WP. 21
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

EDITING IN A TWO-STAGE DESIGN: FROM ELEMENTARY QUESTIONNAIRES TO LOCAL UNITS SCORES

Invited paper

Prepared by Vincent Marcus, INSEE, France

Editing in a two-stage design : from elementary questionnaires checks to local units scores

Vincent Marcus*

August 25, 2006

Abstract

The European Structure of Earnings Survey (SES) is a two-stage survey where local units must fill several individual questionnaires concerning their employees. The editing process of this survey currently used at INSEE (National Institute for Statistics and Economic Studies) is based on a micro editing system with numerous edit checks for the all set of variables displayed in the questionnaire. Priority is given to questionnaires regarding the "importance" of the edits they failed but this importance is assessed on a subject-matter knowledge basis. Moreover, there is no assesment of the impact of the suspected errors on the estimates. Following Hedlin (2003), we first use the standard estimate-related method to build an editing prioritisation process and then its efficiency and potential savings in editing burden. We also reassess an edit-related method in the case of multiple variables of interest to make a comparison. Furthermore, additionnal difficulty in the prioritisation process stems from the fact that edit rules apply to individual questionnaires whereas any follow-up action is based on local units recontact. The problem finds out to be rather similar as the one you may encounter when computing questionnaire score from items scores. Therefore, we design several different rules turning individual questionnaires scores into priority for action at local units levels and assess their efficiency.

Key Words: selective editing, estimate-related, edit-related, local score, global score.

Contact:
Vincent Marcus
Division Salaires et Revenus d'Activité Timbre F240
INSEE
18 bd Adolphe Pinard
75014 Paris
FRANCE
tél : (00) 33 1 41 17 39 53
fax : (00) 33 1 41 17 39 88
mail : vincent.marcus@insee.fr

*INSEE (France)

1 Introduction

The European Structure of Earnings Survey (SES) is a two-stage survey where local units must fill several individual questionnaires concerning their employees. Questionnaires mainly adress earnings topics and working time measurement. Until recently, the European Structure of Earnings Survey (SES) has been conducted on a four-yearly basis. As a consequence, the editing system has been also redesigned almost every four years without taking advantage of previous surveys experience. As seen as a "one shot" survey, little attention was paid on ressources allocation and time consuming tasks. The last survey of this kind was conducted for the year 2002.
Since 2006, the Structure of Earnings Survey is a full yearly survey. As the burden of conducting this survey comes back every year from now on, the different sources of cost of the current data collection process have been heavily scrutinized in search for savings. The efficiency of editing, as one of the most time consuming step within the all process was obviously targeted.
In this article, we present a reassessment of the editing process and explore the potential interest of using selective editing techniques. Following Hedlin (2003), we use the standard estimate-related method and provide assessment of prioritisation efficacy and potential savings in editing burden (section 1). As a by-product, this analysis reveals the rather poor performance of the existing editing model. Building on that,we redesign a new set of edit rules, and develop an edit-related method in the case of multiple variables of interest to make a comparison with the estimate-relatde method (section 2).
Furthermore, for that very specific survey, additionnal difficulty in the prioritisation process stems from the fact that edit rules apply to individual questionnaires whereas any follow-up action is based on local units recontact. The problem finds out to be rather similar as the one you may encounter when computing questionnaire score from items scores. Therefore, we design several different rules turning individual questionnaires scores into priority for action at local units levels and assess their efficiency (section3).

2 Editing process for the SES until 2002

2.1 Main Features of the Survey

The Structure of Earnings Survey (Enquête sur la Structure des Salaires, SES hereafter) performed in France by the National Institute for Statistics and Economic Studies (INSEE) is part of a program initiated in 1966 by the European Statistical Office. More recently, this program has been renewed under the Commission regulation nř1916/2000 (now revised in nř1738/2005). The objective of these surveys is to provide accurate and harmonized data on earnings in EU Member States for policy-making and research purposes. The SES gives detailed and comparable information on the structure and distribution of earnings, as well as individual characteristics of employers and employees. The key variables are :

- the average gross hourly earnings
- the amount of bonuses and other special payments (shift work, overtime...)
- the amount of profit-sharing schemes

- the number of hours paid and hours worked

Statistics on these variables are edited according to several characteristics of employers (principal economic activity of the local unit, size of the entreprise, geographical location) and employees (occupation, gender, age, level of education, type of employment contract...).

