha	946	48	95	142	190	237	284
Cataluña	606	31	61	91	122	152	182
Valencia	519	26	52	78	104	130	156
Extremadura	397	20	40	60	80	100	120
Madrid	204	11	21	31	41	51	62
Murcia	272	14	28	41	55	68	82
Navarra	201	11	21	31	41	51	61
La Rioja	132	7	14	20	27	33	40

IV. CONCLUSIONS

14. The pilot study shows that the selective editing approach using administrative data to edit census data could help us to prioritise follow-up actions for those units with significant discrepancies with administrative data.

References

- 1) Belcher, R.(2003). Application of Hidiroglou-Berthelot method of outlier detection for periodic business survey. Proceeding of survey methods, SSC Annual Meeting, June 2003.
- 2) Granquist, L.(1992).A review of methods for rationalizing the editing of survey data. United Nations Statistical Commission and Economic Commission for Europe, Statistical Data editing Methods and Techniques, Vol I.
- 3) Hidiroglou, M.A. and Berthelot, J.M.(1986). Statistical editing and imputation for periodic business surveys. Survey Methodology, 12, pp.73-83.
- 4) Latouche, M. and Berthelot, J.M. (1992). Use of score function to prioritize and limit recontacts in editing business surveys. Journal of Official Statistics, 8, pp. 389-400.
- 5) Lawrence, D. and MaKenzie, R. (2000). The general application of significance editing. Journal of Official Statistics, 16, pp. 243-255.
- 6) Lawrence, D. and McDavitt, C. (1994). Significance editing in the australian survey of average weakly earnings. Journal of Official Statistics, 4, pp.437-447.