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

Vincent MARCUS[1]

[1]INSEE (France)

Conference of European Statisticians, Bonn 2006

Outline

Main features of the survey

General framework

- European survey initiated in 1966, irregulary conducted until 2002, on a four-yearly basis since then
- gives harmonised data on individual earnings, hours worked and labour cost in EU Member States

Field

- all employees in firms with 10 or more employees
- sections C to K, M, N, O

Key variables (ratios)

- gross hourly earnings
- hours paid and hours worked
- bonuses, profit-sharing schemes...

Design of the Survey

Two-stage design with stratification at both stages

- first-stage : local units (businesses)
- second-stage : employees (1 to 24 per local units)

what matters :

- one questionnaire per employee (individual data are collected)
- questionnaires are completed by the local unit
- only local units could be called back to obtain more information on data related to their employees
- no follow-up action on the employee itself

Design of the Survey

Two-stage design with stratification at both stages

- first-stage : local units (businesses)
- second-stage : employees (1 to 24 per local units)

what matters :

- one questionnaire per employee (individual data are collected)
- questionnaires are completed by the local unit
- only local units could be called back to obtain more information on data related to their employees
- no follow-up action on the employee itself

Editing framework for the SES 2002

- a micro editing system applied to every questionnaire
- numerous edit checks over the whole set of variables
- a questionnaire is edited as soon as it fails at least one edit
- otherwise not

Estimate-related score function

the score function represents the change in the estimate if a raw value is replaced with the edited value (or a proxy of it)

Score function for each questionnaire i

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

where

x and y : the variables of the ratio y/x we are looking at

x_i and y_i : raw values

x^* and y^* : predicted values (mean on D) for true edited values

domains D : defined along the following dimensions (gender, occupation, age, economic activity, size of the entreprise)

w_i : weight for employee i

Assessing potential savings through selective editing

Building graphs

- rank questionnaires by descending scores
- consider the first p % questionnaires (i.e $s_i > (100 - p)^{th}$ percentile of s_i)
- compute domain estimate taking final edited values for the p % edited questionnaires and raw values for the remaining.
- repeat those steps for different values of p