The French SES covers firms with at least 10 employees and economic activity inside NACE sections C to K (i.e. all manufacturing industries, construction, trade, hotels and restaurants, transports, finance, real estate and services supplied to businesses). The 2006 SES will also cover sections M, N and O (i.e education, health and social work, and other social and personal service activities).

2.2 Editing principles

The editing process of the survey currently used at INSEE (National Institute for Statistics and Economic Studies) until 2002 was based on a micro editing system with numerous edit checks for the all set of variables displayed in the questionnaire. Priority is given to questionnaires regarding the "importance" of the edits they failed but this importance is assessed on a subject-matter knowledge basis. Moreover, there is no assesment of the impact of the suspected errors on the estimates.

2.3 Selective editing : estimate-related function

We compute a score function in line with Hedlin (2003) and refinements suggested in Lawrence and McKenzie (2000), in ordre to assess the (absolute) difference in the estimate caused by using the raw observed value instead of the edited value. As we deal with ratio estimates, we use expansion in Taylor series to first order to obtain an approximation of the score function, and finally scale it to the domain final estimate. Formerly, let consider y and x as the components of the ratio y/x we are interested in (gross hourly earnings or bonus rate for instace). Formally, we compute the following score for each questionnaire i that failed at least one edit framed for the 2002 edit system:

$$s_i = w_i \times \left| \frac{1}{\hat{X}_D} \left((y_i - y_i^*) - \frac{\hat{Y}_D}{\hat{X}_D}(x_i - x_i^*) \right) \right| \quad i \in D$$

with x^* and y^* as predicted values for the true edited values. We use the mean of the variable of interest on the domain D the individual i belongs to as predictors. Domains are defined along the following dimensions : gender, occupation (ISCO1), age, economic activity (NACE section), and size of the entreprise.

Scaling scores with the standard error of the estimator would have been better. We are currently handling this task for the gross hourly earning variable as we can produce robust estimation of the standard error.

Once the individual scores have been computed, we can rank the questionnaires by decreasing score so that those with the highest scores should be edited first. Then, it is possible to build a relationship between the amount of editing performed and the effect on the final estimate. We provide examples of such a relation in the following grahps, adressing several (aggregated) domain levels and different variables of interest. Formally, for a given proportion p of edited questionnaires (questionnaires with scores

s_i higher than the p^{th} percentile of the s_i distribution),we compute domain estimate from a composite file taking final edited values for the p edited questionnaires and raw values for the remaining. Please keep in mind that such an estimate does not take into account the correct data (questionnaires that failed no edit checks and thus remained unchanged), so that the estimate could be rather far from the final estimate. But we are more interested here in the variation than in the level per se.

