WP. 34
ENGLISH ONLY

**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)

POST-EDITING OF MICRODATA TOWARD THE ANALYSIS

Supporting Paper

Prepared by Seppo Laaksonen, Statistics Finland and University of Helsinki

I. INTRODUCTION

1. Standard editing in statistical institutes is rather cross-sectional, although some short-term statistics may be exceptions to some extent. This means that each cross-section survey has been edited and imputed during a certain strict process. The target of the editing is to satisfy the certain aggregate statistics requirements. So, the required estimates should be reasonably good after editing and imputation.

2. Later, the same cross-sectional micro files are further used by pooling with other micro files or merging these together both cross-sectionally and longitudinally. At this stage, the statistics unit is no more the main user but a researcher or a planner who wants to get advantage of a number of data files, and thus more perspective for his/her research topic. A post-user cannot automatically trust on such combined data since the merging and pooling discloses gaps that have not been found before. So, post-editing is needed before the data analysis; these two operations are often made interactively. This paper presents the two concrete examples of real situations.

II. EXAMPLE 1: ANALYSIS OF FAMILY BUSINESS DATA[1]

3. The business micro data laboratory of Statistics Finland is about ten years old. It has been regularly used by outside researchers, often with help by inside experts. There are basically all structural business statistics data sets feasible in this laboratory. In addition, some employment data are matched with these business data. This employment data set is the longitudinal census on employed people including an identifier of their two main employers during each reference year. The merging from individual data have been usually realized at enterprise or another appropriate aggregate level.

4. Data sets of the laboratory are cleaned in a usual way in each statistics division, in most cases cross-sectionally, since the time schedule does not give opportunity to look carefully at historical data. The purpose of this editing is to produce material for aggregate statistics, not to check all individual items completely.

5. Laboratory researchers are nowadays almost always using data over several years and so that there are several initial data sources. The data sets are linked routinely in the laboratory following the needs of a client. Technical help is available but any helper cannot know all surveys used by a client.

[1] This example is based on the joint work with Kalevi Tourunen who is a client of the business data laboratory of Statistics Finland.

He/she can ask from the initial conductor but it is guaranteed that this is able to give correct answers even though is working still at office, and even less to edit data further. There are always some documentation deficiencies. For these reasons, the researcher has to start with a new editing that controls even already edited cross-section data but more importantly, he/she will look at problematic points over merged and linked data sets.

6. There are no guidelines for such post-editing but these could be useful. This is an example that is not the most complex one, since there are the structural business statistics data sets first merged together over 5 consecutive years and some variables added from specific surveys such as R&D. Next, one variable has been manually entered, made by the client himself. His work was hard since he has phoned to all businesses of the certain sectors larger than 20 employees and asked their ownership with three alternatives: family owned only over one generation = *Family A*, owned by the same family over a two or more generations = *Family B*, not owned by family = *No Family*. As observed, smallest businesses are not included in the analysis that aims to evaluate performance of those three ownership groups. It should be noted that small businesses are very often owned by families.

7. We constructed a standard multivariate regression model where several performance indicators are estimated. Here just one has been presented without a detailed description of the construction of this variable that is of a ratio type. Table 1 presents our main results.

Table 1. Estimates for performance of enterprises without post-editing and with post-editing (not all variables included); * =significant

Statistic	Without	With
Observations	24584	23834
R-square	4.4%	5.2%
Log_size	0.030*	0.011*
Log_size*Log_size	-0.0029*	-0.0011*
Trade	-.00099	-0.0057*
Construction	-0.0000	+0.0018
Traffic	-0.0105	0.0050
Manufacturing	0	0
Other	-0.0060	0.0073*
Debt rate	-0.313*	-0.144*
Family A	0.019*	0.011*
Family B	0.011*	0.001
No Family	0	0

8. These results give a rather different image. In the beginning we were a bit positively surprised that both types of family businesses succeeded better than non-family businesses. Although there exists little results from this kind of research this was not expected. When we looked at all the variables in more details we found incorrect values. The main editing was to change the negative values of ‘positive variables’ to missing. Another option was to increase these to zero but this was not used for this table although it could be a real alternative. What is your opinion?