Figure: Effect on Gross hourly earnings estimate for Clerks Female

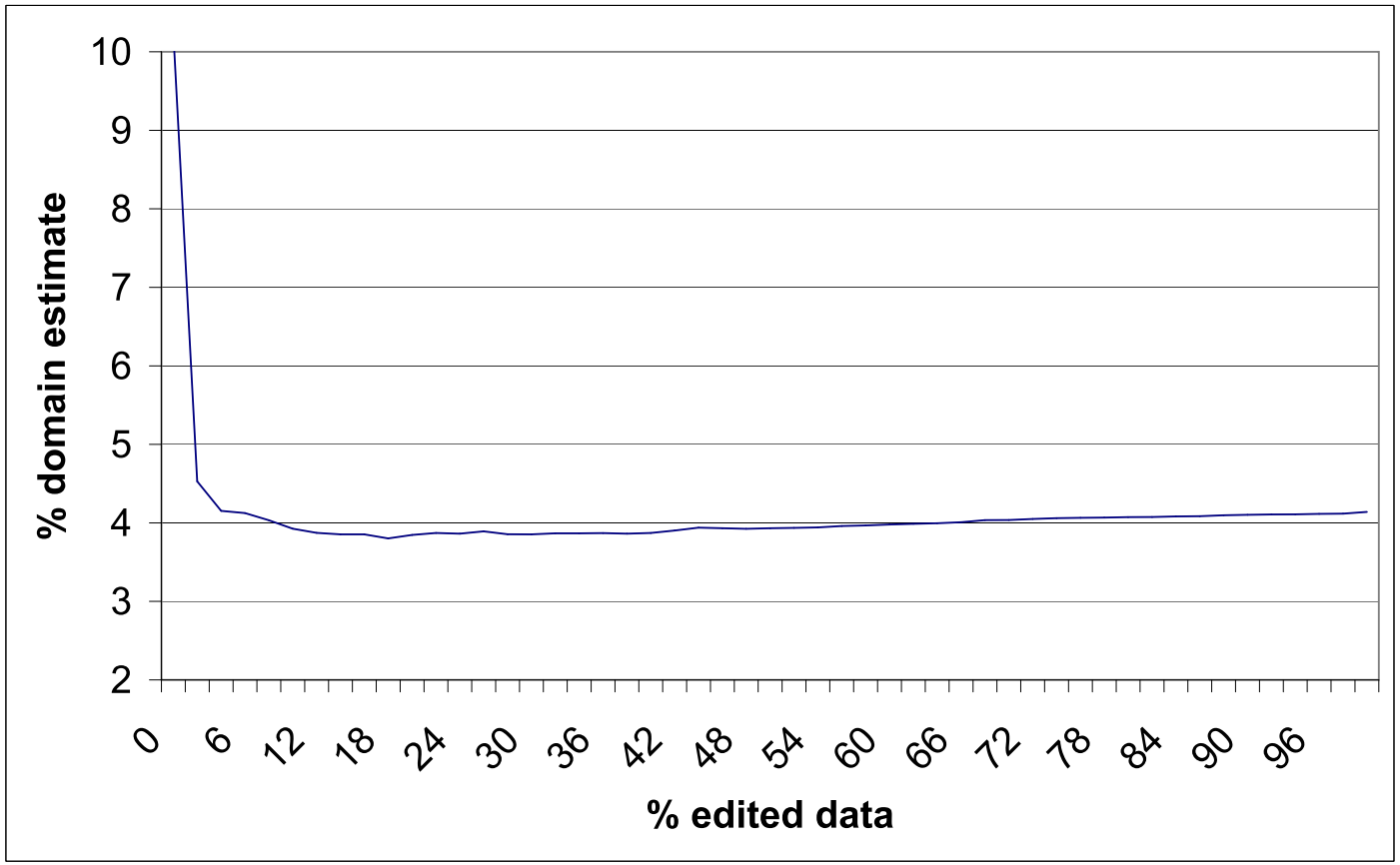