Figure 1: Effect on Gross hourly earnings estimate for Executives Male

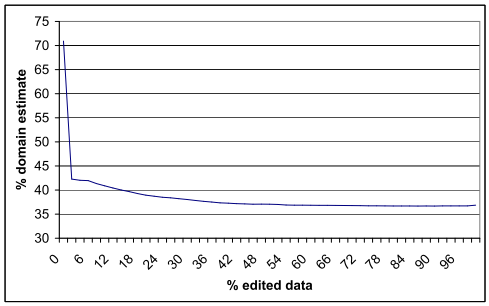


Figure 2: Effect on Gross hourly earnings estimate for Clerks Female

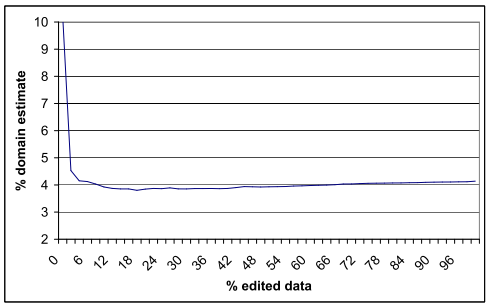
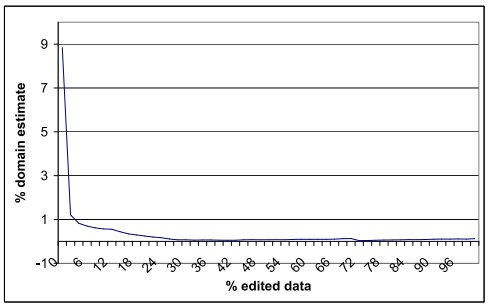


Figure 3: Effect on Gross hourly earnings estimate for Manual Workers Male



Those graphics (Fig 1, 2 and 3) show unambiguously that (i) editing worth it... (ii) but not to much as the curve becomes very flat for rather low percentage of edited questionnaires. Marginal gains from editing more questionnaires turn out to be small once you reach the first $20-30\%$ of the distribution. Results are less clear cut when adressing the bonus rate variable. From Figures 4, 5 and 6, we see that editing almost 50% is needed to reach the flat part of the curve where additionnal work provides marginal gains. Nevertheless, saving about 50% of the current ressources dedicated to data editing is still something.

Figure 4: Effect on Bonus rate estimate for Executives Male

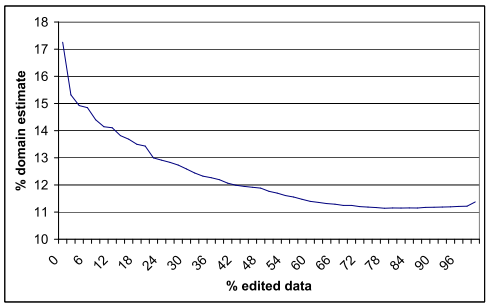


Figure 5: Effect on Bonus rate estimate for Clerks Female

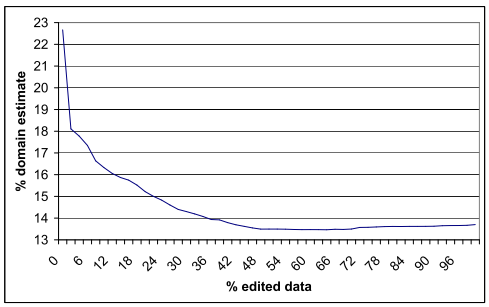
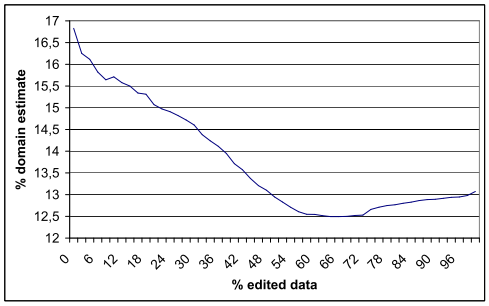


Figure 6: Effect on Bonus rate estimate for Manual workers Male



3 Upgrading the set of edit rules

3.1 Effectiveness of the editing model

The previous analysis partially relies on the relevance of the edit checks. The effectiveness of the editing can be assessed at first glance by calculated the proportion of changes in data for the questionnaires which failed at least one edit. The edit checks seems to underperform as the percentage of changes remain rather low.