9. The second option was to bound the dependent variable into the certain normal acceptable limits, since there were found some high outliers. Note that we ourselves constructed this and the other performance indicators for this analysis. The values are derived from several initial variables.

Presently, 5 February 2008, we are not yet ready. We are still going to look at some new editing before estimating and publishing final results.

III. EXAMPLE 2: CROSS-SECTION AND PANEL ANALYSIS OF INDIVIDUALS FROM HOUSEHOLD SURVEY[2]

10. The European Household Community Survey (EHCP) was conducted under Eurostat's coordination for the years 1994-2001 but in Finland only for 1996-2001. Common guidelines were required to exploit in participating countries but many practices still had to be decided in countries. Finally, the micro data, called ECHP UDB, was created following the same format in each country. Although these data consist of several files, researchers were able to exploit single- or cross-country data rather easily, in principle. But the practice is not simple, especially in the cases when both cross-sectional and longitudinal features, and households vs. household members were interested for a user.

11. The target of our study was to analyze wage formation in Finland between 1996 and 2001. Since wages are paid for individuals, we need individual data but at the same time we wanted to look at households in the sense whether there are some connections between household composition and wages. Hence we had

- to merge together household data and individual data at each panel wave, and then
- both to pool and to merge these longitudinally.

12. To give some idea about these operations, we present two illustrations in Figures 1 and 2 that are naturally much simplified since there are a lot of practical problems required to be solved before the merging is ready.

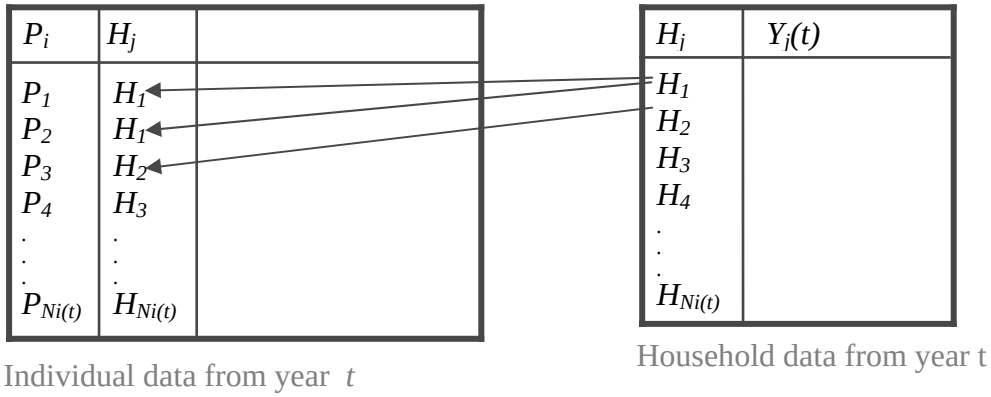
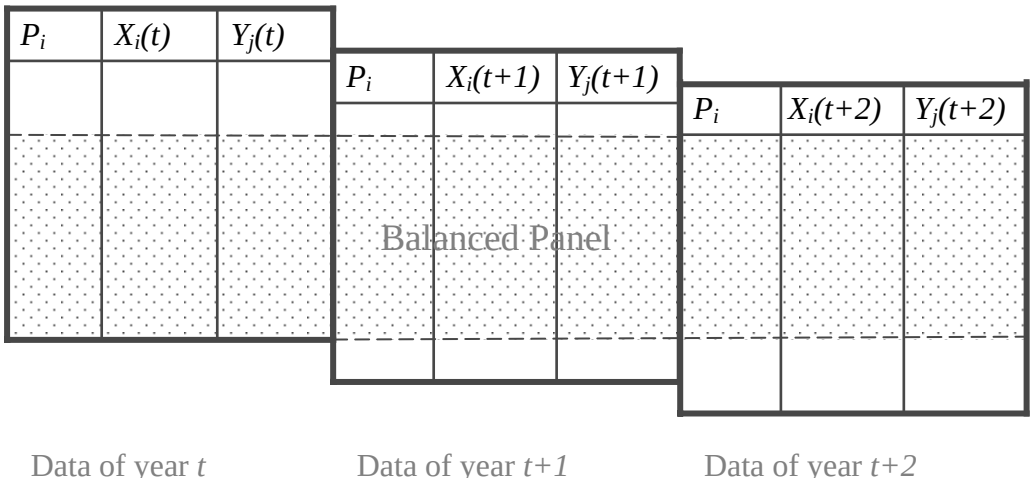


Figure 1: Merging individual and household data of year t ; P = individual identifier, H = household identifier, Y = household variables of interest



[2] This is the joint work with Kirsi Sokura who worked at Statistics Finland for this project.