Figure: Effect on Gross hourly earnings estimate for Manual Workers Male

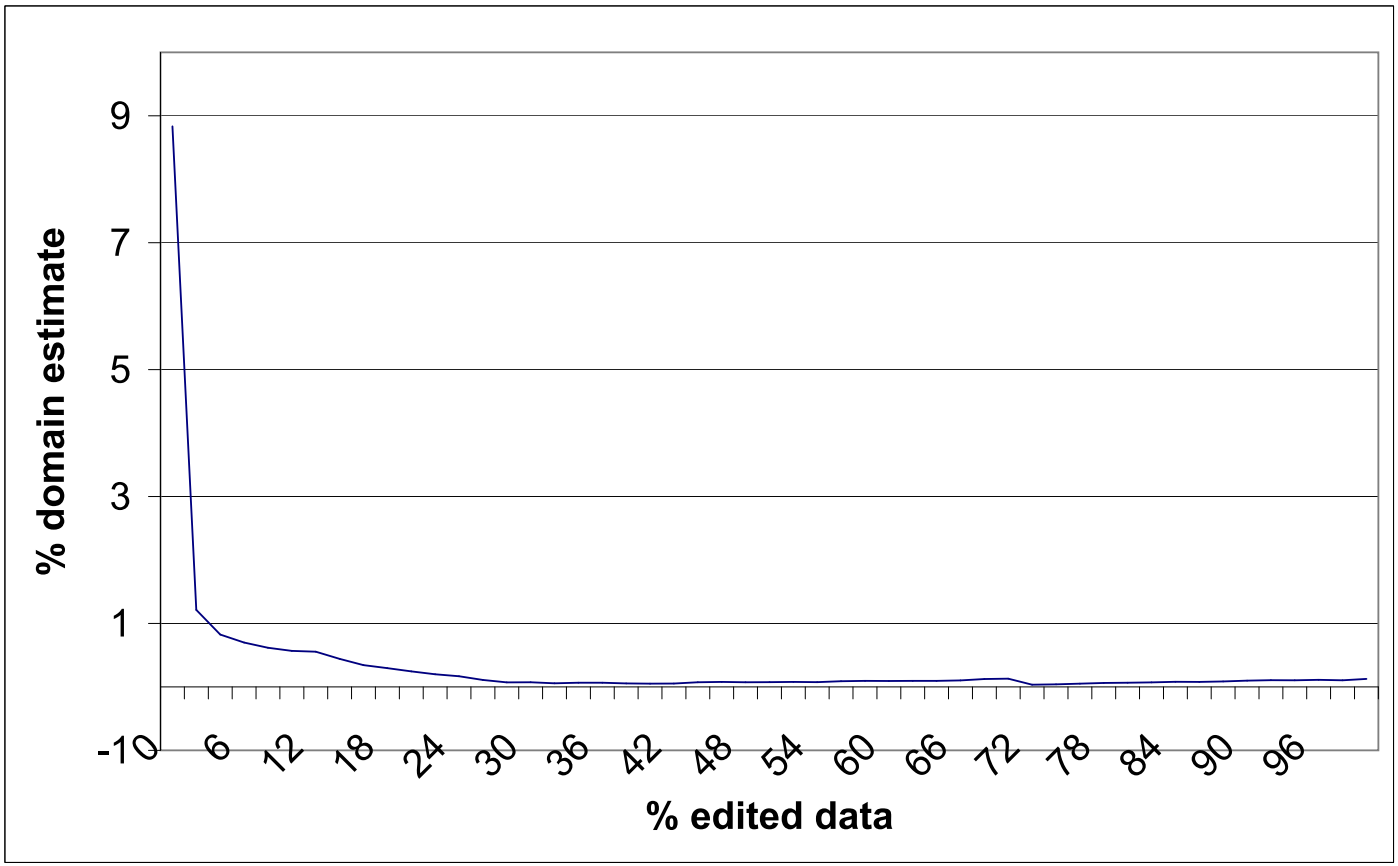


