Working Paper 23

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (i): Selective and macro editing

Method for reviewing selective editing thresholds at ONS, RSI pilot study

Prepared by Sangeetha Gallagher, Ben Graham and Charlotte Gaughan, Office for National Statistics, UK

I. Introduction

1. The Retail Sales Inquiry (RSI) measures the value and volume of retail sales in Great Britain on a monthly basis. The industries included in the RSI are as defined by the UK Standard Industrial Classification 2007 (SIC07)[1]. There are two key variables available from RSI data, they are turnover and employment. The survey has a sample size of 5000 retail businesses taken from the Inter Departmental Business Register (IDBR) which acts as sampling frame for most business surveys at Office for National Statistics (ONS). Not all retailers are asked the question about employment. Every quarter (i.e every three months), a subsample of 2100 retailers is asked to provide information about employment.

2. RSI employs the selective editing method for error detection. Two selective editing scores are calculated separately for the two key variables, turnover and employment. Those businesses which have two scores are then given a single combined overall score, otherwise the score given is just for turnover. The thresholds used in RSI were set in 2010 using the traditionally edited data from previous years.

3. In order to maintain the quality of data and the estimates produced, it is necessary to regularly review the method used in error detection. In this case, the regular review of thresholds will ensure detection of any data issues arising from selective editing. The review of thresholds can be done by sub sampling the businesses which pass selective editing and applying traditional editing to the selected businesses. It was agreed that part of the savings in cost realised by applying selective editing to the surveys would be use to review the thresholds.

4. This paper discusses a sampling strategy developed and used by ONS to review the selective editing thresholds for RSI. This method will be used to review selective editing thresholds used in other business surveys at ONS.

II. Methodology

A. Sampling design

5. Different sample designs were tested such as simple random sampling, stratified sampling and probability proportional to size to identify the best way to select a sample of businesses which had passed the selective editing. The testing of various sample designs was done using past data which was edited using traditional editing. This made it easier to identify which sample design would give a sample of businesses which were predominantly businesses with scores close to the threshold.

[1] http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/standard-industrial-classification/index.html

6. Overall sample size (n) was agreed to be 600 businesses from those which passed selective editing. RSI data is split to 25 domains based on which class they fall into in SIC07. Each of these domains had separate thresholds. A stratified sample was taken using these domains as strata to ensure that each of the thresholds could be reviewed. Four sampling designs were tested:

(a) Stratified Simple Random Sample(SRS)
(b) Stratified probability proportional to size sample(PPS)
(c) Stratified sample with proportional allocation
(d) Stratified sample with percentage allocation

7. The target population was the businesses which passed the selective editing. Sample size for each domain (stratum) was calculated based on formula given by Lewis D (2014)[2]. The sample allocation to each stratum was done in such a way that no stratum had a sample size less than 20 unless the domain size is less than 20.

$$n_h = n \times \frac{\sqrt{n_{h,resp} - n_{h,fail}}}{\sum_h \sqrt{n_{h,resp} - n_{h,resp}}} \tag{1}$$

Where H =1,2,.................,h are the strata

n = overall sample size = 600

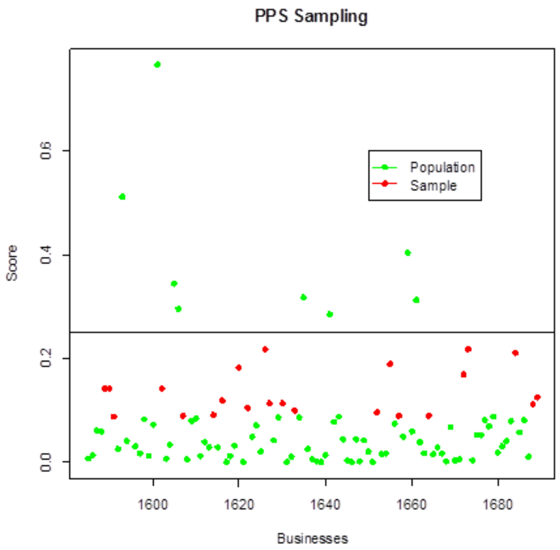
$n_{h,resp}$= number of respondents in stratum h

$n_{h,fail}$= number of respondents in stratum h which has failed selective editing

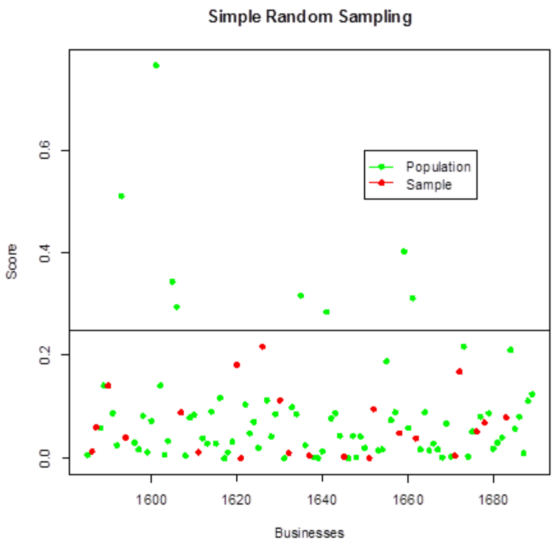
8. For PPS sampling, the sample was selected with scores for each businesses acting as size measures. The inclusion probability for a business (i) was calculated as;

$$\pi_i = n_h \times \frac{score_i}{\sum_h score_i} \tag{2}$$

The PPS method was better than SRS for selecting businesses which passed selective editing but had scores very close to the threshold. This can be seen from graphs 1 and 2. The horizontal line in the graphs represents the threshold value for the domain.