Domains	% changed data Hourly earnings	% changed data bonus rate	% changed data profit-sharing rate
Executives M	9.5	14.6	7.8
Executives F	9	14.6	6.9
Technicians M	8.4	15.3	7.5
Technicians F	8.6	14.9	7.5
Clerks M	7.2	13.2	6.1
Clerks F	7.6	13.5	6.2
Manual Workers M	8.2	15.6	6.5
Manual Workers F	7.6	14.6	6.5

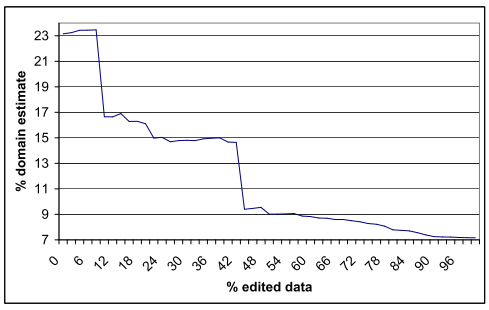
3.2 Refined edit rules

We upgrade slightly a subset of the edit rules for the most important erros and build up an alternative score function based on the magnitude of failures (see Hedlin (2003)). For the Gross Hourly Earnings z_i, we check :

- non-response $z_i \neq .$
- extreme values $z_i > \beta$ or $z_i < \alpha$
- consistence between total gross earnings and sum of the components $z_i \approx \sum_k subz_{ik}$

As shown in the graphic below, this method perfoms rather well as sorting questionnaires by this edit-related score still satisfy the decreasing

Figure 7: Effect on Gross hourly earnings estimate for Clerks Female (edit-related score)



4 From elementary questionnaires to local units

Selective editing methods presented above were used to prioritise editing work on individual questionnaires. But in a two-stage design as the SES, only local units could be called back. Indeed, follow-up actions towards local units can only be made once, as giving a call every time a questionnaire comes out for editing would be unbearable for businesses. In other words, assume you decide to edit every questionnaire scoring above the $80th$ percentile of the score distribution computed from one of the above selective editing methods, and assume that a given local unit has two questionnaires to be edited with very different scores (one score on top, one score close to the $80th$ percentile), it is not possible to call back first the business for the first high scored questionnaire, wait until editing work reach the second, and then give a second call, perhaps a week or even a month later. Therefore, we have to set up priority over local units in a way that preserve as much as possible the original priority obtained on individual questionnaires. For that, we build different rules giving priority over local units from individual scores (computed from the estimate-related method) and discuss their efficiency.

4.1 Rules in line with priority given to individual scores

Let assume that the threshold τ_s is the $80th$ percentile of the score distribution, so that every individual questionnaire scoring above this threshold ($s_i > \tau_s$) would have been edited in the ideal case. For each local unit, we can compute the (weighted) percentage of questionnaires p_j that would need to be edited, and give priority over local units in line with this percentage. Local units with high proportion of questionnaires scoring above the threshold would be edited first. Once a local unit is edited (and called back), it sounds reasonable to edit every questionnaire of the local unit that failed at least one edit (questionnaires scoring under the threshold included). Once again, you can build up a relation between the percentage of edited local units and the effect on the estimates. Furhtermore, you can also assess the efficiency of the priority rule by computing (i) the coverage rate c and (ii) the overediting rate o in relation to the percentage of edited local units. For a given percentage of edited local units j (j such as $p_j > \tau_p$), the coverage rate c is the percentage of edited individual questionnaires belonging to the the selected set of questionnaires (scores above the threshold) among the all set of ideally edited individual questionnaires (as defined above). The overediting rate o is the ratio between the number of edited individual questionnaires and the number of edited individual questionnaires belonging to the selected set of questionnaires (scores above the threshold). Formally, you compute :

$$c = \frac{\sum_i \mathbf{1}_{\{i \in j | p_j > \tau_p \;\; s_i > \tau_s\}}}{\sum_i \mathbf{1}_{\{i | s_i > \tau_s\}}}$$

and

$$o = \frac{\sum_i \mathbf{1}_{\{i \in j | p_j > \tau_p \;\; s_i \geq 0\}}}{\sum_i \mathbf{1}_{\{i \in j | p_j > \tau_p \;\; s_i > \tau_s\}}}$$

The case where you edit all the local units is a particular case of the above formulas, meaning that a local unit is edited as soon as one of its individual questionnaire belongs to the selected set of questionnaires. The following graphs display the relationship between c (respectively o) and the percentage of edited local units, assuming that $\tau_s = p_s^{80th}$.