Figure: Effect on Bonus rate estimate for Clerks Female

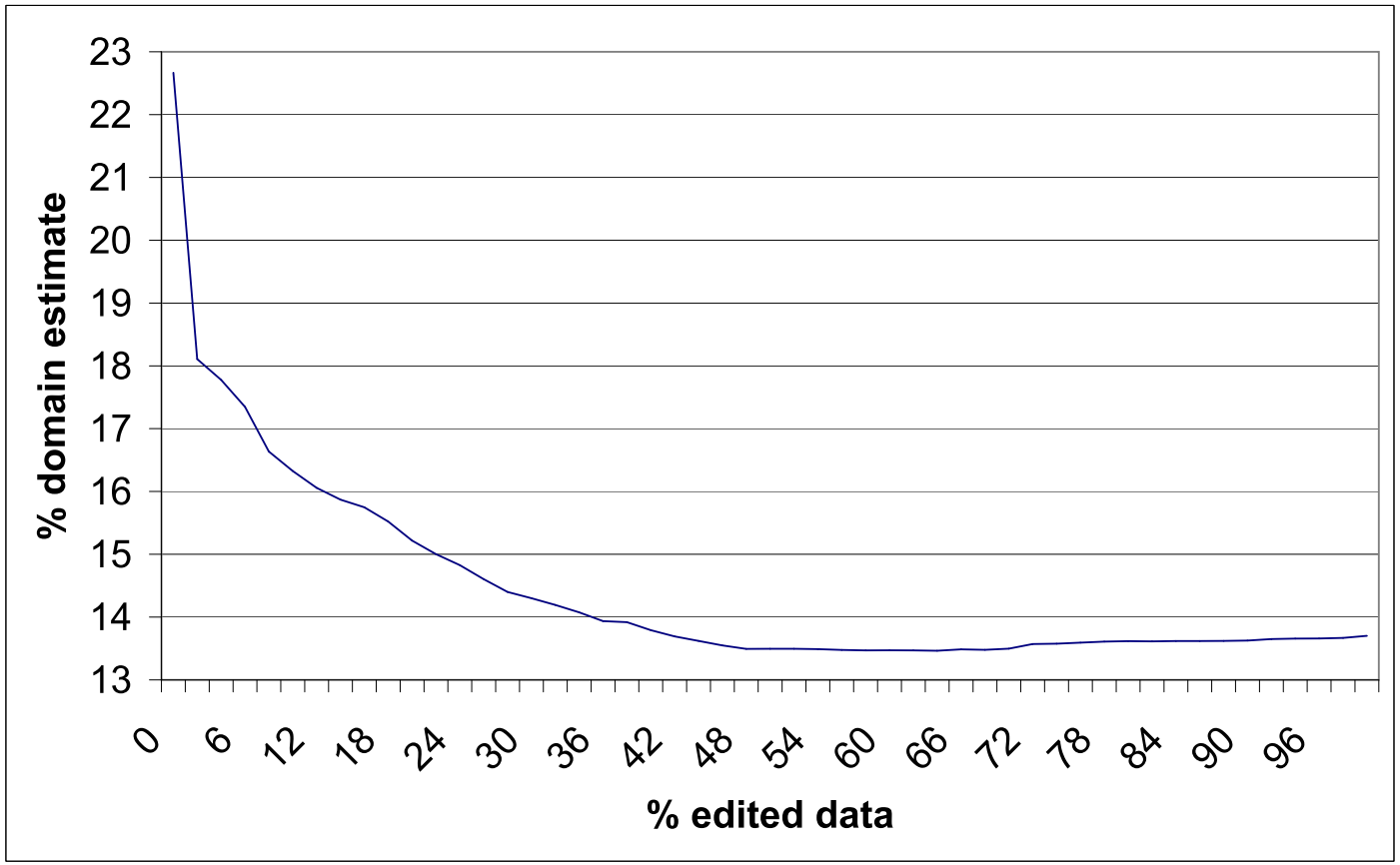
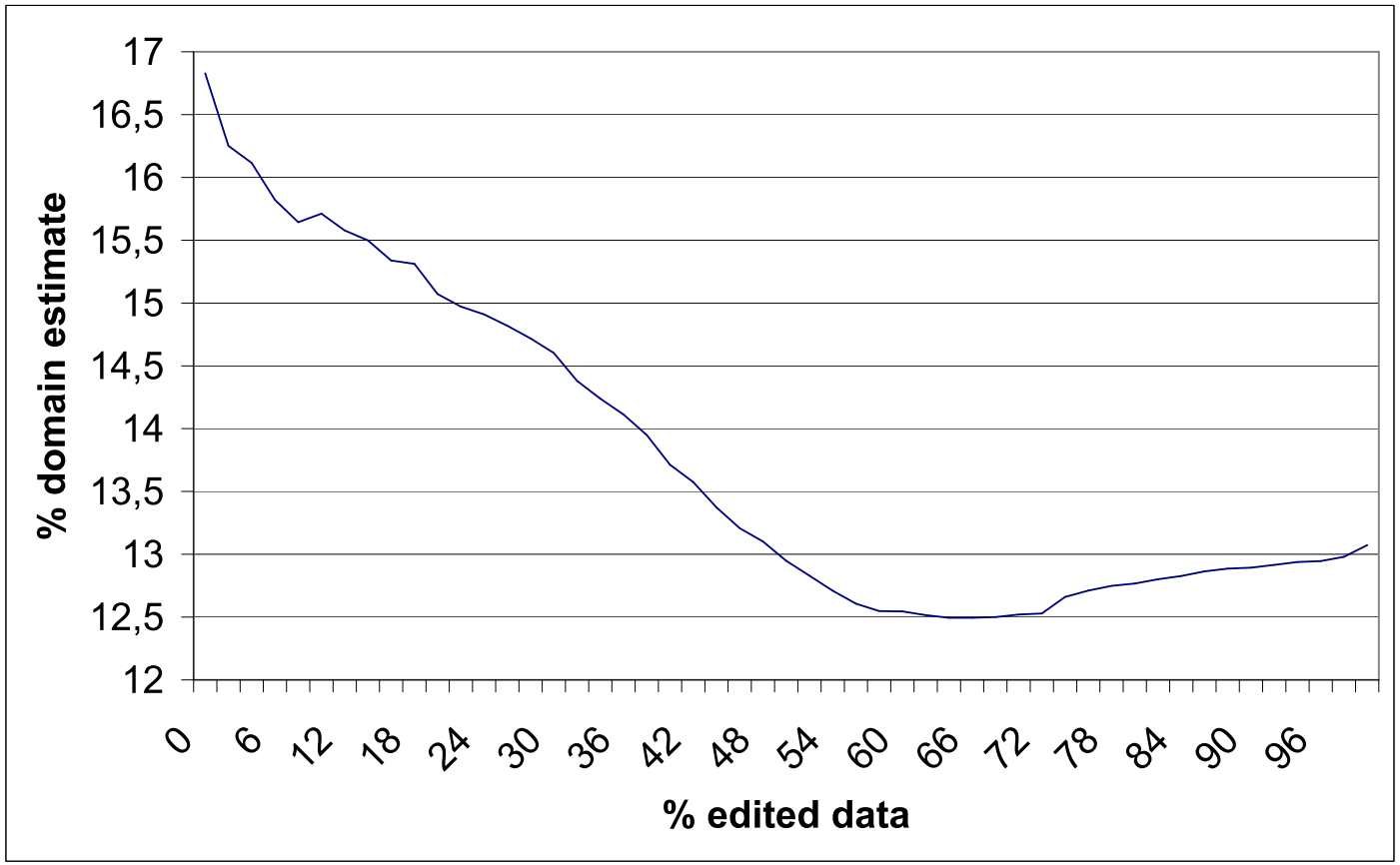