Figure 2: Merging three consecutive cross-sectional files. X = *individual variables of interest*

13. The pooled data are not presented. The data of Figure 2 is thus used in longitudinal analysis. This requires to create new variables that illustrate changes from one year to the next. For continuous variables logarithmic or absolute differences are used whereas change codes for categorical variables. The simplest version for the latter case is to use the dummy that implies whether there is change or not. Figure 3 illustrates combinations of records during the three first years. The number of such combinations naturally increases in the run of the panel duration; in this case it is 54 for the whole 6 years (out of the maximum = 64).

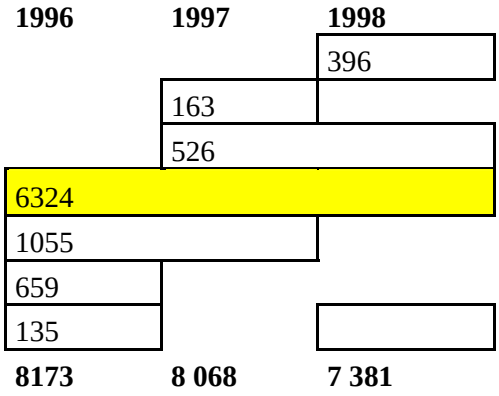


Figure 3: Respondents during the first panel years.

14. The response patterns already illustrate that there will be problems in analyzing these data with our purpose, that is, cross-sectionally and longitudinally. One can imagine that longitudinal analysis could be more awkward since wage changes over all consecutive years are not available always, that is, missing changes will occur. We have not tried to impute these, and hence the number of observations is reduced, respectively.

15. Another, even more awkward problem concerns the definition of wage and salary earners at each wave, and consequently the variables used in the analysis. It is natural that these all should be consistent with each other, that is, the wage of an individual has been earned from a certain job (occupation, full-time vs. part-time, etc) in which he/she has been occupied really at this period. This requirement is not automatically achieved since there are two different reference periods in the ECHP. The reference period for wages is the whole year *t* but the fieldwork has been carried out next spring and once even later. At this occasion the interviewer has been inquired for the employment position but there has been opportunity to look at employment also from the register. Table 2 illustrates these alternatives.

Table 2: The number of wage earners by the different definitions in each wave

Sample	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
All respondents	8173	8068	7381	7109	5614	5637
Employed based on the ILO definition (%)	4702 57.5	4591 56.9	4358 59.0	4246 59.7	3398 60.5	3378 59.9
Wage earner during the fieldwork time based on the information from the respondent	3483	3485	3421	3374	2788	2796
Wage earner at the previous year based on the definition of the respondent (worked 6 months, at minimum)	3436	3309	3250	2663	2697	-

16. The percentage of employed people seems to increase during the panel that does not correspond to the real life; the reason is attrition. The two definitions of wage earners are rather different, the

fieldwork time definition giving a higher figure. The position of a wage earner is not necessarily the same with these two definitions, concerning such variables as occupation group, industry and working hours that are used in our models as explanatory variables. It should be noted that all explanatory variables are taken from the fieldwork data and thus from the interviewing period.

17. The dependent variable, wages, is not either simple since it is not completely known whether this has been earned in similar conditions from one year to the next. We have used monthly wages that has been calculated from the yearly wage divided by worked months. We know as mentioned above the working hours from the fieldwork time but this number can be different than that from the wage-earning period. Hence, all wages are not correctly conditional to explanatory variables. Our only option to robust our data in this point has been to exclude the outliers from the data and to loose observations. Our criterion for outliers has been simple, we have included the individuals who's value lies between 5% and 95% quantiles at each wave. This means that our criterion has been the same for each wave.