Figure 8: Coverage rate when scoring on Gross hourly earnings

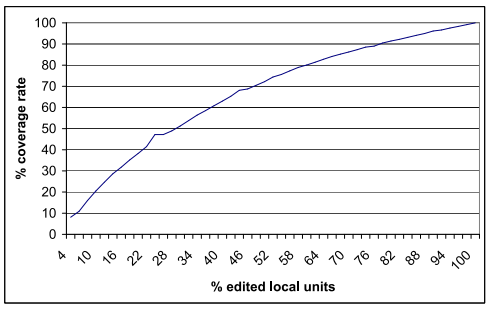
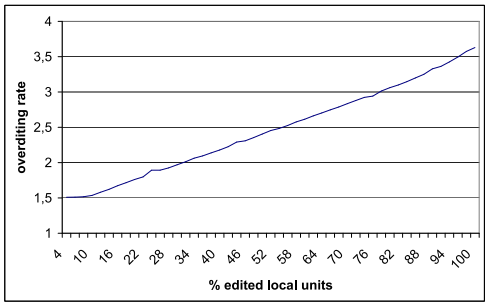
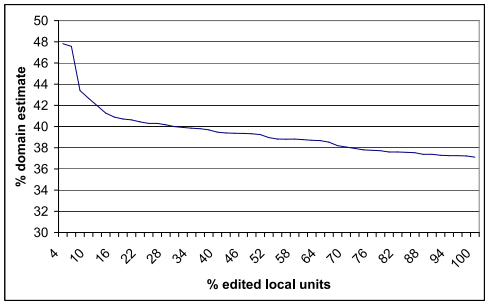


Figure 9: Overediting rate when scoring on Gross hourly earnings



The coverage rate curve is slightly concave, so that marginal gains in terms of fulfilling complete editing over targeted individual questionnaires are rather decreasing. The cost from selecting local units instead of individual questionnaires, namely overediting, is rather linear here and reach almost a 1 to 4 ratio. Following this rule of priority for local units remain consistent with priority in line with effect on the estimates. Indeed, editing local units sorted as such produce very similar curve as the one direcly obtained when dealing with individual scores, as show in Fig 10.

Figure 10: Effect on Gross hourly earnings estimate for Executives Male when editing local units



4.2 Scoring at local units level

From individual scores, one could also compute a local unit score,whatever the distribution of individual scores is. In fact, the problem here is very similar to computing a global score (at the unit level from several item scores (see Lawrence and MacKenzie (2000). Therefore, we compute two different scores for local units j, namely s_j^1 and s_j^2, defined as :

$$s_j^1 = Mean\left\{w_i.s_i \ \ | \, i \in j\right\}$$

$$s_j^2 = Max\left\{w_i.s_i \ \ | \, i \in j\right\}$$

Sorting local units by s_j^1 (resp. s_j^2), we can build up the same relationships as above and compare those three different rules of priority for local units editing.

Figure 11: Coverage rate when scoring local units with the mean of individual scores

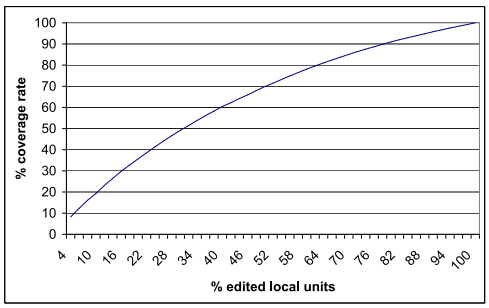
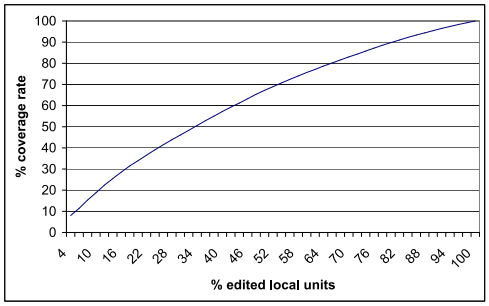