Figure: Effect on Bonus rate estimate for Manual workers Male



Edit-related method

Principle : put selective editing on top of a micro editing system and prioritise raw values by how many edits they fail and by 'how much'

Edit rules r for the gross hourly earnings z_i

- (r1) non-response : $z_i \neq .$
- (r2) extreme values : $\alpha < z_i < \beta$
- (r3) consistence between total gross earnings and sum of the k components $z_i \approx \sum_k z_{ik}$

Edit-related method

'size' of edit failures e_i

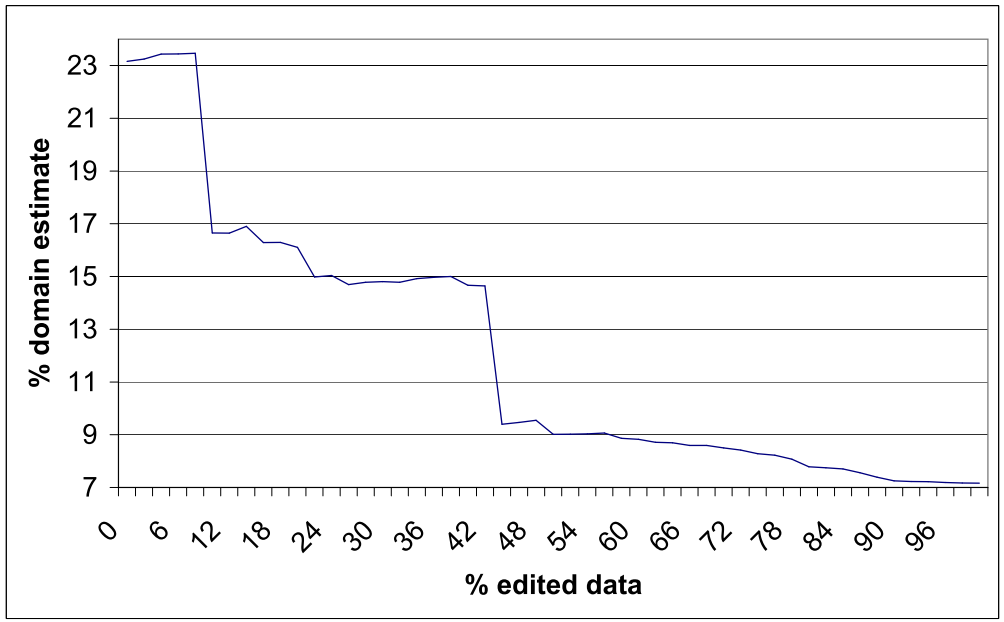
- $e_{i1} = \hat{z}_D$
- $e_{i2} = \frac{(z_i - \beta)}{(z_{max} - \beta)}$ if $z_i > \beta$ or $e_{i2} = \frac{(\alpha - z_i)}{(\alpha - z_{min})}$ if $z_i < \alpha$
- for e_{i3}, same formulas as e_{i2} considering the ratio $\sum_k z_{ik}/z_i$ wiht $[\alpha; \beta] = [0.95; 1.05]$ (5 % tolerance)

edit-related score function

$$d_i = \sqrt{(e_i - \hat{\bar{e}})' \hat{S}^{-1} (e_i - \hat{\bar{e}})}$$

with $e_i' = (e_{i1}, e_{i2}, e_{i3})$

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



from individual questionnaires checks to local units scores

Problem

- You have priority rules over questionnaires (from score functions as above) but you can call back the local unit for further information only once for all the questionnaires.
- what priority rules for these actions towards local units?

Method

- define a target on edited questionnaires (say the first $20^{th}\%$)
- compute a 'score' at the local unit level and sort the local units by descending score
- edit any questionnaire of a local unit (whatever its score) as soon as the local unit is 'edited'
- assess the method through the effect on final estimate, the coverage rate and the overediting rate.

coverage rate c

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

overediting rate o

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

where

s_i is the score for questionnaire i

τ_s is the threshold you set to define your target (say the $80^{th} percentile of the s_i distribution$)

p_j is the score for local unit j

τ_p is the threshold over which a local unit is edited

Local unit scores

score 1 (weighted proportion of questionnaires to be edited)

$$p_j = \frac{\sum_{i \in j} w_i . \mathbf{1}_{\{i | s_i > \tau_s\}}}{\sum_{i \in j} w_i}$$

score 2 (mean of scores)

$$s_j^1 = Mean \{ w_i . s_i \ \mid i \in j \}$$

score 3 (max of scores

$$s_j^2 = Max \{ w_i . s_i \ \mid i \in j \}$$

Figure: Effect on Gross hourly earnings estimate for Clerks Female when editing local units (score 1)