18. We have tested two types of wages over some years, that is, register wages and wages informed by a respondent. The comparison of these wages is not automatic, since the register wages are annual but the interviewed wages are monthly. The comparisons have been made only for those who have received register wages 6 months at minimum. The second trimming has been exclude those who have not answered to this question or have given a very small and unbelievable wage for that month.

19. Figure 4 describes all the conditions that an individual will be included in the wage model of year t. Respectively, these conditions should be satisfied both at year *t* and year *t+1* in order to be included in the pooled panel model over these two years. The whole panel model covers the corresponding 5 couples over all study years.

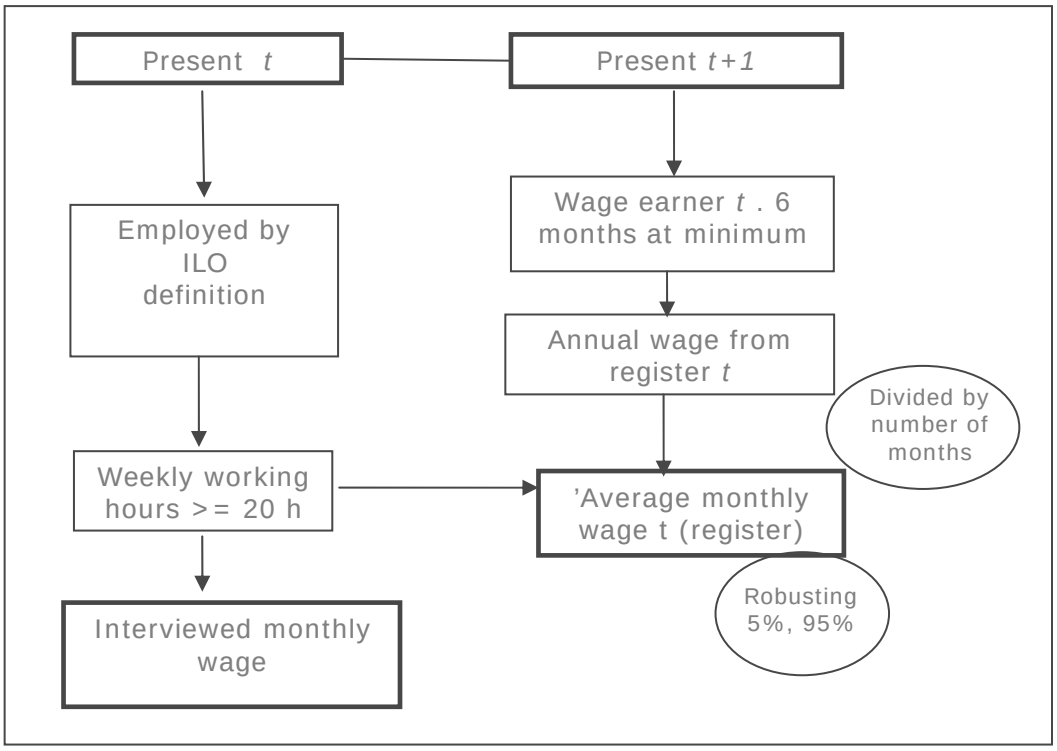


Figure 4: Conditions to be a respondent at year *t*.

20. We thus have explained both wage differences and wage changes. For both types of models a number of explanatory variables are attempted such as gender, age, education level, tenure, marital status, occupation group, size band of enterprise, industry, category of hours worked. These are rather standard in wage modeling. Since we were able to exploit specific ECHP variables, we constructed some models using such explanatory variables as job satisfaction, status, disability or not and how demanding the job

is. There was in some cases hard to construct a good change variable for the panel analysis. An example is the change of the job between the two employers during two consecutive years.

21. The full information from level or change categories was not always available. One strategy would be to exclude these records from the analysis but we avoided this option since we did not like to loose observations. We also wished to analyze these category groups. So, our strategy in several cases was to include an unknown category in important variables. For example, variable industry includes an unknown category as well as change of job that is coded as follows: no change, change voluntary, change mandatory, reason for change was not possible to predict.

22. After several operations our data files were reduced substantially although we thus tried to avoid this. Figure 5 presents this for the cross-section analysis over the five years. The year 2001 is not included since this year was not possible to analyze in each model.