Graph 1: Sample design – PPS



Graph 2: Sample design – SRS

However, the inbuilt procedure in statistical programming language SAS, which was used to select the sample limited the scope of selecting a PPS sample without replacement from every strata. As a result other sampling designs were explored which can be used in the selection of a sample of businesses with scores close to the threshold.

[2] http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_1_ONS_Lewis.pdf

9. The domains were divided in to 4 secondary strata and a simple random sample was selected without replacement from these secondary strata for each domain/primary strata. The sample size for the secondary strata was determined using two methods; proportional allocation and a predetermined percentage of businesses were selected from each stratum (percentage allocation). Since most businesses which passed the selective editing had low scores, proportional allocation did not ensure a sample for each domain with more businesses having scores close to the threshold. As a result, the percentage allocation was chosen as the sample design.

10. The secondary strata were constructed using percentiles from distribution of scores of businesses in each domain, with stratum 1 being the one close to the threshold followed by strata 2, 3 and 4 respectively. Table 1 shows the population in each secondary stratum and the percentage of sample size for each domain selected from them.

Secondary Stratum	Scores of the Population	Percentage of n_h selected
1	75th percentile < Score <= 100th percentile	40%
2	50th percentile < Score <= 75th percentile	30%
3	25th percentile < Score <= 50th percentile	20%
4	Scores <= 25th percentile	10%

Table 1: Scores of the population in each secondary stratum and percentage selected

B. Calculation of bias

11. Once the sample design was finalised, a sample of businesses were selected from the past data available which was used to set the thresholds initially. In order to test whether the sampling method will give a reasonable evidence to change the threshold or not, some of the thresholds were increased and bias was calculated using the following formula:

$$\frac{\sum_{i \in sample} w_i s_i \left| y_{sel} - y_{trad} \right|}{\sum_{i \in above\ threshold} w_i y_{trad} + \sum_{i \in sample} w_i s_i y_{trad}} \qquad \textbf{3}$$

Where w_i is the weight from the survey, s_i is the design weight for the sample selected for the review, y_{sel} is the selectively edited value and y_{trad} is the traditionally edited value.

12. The sample bias was then compared to the full data by calculating a full sample bias. This was calculated as below:

$$\frac{\sum_{\text{full sample}} w_i \left| y_{i,sel} - y_{i,trad} \right|}{\sum_{\text{full sample}} w_i \, y_{i,trad}} \qquad \textbf{4}$$

13. Since not all businesses were asked the employment question, it was necessary to make some adjustments to the sample so that the sample size was big enough to reset the individual and joint thresholds. After calculating the bias for various sample sizes, it was decided to select a sample of 600 businesses of which 500 had joint threshold.

C. Results

14. A sample was selected using this method for month December 2014 and bias measures were calculated. Based on the calculated bias measures, recommendations were made to change the thresholds or not. It was found that most of the thresholds were fit for purpose. It was found that it would be beneficial to lower three of the joint thresholds and two thresholds for the variable turnover.

15. Micro data from September 2011 to June were analysed to measure the impact of increasing the thresholds. The underlying assumption in the calculation of relative bias shown in equation 3 was that the businesses passing selective editing do not introduce any bias. The analysis showed that thresholds could be marginally increased in some cases without adding any bias. Based on this analysis the recommendation was made to increase three joint thresholds.

D. Conclusion

16. The method of traditionally editing a sample of the businesses which passed selective editing to review the selective editing thresholds was tested on the monthly survey RSI. The sample design used was the one which picked more businesses with selective editing scores close to the threshold and ensured fair representation of all the businesses with a range of scores which passed selective editing. The method has proven to be effective in providing enough evidence to reset the thresholds; however some of the influential errors continued to go undetected due to the detailed nature of estimates produced. The method of sampling a proportion of businesses to be traditionally edited from each strata proved simple to implement and efficient.

E. Further considerations

17. From the analysis it was found that selective editing thresholds should also take into consideration the type of estimates produced by the individual survey. For example; there are some estimates produced using the RSI data for subgroups of the domains. This means that the error introduced by certain businesses is inflated when the estimates are calculated for smaller subgroups. The selective editing thresholds are calculated for domains which are larger than the subgroups which are used to calculate some estimates. Taking a sample from a bigger group in the method explained above will not ensure that the businesses which are influential in these subgroups will be picked up.

F. Future Work

18. The method of sample selection needs to be tailored towards the type of survey being reviewed. Even though the aim is to capture more businesses in the sample which are closer to the threshold, the type of estimates being produced and the data or group of data used to produce those estimates should also be considered.

19. When it is a monthly survey, it could be beneficial to add another layer to the sample selection by applying an edit rule which filters out only those businesses which have changed its return value by more than 5%. A sample could be selected from this filtered pool of businesses. Therefore it can be ensured that more businesses which are likely to be edited are picked up in the sample.