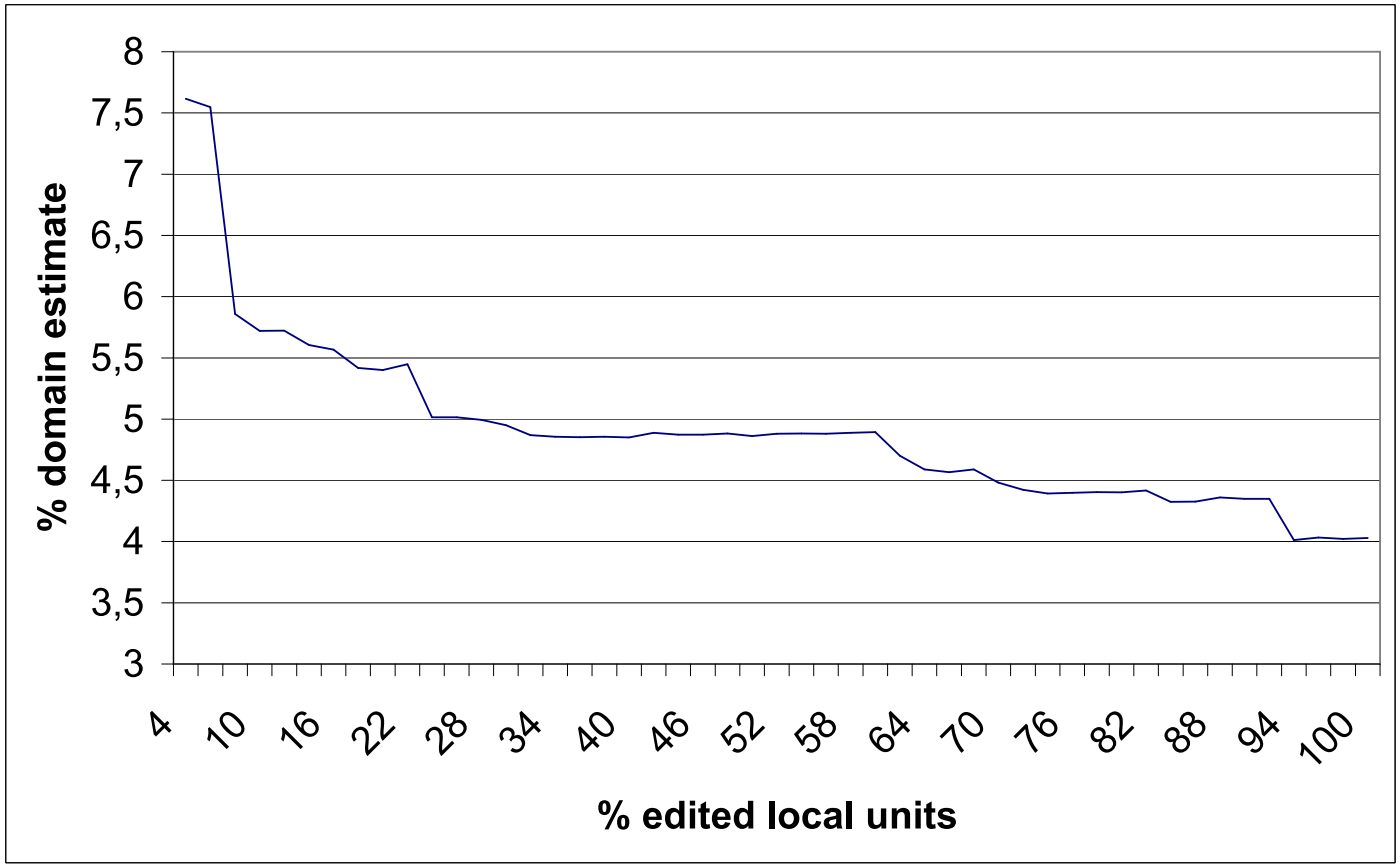


Figure: Coverage rate when scoring on Gross hourly earnings with weighted proportion of questionnaires to be edited (score 1)

