STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 7
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic I: Management and evaluation of editing and imputation procedures

AN ASSESSMENT OF MACRO EDITING METHODS

Submitted by the Central Statistical Office of Ireland[1]

Contributed paper

***Abstract**: The paper reports on the application to Ireland's Annual Services Inquiry of various macro editing techniques: the Hidiroglou-Berthelot method and the Aggregate Method.*

I. INTRODUCTION – THE ANNUAL SERVICES INQUIRY

1. The Annual Services Inquiry measures the principal trading aggregates of the retail, wholesale, real estate, renting, business and other selected services sectors. The survey sample is selected from the Central Statistical Office's (CSO's) Business Register, with enterprises as the target respondents (one return is sought in respect of each enterprise, covering all of its branches). Stratification (4 strata) is based on the "Total Persons Engaged" variable of the Business Register. The selected sample will have a proportion of enterprises "rolled over" (approximately 25%) from the previous year, particularly for enterprises with large numbers of "Total Persons Engaged".

2. The sample for 1996 was 10,184 enterprises (from a population for all sectors of 65,417) and a field force of 25 field officers helped to obtain an overall response rate of 93%.

3. The present editing procedure is the application of a range of checks to the data and the flagging of records that fail these checks for follow-up by 10 people. Responses are categorised by sector (21 sectors) and by size (the 4 strata), giving 84 groups. Annual Services Unit (A.S.) believes that vigorous editing is needed to ensure accuracy at this level of detail of results. Edits are coded in SAS and applied to SAS datasets.

4. To assess different editing techniques, attention is focused on enterprises involved in distribution (i.e. retail and wholesale sectors). About 5,500 enterprises were sampled in 1996 from a population of 36,230 for these sectors and about 4,400 responses were received. Of the 4,400 responses, approximately 1,200 enterprises had been "rolled-over" from 1995 (i.e. data for 1996 and 1995 were available).

[1] Prepared by Jennifer Banim.

II. APPLICATION OF THE HIDIROGLOU-BERTHELOT METHOD

Method Outline

5. The Hidiroglou-Berthelot aims to identify major outliers in data, using the data itself to generate upper and lower bounds. Briefly, for a variable measured in two consecutive periods, $X_i(t)$ and $X_i(t+1)$ the relative change for each observation is:

$$R_i = X_i(t+1)/X_i(t)$$

and transforming for working with increases and decreases gives:

$$S_i = (R_i - R_{median})/R_i \qquad 0 < R_i < R_{median}$$

$$= (R_i - R_{median})/R_{median} \qquad R_i >= R_{median}$$

6. Half of the Si values will be less than zero. Also, as emphasis on the magnitude of the variable may be important, a second transformation is performed giving:

$$E_i = S_i * [MAX(X_i(t), X_i(t+1))]^{**}U$$

7. The method proposes that U be a value between 0 and 1; if U = 0, then no emphasis is placed on the magnitude of the variable.

8. Any E_i values that are too small or too big are considered outliers or errors by the method, as their trend is different from the overall trend of other observations. Upper and lower limits for the E_i values are constructed using:

$$D_{q1} = MAX[E_{median} - E_{q1}, |A*E_{median}|]$$
$$D_{q3} = MAX[E_{q3} - E_{median}, |A*E_{median}|]$$

as

$$Upper\ limit = E_{median} + C*D_{q3}$$
$$Lower\ limit = E_{median} - C*D_{q1}$$

where A is an arbitrary value suggested by Hidiroglou-Berthelot to be 0.05. The $A*E_{median}$ term protects against the detection of too many outliers in situations where the E_i values are tightly clustered around the median. C is a constant that controls the width of the interval.

Application

9. The Hidiroglou-Berthelot method was applied to those enterprises responding to the 1996 survey, “rolled-over” from 1995 and concentrated on the important variable, turnover. Ratios of 1996 unedited turnover to 1995 edited turnover for all “rolled-over” enterprises are shown graphically in Chart 1.

10. Five hundred subsets from the 1,200 “rolled-over” enterprises were selected randomly, each containing between 450 and 500 enterprises and the Hidiroglou-Berthelot method applied. Where the Hidiroglou-Berthelot method flagged an enterprise’s turnover as an error, the mean turnover for an enterprise of that size, from that sector, was used to impute a corrected turnover.