Figure 12: Coverage rate when scoring local units with the max of individual scores



The three rules we designed performed in a rather similar way, but the first one in line with priority given to individual questionnaires provides the best results. For a given percentage of edited local units, it provides the best coverage rate (see Table 2), and cost less in terms of overditing (see Fig. 13 and 14).

Table 2 : Coverage rate for three different rules giving editing priority to local units

	% edited local units		
Scoring with...	20%	50%	80%
$\%i \mid s_i > \tau_s$	38.2	72.2	94.4
Mean	35.7	69.7	91.0
Max	33.9	67.2	89.8

Figure 13: Overediting rate when scoring local units with the mean of individual scores

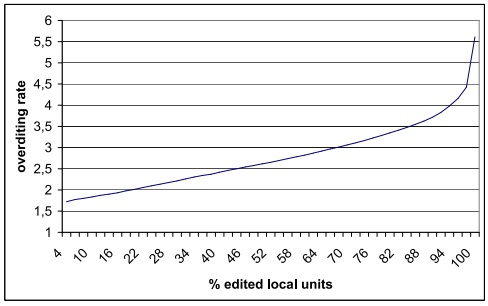
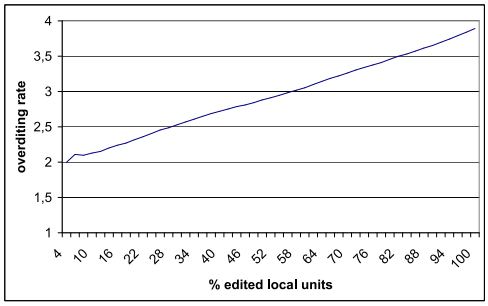


Figure 14: Overediting rate when scoring local units with the max of individual scores



5 Conclusion

Several previous work on data editing has proved that substantial savings could be achived with selective editing methods. Once again, this paper confirms this point. Sorting the questionnaires with respect to the effect of not editing it on the estimations proves to be a powerful means of reducing or reallocating ressources currently dedicated to full data editing. Introducing non-linear constraints stemming a two-stage design, as you must deal with all individual questionnaires of a local unit at the same time whatever the priority you gave, could be handled out by computing local unit "scores"". The best way to do it seems to fit as much as possible with the ideal priority given to individual questionnaires. Nevertheless, for that survey, selective editing is the whole thing. Indeed, we are not only interested in statistics but also in high quality individual data file for econometric studies purposes. Therefore, additionnal work is needed in ordre to ensure that the selective editing tools do not have too much adverse effects on individual data quality.

References

Granquist, L., and J. Kovar (1997): "Editing of Survey Data : How Much is Enough?," in *Survey Measurement and Process Quality*, ed. by L. Lyberg, and ali., pp. 415–435. Wiley, New York.

Hedlin, D. (2003): "Score Functions to Reduce Business Survey Editing at the U.K Office for National Statistics," *Journal of Official Statistics*, 19(2), 177–199.

Latouche, M., and J. Berthelot (1992): "Use of a Score Function to Prioritise and Limit Recontacts in Business Surveys," *Journal of Official Statistics*, 8, 389–400.

Lawrence, D., and C. McDavitt (1994): "Significance Editing in the Australian Survey of Average Weekly Earnings," *Journal of Official Statistics*, 10, 437–447.

Lawrence, D., and R. McKenzie (2000): "The General Application of Significance Editing," *Journal of Official Statistics*, 16, 243–253.

ONS (2005): "A Management Information System for Controlling Editing Quality in a Survey with Multiple Requirements," in *Statistical Data Editing*, ed. by C. of European Statisticians. UNECE.