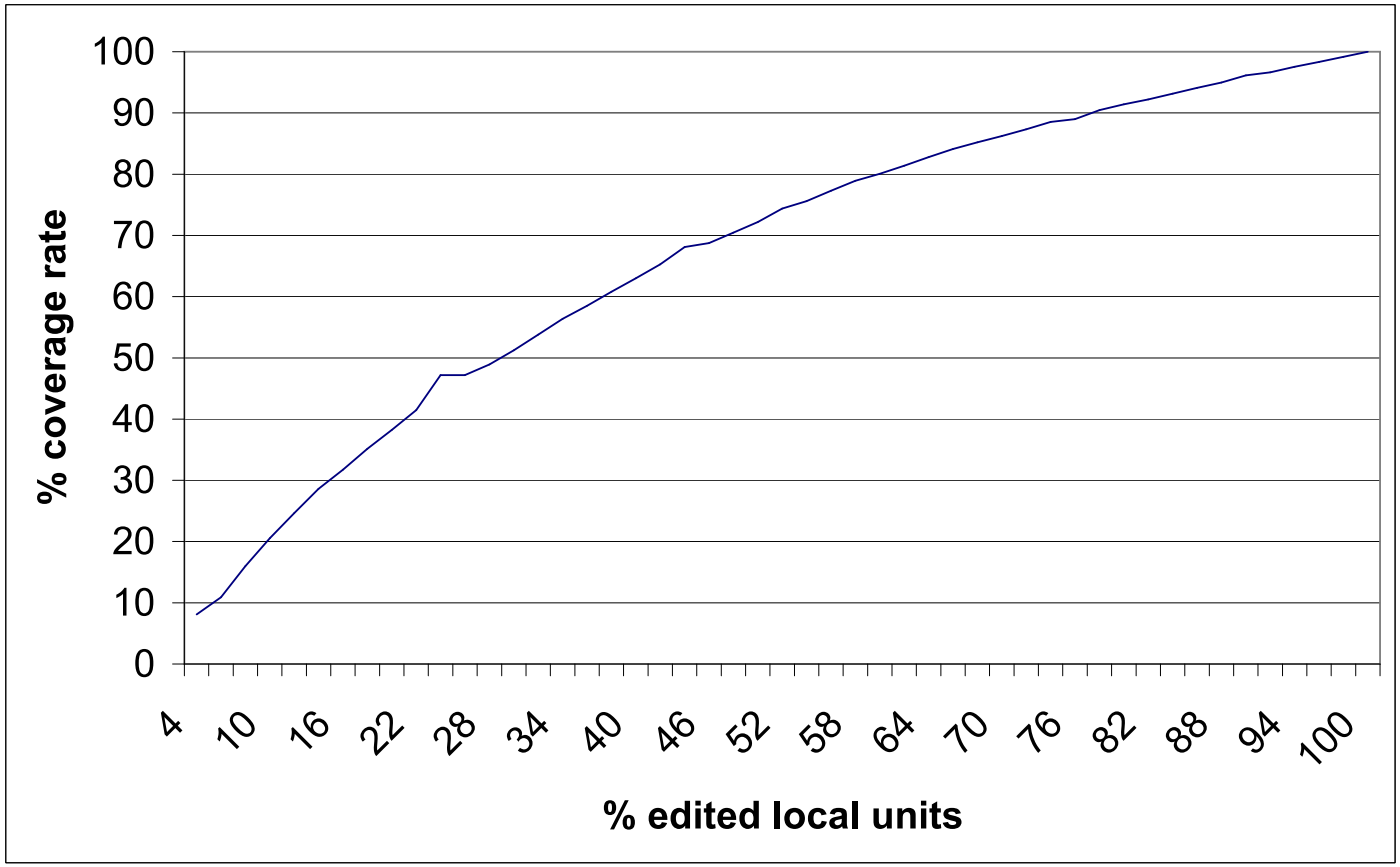


Figure: Overediting rate when scoring on Gross hourly earnings with weighted proportion of questionnaires to be edited (score 1)

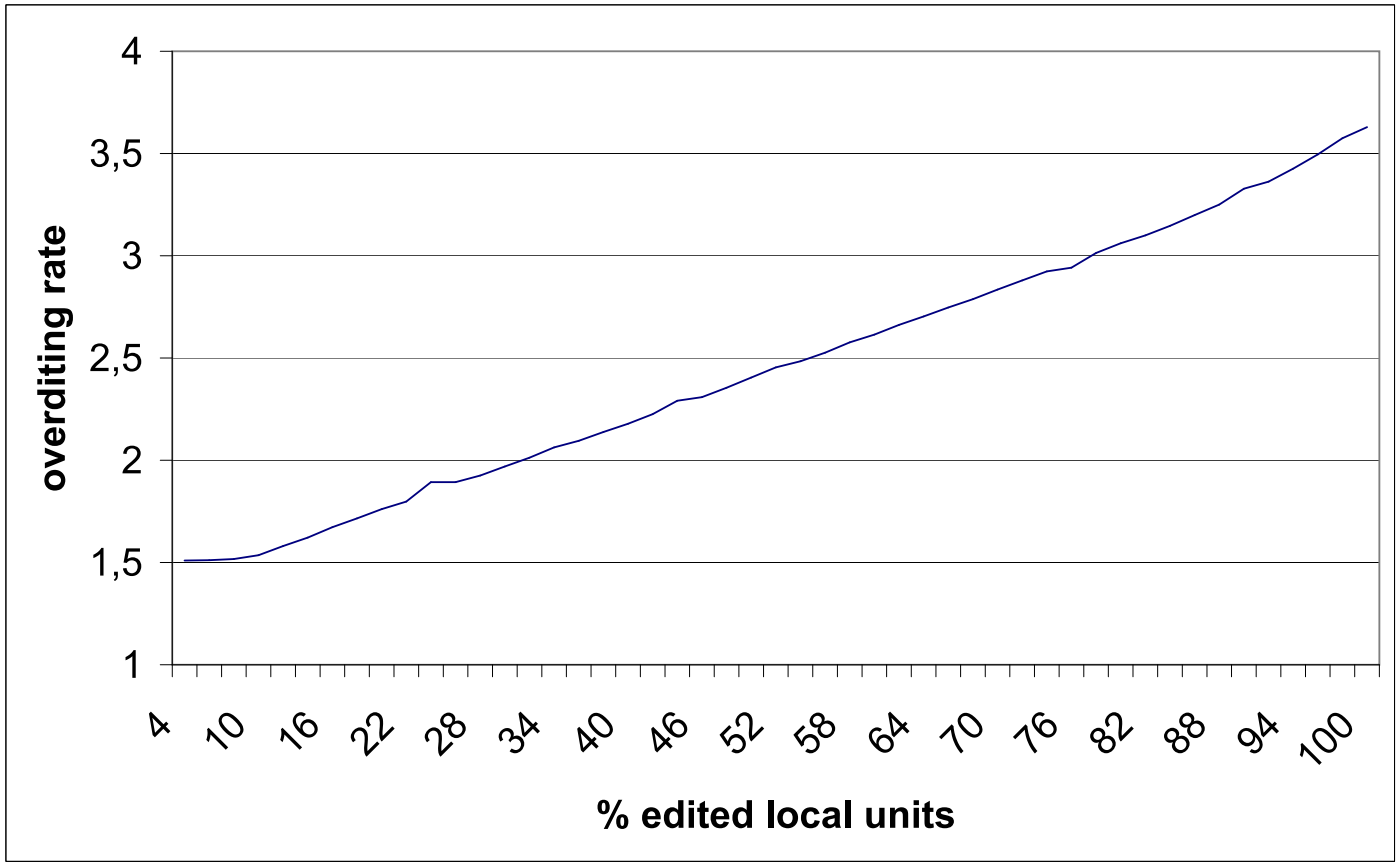


Figure: Coverage rate when scoring local units with the mean of individual scores (score 2)