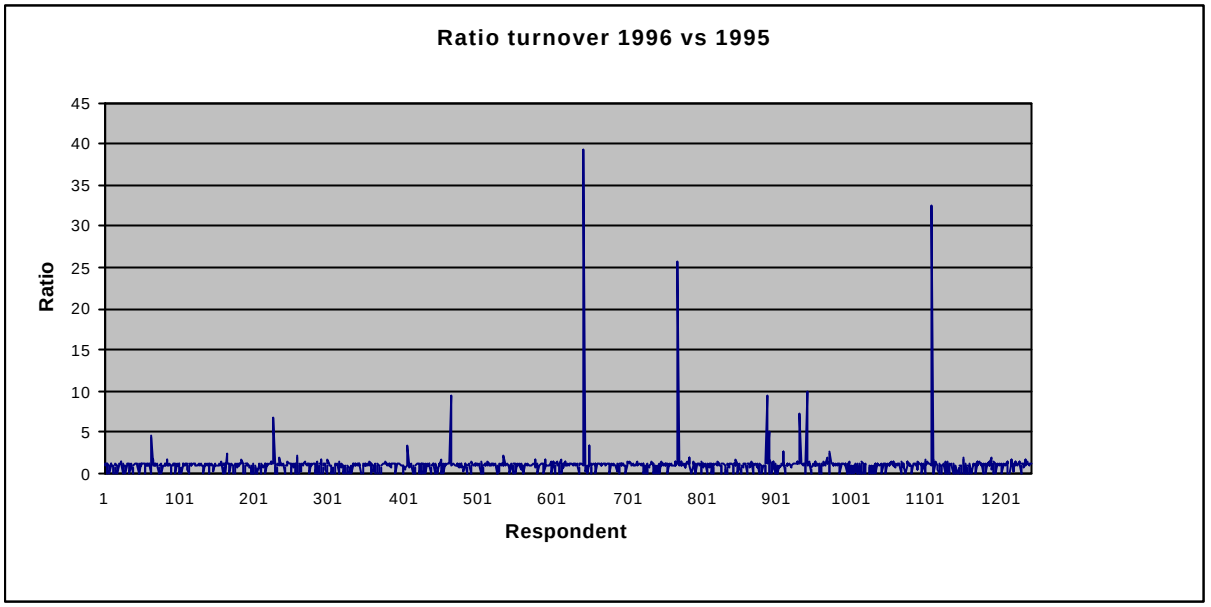

Chart 1

11. The findings from the application of the method to the 500 subsets are shown in Table 1 for two combinations of the Hidiroglou-Berthelot parameters. The averages, over the subsets, of errors identified and corrections made are shown. For each subset, and to facilitate comparisons, the corrections were calculated as the sum of absolute values of the corrections to the turnover totals of the 84 groups.

H-B Method Applied to Ratio Check of Turnover – averages of 500 replications								
H-B parameters	No. errors - A.S.	Correction – A.S. in Euro millions	Total turnover in Euro millions	No. errors - H-B	Correction – H-B. in Euro millions	Total turnover, H-B corrected, in Euro millions	No. errors identified by A.S. and H-B	Correction – for A.S. and H-B errors in Euro millions
U =0.2 A = 0.05 C = 20	*73*	*222*	*2,063*	39	144	1,985	29	176
U =0.2 A = 0.05 C = 10	*73*	*222*	*2,063*	50	91	1,932	32	195

A.S. = Annual Services Unit

Table 1

12. The level of overlap of identified errors by the two approaches was encouraging, particularly as the more significant errors identified by Annual Services Unit were generally identified by the method. For instance, in every subset selected, the method always identified at least 8 of the “Top 10” errors in each size category, as ranked by Annual Services Unit, for various combinations of the parameters U, A and C.

13. Because of the robustness of the technique, variations in findings for different combinations of the Hidiroglou-Berthelot parameters, U, A and C were not marked. The parameter A is set at 0.05 as suggested, and U is set to 0.2 to give (conservatively) less importance to the magnitude of turnover.

14. Mean imputation was chosen as the imputation method for simplicity. (The standard deviation of the average correction to turnover at group level using the Hidiroglou-Berthelot method was about Euro 14 million or 10% of the average correction.) Developing the imputation procedure would be the next step for this project.