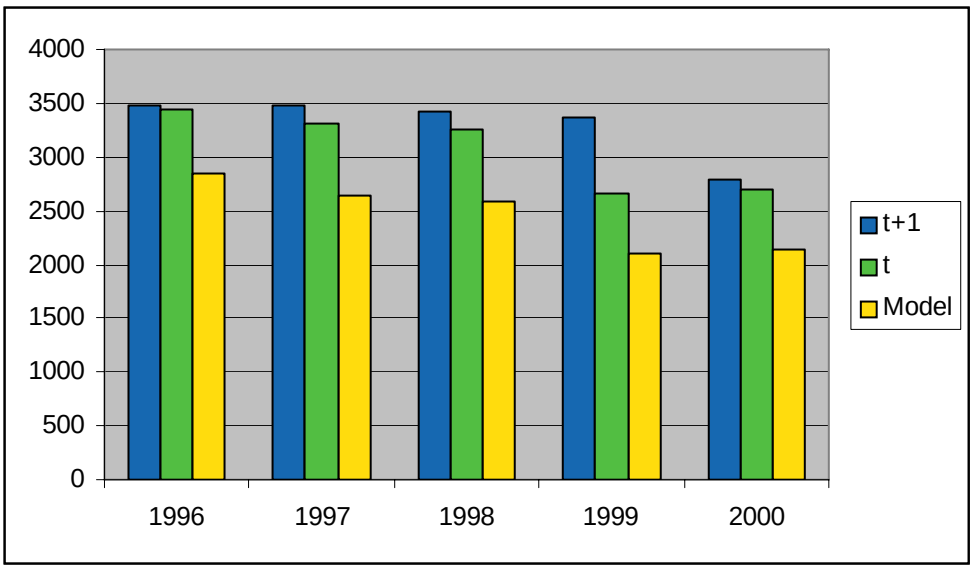


Figure 4. Numbers of observations of the two data alternatives and in the cross-section model

23. Respectively, the data loss in panel analysis is presented in Figure 5. This shows that the reduction in our final analysis is about 50% due to the above mentioned several operations.

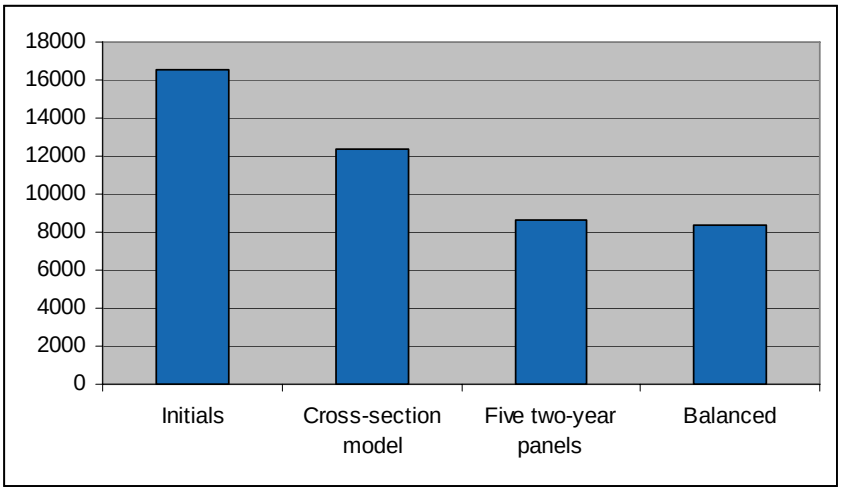


Figure 5. Number of wage earners over 5 years in the four data sets

24. We have not analyzed results against non-post-edited data sets but it is easy to see that such results would not have any sense. We can still discuss whether our all operations were correct. There would have been an option to reduce the data even more.

25. Our estimates from different models are in any case plausible. For example, we see that many estimates for interviewed vs register wages are rather close to each other. An exception is the ‘missing’ category of industry in which the regression coefficient of register wage is significantly positive but that of interviewed wage is negative (we cannot say a good reason). We see also plausibly that wage earners who are little satisfied with their job have a lower wage premium but this does not increase after a medium satisfaction level.