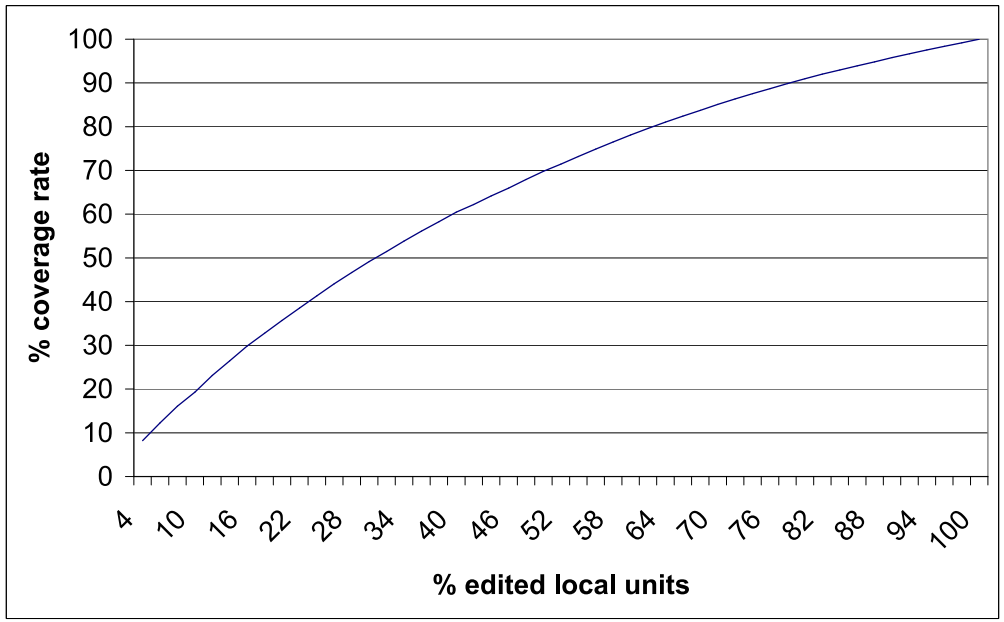


Figure: Coverage rate when scoring local units with the max of individual scores (score 3)

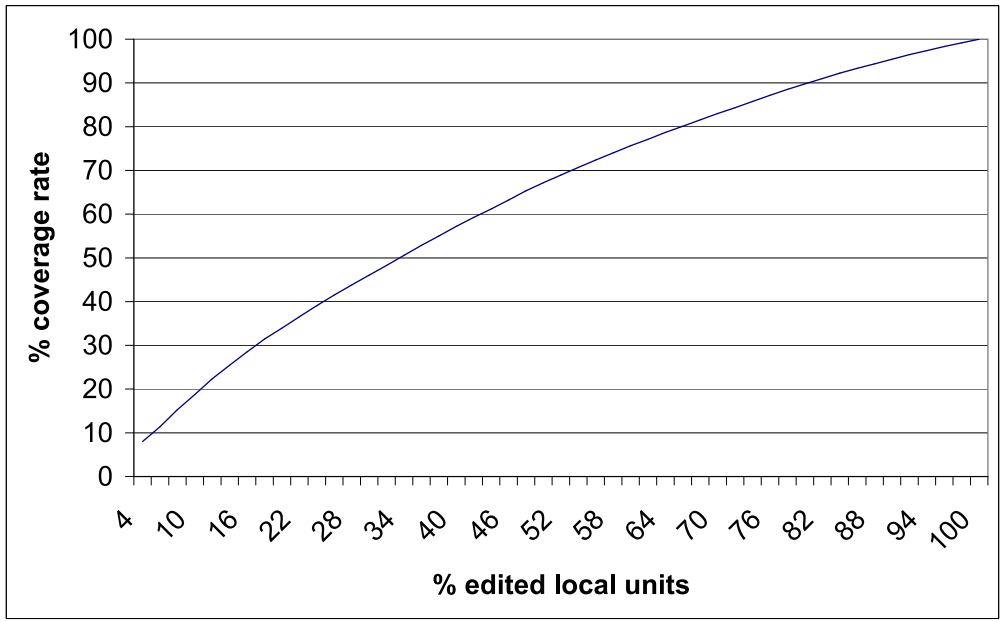


Table : Coverage rate for three different local units scores

	% 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

Concluding remarks

- significant savings could be achieved using selective editing methods (we all know that!)
- in our case : scoring local units in line with priority over individual questionnaires seems the best
- further research : the effects of selective editing techniques on individual data quality as we are also interested in econometric studies (and not only in aggregate statistics)