Summary

15. Overall, the method did reduce the number of enterprises with turnover values identified as errors and the larger corrections made by Annual Services Unit were flagged. The development of a satisfactory imputation system would make this approach very attractive for the Annual Services Unit.

III. APPLICATION OF THE AGGREGATE METHOD

Method Outline

16. The Aggregate Method identifies errors in a two-stage approach. Initially, edits are applied to aggregated or grouped data. Any group failing the checks is flagged. The checks are then applied to all observations in flagged groups and any observations that fail the checks are corrected.

Application

17. Two edits from the Annual Services Unit were used to test the Aggregate Method:

- check the ratio of turnover for 1996 to turnover for 1995 for “rolled-over” companies (similar check to the Hidiroglou-Berthelot method above)
- check the level of closing stock verses opening stock in 1996.

18. Using the Aggregate method, the checks were run initially at the (aggregate) level on the 84 sector by size groups, and then on all enterprises in any groups that failed the checks. The correct values discovered by Annual Services Unit were used to correct the turnover for those enterprises failing the checks in flagged groups. If Annual Services Unit had not edited the enterprise, no change was made.

19. For the ratio check of 1996 turnover to 1995 turnover, the Aggregate method was applied to the “rolled-over” enterprises referred to in Section 1 above. Repeated subsets of approximately 500 enterprises were selected 100 times and edited at the group level and then at a micro level. The findings are shown in Table 2 for various acceptance limits of the check function.

Aggregate Method Applied to Ratio Check of Turnover – averages of 100 replications					
Acceptance limits for ratios:	No. errors identified by A.S. Unit	Total turnover, in Euro millions	No. errors identified by Agg. Method	Total turnover, Agg. Method corrected, in Euro millions	Value of errors not identified by Agg. Method in Euro millions
Upper: 1.5 Lower: 0.7	*13,356*	*2,131*	8,954	2,062	69
Upper: 1.3 Lower: 0.7	*13,356*	*2,131*	9,484	2,065	66
Upper: 1.2 Lower: 0.8	*13,356*	*2,131*	10,846	2,069	62

Table 2

20. From the results of the repeated applications, the Aggregate method, on average, missed one big error in every subset. A big error was defined as any correction made by Annual Services Unit greater than Euro 13 million.

21. Findings indicate that reductions in the level of editing without great impact on results are possible for the above check.

22. The second check function tested was that (from Annual Services Unit experience) opening stock and closing stock should not differ greatly in a year. This function was internal to 1996 responses, and was run on 100 randomly selected subsets of all replies to the 1996 Annual Services Inquiry from the retail and wholesale sectors. Each subset contained approximately 1,500 enterprises. Results are shown in Table 3 below.

Aggregate Method Applied to Check of Opening & Closing Stock – averages of 100 replications				
Acceptance limits for ratios:	No. errors identified by A.S. Unit	Total opening+closing stock for subset, in Euro millions	No. errors identified by Agg. Method	Total opening+closing stock, Agg. Method corrected, in Euro millions
Upper: 1.4 Lower: 0.7	*129*	*853*	59	829
Upper: 1.2 Lower: 0.8	*129*	*853*	146	833
Upper: 1.15 Lower: 0.85	*129*	*853*	235	835

Table 3

23. Annual Services Unit imposes upper and lower limits of 1.4 and 0.7 respectively on the ratio of closing stock verses opening stock. The variables seem to be over-edited in the sense that reducing the number of corrections made (under the Aggregate method) has no great impact on results.

Summary

24. Given the simplicity of the Aggregate method, application of this method by Annual Services Unit would not lead to major re-coding of their present editing system and given the clear benefits of the method should be attractive to the Unit. Upper and lower limits for the different variables would need to be determined with the Unit to meet their accuracy requirements.

IV. CONCLUSIONS

25. From the tests above, adoption of macro-editing methods would lead to savings in resources assigned to data editing by Annual Services Inquiry Unit. Opening and closing stock does appear to be over-edited by the Unit. Application of either the Hidiroglou-Berthelot method or the Aggregate method to the turnover variable would also lead to a reduction in resources assigned to editing.

26. One issue that arose when applying the methods was the high level of item non-response for turnover, particularly among smaller enterprises. Many of the corrections under the present editing system are due to this “missingness” in enterprises’ responses. Focus on this issue (for example, through the field force mentioned earlier) would reduce the number of corrections, even before the introduction of a macro-editing technique.